Reliability modelling for fatigue life prediction with application to components in dynamic systems of rotorcraft

Dekker, Sam

DOI
10.4233/uuid:88d76fda-0cbb-402e-a834-aa76b88a4e3d

Publication date
2018

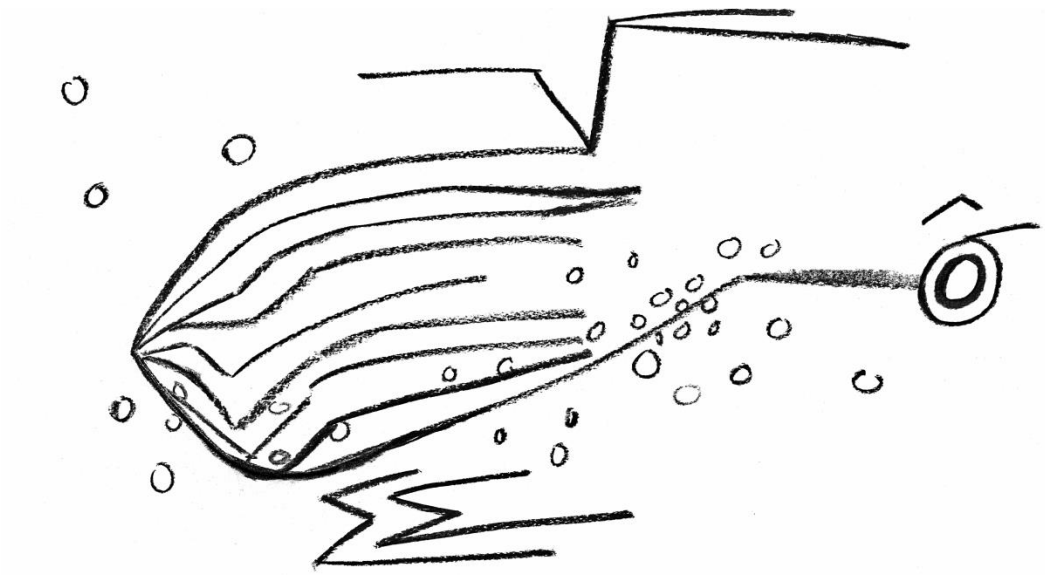
Document Version
Final published version

Citation (APA)
Dekker, S. (2018). *Reliability modelling for fatigue life prediction: with application to components in dynamic systems of rotorcraft*. [Dissertation (TU Delft), Delft University of Technology].
https://doi.org/10.4233/uuid:88d76fda-0cbb-402e-a834-aa76b88a4e3d

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Reliability modelling for fatigue life prediction

with application to components in dynamic systems of rotorcraft



Sam Dekker

Illustration front page: Ike van Cleeff ©2017

Reliability modelling for fatigue life prediction

with application to components in dynamic systems of rotorcraft

Dissertation

for the purpose of obtaining the degree of doctor at Delft University of Technology
by the authority of the Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen;
Chair of the Board for Doctorates to be defended publicly on
Monday, 12 March 2018 at 12:30 o'clock

By

Sam Hiawatha DEKKER
Master of Science in Aerospace Engineering, Delft University of Technology, the Netherlands
born in Amsterdam, the Netherlands

This dissertation has been approved by the:

promotor: Prof.dr.ir. R. Benedictus and
copromotor: Dr.ir. R.C. Alderliesten

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof.ir. R. Benedictus,	Delft University of Technology, promotor
Dr.ir. R.C. Alderliesten,	Delft University of Technology, copromotor

Independent members:

Prof.dr. F.H.J. Redig,	Delft University of Technology
Prof.Dr.-Ing. M. Hajek,	Technical University of Munich
Prof.dr.ir. T. Tinga,	University of Twente
Dr.Dipl.-Ing. O. Fink,	Zurich University of Applied Sciences

Other member:

Dr.Dipl.-Ing. G. Wurzel,	Airbus

Reserve member:

Prof.dr. R. Curran	Delft University of Technology

ISBN: 978-94-6295-865-4

Printed and published by ProefschriftMaken || www.proefschriftmaken.nl
A digital version of this document can be obtained from: https://repository.tudelft.nl/

Present work was prepared with a significant contribution as supervisor by Dr.Dipl-Ing. G. Wurzel. The contribution primarily focussed on the structure and presentation of the work.

Present work was carried out at Airbus in Munich and Donauwörth, Germany and was funded with the support of Airbus and the German Federal Ministry of Economics & Energy in the framework of the German federal research programme LuFo-IV.

Acknowledgements

First and foremost, I am extremely happy to be able to present a finished and complete dissertation. It has been a long and deep commitment and I hope the result does bring a worthwhile advancement to those who endeavour to read it. Doing the work brought me many new skills and insights that I would otherwise never have obtained. Nevertheless, I realise that many friends have had to exercise patience with my limited scope of conversation and leisure.

I wish to express sincere gratitude to Airbus, and in particular Falk Hoffmann, for developing such an exciting research program and for the opportunity to be part of it. Admiration for their bravery of taking me onboard for this project and letting me poke around in long-established methods is in order too. Also, the patience and flexibility of both Deft University of Technology and Airbus in supporting and guiding me was much appreciated.

This dissertation would not have taken its current form without the support, suggestions, and motivation from (in random order): Falk Hoffmann; Stefan Bendisch; Georg Wurzel, Daniel Reber, René Alderliesten, Kamil Kaczmarczyk, and many other colleagues from Airbus. In addition, the feedback from (anonymous) reviewers committing their time to suggest improvements was indispensable.

Present work also builds on outlier detection and filtering performed by a custom and proprietary tool for semi-automated data analysis and developed and validated by Kamil Kaczmarczyk, Alexandre Neureiter, Michael Proff and Inge Hoffgärtner, in addition to diligent pre-processing and quality control processes by the flight test department of Airbus in Donauwörth and Ottobrunn. Although these contributions have not been referenced in detail, they were of importance nevertheless.

Also, I wish to explicitly acknowledge the work and organisation performed by Falk Hoffmann, Georg Wurzel and an anonymous but renowned helicopter operator to collect the in-service data that was invaluable in developing and validating new methods for usage-based fatigue life predictions.

Sam Dekker, 16 April 2017

Abstract

A mechanical component can break due to repeated load cycling, even if these loads remain well below the component's regular static strength. In a simplified fashion, a component's fatigue life depends on the loads that it has to endure during its service life, as well as its fatigue strength to resist the formation of cracks. Since both of these factors can be considered as random variables, the time until a fatigue-induced rupture occurs can be considered as a random variable as well. Airworthiness regulations require that aircraft manufacturers show by numerical analysis that the probability that a fatigue failure occurs during a critical part's maximum allowable service life does not exceed a specified probability.

Classic fatigue life predictions depend on a-priori and conservative assumptions about the usage of aircraft. The use of modern flight data recorders and digital infrastructure enables continuous processing of recorded flight data. This makes it possible to derive the individual usage of aircraft components and to adapt the permissible service life of individual components to their actual usage. This approach is referred to as Virtual Fatigue Life Monitoring (VFLM). However, unless elaborate load measurement equipment is added to aircraft, recorded flight data does often not enable the reconstruction of in-flight loads and induced fatigue damage with high accuracy and precision. Present work introduces new methods to implement VFLM by machine learning and to statistically predict and mitigate the effect of random load prediction errors on the reliability of derived and part-individual fatigue life predictions.

Many classic fatigue life prediction methods treat and mitigate the effects of random loads and random fatigue strength independently from each other. The effect of their combined randomness is usually not modelled and assumed to be negligible. Also, the individual effect of random loads is mostly not quantified but rather mitigated by using top-of-scatter or similar loads, or even by assuming that average loads may be used and that the effect of load scatter is small. The present work aims to substantiate the reliability of VFLM-determined service lives without assuming that the effect of random load prediction errors is small; without assuring that predicted loads are always conservative, and without assuming that the effect of simultaneous uncertainty about fatigue strength and loads is small. These objectives represent a significant change in the methodology to predict fatigue. Therefore, after giving an introduction in chapter 1, present work starts with the introduction and benchmarking of new methods that enable these goals for classic fatigue life prediction in chapters 2 and 3, before modifying and validating these methods for VFLM in chapter 4.

Present work introduces a new and significantly improved (semi-)synthetic simulation framework in chapter 2. With this framework, it is possible to better quantify and compare the accuracy and precision of different methods to model the maximum allowable service life of a component. This framework is used in the same chapter to compare the accuracy and precision of two reliability substantiation models for classic fatigue life prediction which both assert that the usage of a helicopter is known in advance in the form a conservative design usage assumption.

The first method is a classic and simplified method that assumes that the effect of randomly distributed manoeuvre loads upon the distribution of fatigue life is not significant. This method thus simplifies the numeric evidence of reliability by predicting the allowable service life of a component using a conservative value of fatigue strength. The reliability of the employed value of fatigue strength is asserted to equal the level of reliability of the entire fatigue life prediction. Simulations in present work demonstrate that there are indeed circumstances under which this method yields accurate and precise results.

However, as an alternative and improved solution, present work introduces and validates a new and more generically applicable modelling framework in chapter 2 as well that can simulate, quantify and mitigate the effect of uncertainties coming from random manoeuvre loads and fatigue strength simultaneously. This new method is designed to be suitable for VFLM adaptation, in particular, due to its novel method to model in-

flight loads; which is more suitable for machine learning and integration in numerical reliability substantiation models. This method is the basis for the two new VFLM models that are introduced in chapter 4.

Before continuing to VFLM, present work first introduces a new and improved method to model random fatigue strength in chapter 3. The simulation and benchmarking work in chapter 2 revealed that the application of explicit statistical methods to substantiate the reliability of fatigue life predictions is hindered by the availability of test data to build an accurate statistical model for fatigue strength. In many cases, conducting enough fatigue tests to precisely quantify the scatter of the fatigue strength that a component possesses is too expensive.

The simulation-based methods introduced in chapter 2 to provide numeric evidence of the reliability of fatigue life predictions do however require the explicit definition of a scattering model for fatigue strength. If only a few, or even none at all, fatigue test results are available, then the use of explicit statistical scatter models is either impossible or results in highly conservative fatigue life predictions that are not realistic.

Therefore, present work introduces a new and easy-to-apply Bayesian statistical model that allows to bound conservativeness of predicted fatigue strength by taking into account traceable and objective alternative sources of information as well, while still enabling the use of numeric statistical models. It is demonstrated in chapter 3 that explicit numerical models yield realistic fatigue life predictions when using a generic database of fatigue test results to formulate a Bayesian prior expectation about the scatter of fatigue strength. Not only can this method enable more accurate fatigue life predictions in general, it also specifically enables the generic use of the advanced numerical reliability substantiation models that present work introduces for VFLM in chapter 4 and for classic fatigue life prediction in 2 alike.

For VFLM, present work first introduces and tests in chapter 4 a simplified approach, called Direct Load & Damage Modelling (DLDM) which assumes that the effect of random prediction errors can be neglected. Using more than one thousand hours of flight data collected from three commercially operated helicopters specially equipped with strain-gauges to independently compare predicted loads with actually measured loads, it could be demonstrated that DLDM can enable large fatigue life extensions in comparison to standard predictions using conservative design usage assumptions. However, it is also demonstrated that its simplified numerical reliability substantiation model does not generally hold and that the effects of random load prediction errors cannot always be neglected.

Therefore, present work also introduces Probabilistic Load & Damage Modelling (PLDM) as a more accurate and generic model for VFLM. PLDM predicts the influence of random load prediction errors and uses a simulation-based statistical model to quantify and mitigate uncertainties from unknown fatigue strength and load prediction errors. It is demonstrated that PLDM yields highly accurate results and can generally be applied to substantiate large fatigue life extensions too.

Samenvatting

Mechanische onderdelen kunnen kapot gaan onder invloed van cyclische belasting, zelfs door krachten die lager zijn dan de normale treksterkte. Gebruikmakend van een gesimplificeerd model kan de levensduur van een onderdeel dat blootstaat aan zulke cyclische belasting worden bepaald door het belasting profiel en de kracht waarmee een onderdeel een vermoeiingsbreuk kan tegengegaan. Omdat beide factoren als willekeurig kunnen worden beschouwd, is ook de tijd totdat een vermoeiingsbreuk ontstaat willekeurig verdeeld. Luchtrecht verplicht fabrikanten van luchtvaartuigen om door middel van numerieke en statistische berekeningen te bewijzen dat kritische onderdelen vervangen worden voordat de kans dat een vermoeiingsbreuk optreed groter is dan toegestaan.

Klassieke methoden voor de voorspelling van vermoeiingslevensduren maken a-priorische en conservatieve aannames over het toekomstige gebruikt van onderdelen en de manier waarop met deze onderdelen gevlogen zal worden. Tegenwoordig kan er echter gebruik gemaakt worden van vluchtdatarecorders en uitgebreide digitale infrastructuur om continu vluchtdata te verwerken en om af te leiden hoe individuele onderdelen zijn gebruikt en gevlogen. Toch kan er zonder dat er gebruik gemaakt kan worden van uitgebreide en in-situ meetuitrusting, bijvoorbeeld rekstrookjes, niet vanuit worden gegaan dat afgeleide waardes van gevlogen belasting en ondergane vermoeiingsschade volledig accuraat en precies zijn. Het gebruik van alternatieve methodes waarbij de belasting en ondergane vermoeiingsschade alleen geschat kan worden, wordt Virtuele Monitoring van Vermoeiingsschade (VMV) genoemd. Het huidige werk introduceert nieuwe manieren om VMV te implementeren door middel van machinaal leren, evenals nieuwe manieren om de effecten van onvermijdbare maar willekeurige voorspellingsfouten voor belasting op voorspelde levensduren te modelleren en te voorspellen. Hierdoor kan de betrouwbaarheid van VMV-gebaseerde voorspellingen voor de levensduur van individuele onderdelen statistisch onderbouwd worden.

De meeste klassieke methodes voor het voorspellen van vermoeiingslevensduren behandelen de effecten van onzekerheid die voortkomt uit de onbekende en willekeurige belasting gedurende de levensduur van onderdelen en de willekeurige vermoeiingssterkte van deze onderdelen onafhankelijk van elkaar. Het effect van gecombineerde onzekerheid over belasting en vermoeiingssterkte wordt meestal niet geanalyseerd bij het voorspellen van de maximaal toelaatbare levensduur van onderdelen. Daarbij komt nog dat onzekerheden door willekeurige belasting meestal niet berekend worden maar in plaats daarvan worden tegengegaan door het gebruik van maximaal gemeten belasting waarden. Ook wordt er vaak aangenomen dat er gebruik gemaakt kan worden van gemiddelde belastingswaarden en dat het kan worden aangenomen dat het effect van onzekerheid door willekeurige belasting verwaarloosbaar is.

Het huidige werk heeft als doelstelling om te laten zien dat de betrouwbaarheid van voorspellingen over de maximaal toelaatbare levensduur van onderdelen kan worden onderbouwd zonder deze gebruikelijke aannames. Dit betekent dat het niet meer nodig is om aan te nemen dat het effect van willekeurige belasting klein is. Of dat het niet noodzakelijk is om zeker te stellen dat er altijd gebruik gemaakt wordt van conservatieve belasting waarden. Ook is het niet meer nodig om aan te nemen dat de onzekerheid door gecombineerde en gelijktijdige onzekerheid over belasting en vermoeiingssterkte verwaarloosbaar is. Deze doelstellingen betekenen een significante verandering in de manier waarop maximaal toelaatbare levensduren worden bepaald. Daarom begint het huidige werk, na een introductie in hoofdstuk 1, met het introduceren en testen van nieuwe manieren voor het bepalen van klassieke en maximaal toelaatbare levensduren in hoofdstukken 2 en 3. Daarna worden deze nieuwe manieren aangepast en gevalideerd voor gebruik voor VMV in hoofdstuk 4.

Het huidige werk begint in hoofdstuk 2 met het introduceren van een nieuwe en significant verbeterde methode voor het door middel van (semi-)synthetische simulaties voorspellen van de accuraatheid en nauwkeurigheid van verschillende modellen voor het voorspellen van de maximaal toelaatbare levensduur van onderdelen. Door gebruik te maken van deze methode, kan in het huidige werk in hetzelfde hoofdstuk de

nauwkeurigheid en accuraatheid van twee methodes voor het demonstreren van de betrouwbaarheid van klassieke voorspellingen van levensduren vergelijken. Een klassieke methode wordt hier gekenmerkt door het gebruik van een conservatieve aanname ten aanzien van het gebruikt ven een helikopter.

De eerste klassieke methode die getest wordt is een gesimplificeerde methode die aanneemt dat de effecten van willekeurig verdeelde belastende krachten op de verdeling van vermoeiingslevensduur verwaarloosbaar zijn. Deze methode vergemakkelijkt hiermee het numerieke bewijs voor de veiligheid van maximaal toelaatbare levensduren. Dit numerieke bewijs kan nu geleverd worden door alleen gebruik te maken van de numerieke betrouwbaarheid van de vermoeiingskracht waarmee de voorspelling wordt uitgevoerd. De methode neemt aan dat deze betrouwbaarheid gelijk staat aan de betrouwbaarheid van de voorspelde levensduur. Simulaties in dit werk laten zien dat deze aanpak onder bepaalde omstandigheden inderdaad tot accurate en nauwkeurige resultaten leidt.

Als een alternatieve en verbeterde manier voor de klassieke voorspelling van vermoeiingslevensduren introduceert het huidige werk in hoofdstuk 2 ook een nieuwe en meer generieke methode. Met deze methode is het mogelijk om de effecten van zowel willekeurige onzekerheden over de belasting tijdens vlucht elementen evenals onzekerheid over de vermoeiingskracht gelijktijdig te simuleren, te kwantificeren en te compenseren. Deze methode is ook de basis voor twee nieuwe VMV modellen. Vooral doordat de methode gebruik maakt van een nieuw model voor het beschrijven van belasting tijdens een vlucht, is dit model relatief simpel te integreren in de methodes voor machinaal leren en numerieke betrouwbaarheidsberekeningen die in hoofdstuk 4 worden geïntroduceerd voor VMV.

Voordat het huidige werk verder gaat met VMV, introduceert het huidige werk eerst een nieuwe en verbeterde manier voor het modelleren van willekeurige vermoeiingsschade in hoofdstuk 3. De uitgevoerde tests en simulaties in hoofdstuk 2 laten zien dat de toepassing van expliciete stochastische modellen voor het berekenen van de betrouwbaarheid van vermoeiingslevensduren in de praktijk wordt beperkt doordat er vaak een gebrek is aan voldoende data voor het maken van een model voor de onzekerheid over vermoeiingsschade. In veel gevallen is het uitvoeren van genoeg vermoeiingstesten voor het accuraat kunnen schatten van de strooiing van de vermoeiingssterkte van een onderdeel te duur.

De methodes die zijn geïntroduceerd in hoofdstuk 2 voor het numeriek aantonen van de betrouwbaarheid van voorspelde levensduren zijn gebaseerd op simulatiemodellen die alleen toepasbaar zijn als het mogelijk is om een expliciet model te definiëren voor de strooiing van vermoeiingsschade. Als er alleen een paar, of helemaal geen, vermoeiingstesten beschikbaar zijn, dan is het gebruik van deze expliciete en stochastische modellen of niet mogelijk, of resulterend in zeer conservatieve voorspellingen die niet realistisch zijn.

Daarom introduceert het huidige werk een nieuw statistisch model om op een eenvoudige manier en door middel van Bayesiaanse statistiek een realistische verdeling voor de vermoeiingsschade op te kunnen stellen. Het huidige werk demonstreert hoe op een expliciete en traceerbare manier met gerelateerde kennis en data een statistische verwachting over de verdeling van de vermoeiingsschade kan worden geformuleerd. Daarmee is het mogelijk realistische en meer accurate resultaten te bereiken voor onderdelen waarvoor er weinig of geen direct toepasbare test resultaten beschikbaar zijn voor de vermoeiingsschade. Het huidige werk laat in hoofdstuk 3 zien dat het door het gebruik van een generieke dataset met vermoeiingstesten mogelijk is een Bayesiaanse verwachting te formuleren over de verwachte spreiding van vermoeiingsschade, en dat dit resulteert in realistische voorspellingen voor de maximaal toelaatbare levensduur, ook als er gebruik wordt gemaakt van een expliciet statistische model voor het aantonen van de betrouwbaarheid van de voorspelling. Hiermee kan niet alleen de beschrijving van de willekeurige verdeling van vermoeiingsschade worden verbeterd. Hiermee is ook generieke toepasbaarheid van de geavanceerde simulatie modellen en numerieke betrouwbaarheidsmodellen die het huidige werk introduceert voor de klassieke voorspelling van levensduren in hoofdstuk 2 en voor VMV in hoofdstuk 4 gewaarborgd.

In hoofdstuk 4 introduceert het huidige werk een gesimplificeerde methode voor de implementatie van VMV. Deze methode wordt aangeduid met Directe Kracht & Schade Beschrijving (DKSB) en gaat er vanuit dat het effect van willekeurige onnauwkeurigheden in voorspelde krachten verwaarloosbaar is. Het huidige werk maakt gebruik van meer dan duizend uur aan vluchtdata van drie commercieel gevlogen helikopters die speciaal zijn uitgerust met apparatuur voor het meten van daadwerkelijke krachten om de geïntroduceerde methodes voor VMV te testen. Hiermee was het mogelijk om voorspelde krachten te vergelijken met de krachten die daadwerkelijk optraden. Het huidige werk laat zien dat VMV in staat is om de veiligheid van grote verlengingen van de maximaal toegestane levensduur aan te tonen. Uitgebreide simulaties ter verificatie laten echter ook zien dat het gesimplificeerde betrouwbaarheidsmodel van DKSB niet onder alle omstandigheden adequaat is en dat de effecten van onnauwkeurig voorspelde krachten niet altijd verwaarloosbaar zijn.

Daarom introduceert het huidige werk ten slotte Stochastische Kracht en Schade Beschrijving (SKSB) als een meer nauwkeurig en generiek alternatief voor VMV. SKBS is in staat om de effecten van willekeurige fouten in voorspelde krachten en onzekerheid over willekeurige vermoeiingskracht gelijktijdig te voorspellen en te compenseren. Test resultaten in het huidige werk laten zien dat SKSB accuraat en precies is onder alle geteste omstandigheden en dat SKSB ook in staat is de veiligheid aan te tonen van grote verlengingen van de maximaal toegestane levensduur van onderdelen.

Contents

List of figures

List of tables

1 Introduction

Helicopters contain many primary parts that are subject to repetitive loading and for which it must be considered that they may develop fatigue cracks that are difficult or expensive to detect. For susceptible parts, such cracks may grow extremely rapidly and suddenly cause rupture or breakage. Fatigue failures shall never occur in flight-critical aircraft components and airworthiness regulations, therefore, require that rotorcraft manufacturers show, by explicit statistical analysis, that the probability of a fatigue failure to occur is at most extremely remote. Classically, and as recommended by aviation authorities, this can be done by modelling and mitigating the effects of three factors influencing the time until a fatigue failure can occur: a component's fatigue strength, the loads that occur while flying manoeuvres, and the type of flight operations that a helicopter actually flies. This work introduces new methods for the modelling and statistical analysis of all these three elements and demonstrates how these methods can be used to improve the accuracy of fatigue life predictions and to justify extensions to permissible service lives of fatigue-susceptible parts.

1.1 Definition of a Service Life Limit

For many critical components in a helicopter, the time in which a crack or flaw can grow from a detectable size to ultimate component failure is too short to be covered by a practical inspection interval. This relatively rapid growth of damage has several primary causes. First, rotorcraft components are inherently subjected to vibratory and high-frequency cycling load. Therefore, once a large crack has developed in a part, this crack can grow rapidly. Second, component weight must generally be minimized, often resulting in components without generous strength margins and thus susceptibility to fatigue damage. Third, helicopters are highly manoeuvrable and can be flown in many ways, potentially leading to a relatively high number of load cases which could cause fatigue damage. Therefore, once a crack has developed, its further growth can be expected to be accelerated and rapid.

Due to this potential rapid crack growth, rotorcraft manufacturers must predict a fatigue life, and a component operating limit, during which a component can be used safely. Beyond this limit, the probability of a fatigue failure may rise to an unacceptable level. Traditionally, such a limitation is expressed in the form of a maximum number of flight hours that a component may be used, and is referred to as a Service Life Limit (SLL). The time until fatigue failure is a random variable. As illustrated in Figure 1.1, the SLL can be seen as a quantile of the distribution of fatigue life. In practise, the distribution of in-service component fatigue life can be determined by simulation and with the use of modelling assumptions, as introduced in detail in chapter 2. Testing the fatigue life of a large number of full-scale components under representative in-service loads is generally too expensive and would still require several modelling assumptions, including service load spectra and the distribution model for fatigue life.

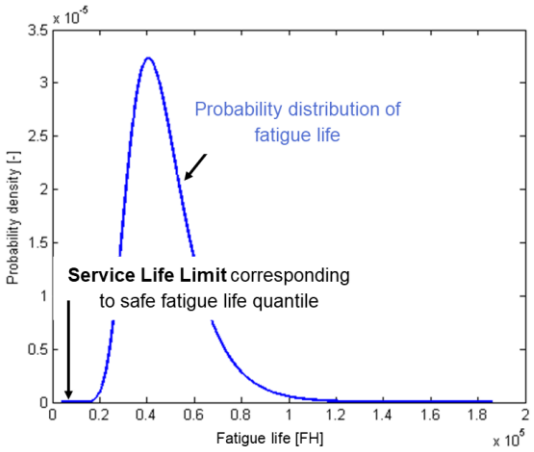


Figure 1.1: Illustration showing how a Service Life Limit is a quantile of a statistical fatigue life distribution

1.2 Modelling of fatigue damage

According to the Palmgren-Miner linear damage accumulation hypothesis [1], the development of material fatigue is expressed on a normalised scale ranging from zero, corresponding to a part that has never been subjected to any damaging load cycle, to one, corresponding to a part that has developed a fatigue crack causing failure. It is common to express progress on this scale as the development or accumulation of fatigue damage, although this progress may not correspond to the development of actual physical damage.

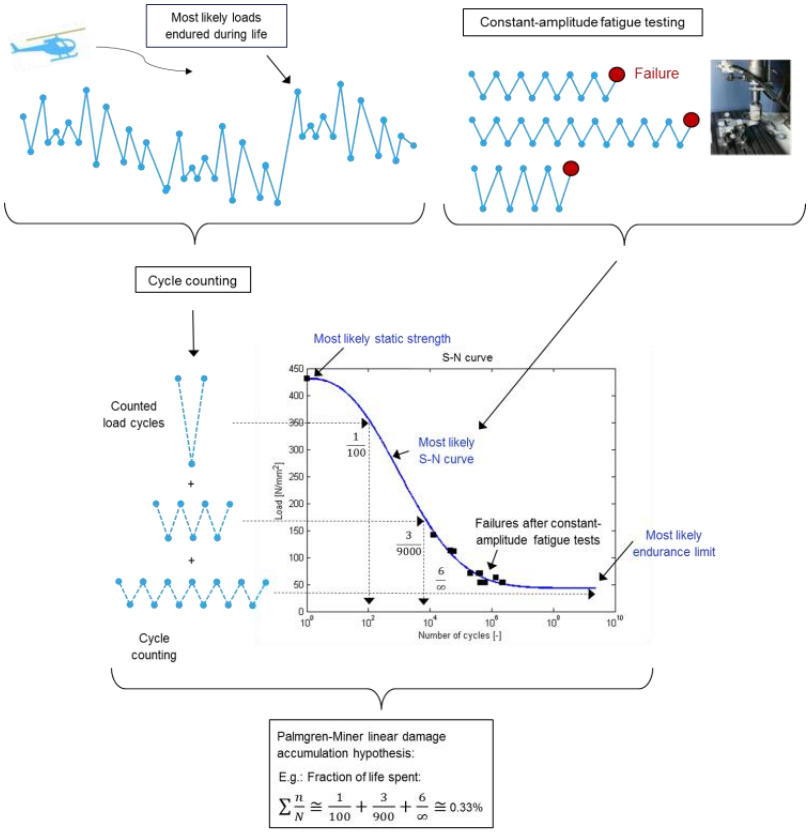


Figure 1.2: Schematic illustration how a load spectrum can be computed by cycle counting of a load history, and how an S-N curve can be generated from constant-amplitude fatigue tests, and how an S-N curve and the Palmgren-Miner linear fatigue damage hypothesis can be used to compute accumulated fatigue damage.

As graphically summarised in Figure 1.2, the fatigue strength of a part can be modelled by an S-N curve, which is also known as a Wöhler curve. An S-N curve describes how many load cycles a part can endure until fatigue failure. Points on the S-N curve can most easily be determined by testing parts under constant amplitude loading until they fail. The complete curve can then be determined by curve fitting. In reality, parts are seldom subjected to constant amplitude loading throughout their life. Real load signals must, therefore, be decomposed into constant amplitude load cycles by cycle counting, before their effect on fatigue damage accumulation can be modelled by an S-N curve. After cycle counting, i.e. load signal decomposition into constant amplitude loading blocks, their proportional effect on accumulated fatigue damage can be modelled by simple summation. In the illustrated case in Figure 1.2, the load signal 'consumes' about 0.33% of the part's fatigue resistance. Also illustrated is a case where cycles do not cause any, or only negligible, fatigue damage because these load cycles are too weak to cause any fatigue damage according to the S-N curve model. In such a case, it can be said that the loads are below the part's endurance limit.

1.3 Determining Classic Service Life Limits for helicopter dynamic components

A classic and commonly used method to numerically substantiate the reliability of SLL predictions of how long a part can be used before the probability of a fatigue failure exceeds an acceptable limit is summarised in Figure 1.3. This simple and analytical prediction method is analysed in more detail in chapter 2. In principle, when using this method it is assumed that fatigue life is determined by three major contributors: fatigue strength, manoeuvre loads, and mission profile. Although these last two contributors are actually random variables, it is nevertheless assumed that the mission profile is known and can be determined conservatively and that the influence of variability of manoeuvre loads can be neglected. All the uncertainty concerning a part's fatigue life is then determined by uncertainty about its fatigue strength. A simplified way to predict a part's fatigue life conservatively, and to comply with airworthiness requirements, is to predict fatigue life while assuming very low and unlikely fatigue strength. The probability that the part actually has such low fatigue strength can then be set as extremely remote and it can then be regarded that the probability that the part experiences a fatigue failure during its operational life will be equally low.

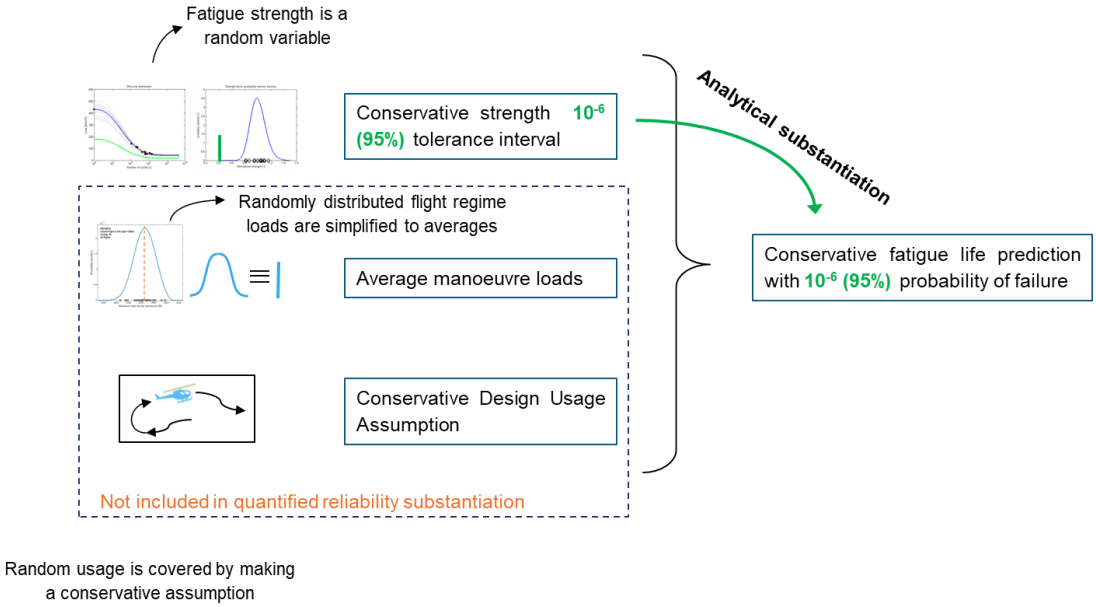


Figure 1.3: Schematic summary how the standard analytical method to predict fatigue life detailed in section 2.2 predicts a Service Life Limit. The standard classical method uses only the reliability of the conservative working S-N curve to numerically substantiate the reliability of the SLL prediction.

A slightly more detailed overview of the method to compute an SLL is schematically outlined in Figure 1.4. It is again summarised that the SLL principally depends on a conservative Design Load Spectrum and a working S-N

Curve corresponding to a specific reliability level. The working curve is derived from a probabilistic model for fatigue strength, such as the model discussed in detail in chapter 3, and satisfies an explicit reliability requirement, for example, that the probability of a fatigue failure during the service life of the component may not exceed 10^{-6} with 95% confidence (hence further also written as 10^{-6} (95%)). The Design Load Spectrum is the result of a flight test campaign and a conservative Design Mission Profile (DMP) or Design Usage Assumption. The flight test campaign provides information on the loads that occur during all flight regimes and the DMP sets a conservative assumption about how often and in which order these flight regimes occur.

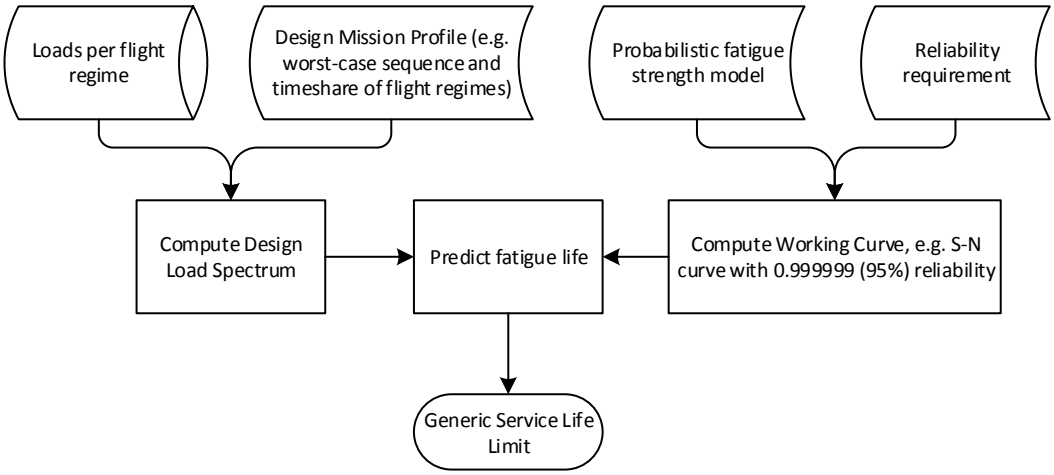


Figure 1.4: Schematic summary of the classic process to set a generic Service Life Limit (SLL) based on a working S-N curve and a design mission profile.

1.3.1 Flight regime loads from Load Classification Flights

The flight test campaign during which the in-flight component loads corresponding to all usage elements are measured is referred to as the Load Classification Flight (LCF) test campaign. During this test campaign, a representative helicopter is equipped with specially installed strain gauges to measure the in-flight loads on selected components and locations. Ideally, this helicopter then repeatedly flies through the entire allowable flight envelope. In practise though, the scope of the campaign must be limited and not all elements can actually be flown. Usually, only the most extreme load cases, gradually less severe variants thereof, and other common flight elements are covered. In many cases, the flight elements are repeated in order to cover variances too. If no flight data is available on a flight element, its loads are normally conservatively assumed to equal a more severe or similar case for which flight data is available.

1.3.2 Conservative usage assumption by the Design Mission Profile

In the practical case that is considered for this work, the flight envelope is discretized into about 150 flight conditions, e.g. a high-speed left turn with a load factor of 2.5g, and around thirty different configurations, each corresponding to a combination of weight, centre-of-gravity and pressure altitude classes. For clarity, this work distinguishes in its wording between a flight state, a flight manoeuvre and a flight regime. A flight state indicates an instantaneous or momentary condition of flight, e.g. a one second period of flight. Whereas a flight manoeuvre designates an entire manoeuvre, e.g. a flare. A flight regime finally designates a flight manoeuvre flown with a specific configuration, e.g. a flare flown with a specific weight and centre-of-gravity and at a specific altitude.

Well known examples of a DMP that often serve as a basis for specialized profiles set by individual rotorcraft manufacturers have been published by the FAA in AC 27-1B MG-11 [2] or by Edwards & Darts in the form of HELIX and FELIX [3]. Figure 1.5 is a simplified illustration of a DMP. The spectrum distinguishes between high and low-frequency spectra, a difference that will be discussed in more detail in sections 2.2.3 and 4.2.1.

Manoeuvre distribution for high-frequency spectrum	
Manoeuvre	% occurrence
Rotor run-up	5
Take-off	5
Hover	10
Level flight	60
Turn	15
Landing	5

Configuration distribution	
Configuration	% occurence
90-100% MTOW	25
<90% MTOW	75
Aft C.G.	30
Forward C.G	70
High altitude	75
Low altitude	25

Manoeuvre sequence / GAG-cycle for low frequency spectrum [10 sequences / FH]
Rotor run-up
Take-off
Hover
Landing
Take-off
Turn
Landing
Take-off
Hover level flight
Turn
Level flight
Turn
Level flight
Turn
Hover
Turn
Hover
Landing

Figure 1.5: Simplified example of a Design Mission Profile defined by a high-frequency manoeuvre distribution, a corresponding configuration distribution and an independent low-frequency manoeuvre sequence

The DMP must be conservative for all helicopters in the fleet. The reliability of a Service Life Limit (SLL) following from the process in Figure 1.4 can be substantiated using methodologies elaborated in chapter2. Helicopter usage can differ significantly within the fleet, e.g. ranging from aerial work, search & rescue, passenger transport, to law enforcement. Thus, to make sure that the usage assumption is conservative for all helicopters in the fleet, the usage assumption is considerably over-conservative for most of them. SLLs for these helicopters can thus be extended, if their actual usage would be known.

The use of a conservative DMP adds extra reliability to the classic SLL prediction. However, this extra reliability is not taken into account in the numerical reliability substantiation of a classic SLL prediction. The reason for not taking into account this extra source of reliability is that the numerical reliability requirement which must be substantiated for an SLL must be valid for all helicopters in the fleet. When using a generic and 'worst-case' DMP then it can be assumed that the DMP is valid for all helicopters. Thus, an operator only flying demanding missions can be sure that the reliability of his helicopter meets the same minimum reliability requirement as the helicopter from another operator with only light usage. This would however not be the case if, for example, usage variance would be added as a random variable in a simulation of the type introduced in section 2.5. In that case, the numerical reliability substantiation would only be valid for a helicopter with an at most average usage profile. If that would be the case, and once an operator knows that its operations are more severe than average, then this operator would also know that his helicopter is expected to have lower-than-average reliability. Ultimately that would even imply that due to his more-than-averagely severe mission profile his helicopter would actually not meet the minimum airworthiness and reliability requirements. This is the case if SLLs have only been computed and numerically substantiated by the OEM for an at most averagely used helicopter.

1.4 Introduction to Virtual Fatigue Life Monitoring

It is conceivable to install actual load measurement equipment on individual components and helicopters, for example in the form of strain gauges, just as during the Load Classification Flights. This would allow very accurate recording of actual in-flight loads. However, in the case of common strain gauges for load measurements, this would require the installation of dedicated on-board signal processing equipment,

complex calibration, and significant extra maintenance effort. Strain gauges can have a relatively low fatigue life, are susceptible to environmental effects, and require special expertise to calibrate and install. Such direct load measurements are therefore not attractive for fleet-wide serial application and their extra maintenance and equipment cost may offset any benefits gained from customized SLL extensions.

Instead, with the advent of flight data recorders and their installation in helicopters, possibly as part of an onboard Health & Usage Monitoring System (HUMS), it has become possible to track the actual usage of an individual helicopter in detail. The term Virtual Fatigue Life Monitoring (VFLM) in this work refers to adjusting the fatigue life prediction of individual components according to the actual usage of the helicopter(s) they are installed on. Since the classic DMP is asserted to be conservative, and since its conservatism is not taken into account by the numerical reliability substantiation for an SLL, replacing conservatively assumed usage by actual usage should result in elongated Service Life Limits of individual components, as in Figure 1.6. Preferably, VFLM recordings only have to comprise data from generally available onboard sensors to derive actual usage, e.g. airspeed, engine torque and bank angle.

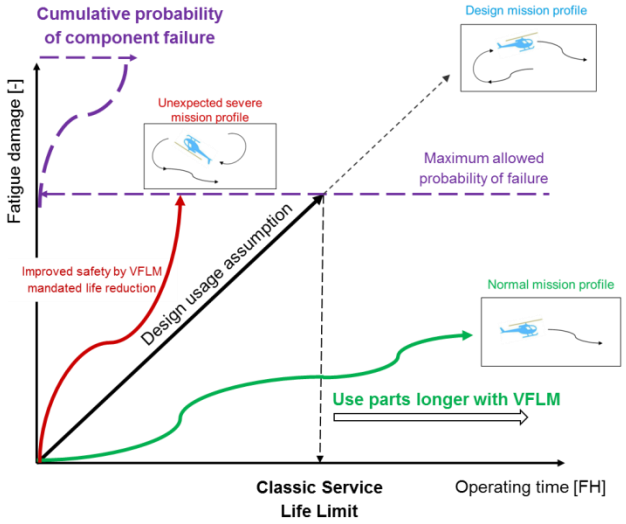


Figure 1.6: Schematic illustration how Virtual Fatigue Life Monitoring adjusts the maximum allowed operating time of a component according to its actual usage and differs from a generic SLL set by DMP.

The approach to VFLM analysed in this work assumes the presence of a complete flight data record of the entire flight history of a component. If there is sufficient correlation between recorded flight parameters and in-flight loads, then statistical correlation learned from Load Classification Flights can be used to estimate in-service loads, which can, in turn, be used to estimate the actual fatigue damage that a component has so far endured. Ideally, automatic analysis of recorded flight data in the fashion of Figure 1.7 to Figure 1.9 should thus enable a very accurate derivation of in-flight loads and thereby the determination of the true in-service load spectrum.

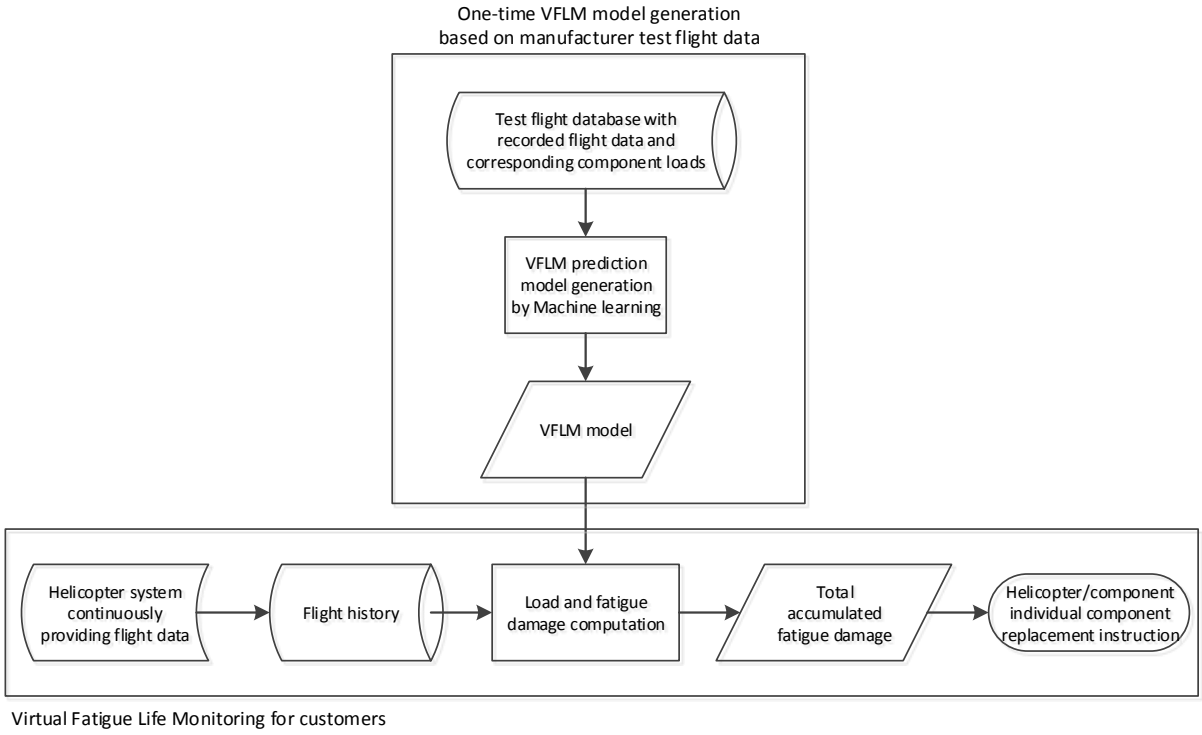


Figure 1.7: Simplified process overview summarizing how a model for Virtual Fatigue Life Monitoring is generated and used in practise.

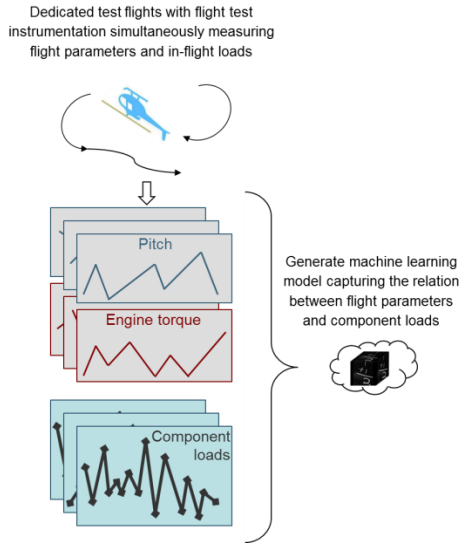


Figure 1.8: Schematic summary of how Load Classification Flights can be used to generate a predictive model that correlates flight parameters with in-flight component loads.

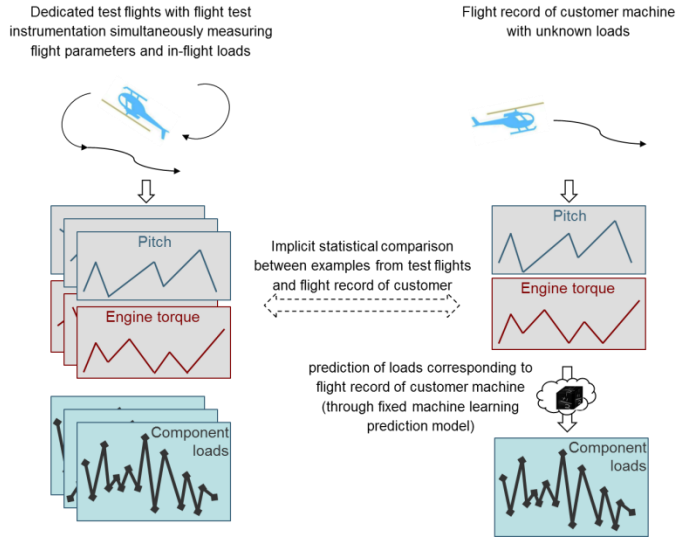


Figure 1.9: Diagram summarising how Machine Learning can be used to estimate otherwise unknown loads comparing recorded flight parameters with in-flight load examples recorded during Load Classification Flights.

1.5 Examples of the practical significance of VFLM

A successful example of VFLM was published for the Westland Lynx operated by the Royal Netherlands Navy (RNLN). The Lynx multi-purpose military helicopter has permissible airframe life of seven thousand flight hours. To prolong the lifetime, the Dutch National Aerospace Laboratory (NLR), the RNLN and the Royal Netherlands Air Force (RNLAF) initiated research efforts [4, 5] to show that the airframe life of their Lynx fleet could be extended. The objective was to show that the RNLN did not use their Lynx machines as severely as assumed by

the manufacturer in its airframe life prediction. All machines were equipped with flight data recorders and their flight data was then continuously analysed. With the analysis results, it was possible to justify airframe lifetime extensions of about 25% and to achieve annualised cost reductions in the order of 2.5 million euros.

Results obtained from surveys, mainly by the United States Military, among a wide range of rotorcraft and fixed-wing aircraft gave an example of the potential of determining the true usage of an individual component. One particular example is the design usage spectrum for the Bell AH-1W Iroquois. It assumes that a helicopter spends 3.5% of its flight time in level turn. An actual usage survey showed however that this percentage can actually vary between 0.1 and 22.4% [6]. Bos and Oldersma published another notable case where true usage significantly differed from the OEM's design usage. Flight Regime Recognition results showed that the Boeing CH-47D "Chinook" operated by the Royal Netherlands Airforce spent twice as much time in manoeuvring compared to the OEM's design assumptions [7]. Another example with analysis results of in-flight recorded data and comparison with an OEM DMP is a survey by Cronkhite *et.al.* [8]. Adams *et al.* report an almost one-quarter lifetime extension potential for the main rotor shaft of the S-61 "Sea King" helicopter by virtue of knowing the true machine usage [9]. More recent examples include a reported 240% lifetime extension potential for the S-92 main rotor swashplate [10, 11] and similar prospects for the V-22 Osprey [12]. These examples clearly underline the potential of tracking true and individual usage instead of using fleet-wide usage assumptions.

1.6 Summary of research strategy

VFLM is an actively researched field in the rotorcraft industry over 25 years. A comprehensive overview of different approaches and their properties is given in section 4.1. From this overview, it can be concluded that most current models are either highly simplified, limiting their potential for generic applicability, or prone to significant prediction inaccuracies. In addition, practically all of the existing models to substantiate the reliability of VFLM-based fatigue life predictions assume that VFLM models make perfect predictions. This means that these reliability substantiation models assert that VFLM predictions of flown manoeuvres and in-flight loads are free of errors and scatter. Such accurate prediction is however not always attainable and, therefore, inhibits generic applicability and may introduce significant inaccuracies in the reliability substantiation.

Present work aims to introduce a VFLM model and an associated statistical reliability substantiation model that enables the substantiation of high reliability levels, e.g. in the order of 0.999999 with 95% confidence, and that is universally applicable. In particular, it aims to do so without assuming that the effect of random load prediction errors is small; without assuring that predicted loads are always conservative, and without assuming that the effect of simultaneous uncertainty about fatigue strength and loads is small. These objectives represent a significant change in the methodology to predict fatigue. Therefore, after giving an introduction in chapter 1, present work starts with the introduction and benchmarking of new methods that enable these goals for classic fatigue life prediction in chapters 2 and 3, before modifying and validating these methods for VFLM in chapter 4.

Present work introduces a new and significantly improved (semi-)synthetic simulation framework in chapter 2. With this framework, it is possible to better quantify and compare the accuracy and precision of different methods to model the maximum allowable service life of a component. The framework makes use of synthetically generated fatigue life prediction problems and represents a significant improvement in the depth and rigour by which existing substantiation frameworks can be reviewed and benchmarked. Especially the ability to repeatedly simulate and verify an entire fatigue life prediction process expands the scope of verification with respect to previous work and uniquely enables the verification of confidence level estimations as well. The new framework is first used in the same chapter to compare the accuracy and precision of two reliability substantiation models for classic fatigue life prediction which both assert that the usage of a helicopter is known in advance in the form a conservative design usage assumption.

The first method that is introduced in chapter 2 is a classic and simplified method that assumes that the effect of randomly distributed manoeuvre loads upon the distribution of fatigue life is not significant. This method thus simplifies the numeric evidence of reliability by predicting the allowable service life of a component using a conservative value of fatigue strength. The reliability of the employed value of fatigue strength is asserted to equal the level of reliability of the entire fatigue life prediction. Simulations in present work demonstrate that there are indeed circumstances under which this method yields accurate and precise results.

However, as an alternative and improved solution, present work also introduces and validates a new and more generically applicable modelling framework for classic fatigue life prediction in chapter 2 that can simulate, quantify and mitigate the effect of uncertainties coming from random manoeuvre loads and fatigue strength simultaneously. This new method is designed to be suitable for VFLM adaptation, in particular, due to its new method to model in-flight loads; which is more suitable for machine learning and integration in numerical reliability substantiation models. This method is the basis for the two new VFLM models that are introduced in chapter 4.

Before continuing to VFLM, present work first introduces a new and improved method to model random fatigue strength in chapter 3. The simulation and benchmarking work in chapter 2 revealed that the application of explicit statistical methods to substantiate the reliability of fatigue life predictions is hindered by the availability of test data to build an accurate statistical model for fatigue strength. In many cases, conducting enough fatigue tests to precisely quantify the scatter of the fatigue strength that a component possesses is too expensive.

The simulation-based methods introduced in chapter 2 to provide numeric evidence of the reliability of fatigue life predictions do however require the explicit definition of scattering models for fatigue strength. If only a few, or even none at all, fatigue test results are available, then the use of explicit statistical scatter models is either impossible or results in highly conservative fatigue life predictions that are not realistic.

Therefore, present work introduces a new and easy-to-apply Bayesian statistical model that allows to bound conservativeness of predicted fatigue strength by taking into account traceable and objective alternative sources of information as well, while still enabling the use of numeric statistical models. Not only can this method enable more accurate fatigue life predictions in general, it also specifically enables the generic use of the advanced numerical reliability substantiation models that present work introduces for VFLM in chapter 4 and for classic fatigue life prediction in 2 alike.

For VFLM, present work first introduces and tests in chapter 4 a simplified approach, called Direct Load & Damage Modelling (DLDM) which assumes that the effect of random prediction errors can be neglected. Present works uses more than one thousand hours of flight data collected from three commercially operated helicopters specially equipped with strain-gauges to independently compare predicted loads with actually measured loads. With this data, present work demonstrates that DLDM's simplified numerical reliability substantiation model does not generally hold and that the effects of random load prediction errors cannot always be neglected.

Therefore, present work finally also introduces Probabilistic Load & Damage Modelling (PLDM) as a more accurate and generic method for VFLM. PLDM predicts the influence of random load prediction errors and uses a simulation-based statistical model to quantify and mitigate uncertainties from unknown fatigue strength and load prediction errors. It is demonstrated that PLDM yields highly accurate results and can generally be applied to substantiate large fatigue life extensions.

1.7 Definition of research questions

Following the research strategy outlined in the previous section 1.6, the current section introduces the scientific research questions that are addressed by the present work. Concluding discussions for these research questions are presented in chapter 5.

1. **Which additional uncertainties about predicted fatigue life are introduced if it is not assumed that the variance of a component's fatigue strength can be estimated without uncertainty?**

The fatigue strength of a component can be subject to significant scatter. The variation of the fatigue strength can be modelled by a distribution model. This model can then be used when substantiating the reliability of a fatigue life prediction.

A distribution model for a component's fatigue strength generally needs the mean and variance, or standard deviation, of fatigue strength. These distribution parameters can be estimated by fatigue testing. In many aerospace applications, it is common to assume that the variance of an estimated distribution of fatigue strength is known without any residual estimation uncertainty [13, 14, 15]. This means that it is asserted, that the standard deviation of the fatigue strength of full-scale components can be known conservatively from observed scatter from coupon tests, from tests results from similar components, or by other statistical or analytical methods.

Present work, however, asserts that the variance of a component's fatigue strength must be estimated while considering the probability of estimation errors. The fatigue life substantiation methodology utilised in present work is based on chapter 4.1 in NATO AGARD-AG-292 [16]. This guideline assumes that an estimate of the variance of the distribution of fatigue strength can be obtained from the results of full-scale fatigue tests, provided that more than three to six components have been tested.

However, the NATO AGARD-AG-292 guideline does not specify means to prevent or account for possible estimation errors of the distribution variance of a component's fatigue strength. Therefore, present work includes simulations in chapter 2 and 3 to indicate the potential inaccuracies from assuming that the standard deviation of a component's fatigue strength has been estimated with full precision and accuracy.

All the newly introduced reliability substantiation models and verification simulations in present work generally do take the existence of such estimation inaccuracies into account. In doing so, present work does thus potentially enable significant accuracy improvements for numeric reliability substantiation of fatigue life predictions and associated verification methods.

2. **Can the accuracy of fatigue life predictions be improved by accounting for the effects of combined randomness of fatigue strength and flight regime loads?**

It is common in aerospace [17] to model a component's fatigue life as a function of three variables: its fatigue strength, the loads that a component experiences during a manoeuvre, and the type, sequence and time of these manoeuvres. The last is commonly assumed in the form a conservative Design Mission Profile. However, fatigue strength and manoeuvre loads can both be considered as random variables whose variability must be accounted for by statistical analysis.

Normally, the influence of randomness of fatigue strength and manoeuvre loads on the fatigue life of a component is treated separately, e.g. as in AGARD-AG-292 [17] or by Thompson and Adams [18]. It is for example common to assume that manoeuvre loads can be estimated conservatively from flight test results, e.g. by top-of-scatter modelling, or that their variance may be neglected. Uncertainty in fatigue strength is however generally fully modelled and a conservative quantile is commonly used for fatigue life prediction [17]. A numerical estimate of the probability of a non-conservative fatigue life prediction is usually based on the

conservative quantile of fatigue strength only [17], or some multiplication of estimated probabilities of failure of individual and independent factors [18].

The validity of these modelling approaches can be challenged by simulation results presented by Tong *et.al.* [19]. In addition, analysis of flight data in Appendix F demonstrates significant variability in manoeuvre loads that may not be neglected nor covered by top-of-scatter modelling based on few flight test results. The generic use of a top-of-scatter provides a level of conservatism that is proportional to the size of the sample from which the top-of-scatter value is computed. In aerospace applications, this sample size is often small and the top-of-scatter must, therefore, be expected to lie close to the mean of the load distribution and thus cannot be expected to add major conservatism.

Present work introduces a unique statistical verification framework in chapter 2 enabling to test the accuracy and precision of statistical modelling assumptions for the reliability substantiation of fatigue life predictions. The introduced simulation framework allows modeling with more accuracy how fatigue life is a probabilistic and a non-linear function of in-flight loads and component fatigue strength. The framework can also specifically be used to simulate the effects of the assumption in NATO AGARD-AG-292 [17] that the influence of variance of manoeuvre loads on predicted fatigue life may be neglected.

In order to improve the accuracy and generalised applicability of fatigue life predictions, chapter 2 also introduces a new prediction method which is capable of simulating and mitigating the effects of simultaneous and combined uncertainty of fatigue strength and manoeuvre loads. This new modelling framework is summarised in Figure 1.10. Simulations in section 2.6 demonstrate that this new simulation-based method can improve the accuracy of the predicted reliability of conservative fatigue life predictions.

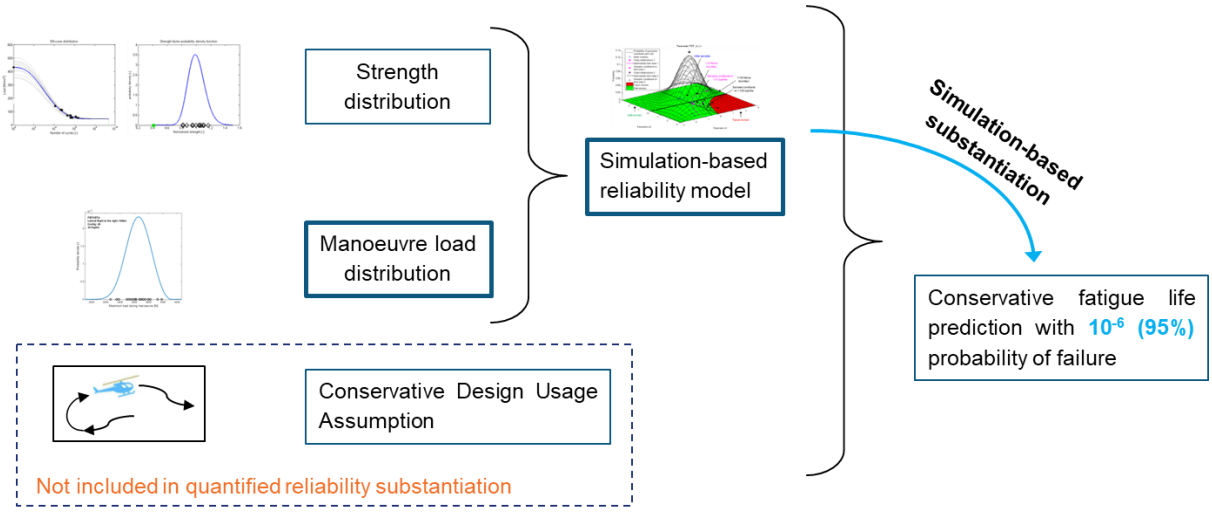


Figure 1.10: Schematic overview of how the new simulation-based method for classic fatigue life prediction introduced in section 2.5 uses a statistical simulation model to numerically substantiate the reliability of a predicted fatigue life and how the model takes into account simultaneous uncertainty about fatigue strength and manoeuvre loads.

3. What is the importance of confidence level analysis for fatigue life prediction?

If life-limited components on a rotorcraft must be replaced before their Service Life Limit is reached to be able to guarantee that the probability of an in-service fatigue failure remains below an acceptable low limit. Numerical and explicit substantiation that these limits are not exceeded is legally required for critical rotorcraft components [17]. The statistical models that substantiate the reliability of these SLLs generally depend on estimated distributions. Airworthiness regulations, however, do not explicitly require to account for the possibility of estimation errors for these distributions.

Distribution estimates for fatigue strength and manoeuvre loads are often based on a limited amount of test data. Full-scale fatigue tests and instrumented flight tests are costly and can only be conducted sparsely. As a result, distributions for fatigue strength and manoeuvre loads must be estimated with limited accuracy and precision and may inadvertently be inaccurate. This can in turn cause uncertainties to derived statistics, such as the expected distribution of fatigue life, its quantiles, and the predicted reliability of conservative fatigue life estimates.

Present work introduces simulation models that can determine the potential effect of estimation uncertainty due to distribution estimates based on small sample sizes. The framework thus enables the explicit quantification of the precision of simplified fatigue life prediction approaches in which no use is made of confidence level analysis. Several simulations using this model illustrate the potential effects of small sample-size induced uncertainties on the reliability of predicted fatigue life limits.

In particular, chapter 2 exemplifies the importance of taking into account that especially fatigue strength can generally only be estimated from a limited number of fatigue test results. If there are only a few of these test results, then they can randomly misrepresent the true distribution of fatigue strength and cause inadvertent but significant prediction errors.

4. **When substantiating the reliability of a fatigue life limit under the consideration that the variation of fatigue strength is a random variable, can the use of Bayesian statistics prevent over-conservative fatigue life predictions and enable more economical test requirements?**

The prediction and simulation models in present work do not incorporate the common design simplification that estimated scatter of fatigue strength equals true scatter. Following AGARD-AG-292 [17], the models instead consider that the variance of a component's fatigue strength can principally be estimated based on a limited number of full-scale fatigue tests.

Simulations in present work, particularly in chapter 4, demonstrate that regarding the variance of fatigue strength as a statistically distributed and uncertain parameter increases the likelihood that simplified reliability substantiation models can be used for the reliability substantiation of fatigue life predictions. The simulations demonstrate that the more significant uncertainty about fatigue strength is, the more the effect of uncertainty about in-flight loads can be neglected.

As a consequence of removing the simplification that the sample estimate of the standard deviation of fatigue strength equals the true standard deviation, the statistical evaluation models must statistically cover the possibility that actual scatter might be higher than estimated. In general, the fewer test results are available, the higher the probability that scatter has been estimated inaccurately. To mitigate this effect, it is necessary to apply additional conservatism, increasingly severe with decreasing sample sizes. Common statistical methods consider that the availability of very few samples makes the likelihood of making a correct estimate of the scatter of fatigue strength very low. These common methods, therefore, prescribe the use of highly conservative design values for fatigue strength.

Most aerospace applications currently do not explicitly model the likelihood of estimation errors of fatigue strength scatter. These implementations rely on other methods to ensure that conservative values for scattering are used during design computations. The use of new and explicit statistical models instead may thus result in fatigue life predictions that are significantly different from common legacy applications.

For many applications, there is a significant amount of pre-existing data available that can be used to make a well-informed initial estimate of the variance of a component's fatigue strength, even before any full-scale fatigue test has been performed on the component itself. Present work introduces Bayesian statistics to use such prior data and experience to bound possible fatigue strength variances to a realistic domain. This method allows the explicit statistical consideration of uncertainty about estimated variance of fatigue strength. Present

work also tests if Bayesian statistics can be used to form a generic and easy-to-use method that allows considering estimates of fatigue strength variance as error-prone, without forcing unnecessary conservatism or complexity.

In particular, chapter 3 illustrates that commonly used methods to mitigate the effect of estimation errors due to the use of a small sample size can yield inaccurate results. Most of these methods rely on assumptions that are only accurate for large to medium sample size conditions. Their use under small sample-size conditions can thus give rise to significant inaccuracies. Otherwise, other methods that are theoretically suitable for small sample-sizes are found by present work to generally yield impractical and over-conservative results and their use is thus considered to be economically unattractive. Therefore, a new method for modelling estimation uncertainty of fatigue strength under small sample-size conditions is introduced in chapter 3. Although this method still contains considerable modelling assumptions, a generic application example demonstrates that its results are in line with industry accepted practise. In addition, its simplicity and adaptability should make it useful for practical application.

5. **What are suitable and generic reduction factors for S-N working curves for classical SLL substantiations when these are based on few or no results from directly applicable full-scale fatigue tests?**

FAA airworthiness guidance material AC-27 MG-11 [17] prescribes the use of a generic reduction of expected fatigue strength by a factor 3 if this expectation has been formed without fatigue testing the applicable component. The application of this standard factor does not clearly correspond to a numerical reliability requirement.

Many rotorcraft OEMs, in addition, make use of generic reduction factors when too few fatigue test results for the applicable component are available to apply classical statistical methods to derive useful reduction factors. The factors can, for example, be derived by interpolation between the FAA recommended upper limit of 3 and statistically derived reduction factors computed from cases where more samples are available. The traceability and statistical substantiation of such generic reduction factors, as well as their consistency, may be improved by the use of Bayesian statistics.

In chapter 3, present work uses the newly introduced Bayesian statistical model and associated dataset to numerically investigate how conservative the application of a 1/3 reduction factor for fatigue strength can be expected to be.

6. C**an the reliability of Virtual Fatigue Life Monitoring by Direct Load & Damage Monitoring be substantiated without accounting for the influence of regression or recognition errors?**

When performing fatigue life predictions, it is industry standard practise to make a conservative assumption about the manoeuvres and loading profiles that a component will be subjected to during its service life. Monitoring the actual and component-individual fatigue loading can thus be expected to lead to significant extensions of permissible service lives, and thus to less scheduled replacements of components.

Monitoring and tracking real component-individual usage can be done by recording and analysing typical flight data. Direct Load & Damage Modelling (DLDM) introduced by Dekker *et.al.* [20, 21] presented a methodology using non-linear statistical regression models to estimate in-flight loads and accumulated fatigue damage. DLDM is a generic modelling framework that can be applied to any fatigue life prediction problem and that should enable a significant improvement over other industry-common methods for Virtual Fatigue Life Monitoring [22, 23, 6]. However, DLDM's application of machine learning based models cannot be expected to result in error-free estimates. Most models introduced by industry to implement Virtual Fatigue Life Monitoring make use of simplified reliability substantiation models and either assume that their estimates of in-service usage and loads are error-free, or that the effects of their estimation errors are not significant.

A new statistical and numerical method is introduced in present work to explicitly test the potential effect of DLDM regression errors upon the reliability of predicted fatigue lives. The tests are conducted using a new dataset with long-term load measurements that have been collected on multiple commercially operated helicopters.

The modelling framework in chapter 2 introduces and validates a new method to model the effect of in-flight loads on accumulated fatigue damage. This method consists of dividing a continuous load history into discrete segments. The effect on accumulated fatigue damage of loads occurring within each segment can be described with high accuracy and at the same time with only three parameters. Because of its verified high accuracy, this model comprises the core of the subsequent DLDM approach for VFLM.

DLDM estimates accumulated fatigue damage under the assumption that random errors from prediction errors are negligible as summarised in Figure 1.11. This approach is based on the simplified analytical method for fatigue life prediction that is analysed in chapter 2 and which is summarized in Figure 1.3. In chapter 4, simulations based on the validation framework introduced in chapter 2 are used to test the reliability of DLDM.

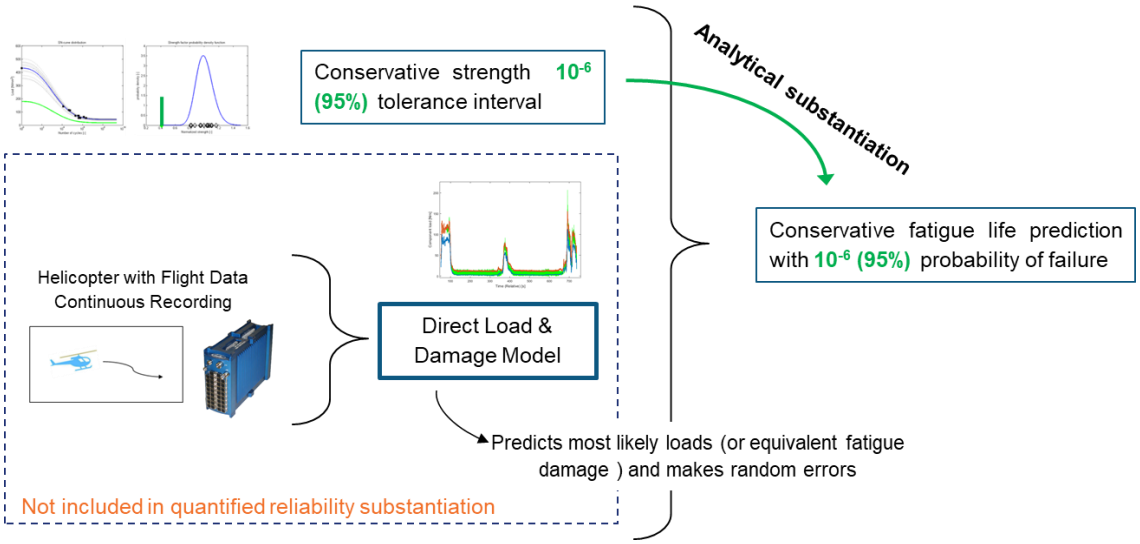


Figure 1.11: Overview how Virtual Fatigue Life Monitoring by Direct Load & Damage Modelling introduced in chapter 4.2 (and using a random model for fatigue strength in chapter 3) predicts the reliability of its estimates of accumulated fatigue damage using the reliability of the employed working S-N curve only and without accounting for random errors from its virtual Direct Load & Damage Sensor.

7. **When using uncertain estimates of in-service loading, and resulting fatigue damage accumulation, can the reliability of derived fatigue life limitations still be predicted accurately?**

Some test results in this work revealed cases for which the simplified reliability substantiation method for DLDM is not accurate enough. Therefore, another new simulation-based methodology is introduced and tested in chapter 4 in the form of Probabilistic Load & Damage Modelling (PLDM). The validity and accuracy of this simulation-based method are demonstrated in chapter 4 by means of independent testing data from two specially instrumented commercially operated helicopters. As summarised in Figure 1.12, PLDM builds on the simulation-based reliability substantiation method introduced and tested before in chapter 2 and can model and mitigate the effects of combined uncertainty from regression errors for in-flight loads as well as fatigue strength.

PLDM is introduced to estimate and mitigate the effect of DLDM-based estimation errors on resulting estimates of accumulated fatigue life. PLDM enables the deployment of VFLM for any target reliability level and regardless of achievable regression precision and accuracy. PLDM achieves this by explicitly predicting and

then mitigating the influence of expected random errors. PLDM is thus designed to automatically assess and mitigate the effects that poor accuracy and precision of predicted loads and fatigue strength can have on the reliability of predicted fatigue life.

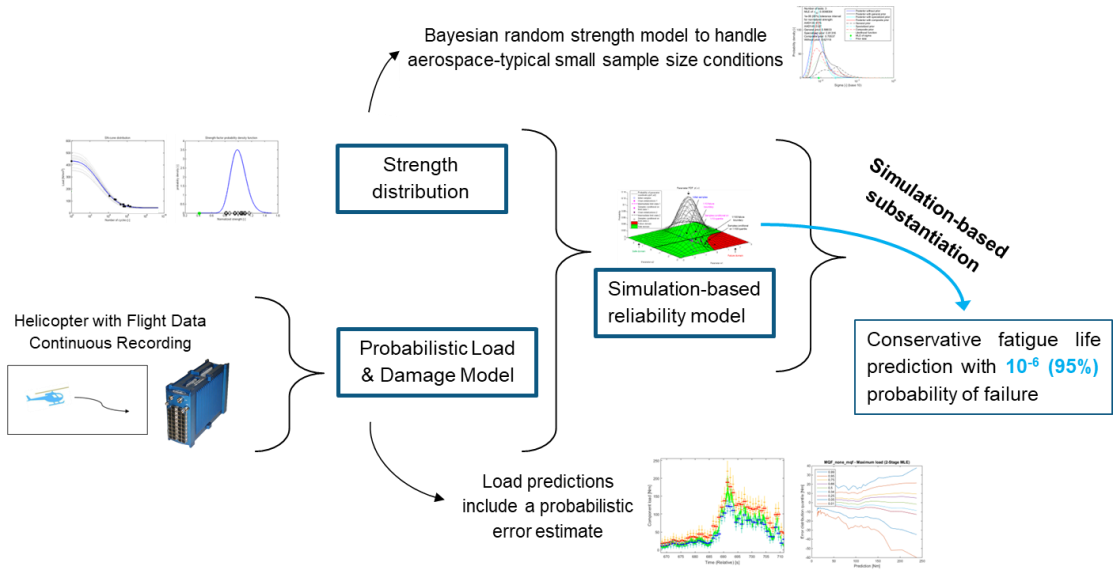


Figure 1.12: Summarising overview how Virtual Fatigue Life Monitoring by Probabilistic Load & Damage Modelling in introduced in chapter 4.3 uses the random model for fatigue strength from chapter 3, and a new statistical simulation model to substantiate the reliability of its usage-based fatigue life predictions.

8. **Can Probabilistic Load & Damage Monitoring accurately and usefully predict and substantiate component-individual and usage-based fatigue damage accumulation?**

PLDM incorporates new methods to model and mitigate the combined influence from multiple sources of uncertainty. In particular, PLDM uniquely incorporates the prediction and proportional mitigation of simultaneous uncertainty from loads and fatigue strength, estimation uncertainty for the scatter of a component's fatigue strength, as well as random in-service regression errors for in-service load and usage.

The mitigation of these sources of random error may significantly limit the gain that could be obtained by implementing Virtual Fatigue Life Monitoring. Before, rotorcraft OEMs were not required to analyse the effects of these sources of uncertainty using explicit numerical and statistical analysis. It could thus be that PLDM-based maintenance requirements end up being more demanding than the requirements imposed by classic, and authority-approved, fatigue life prediction methods. Present work, therefore, tests the economic application potential that PLDM presents by application to more than one thousand hours of recorded flight data of commercially operated in-service helicopters.

2 Reliability modelling for fatigue life prediction with assumed usage

Fatigue life is a random variable. The reliability of a fatigue life prediction for a component in the helicopter dynamic system thus needs to be substantiated. A standard analytical substantiation method simplifies manoeuvre loads to their averages instead of modelling manoeuvre loads as random variables whose distribution is estimated with limited precision. This simplification may lead to inaccuracies. A new simulation-based method is developed to improve the prediction accuracy of fatigue life by also accounting for the full random distribution and uncertainty of manoeuvre loads. Both the simulation-based and analytical methods fully account for uncertain fatigue strength but assume that the mission profile is known or can at least be conservatively estimated. A new validation process to simulate the process of fatigue life prediction and the true distribution of fatigue life was used to validate the analytical and simulation-based models. Using this validation framework, it was demonstrated that the use of a new simulation-based prediction model can improve prediction accuracy. However, it was also demonstrated that these improvements are not always significant or practically attainable.

An adapted version of this chapter was published by Dekker *et.al.* In "The Aeronautical Journal" [24].

2.1 Introduction

Failure of components in the helicopter dynamic system, such as the main rotor mast or the levers that control the angle of attack of main rotor blades, may have catastrophic consequences. For many of these components, the time between detectable crack initiation and component failure is usually too short to be covered by inspection intervals. Such components thus need to be replaced before there is a too high probability that there may be a crack that could reduce the component's static strength. Rotorcraft certification according to FAR 27.571 [2], FAR 29.571 [25], CS-27.571 [26] or CS-29.571 [27] and all by means of AC 27-1B MG11 [17] requires providing an appropriate fatigue life substantiation for each of these components. If necessary, an upper limit to the time a component can be used is set by a fixed Service Life Limit (SLL).

Fatigue life of a component can be predicted when one knows the following three elements:

- How fatigue damage accumulates, i.e. by the Palmgren-Miner linear damage accumulation hypothesis
- The component's fatigue strength, i.e. the S-N curve
- The loads during life, i.e. the load spectrum

Figure 2.1 shows a high-level process overview for common fatigue life prediction.

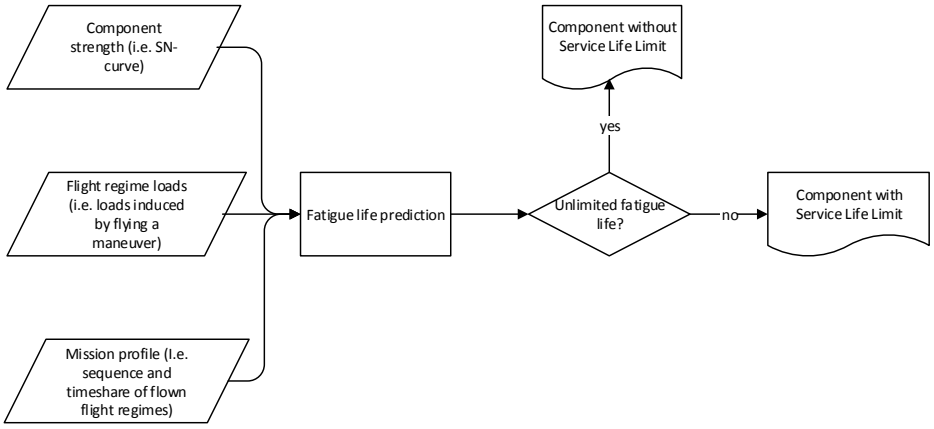


Figure 2.1: Process summary of how a classic fatigue life prediction results from an S-N curve, flight regime loads and a mission profile.

The exact fatigue strength of a specific component is never known in advance. Scatter in, for example, material properties, dimensioning, machining or other manufacturing processes demands that fatigue strength is considered as a random variable.

The loads that a component experiences during its life depend on numerous variables. Examples include the type of missions that are flown, how these missions are executed, i.e. speed, duration, number and type of manoeuvres etc., the precise technique of the pilot(s) executing the manoeuvres, and even the meteorological conditions. Therefore, the loads that occur during life can be regarded as a random variable as well. In more detail, the load spectrum that a component is subjected to during its life can be decomposed into two random variables:

- The mission profile, i.e. the sequence and timeshare of turns, hovers, landings, etc.
- The loads that occur when flying each type of manoeuvre

The models that can be used to describe how a combination of in-service loads and fatigue strength results in an expected time-to-failure can contain modelling errors too and can, therefore, add uncertainty to forecasted fatigue lives as well. However, in present work, the influence of such systematic modelling errors is neglected.

Clearly, the fatigue life of a specific component cannot be predicted exactly but must also be considered as a random variable. For certification, it is common to show that the probability of a fatigue failure during the specified maximum service life any component in the fleet is not higher than a certain probability, e.g.10^{-6}.

A common standard analytical method to predict a conservative fatigue life simplifies the full distribution of the loads during a flight regime[1] to a single averaged load spectrum and only uses the average manoeuvre minimum and maximum loads to form a low-frequency Ground-Air-Ground load spectrum. Its reliability substantiation is fully derived from the distribution of component strength. Such a method thus assumes that uncertainty in flight regime loads is negligible with respect to uncertainty in fatigue strength. The validity of this assumption is however not obvious and may not be general. For example, flight test results in Figure 2.2 clearly demonstrate significant variance in the maximum load when a lateral flight manoeuvre is repeatedly flown with a similar weight, center-of-gravity, and altitude.

This chapter, therefore, introduces a new simulation-based method to predict fatigue life while also accounting for the full random distribution and uncertainty of loads.

Both the standard analytical and the new simulation-based methods make two core assumptions:

- The mission profile is known or can at least be conservatively estimated
- All the modelling assumptions about the modelling of fatigue damage outlined in sections 2.2 and 2.3 are valid and accurate.

The two methodologies were applied to a simulated fatigue life prediction problem. Their accuracy and applicability were investigated using a controlled and synthetic reference problem simulating a realistic fatigue life prediction case.

[1] A flight regime is defined as a manoeuvre flown under specific conditions, i.e. aircraft weight, centre-of gravity and environmental conditions. The simulations in chapter 2 do however not model the difference and the terms ‘flight regime’ and ‘manoeuvre’ may thus be used interchangeably in this chapter.

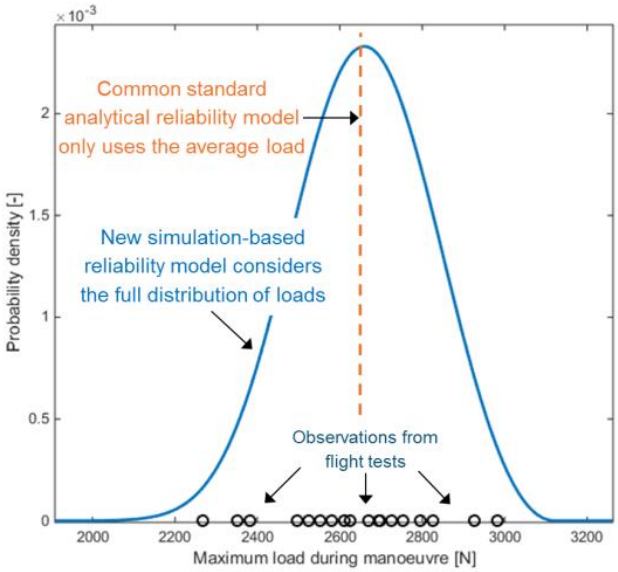


Figure 2.2: Flight test observations illustrating the distribution of the maximum load on a component in the dynamic system when executing a lateral flight to the right under similar conditions. More examples are included in Appendix F.

2.2 Fatigue life prediction by an analytical model

A baseline standard analytical fatigue life prediction methodology is outlined first. This analytical method is similar to approved and current industry practise and to chapter 4.1 in NATO AGARD-AG-292 [16] and FAA AC-27-1B MG-11 [2]. Section 2.5 later introduces a simulation-based methodology that features more complexity but aims for higher accuracy. The simulation-based methodology generally will make use of the same basic model for fatigue life prediction as outlined in sections 2.2.1 - 2.2.1.4. This basic model will be retained throughout the entire work and will thus also be used in chapters 3 and 4.

2.2.1 Definition of fatigue damage accumulation model

A fatigue damage accumulation model is needed to predict fatigue life for given component strength and loads during life. The model employed here consists of four main components: a Weibull-type S-N curve, the Goodman relation, the Palmgren-Miner linear damage accumulation hypothesis, and a specific cycle counting method.

2.2.1.1 Fatigue strength modelling by an S-N curve

A Weibull-type S-N curve defines the number of load cycles until fatigue failure under constant amplitude loading:

$$\sigma_a(N)|_R = \sigma_{a_\infty} + \frac{\sigma_{a_{ult}} - \sigma_{a_\infty}}{\exp\left\{\left(\frac{\log_{10} N}{\alpha}\right)^{\beta}\right\}} \tag{2.1}$$

where: σ_a is the applied stress amplitude (at stress ratio *R*); N is the number of load cycles (until failure); σ_{a_∞} is the stress amplitude of the endurance limit or fatigue limit (at stress ratio *R*); $\sigma_{a_{ult}}$ is the ultimate stress amplitude determined by: $\sigma_{a_{ult}} = \sigma_{ult} \cdot \frac{1-R}{2}$ where σ_{ult} is the ultimate strength; R is the stress ratio $\frac{\sigma_{\min}}{\sigma_{\max}}$; $\{\alpha_w, \beta_w\}$ are component specific Weibull curve parameters.

Alternatively, many rotorcraft manufacturers use a two-parameter exponential function to approximate an S-N function around $N = 10^5$ [28]. Although such a model is less prone to overfitting, it generally provides over-optimistic estimates for low-cycle fatigue. A four-parameter Weibull curve instead, can also accurately model low cycle fatigue. A Weibull type S-N curve is expected to provide results that are more realistic when fatigue lives are simulated for very low strengths, as is done by the simulation-based model presented in section 2.5 and during the Monte-Carlo simulations in section 2.6.

2.2.1.2 S-N curve generalisation by the Goodman relation

The Goodman-relation to translate load cycles to the stress ratio of the S-N curve is given by:

$$\sigma_a(R) = \frac{\sigma_{a_{ult}} \cdot \sigma_a|_{R_i}}{\sigma_{a_{ult}} - \sigma_m|_{R_i} \cdot \frac{1+R}{1-R}} \tag{2.2}$$

where: $\sigma_a|_{R_i}$ and . $\sigma_m|_{R_i}$ .are the stress amplitude and mean stress of the i^{th} load cycle class respectively. This relation is considered to be conservative for metallic parts, except for high-strength but low-ductility alloys [1].

2.2.1.3 Load spectrum determination by cycle counting

Rainflow counting according to ASTM E1049-85 [29], preceded by basic Peak-Valley (PV) filtering is used to determine the number of cycles in each load cycle class[2]. Rainflow counting is generally regarded as an accurate method, e.g. Schijve [1] - however, other methods for cycle counting are common in the industry as well.

2.2.1.4 Definition of damage accumulation model

The Palmgren-Miner linear damage accumulation hypothesis to define fatigue failure under spectrum loading is given by:

$$\text{Fatigue failure} \equiv \sum \frac{n_i}{N_i} = 1 \tag{2.3}$$

where: n_i is the number of load cycles in the i^{th} load cycle class; N_i is the number of cycles until fatigue failure under constant amplitude load defined by the i^{th} load class. This model is generally considered valid under conditions where loads are random and non-periodic [2]. Fatigue tests under these conditions show that a damage accumulation model such as equation (2.3) is on average accurate, e.g. as presented by Schijve [1].

2.2.2 Definition of probabilistic fatigue strength model

As fatigue strength is a random variable, both the shape and vertical translation of an S-N curve can be considered as uncertain. While neglecting shape variations, the following random fatigue strength model is used to define an S-N-P curve: (where the extension '-P' denotes that the S-N curve is randomly distributed)

$$\sigma_a(N)|_R = SF|_{\hat{\sigma}} \cdot \left\{ \hat{\sigma}_{a_\infty} + \frac{\hat{\sigma}_{a_{ult}} - \hat{\sigma}_{a_\infty}}{\exp\left\{ \left(\frac{\log_{10} N}{\hat{\alpha}_w} \right)^{\hat{\beta}_w} \right\}} \right\} \tag{2.4}$$

The strength factor *SF* herein is a random variable distributed according to a lognormal distribution (as a transformation of an associated standard normal distribution $\mathrm{N}(0,1)$):

[2] Implemented by an adapted and performance-optimized version of a software package provided by A. Nieslony [146]

$$p(SF \mid \hat{\mu}, \hat{\sigma}) = \exp\left\{\hat{\sigma} \cdot \mathrm{N}(0,1) + \hat{\mu}\right\} \tag{2.5}$$

In equation (2.4) $\left\{\hat{\sigma}_{a_{ult}}, \hat{\sigma}_{a_\infty}, \hat{\alpha}_w, \hat{\beta}_w\right\}$ are Maximum Likelihood Estimates (MLEs) of the S-N curve parameters, given component static test results and/or component constant amplitude fatigue tests. The median of the strength distribution, i.e. the distribution of SF, should have its median equal to one, i.e. $\hat{\mu} = 0$, such that the expected S-N curve remains unaltered. Nevertheless, $\hat{\mu}$ is only a sample estimate and its value can be biased and unequal to zero, i.e. offset with the true mean.

The scatter of the strength factor *SF* is assumed to be independent of N, i.e. uncertainty is assumed to be homoscedastic. Therefore, it is allowed to translate all fatigue test results used to fit the S-N curve to an arbitrary *N*. A straightforward one-dimensional distribution fit can then provide $\hat{\sigma}$, the MLE of the standard deviation of strength. The assumption of homoscedasticity is not generally applicable since scatter can positively correlate with *N*, e.g. as demonstrated by test results on 7075-T6 aluminium specimens summarised by Schijve [1]. However, this engineering assumption is acceptable to aviation authorities and is general engineering practise in the rotorcraft industry - where scatter is often estimated in the load dimension based on test results falling in the important region around $N=10^5$.

With the full S-N-P curve defined, a conservative working curve can be derived. For example, if a working curve should represent the fatigue strength of the (on average) weakest component out of one million randomly selected components, then SF_{work} can be computed according to:

$$SF_{work}(P_{fail} = 10^{-6}) = \exp\left\{\hat{\sigma} \cdot \mathrm{N}^{-1}\left(0,1,P_{fail}\right) + \hat{\mu}\right\} \tag{2.6}$$

with $\mathrm{N}^{-1}\left(0,1,P_{fail}\right)$ denoting the inverse Cumulative Distribution Function (CDF) of the normal distribution. An example of an S-N-P-curve and associated working curve is included in Figure 2.3 and Figure 2.4.

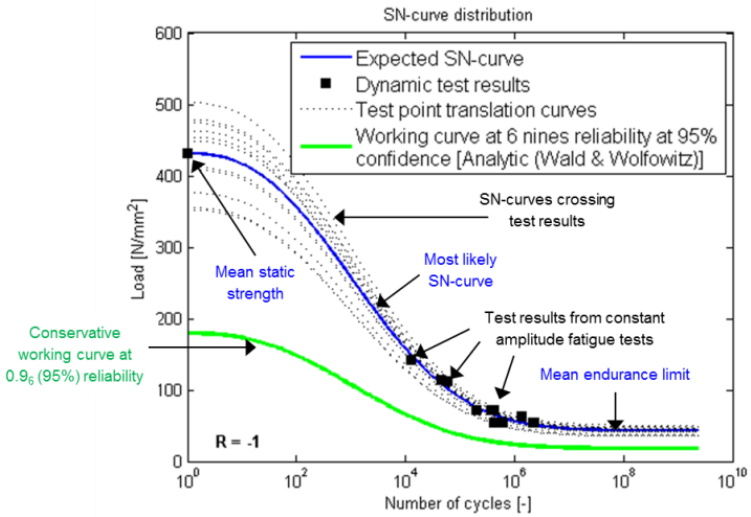


Figure 2.3: Example of constant amplitude fatigue test results for a component from the dynamic system, the resulting Maximum Likelihood estimation of the S-N curve and the associated conservative working curve with a reliability of 0.999999 (95%).

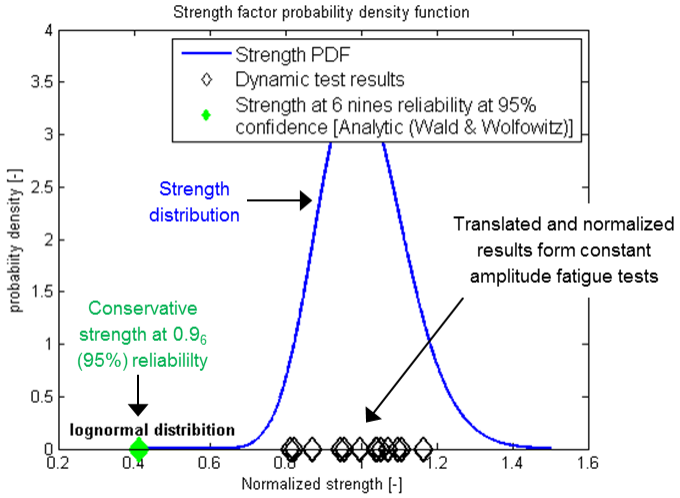


Figure 2.4: Exemplary fatigue test results (normalised by the MLE S-N curve), the derived MLE estimate of the PDF of normalised fatigue strength (*SF*), and the strength factor corresponding to the conservative working curve.

Airworthiness regulations, i.e. AC 27-1B MG11, do not explicitly prescribe the use of tolerance interval analysis for fatigue life substantiation. It is common among rotorcraft manufacturers to assume for simplification that S-N relationships and associated scatter observed from large numbers of coupon tests are sufficient to make a perfect estimate of an S-N-P diagram for a specific component. However, NATO AGARD-AG-292 [16] considers that the scatter in fatigue properties of a component mainly depends on the variability of tolerances, surface finishing and other properties affecting the component-level manufacturing quality and that these influences cannot be predicted accurately by coupon tests. Especially the scatter in S-N relationships must then be derived from fatigue tests of full-scale components representative for serial production.

Since only a limited number of such component level fatigue tests can be done, it is considered to be impossible to make a perfect estimate of the S-N-P curve, especially concerning its variability. Therefore, it is considered that any estimate of the Probability Density Function (PDF) of *SF* itself, and thereby also a conservative strength quantile SF_{work} estimated by equation (2.6), is imperfect. To account for this uncertainty,

a confidence interval for the conservative SF_{work} must be computed. To require a 95% upper single sided confidence level here means that, if a set of fatigue tests would be repeated many times, then 95% of the conservative SF_{work} estimates, one for each new set of fatigue test results, would really meet a 0.999999 reliability requirement. The remaining 5% conservative SF_{work} estimates would, in fact, correspond to a probability of failure that would be higher than 10^{-6}. Hahn & Meeker [30] or section 3.2.3 may be consulted for further explanations on confidence intervals.

Both the mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ (of the associated normal distribution) of the strength factor *SF* (2.5) must thus be considered as random variables and are distributed according to Hahn & Meeker [30]:

$$p\left(\hat{\mu} \mid \hat{\mu},\hat{\sigma},n_{test}\right) = t\left(\mu = \hat{\mu},\sigma = \frac{\hat{\sigma}}{\sqrt{n_{test}}},\nu = n_{test} - 1\right) \tag{2.7}$$

$$p\left(\hat{\sigma} \mid \hat{\sigma},n_{test}\right) \propto \hat{\sigma} \cdot \sqrt{\frac{n_{test} - 1}{\chi^2\left(\nu = n_{test} - 1\right)}} \tag{2.8}$$

where: $t\left(\mu,\sigma,\nu\right)$ denotes the Student t-distribution; $\chi^2\left(\nu\right)$ is the Chi-squared distribution; both with ν degrees of freedom; n_{test} denotes the number of test results that are available to fit the S-N-P curve.

A conservative strength factor for the working curve at a reliability level $1-\gamma$ (e.g. 1-10^{-3}) and a lower single sided confidence level α (e.g. 0.95 for 95%) can be computed according to Wald & Wolfowitz [31]:

$$SF\left(\gamma,\alpha \mid \hat{\mu},\hat{\sigma},n_{test}\right) = \exp\left\{\hat{\mu} - \sqrt{\frac{n_{test} - 1}{inv\chi^2\left(P_{fail} = 1-\alpha \mid \breve{\nu} = n_{test} - 1\right)}} \cdot r\left(\gamma,n_{test}\right) \cdot \hat{\sigma}\right\} \tag{2.9}$$

with:

$$r\left(\gamma,n_{test}\right) = \frac{1}{\sqrt{n_{test}}} - \mathrm{N}^{-1}\left(P_{fail} = \gamma \mid \mu = 0,\sigma = 1\right) \tag{2.10}$$

The validity of equations (2.7)-(2.10) is confirmed by extensive simulations in section 3.3.

2.2.3 Definition of load spectrum model

The loads during a service life are represented by a load spectrum that is cycle counted from a load sequence. Ideally, this load sequence would be the continuous load signal measured on the component during its life. In practise though, a conservatively estimated load spectrum is used instead.

The first step in obtaining this load spectrum is to define a set of manoeuvres that cover how the helicopter can be flown. For example, A: take-off; B; level flight; C: hover; etc. Using these regimes, a mission profile can be made. This mission profile sets how much time, as a percentage, the helicopter spends in each manoeuvre, e.g. [A: 3%; B: 80%; …], and in which sequence the manoeuvres are flown per unit of time, e.g. [A C B F B …] every 100 flight hours (FH).

In practise, this mission profile is generally based on pilot and operator surveys as well as experience. In any case, it must be conservative for all helicopters in the fleet for which fatigue life is predicted. The use of such a mission profile is standard industry practice, e.g. Darts & Schütz [32], FAA AC27-1B MG11 [2] or AGARD-AG-292 [16].

Test flights with a specially instrumented helicopter can in practise provide continuous recordings of component loads during the manoeuvres. The same flight regimes are generally flown multiple times, for example to cover variations in manoeuvre execution.

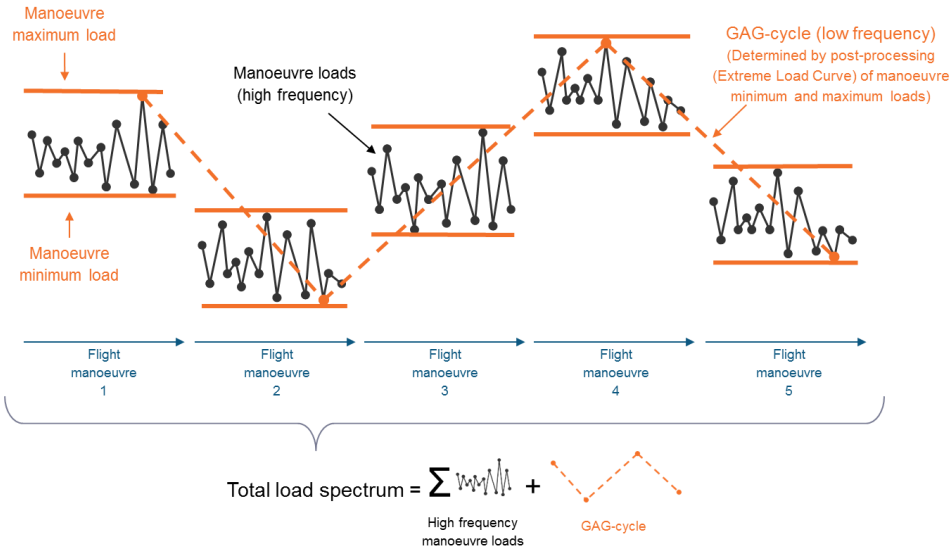


Figure 2.5: Schematic overview how high-frequency flight regime loads and Ground-Air-Ground loads together determine the full load spectrum.

The fatigue damage that is accumulated during a flight is computed with a load spectrum of the type as in Figure 2.5. The total fatigue relevant load spectrum for a flight is thus the summation of the load spectra of each flight regime and the load spectrum from the Ground-Air-Ground (GAG) load sequence. The GAG load sequence accounts for the transitions between the manoeuvres and is the most severe load signal that goes through the extreme (i.e. minimum or maximum) load in each manoeuvre.

There is uncertainty regarding manoeuvre loads and manoeuvre extreme loads when predicting the loads during the full fatigue life. In the case of manoeuvre loads, the measured load spectra, one for each time the flight regime was flown during test flights, are averaged and scaled by linear weighting to a reference time, i.e. 100FH. Extreme loads from multiple manoeuvre load tests are simply averaged. Inserting these averaged loads into the conservative mission profile and according to the model in Figure 2.5, leads to an average load spectrum per unit of time. Thus, variations in the time duration of flown manoeuvres are accounted for by normalizing obtained cycle-counted spectra to a single time and the minimum and maximum loads that occur during a manoeuvre are assumed to be uncorrelated to manoeuvre duration.

2.2.4 Reliability substantiation for fatigue life prediction

Commonly, a Service Life Limit (SLL) is set according to a maximum allowed probability of fatigue failure during the service life, e.g. P_{fail}(SLL) = 10^{-6}. However, most, general safety analysis, such as SAE ARP4761 [33], works with reliability requirements expressed as a probability of failure per flight hour and not per service life.

When it must be substantiated that the probability of failure in a next flight hour will on average never exceed a required P_{fail}, for example 10^{-9}, and when this requirement is not specified while assuming a constant failure rate, then the SLL follows from the following optimisation problem:

$$SLL = \underset{\{L \in \square : L > 0\}}{\arg\min} \left(P_{fail_{nextFH}}(L) - \gamma_{FH} \right)^2 \tag{2.11}$$

where: γ_{FH} is the maximum allowed average probability of failure per flight hour and $P_{fail_{nextFH}}(L)$ denotes the average probability of failure during the next flight hour after *L* flight hours have been accumulated. $P_{fail_{nextFH}}(L)$ can be computed using the SLL reliability estimator $P_{fail}(SLL)$:

$$P_{fail_{nextFH}}(L) = \frac{P_{fail}(L+1) - P_{fail}(L)}{1 - P_{fail}(L)} \tag{2.12}$$

For simplicity, this work will further only consider the reliability estimator $P_{fail}(SLL)$, i.e. the estimator of a probability of failure per service life.

In either case and in line with AGARD-AG-292 [16], the analytical method assumes that the reliability of a working curve only can substantiate overall reliability. E.g.. the standard analytical method substantiates an *SLL* with a probability of failure of 10^{-6}/life at a 95% single sided upper confidence level by:

- a working curve with $\gamma = 10^{-6}$ and $\alpha = 0.95$ in equation (2.9)
- a load spectrum according to a conservative mission profile and average manoeuvre (extreme) loads.

A summary overview of this exemplary analytical reliability substantiation model is shown in Figure 2.6. There is no reliability derived from the conservative mission profile. The reliability requirement must be met for all helicopters and for all flight hours. If the conservatism that is incorporated in the conservatively estimated design mission profile would be used to substantiate additional reliability, then this would only be valid for at most averagely demanding operators, i.e. this additional reliability would be valid for VIP operators but significantly less to Search & Rescue operators, as also argued by Adams & Zhao [34].

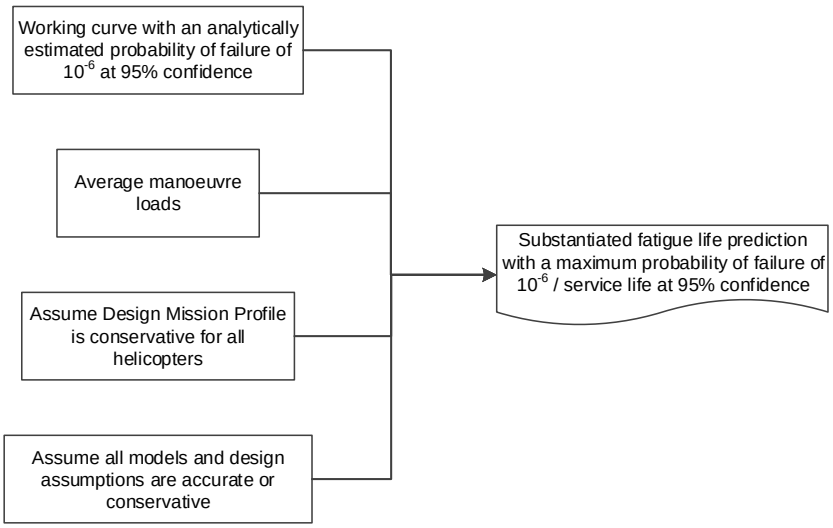


Figure 2.6: Overview how the analytical reliability substantiation model incorporates several design assumptions and uses the reliability of the working S-N curve only to numerically substantiate the reliability of fatigue life predictions.

2.3 Coverage of miscellaneous modelling assumptions

Throughout all analysis it is assumed that the outlined models for fatigue damage accumulation, random fatigue strength and loads are perfect, i.e. do not introduce any errors or additional uncertainties. Assuming the use of perfect models is in line with standard practise in rotorcraft industry and in compliance with AC 27-1B MG11. Nevertheless, different organizations generally make use of different models and design assumptions to comply with airworthiness regulations. Everett [35] observed that fatigue life predictions by different manufacturers for the same component can vary significantly. The generality of the tests and analysis

presented in section 2.6, but also in chapters 3 and 4, may therefore be limited due to the use of particular modelling assumption. Modifying or removing one or more of the adopted modelling assumptions may significantly alter the outcome of the analysis.

2.4 Overview of state-of-the-art in probabilistic fatigue life prediction

Questions have been raised during the last decades on the accuracy of the reliability substantiation in standard fatigue life predictions, for example by Lombardo & Fraser [36]. They specifically drew attention to uncertainties coming from mission profile and design load spectrum estimation but also to possible inaccuracies in standard models used to estimate fatigue damage, e.g. the Palmgren-Miner linear accumulation hypothesis. To the best of the authors' knowledge, there has so far been no systematic attempt to develop numerical error models for such standard fatigue damage models. This is also outside the scope of this analysis. The influence of uncertainties from the estimation of regime loads and design load spectra on predicted fatigue life has however been researched before, as discussed in the next sections.

Thompson & Adams [18] were one of the first in the rotorcraft industry to extensively model the reliability of SLLs. They included the combined uncertainty from variance in component strength, regime loads and mission profiles in a reliability substantiation model by using a Basic Monte Carlo (BMC) simulation and models for random strength, loads, and usage. For their random load model, the average load spectrum per manoeuvre and also the statistical distribution of manoeuvre maximum loads were computed from results of dedicated flight tests. Distributions on load and usage were set by distribution fitting and by requiring that the 0.95 and 0.99 quantiles of the usage and load distributions respectively coincided with the values used in their classic design mission profile and top-of-scatter load observations respectively. Their (random) strength model was similar to the model described in section 2.2.2. Due to the low efficiency of BMC simulation for aerospace typical low failure probabilities, it was necessary to assume that fatigue life quantiles below 10^{-3} could be estimated by using the quantile of the fatigue strength distribution only, and while effectively using average loads and usage.

This work was extended by Zhao & Adams [37, 38] who used Importance Sampling preceded by First and Second Order Reliability Modelling (FORM/SORM) to first estimate the critical failure region in the parameter space. Appendix A contains more details on these mathematical reliability modelling techniques. Their results verified the results from the initial reliability simulation model from Thompson & Adams.

Benton [39] and others [40, 41, 42, 43, 44] have all introduced (semi-)analytical fatigue life reliability substantiation models. Each of these models requires specifying a PDF for the amplitude and number of cycles of every load case to be considered. The load cases themselves are modelled as constant amplitude, single frequency, loading blocks. The model for random strength is similar as in section 2.2.2. The probabilistic modelling framework is displayed in Figure 2.7.

All previous work on reliability substantiation for fatigue life prediction confirmed the importance and value of explicit and combined modelling of uncertainty in strength, loads and usage. Thompson & Adams used their simulation model to re-confirm their standard analytical fatigue life design methodology. They argued that a combination of a conservative working curve with $\gamma = 10^{-3}$, top-of-scatter flight regime loads and a conservative design mission profile results in an overall reliability approaching $\gamma = 10^{-6}$ [45, 18]. In addition, they estimate that total prediction reliability can be modelled in a highly simplified fashion by assuming that the fatigue failure would only occur before the SLL would be reached if the following three assumptions would fail simultaneously:

- The working curve is not conservative, an event which Thompson & Adams attribute with a $\gamma = 10^{-3}$ probability

- The top-of-scatter flight regime loads are not conservative, an event which they attribute with a $\gamma = 10^{-2}$ probability
- The design mission profile is not conservative, an event, an event which they attribute with a $\gamma = 10^{-1}$ probability

However, Tong *et.al.* [19] have challenged the accuracy of the method presented by Thompson & Adams and argue that the conservative treatment of loads by Thompson & Adams does not add significant reliability to the overall fatigue life prediction. Tong *et.al.* used a modelling framework as in Figure 2.7 to indicate that the conclusions from Thompson & Adams may not generally hold.

Challenges to simplified reliability models for fatigue life predictions can also be supported by noting that fatigue life is a non-linear function of both a component's in-service loads and a component's fatigue strength. Due to the complexity of this function, the relative influence of changes in loads or fatigue strength on fatigue life is not readily known. A small deviation in fatigue strength may thus have a much larger effect on a fatigue life than a large change in loads, or vice versa.

The simulation results from both Tong *et.al.* and Thompson & Adams do however encourage the assumption that all reliability may be substantiated by a conservative working curve, an assumption that also underlies the simplified analytical method in section 2.2. Their simulation results suggest that fatigue strength is a dominating factor in the reliability model for fatigue life prediction.

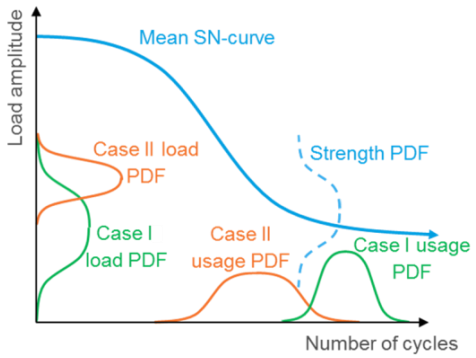


Figure 2.7: Schematic of the modelling framework that many recent (semi-)analytical SLL reliability model use. The example includes two randomly distributed load cases and a randomly distributed S-N curve, which together cause fatigue life to be randomly distributed as well.

The following challenges were identified based on previous work discussed in this section:

- Modelling of situations of complex spectrum loading, i.e. as in Figure 2.5, in the framework of current (semi-)analytical methods (i.e. as in Figure 2.7) and other previous work discussed in this section
- Modelling of tolerance intervals, i.e. providing confidence intervals on estimated quantiles, despite the high uncertainty associated with probabilistic fatigue life predictions derived from few statistical samples
- Methodology verification using end-to-end reference simulation involving the simulation of sampling uncertainty

2.5 Fatigue life prediction by a simulation-based model

A new simulation-based methodology to substantiate fatigue life predictions for critical components in the helicopter dynamic system is introduced in this section. This new method aims to meet the following main requirements:

- Modelling of the effects of combined uncertainty from loads and strength on predicted fatigue life
- Applicability to problems of very high dimension, i.e. mission profiles with many flight regimes
- Provide accurate results up to very low failure probabilities, i.e. 10^{-9}
- Provide tolerance intervals and accurately account for uncertainties from load and strength statistics estimated from small-sample sizes
- Computational efficiency allowing to run the model on a regular PC
- Make use of a generic load spectrum compatible with chapter four of AGARD-AG-292 [16].

One of the distinct features of the new simulation-based substantiation methodology compared to the analytical methodology is the addition of explicit simulation of random manoeuvre loads, also illustrated by a comparison between Figure 2.8 and Figure 2.6.

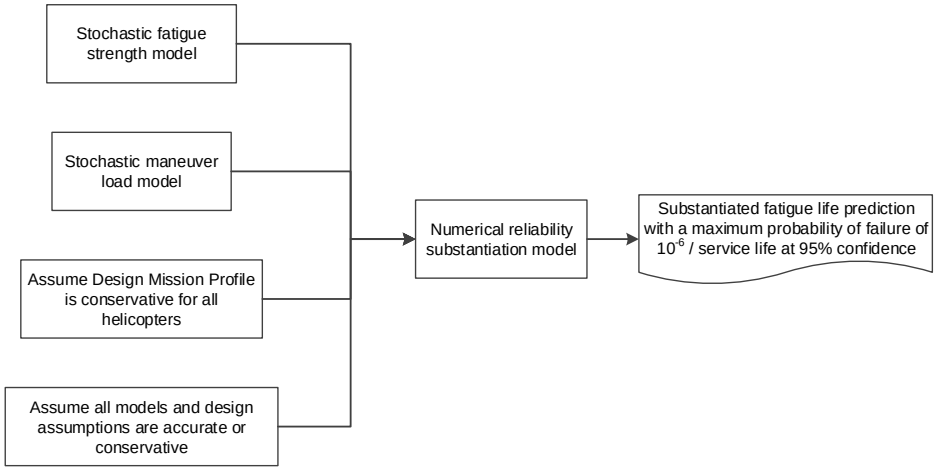


Figure 2.8: Process summary of how the simulation-based substantiation model for classic fatigue life prediction takes into account both randomly distributed fatigue strength and randomly distributed manoeuvre loads to predict the reliability of a fatigue life

2.5.1 Introduction of modelling assumptions

The following fundamental assumptions are made in the development of this model:

- The following modelling assumptions are correct: (see also section 2.3)
  - Fatigue strength modelling as described in section 2.2.1.1
  - Load spectrum modelling as described in section 2.2.1.3
  - Fatigue damage accumulation modelling as described in section 2.2.1.4
- The helicopters' mission profile is known, or can be conservatively assumed, and can be modelled as described in section 2.2.3
- Flight regime loads are independent. For example, an abnormally high load in a turn to the left is uncorrelated to the load in a next right turn

The practical implementation of the model also assumes that regime loads are identical throughout a fatigue life, e.g. all turns are flown identically. This practical assumption is expected to promote variance in lifetime, and thus to be conservative, because the influence of load scatter does not average out during life. This assumption can however easily be lifted and is not a necessary condition for the practical use of the introduced model. This is important as it may also be argued that a small change in a single load can have a

major effect on fatigue life due to the non-linearity of the S-N curve. The presence of a single high load 'outlier' may thus have a dominating effect on fatigue life. Following this argument, it is reasonable to conclude that the rate of occurrence of a high load 'outlier' may be too much restricted if loads are only sampled once for each manoeuvre type, instead of once per occurrence of the manoeuvre. For future work, it is therefore recommended to analyse the influence of assuming identical regime loads throughout a fatigue life by conducting comparative simulations. The difference between load sequences from the two modelling assumptions is illustrated in Figure 2.9.

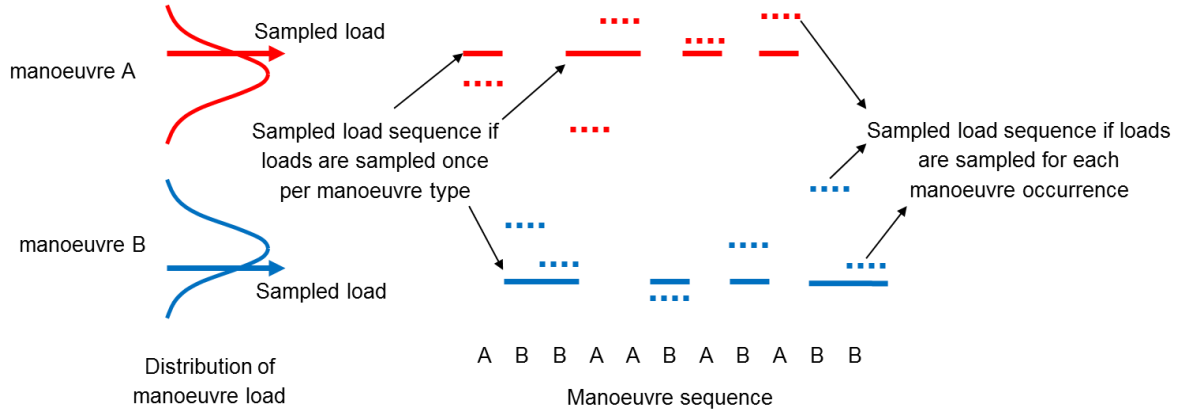


Figure 2.9: Schematic outlining the modelling difference between sampling manoeuvre loads once per manoeuvre type or once per occurrence of the manoeuvre.

2.5.2 Statistical modelling of random variables determining fatigue life

The simulation-based substantiation model features an independent probabilistic strength model and a strength-dependent combined probabilistic manoeuvre load and fatigue damage model which is similar to the model used by the virtual fatigue damage accumulation sensor from Dekker *et.al.* [20].

2.5.2.1 Definition of stochastic fatigue strength model

The implemented random fatigue strength model is equal to section 2.2.2. Note that as the proposed substantiation model is simulation-based, the new methodology may easily be adapted to accommodate other strength models.

2.5.2.2 Definition of stochastic load spectrum model

Ideally, flight regime loads can be modelled in full and with only a small number of random parameters, e.g. by means of Fourier decomposition and/or Principle Component Analysis, as for example implemented by Khibnik *et.al.* [46]. Nevertheless, experiments using flight testing data further introduced in chapter 4 indicated that, especially in complex and dynamic manoeuvres, the high-frequency content of load signals contains many load spikes that may have a significant influence on modelled fatigue life. Also, the dataset from chapter 4 often does not provide enough flight data to reliably derive the high number of model parameters that would be necessary to properly represent the high-frequency content of a load signal. As with any regression model, more detailed and accurate prediction, i.e. accurate modelling of load signals with frequency contents up to 100 Hz, requires the use of more complex regression models employing more variables that need to be set. In the case of the data available in chapter 4, precise load prediction models would be challenging to implement and likely susceptible to overspecialization of the statistical models.

Instead, after using the load model from Dekker [47] and the flight test data introduced in chapter 4 it was concluded that modelling of the fatigue damage that is equivalent to the full load signal during a flight regime is easier than attempting to model the full load signal. Distribution fits through available test flight data and large samples with synthetically generated flight manoeuvre load sequences demonstrated that, for a given S-N curve, and given that there is at least one half-cycle above the endurance limit, the fatigue damage of a flight regime follows a generalized extreme value (GEV) distribution with reasonable accuracy, as illustrated in

Appendix F. The magnitude of the minimum and maximum load that occurs within a flight regime can also be described by a generalized extreme value distribution. Again, distribution fits through large samples with synthetically generated manoeuvre load sequences, but as well as through available test flight data, are in reasonable agreement with this design choice.

Often, there is not enough test flight data available from instrumented test flights that it can reasonably be expected that these few manoeuvre trials statistically include worst-case loading conditions. The use of a load model based on fixed design spectra based on average or top-of-scatter loads may thus lead to non-conservative results. Therefore, the use of a top-of-scatter load model, as for example implemented by Thompson & Adams [18], is not followed in present work.

Based on analysis of test flight data presented in Appendix F it is chosen to model manoeuvre extreme loads by an unbounded Generalised Extreme Value distribution (GEV). Effectively, this choice accepts the existence of potentially infinitely high or low loads, however rare their occurrence might be. However, as argued in Appendix F, the load model may not require incorporating an accurate tail model. In addition, considering manoeuvre extreme load as an unbounded variable is expected to lead to conservative results.

The GEV distribution of a parameter x is defined as follows:

if $k \neq 0$ then:

$$p(x \mid k,\mu,\sigma) = \frac{1}{\sigma}\exp\left[-\left(1+k\frac{x-\mu}{\sigma}\right)^{-\frac{1}{k}}\right]\left(1+k\frac{x-\mu}{\sigma}\right)^{-1-\frac{1}{k}} \tag{2.13}$$

else:

$$p(x \mid k,\mu,\sigma) = \frac{1}{\sigma}\exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)-\frac{x-\mu}{\sigma}\right]$$

where $[k,\mu,\sigma]$ are distribution parameters: k is a shape parameter, μ is the distribution mean, and σ its standard deviation.

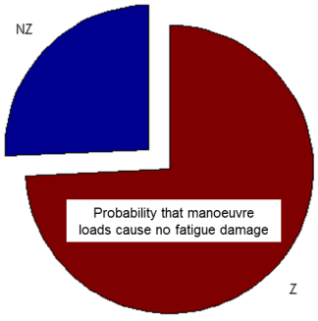


Figure 2.10: Pie chart showing an example of how probable it is that there are load cycles within a particular flight regime above the endurance limit (Z) or not (NZ).

For a given fatigue strength, a random model that implements the load model as in Figure 2.5 can now be established by defining for each manoeuvre:

- the probability that load cycles within the flight manoeuvre cause fatigue damage. This can be estimated by computing the fatigue damage for each available manoeuvre loading sample and by computing the ratio between the number of times the manoeuvre was flown with and without

causing damage. A visualisation of a resulting binomial distribution is shown in Figure 2.10. This approach circumvents a discontinuity in the manoeuvre damage distribution. Due to the endurance limit, many manoeuvre instances may not cause any manoeuvre damage at all, whereas the damage of the damaging instances is GEV distributed.

- If there is no regime damage, a multivariate Probability Density Function (PDF) for the minimum and maximum load during the manoeuvre. Such a distribution is shown in Figure 2.11.
- or, if there is manoeuvre damage, a multivariate PDF for manoeuvre damage and extreme loads. Figure 2.12 shows an example of such a distribution.

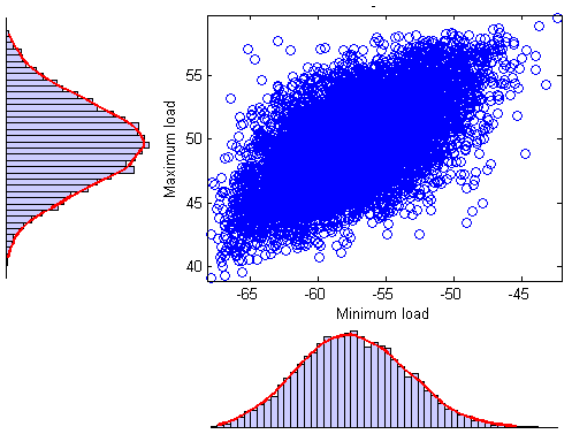


Figure 2.11: An example of a large sample from a fitted multivariate manoeuvre minimum and maximum load distribution and its corresponding marginal distributions where manoeuvre damage is zero.

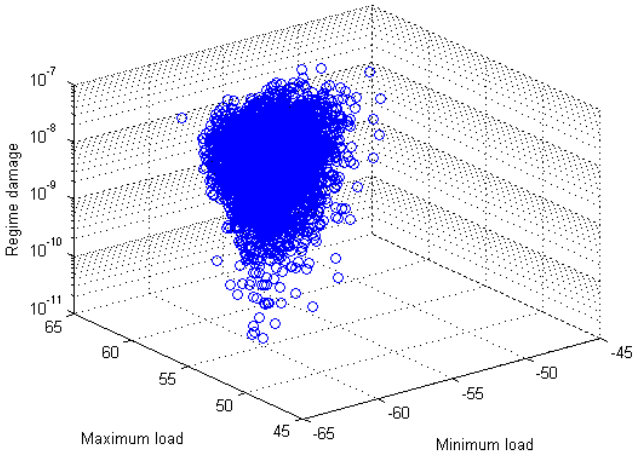


Figure 2.12: Example of a large sample from a fitted multivariate manoeuvre damage and extreme load distribution.

The multivariate distributions in the practical implementation of the model are realised by *t*-copulas [48, 49]. An alternative implementation[3] by means of NATAF transformation following Hurtado [50] resulted in non-conservatively biased and inaccurate results according to an idealised and synthetic verification test. The method of this verification test will be detailed in section 2.6.3.2. This verification test demonstrated that a true probability of failure of 10^{-3} was over-optimistically estimated as 5.9·10-4, whereas using t-copulas resulted in a virtually error-free estimate. According to Lebrun & Dutfoy [51], NATAF's limitations in modelling

[3] An adapted version of the FERUM 4.1 reliability-modelling package was used for this initial test. All subsequent results are obtained with newly developed software.

(tail) dependence of correlated multivariate distributions may provide an explanation for NATAF's inaccurate and non-conservative results.

Having defined stochastic models for fatigue strength and manoeuvre load spectra, it is possible to simulate a fatigue life distribution. This can be done by Basic Monte Carlo (BMC) simulation using the process in Figure 2.13. Although easy and transparent to implement, BMC simulation is highly inefficient and therefore only used to generate reference distributions for benchmarking purposes. The BMC simulation process of the stochastic load spectrum model is the basis of the simulation-based substantiation model that is introduced in the next sections.

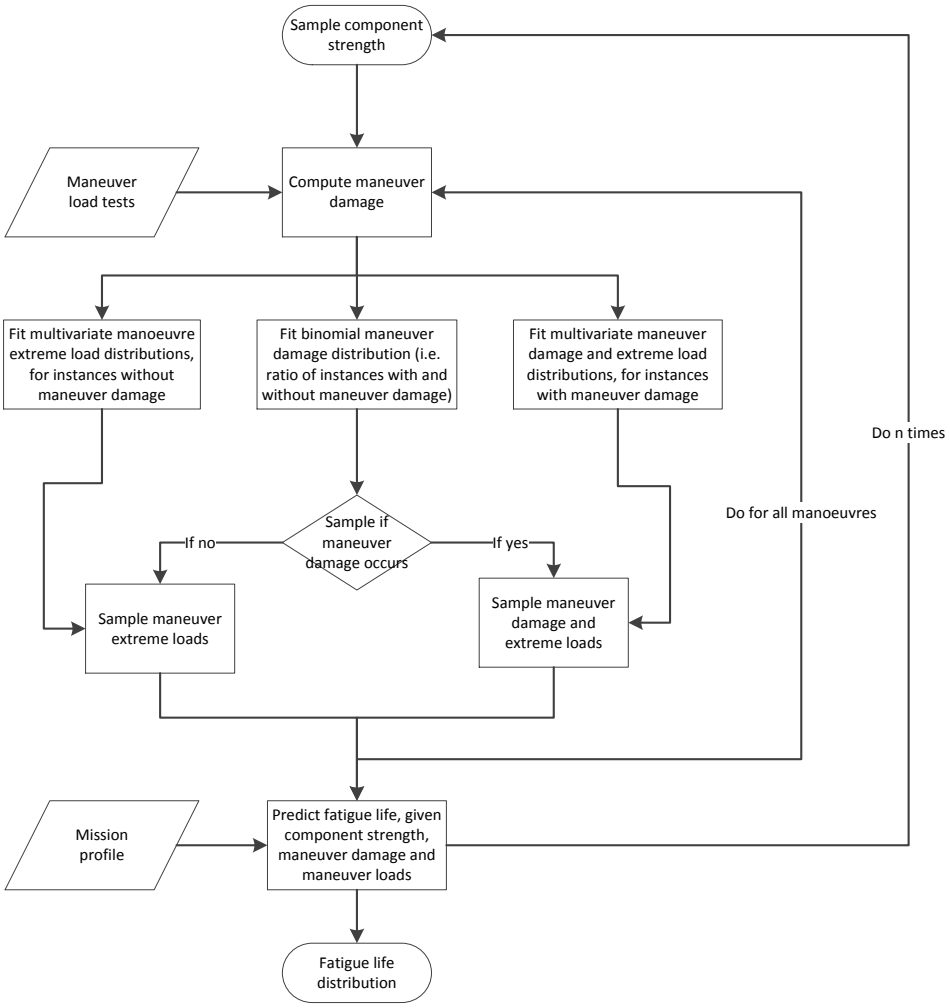


Figure 2.13: Diagram with the sampling process of the simulation-based substantiation model and how this model uses Basic Monte Carlo simulation to simulate a fatigue life distribution and to estimate its quantiles. ("n" denotes the number of BMC samples)

2.5.3 Introduction of numerical reliability estimation methods

The reliability R of a Service Life Limit SLL is one minus the probability P_{fail} that a component experiences a fatigue failure before it reaches the SLL:

$$R(\mathrm{SLL}) = 1 - P_{fail}(\mathrm{SLL}) \quad \text{with} \quad \text{failure} \equiv L < \mathrm{SLL} \tag{2.14}$$

Considering that the fatigue life L of a specific component is a function of a random parameter vector ω, containing the sampled strength factor and sampled loads and damages of the manoeuvres, the following indicator function $\mathrm{I}[...]$ can be defined:

$$\mathrm{I}\left[L(\omega)\right] = \begin{cases} 1 \text{ if } L(\omega) < \mathrm{SLL} \\ 0 \text{ otherwise} \end{cases} \tag{2.15}$$

Analytically, P_{fail} can then be computed as:

$$P_{fail}(\mathrm{SLL}) = \int_{\Omega} \mathrm{I}\left[L(\omega) \mid \mathrm{SLL}\right] \cdot p(\omega) \cdot d\omega \tag{2.16}$$

where $p(\omega)$ denotes the PDF of the of the parameter vector ω.

However, such an integral over the parameter space Ω is not expected to be mathematically tractable for the model in section 2.5.2 and numerical approximation techniques must therefore be used.

The following two sections introduce some numerical reliability methods and especially Subset Simulation, which is the prime numerical reliability tool in the simulation-based method. In addition, Appendix A gives a more elaborate review of different reliability modelling methods, including a more in-depth description of Subset Simulation.

2.5.3.1 Introduction to practical numerical reliability estimators

The most intuitive way to estimate $P_{fail}(\mathrm{SLL})$ is by a BMC estimator:

$$P_{fail}(\mathrm{SLL}) = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \left\{ \mathrm{I}\left[L(\omega_i) \mid \mathrm{SLL}\right] \right\} \quad \text{as} \quad n_{sim} \to \infty \tag{2.17}$$

which is simply drawing a large number, n_{sim}, of parameter vectors from the parameter PDF $p(\omega)$, computing the corresponding fatigue lives and then the fraction of parameter vectors that produce a fatigue life lower than the SLL.

The coefficient of variation (CoV) of a BMC estimate of P_{fail} approximately approaches:

$$CoV_{P_{fail}} = \frac{\sigma_{P_{fail}}}{\mu_{P_{fail}}} = \sqrt{\frac{1 - P_{fail}}{P_{fail} \cdot n_{sim}}} \tag{2.18}$$

Where n_{sim} denotes the number of BMC samples and $\sigma_{P_{fail}}$ and $\mu_{P_{fail}}$ are the standard deviation and mean of the n_{sim} BMC samples of the probability of failure P_{fail}.

The estimation error is thus proportional to $1/\sqrt{n_{sim}}$ and independent of the dimension of Ω. This is a highly advantageous feature as the dimension of the parameter vector according to the model in section 2.5.2 is generally high. However, using equation (2.18) it can be computed that when the precision of the estimate needs to have a CoV of 30%, then it is required to evaluate at least approximately $10/P_{fail}$ BMC samples. This means that estimating an aerospace-typical small P_{fail} becomes highly impractical due to the very large number of samples that need to be evaluated.

Traditionally, reliability problems have been solved semi-analytically by First and Second Order Reliability Methods [50]. These methods are however only accurate under strict conditions, they require transformation of the parameter space to a multivariate standard normal distribution, e.g. by transformation of the marginal distributions into Gaussians and by NATAF transformation, and their computational costs are strongly dependent on the dimension of Ω. Utilisation of FORM/SORM to handle the high-dimensional and potentially discontinuous parameter space that the model in section 2.5.2 stipulates was attempted using an adapted

version of the FERUM 4.1 reliability modelling package. However, this attempt was abandoned due to persistent accuracy and convergence problems.

Importance Sampling [50] is another common technique to improve the efficiency of the BMC estimator. However, this requires defining a special sampling distribution around the critical region, i.e. where $L(\omega) \approx \mathrm{SLL}$, which is commonly obtained following FORM/SORM solutions. Improperly setting this special sampling distribution may cause large errors in the estimate of P_{fail}. The model in section 2.5.2 dictates a high dimension and complexity of the parameter space. Setting a proper sampling distribution is thus difficult, even more so given the discouraging results from FORM/SORM for the simulation-based model. Therefore, importance sampling was not pursued as a solution method.

Other methods were studied as well, including BMC acceleration by statistically 'learned' indicator functions, e.g. by Kriging [52] or Support Vector Machines [50]) or recent Particle Algorithms [53]. However, application of these methods was considered unpractical for the particular problem at hand, mainly due to their complexity and expected difficulties due to the high dimensionality and complexity of Ω that the model in section 2.5.2 dictates.

2.5.3.2 Introduction to Subset Simulation for reliability estimation

The method of choice that is implemented to estimate P_{fail} is Subset Simulation (SS) as developed by Au & Beck [54]. The core concept is to divide a difficult problem of estimating a total probability of failure into multiple sub-problems that are by themselves easy to solve. Considering the CoV of the BMC estimator (2.18), it shows that estimating, for example, a 1/10 probability of failure can be done with reasonable accuracy while using 'only' one hundred samples, independent of the dimension of the parameter space. Subset Simulation exploits this benefit by estimating the total probability of failure by multiplication of a sequence of conditional high failure probabilities.

A set of intermediate failure events can be defined such that:

$$F_1 \supset F_2 \supset \ldots \supset F_m = F \tag{2.19}$$

This means that the failure event $F_m \equiv L < \mathrm{SLL}_m$ is a subset of the more probable intermediate failure event $F_{m-1} \equiv L < \mathrm{SLL}_{m-1}$, which is, in turn, a subset of the even more probable intermediate failure event $F_{m-2} \equiv L < \mathrm{SLL}_{m-2}$, and so forth.

The total probability of failure is now:

$$P_{fail} = P_{fail,1} \cdot \prod_{j=2}^{m} P_{fail,j}\big|_{F_{j-1}} \tag{2.20}$$

Here, $P_{fail,1}$ is the probability of the first intermediate failure event F_1. And $P_{fail,j}\big|_{F_{j-1}}$ is the probability of failure event F_j, given that the more probable failure event F_{j-1} occurs.

Computation of $P_{fail,1}$ can be done straightforwardly by a BMC estimator, especially when the first intermediate failure event F_1 is set such that $P_{fail,1}$ equals an easy to compute probability γ, i.e. 1/10. Now, a limited number of samples are drawn, i.e. one hundred, and the fatigue life is predicted for each of these samples. The intermediate failure event F_1 is then defined such that $P(\mathrm{SLL}_1 > L) = \gamma$. For example, the first intermediate limit state SLL_1, or intermediate failure boundary, an implicit hypersurface in Ω, is set such that ten out of one hundred of the initial samples lie in the first intermediate failure domain.

A similar procedure can be followed for the subsequent intermediate failure events. Again making use of a simple BMC estimator, it is now, however, necessary to generate samples that are part of the intermediate failure domain F_{j-1}. Generation of a random sample that is conditional on the domain F_{j-1} can be done with Modified Metropolis-Hastings Markov Chain Sampling, see [54] for a detailed description.

Additional intermediate failure events are added until the actual SLL for which P_{fail} needs to be known is reached. Figure 2.14 to Figure 2.16 show an example of computing $P_{fail}(\mathrm{SLL}, s_i)$ by subset simulation.

2.5.4 Numerical estimation of the reliability of an SLL

The load model from section 2.5.2.2 causes that the PDFs for regime damage and extreme load are dependent on the fatigue strength *s*, which is itself a random variable. Therefore, P_{fail} should be computed according to:

$$P_{fail}(\mathrm{SLL}) = \int p_{fail}(\mathrm{SLL}, s) \cdot p(s) \cdot ds \approx \sum_{i}^{n_{bin}} \left[P_{fail}(\mathrm{SLL}, s_i) \cdot P(s_i) \right] \tag{2.21}$$

The integral is approximated by discretizing the strength distribution into *i* intervals (bins) and while assuming that within each strength interval:

- Regime damage is constant and according to the lowest strength value in the interval
- Correlations between regime extreme loads (and regime damage) are invariant

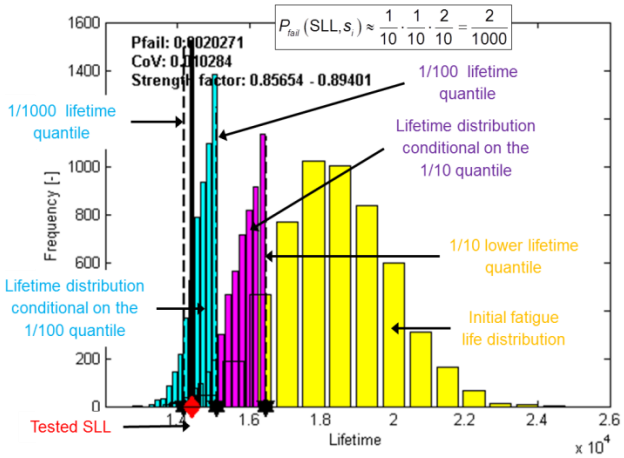


Figure 2.14: Example of Subset Simulation where it takes three intermediate failure events (black stars) to reach the SLL under evaluation (red diamond). The initial lifetime sample is in yellow, the lifetime distribution conditional on F_1 is purple and the lifetime distribution conditional on F_2 is light blue. $P_{fail}(\mathrm{SLL}, s_i) \approx 0.1 \cdot 0.1 \cdot 0.2 = 0.002$

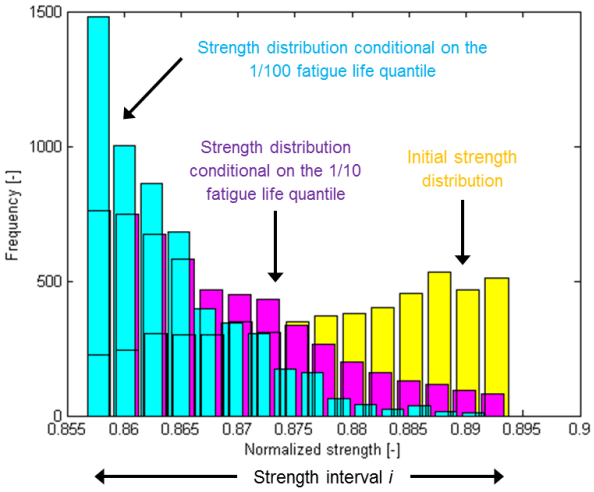


Figure 2.15: Distribution of strength samples from Subset Simulation for the example in Figure 2.14. The example illustrates that sampled strength generally decreases as the intermediate failure events become less probable.

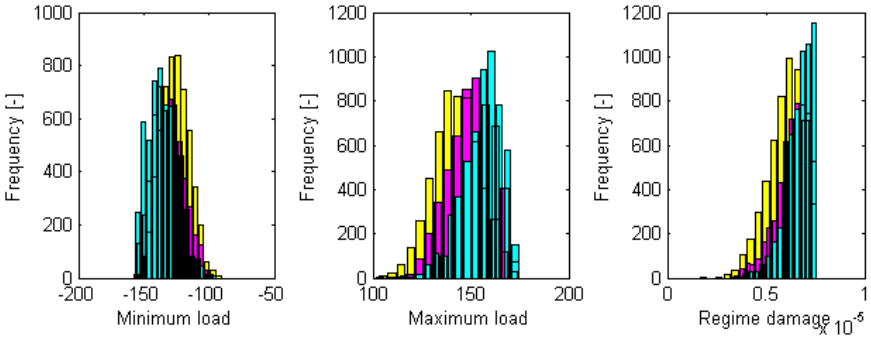


Figure 2.16: SS Distribution of samples of the minimum load, maximum load and regime damage of a flight regime for the example in Figure 2.14. The example illustrates that the maximum load (in the middle graph) generally increases with less likely intermediate failure events.

The parameter PDFs are now fixed for each strength interval. The strength PDF in one such interval is exemplified in Figure 2.17. Note that in general, the coarser the strength discretization grid, the more conservative the estimates of P_{fail}, as regime damage is consistently overestimated. This was confirmed by simulations under both ideal and small sample size conditions. High imprecision may arise though if too few samples per subset are used in combination with a very coarse strength grid.

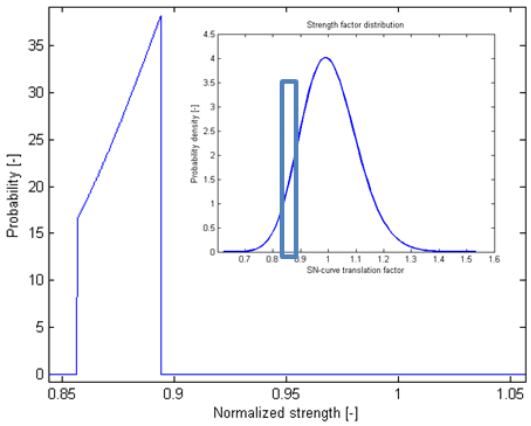


Figure 2.17: Example of a strength PDF that is conditional on a strength interval in the upper right thick blue box.

2.5.5 Introduction of confidence level analysis for SLL reliability estimations

In practise, the number of fatigue tests and flight tests that can be done is limited. Also, computational resources are generally limited so that the sample sizes used in Subset Simulation must be limited. This means that both the parameter distributions themselves, as well as computational results from the quantile estimator (2.20), are actually subject to significant uncertainty. It is assumed that other sources of uncertainty (i.e. establishing of the copulas) can be neglected or are conservatively hedged (see Appendix G for implementation details).

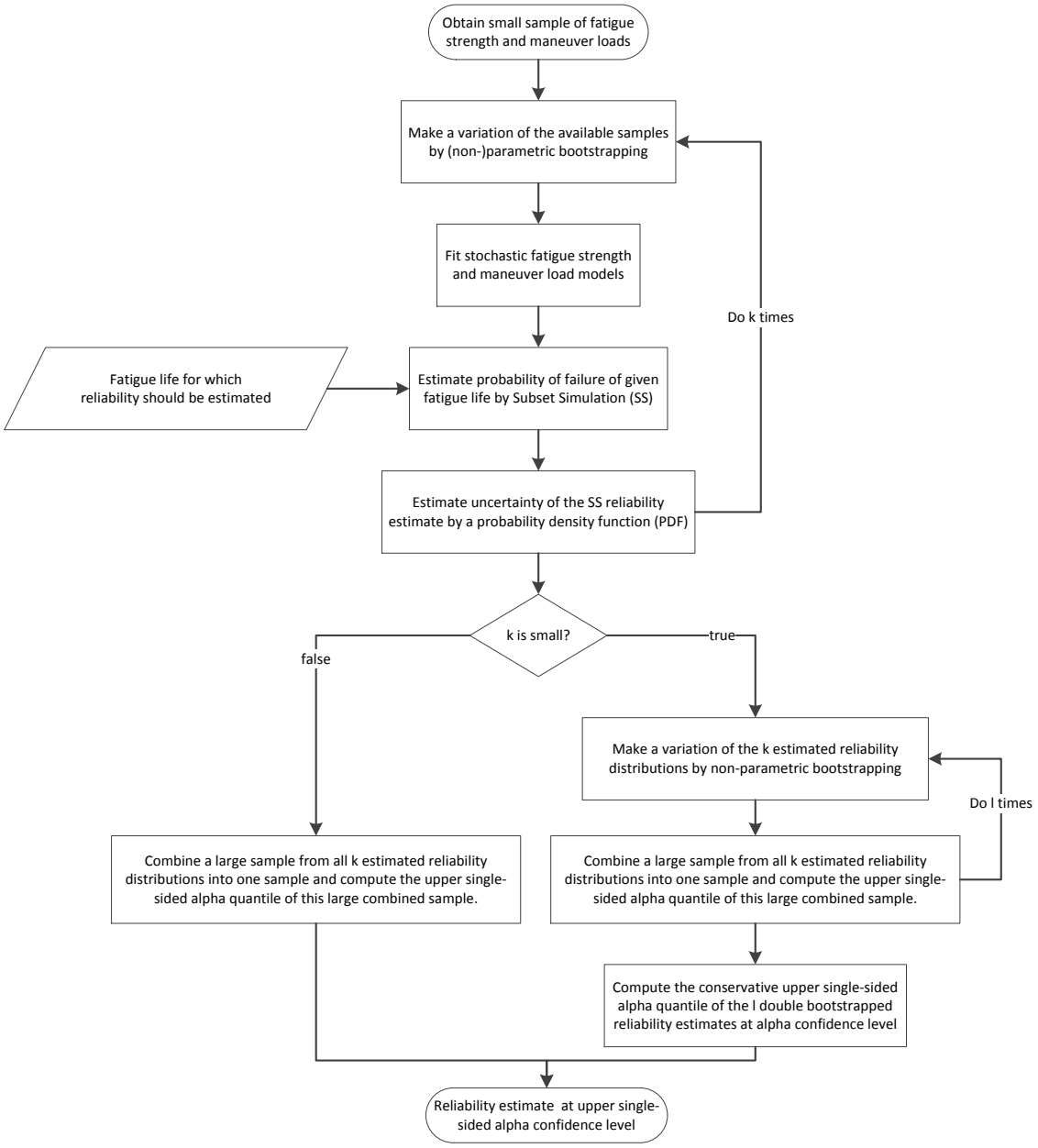


Figure 2.18: Process overview how the simulation-based method estimates the reliability of an SLL under small sample size conditions, i.e. at a α level of confidence.

Confidence intervals on P_{fail} are computed by parametric and non-parametric bootstrapping [55]. Essentially, this means that P_{fail} is computed for several alternative variants of the strength, regime extreme load, and regime damage distributions, and for several alternative SS estimates. Thus, a distribution for P_{fail} can be estimated and, for example, the upper 95th percentile of P_{fail} can be selected for an upper single sided 95% confidence interval. A corresponding process overview is given in Figure 2.18 and an application example is shown in Figure 2.19.

Au and Beck [54] provide an algorithm to estimate the coefficient of variation $CoV_{P_{fail},i}$ for $P_{fail}\left(\mathrm{SLL},s_i\right)$ in equation (2.21), while assuming that $P_{fail}\left(\mathrm{SLL},s_i\right)$ is normally distributed. The standard deviation of P_{fail} can then be estimated as:

$$\hat{\sigma}_{P_{fail}} = \sqrt{\sum_{i}^{n_{bin}} \left[CoV_{P_{fail},i} \cdot P_{fail}\left(\text{SLL}, s_i\right) \cdot P\left(s_i\right) \right]^2} \tag{2.22}$$

This feature is important as it allows to use small sample sizes during SS and to keep computational costs low while still being able to ensure conservatism.

Alternative regime loads are determined by non-parametric bootstrapping (i.e. random 'reshuffling' with allowing duplicates) of the available manoeuvre load test results. Note that standard literature indicates that non-parametric bootstrapping is inaccurate and generally not conservative for small sample sizes. Simulations in section 3.3 also confirm this. Nevertheless, it is assumed that this inaccuracy is negligible, i.e. small in comparison to variance due to parametric bootstrapping of the estimated strength distribution. Previous sensitivity studies, e.g. by Zhao & Adams [38], show that fatigue strength is significantly more influential than manoeuvre loads in fatigue life prediction and thereby support this assumption. The following simulation results in section 2.6 confirm this as well.

Alternative strength factor distributions are simply drawn from the parameter PDFs (2.7) and (2.8). This method of parametric bootstrapping was confirmed to be accurate by means of extensive simulations, see section 3.3.

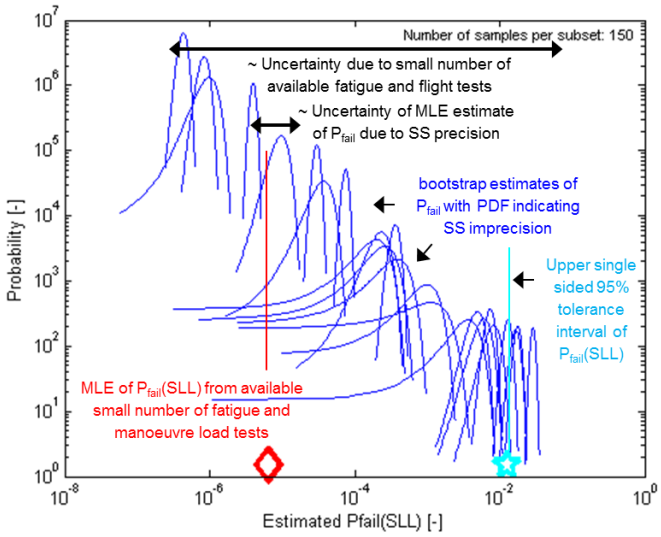


Figure 2.19: Example of the PDFs of bootstrap estimates of P_{fail}(SLL). The width of a PDF represents uncertainty due to limited SS accuracy and the variance in the mean of the different PDFs represents uncertainty due to a low number of fatigue- and manoeuvre load tests. The example demonstrates that imprecision from SS is small with respect to uncertainty due to a low number of fatigue and manoeuvre load tests. The result is obtained for seven available fatigue tests and fifteen instances per manoeuvre.

2.5.6 Introduction of Reliability Based Design Optimisation

The practical engineering problem is often not to predict P_{fail} of a given lifetime but rather to predict a lifetime that meets a reliability requirement (i.e. 1-10^{-3}). Hence, a custom Reliability Based Design Optimisation (RBDO) application was developed that uses the simulation-based lifetime substantiation model to 'design' lifetimes that meet a reliability requirement.

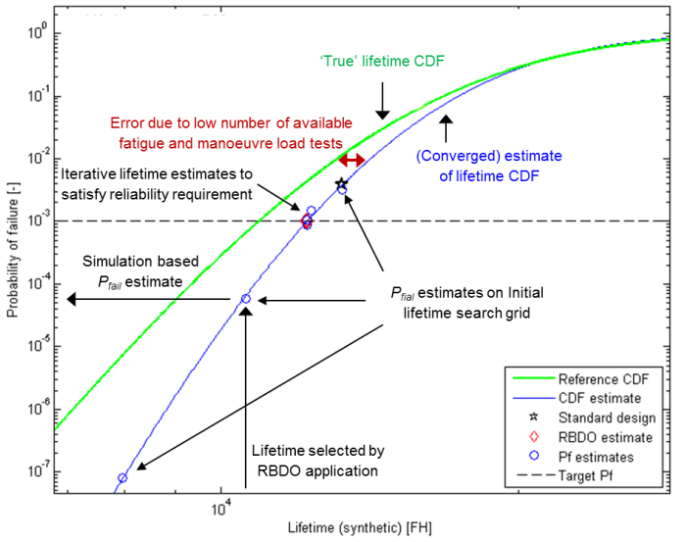


Figure 2.20: Illustrative result from a custom developed RBDO application to predict fatigue life using the simulation-based fatigue life substantiation model. The example illustrates the high precision of SS in the newly proposed method by the small scatter of P_{fail} estimates around the same lifetime, (The example is generated with 150 samples per subset).

A process overview of the RBDO application is included in Figure 2.21, with Figure 2.20 showing an illustrative result. The RBDO application starts making an estimate of the required fatigue life quantile using the analytical method. Then it uses the simulation-based model to compute the probability of failure of several lifetimes in the neighbourhood of the initial quantile estimate. As fatigue life is observed to follow a Generalised Extreme Value distribution, a GEV CDF can then be fitted through the initial simulation-based estimates.

$$P_{fail}\left(L \mid k,\mu,\sigma\right) = \exp\left\{-\left[1+\left(\frac{L-\mu}{\sigma}\right)\cdot k\right]^{-\frac{1}{k}}\right\} \tag{2.23}$$

where $[k,\mu,\sigma]$ are distribution parameters and where L represents fatigue life.

An updated estimate of the lifetime that satisfies the reliability requirement can then be made using this fitted CDF. If the updated quantile estimate is within 75FH of the previous estimate, the quantile 'design' is considered complete. (i.e. smaller differences in fatigue life are not considered relevant or significant) Otherwise, the probability of failure for this new lifetime is checked by the simulation-based model and the design curve (i.e. GEV CDF fit) is fitted again to make a new quantile estimate.

In practise, in can occur that the GEV CDF cannot be fitted successfully. It is suspected that this is mainly caused by improper selection of the lifetime points taken into consideration for the fitting of the CDF (i.e. only lifetimes near the 'design' quantile and not at higher quantiles) and too low precision of simulation-based reliability estimates. To solve this problem, each time an unsuccessful GEV CDF fit is detected, the average of at most four lifetimes surrounding the target quantile is used to update the quantile estimate.

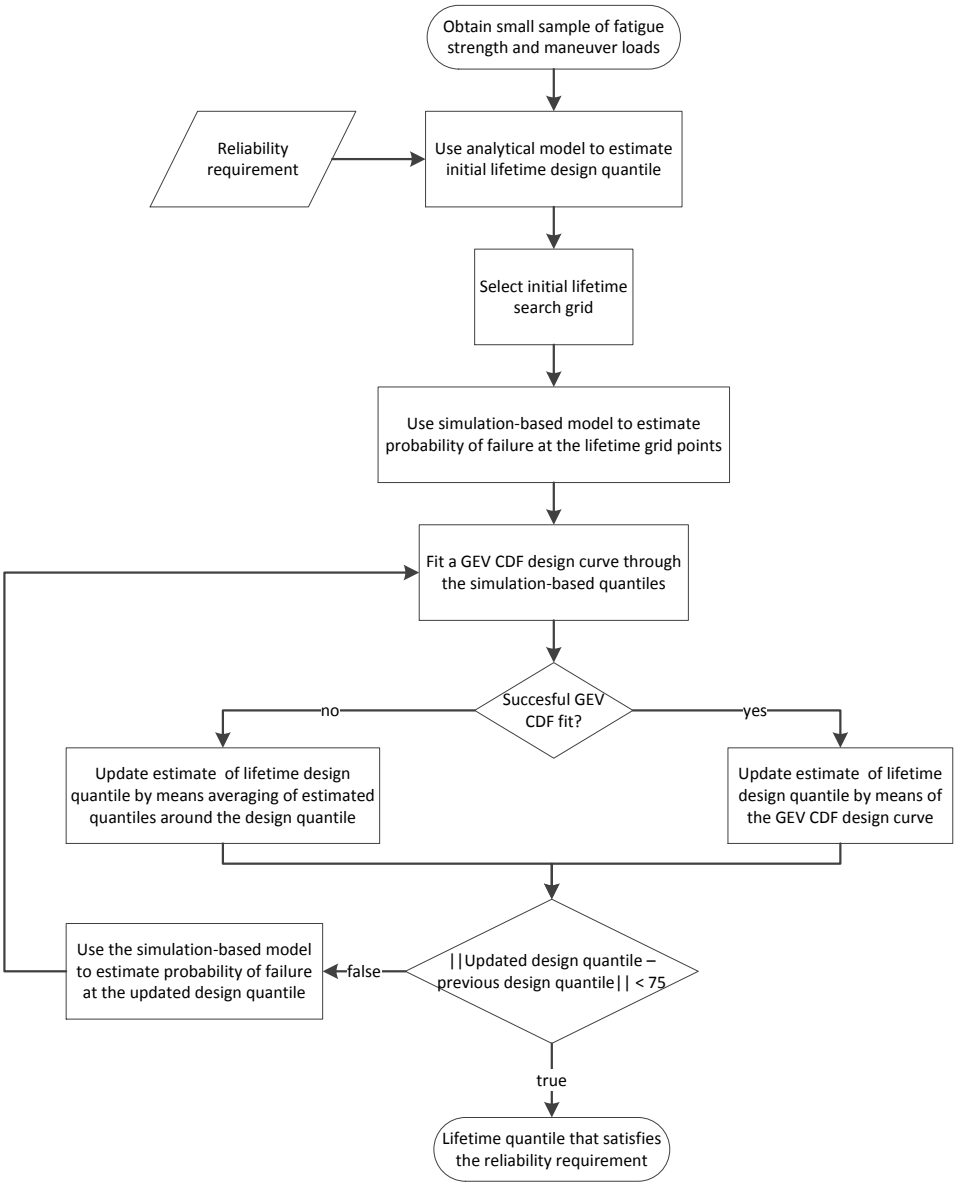


Figure 2.21: Process summary how the simulation-based RBDO application to search for a lifetime quantile satisfies a reliability requirement.

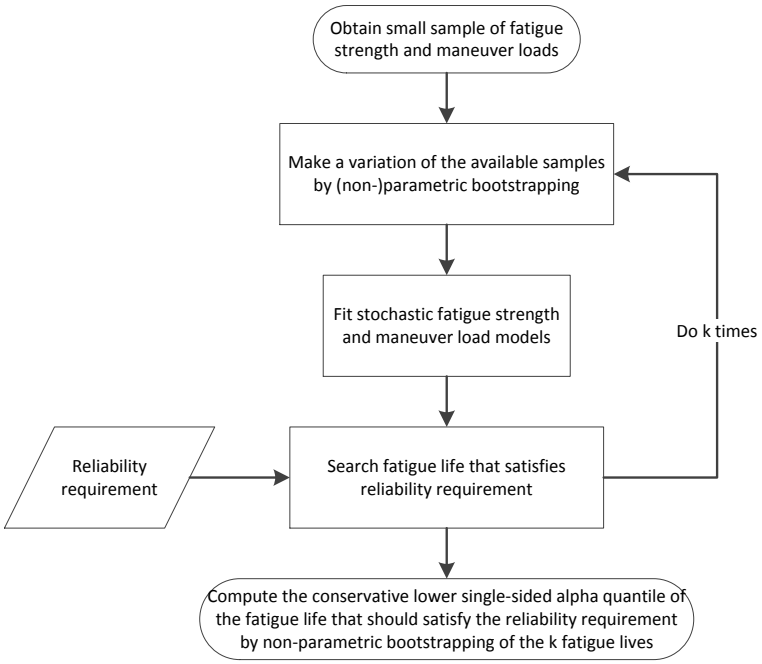


Figure 2.22: Process summary for estimating a fatigue life that satisfies a given reliability requirement at a required level of confidence. In practise, the number of bootstraps k must be kept small (e.g. 20) due to the high computational costs of searching for the required fatigue life quantile. Therefore, overview outlines how the k lifetime estimates are bootstrapped themselves again, i.e. a double-bootstrap is applied, to hedge the probability that the required confidence level is not met due to an insufficient number k bootstraps.

Confidence intervals on the quantile estimate by the simulation-based RBDO application are computed by repeated (non-)parametric bootstrapping, as by the process summarised in Figure 2.22. The single bootstrapping loop over the simulation-based RBDO application accounts for imprecisions due to a small number of fatigue strength and manoeuvre load tests as well as imprecision of the simulation-based model itself.

2.6 Testing of numerical reliability substantiation models for fatigue life prediction

2.6.1 Introduction to testing strategy

Straightforward validation of the analytical and simulation-based fatigue life prediction methods on a real fatigue life prediction case is fundamentally impossible due to the extremely large sample sizes that would be required. To validate the accuracy and precision of a fatigue life prediction that should meet a 0.999999 reliability requirement using model-free statistics, e.g. Monte Carlo simulation, it would be necessary to test if indeed one out of 10^6 components would experience a fatigue failure before reaching its SLL In order to also validate the accuracy and precision of the confidence level prediction, it would even be necessary to repeat this test at least 100 times to see if in 5 out of 100 tests more than 1 out of 10^6 components would fail in fatigue before reaching their SLL. Such a test procedure cannot be carried out in practise, and can also not be performed in retrospect based on in-service failure data.

Therefore, the analytical and simulation-based fatigue life prediction substantiation models are both tested using a synthetic reference problem for which the 'true' fatigue life distribution can be simulated. The synthetic reference case that is used is defined in section 2.6.2. This reference case is designed to be realistic but is not specific for any particular helicopter component.

The synthetic reference problem defines a 'true' and known random population for fatigue strength and flight manoeuvre loads, as well as a known usage profile. It is thus possible to perform a virtually infinite number of flight- and fatigue tests. For such a synthetically generated fatigue life prediction problem, it is possible to simulate a virtually infinite number of component fatigue lives and to very accurately simulate the 'true' distribution of fatigue life by simple BMC simulation. Figure 2.29 shows an example of a simulated reference fatigue life distribution. In addition, it is possible to simulate that only a limited number manoeuvre load flight tests and component fatigue strength tests are available to predict a quantile of the fatigue life distribution, as is generally the case in practise. Predicted quantiles from the standard analytical and simulation-based fatigue prediction methods can then be compared with 'true' reference quantiles, see also Figure 2.23.

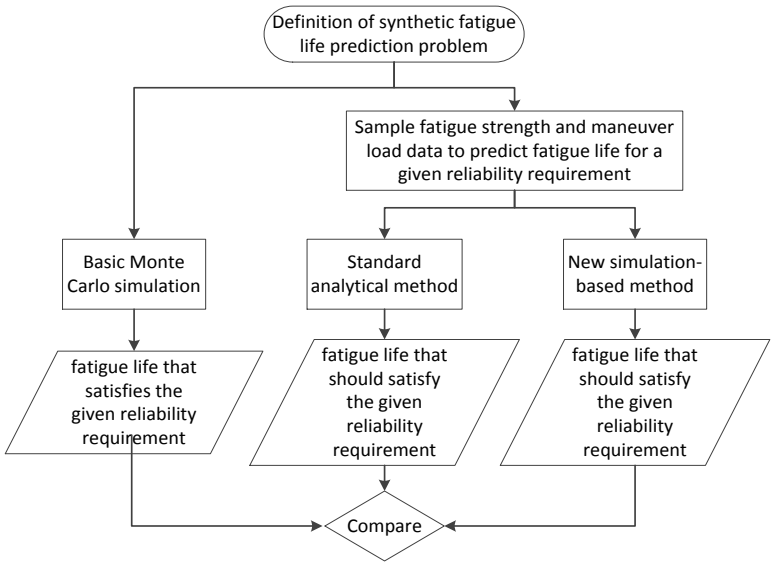


Figure 2.23: Overview of the validation procedure to test the reliability of the analytical and simulation-based fatigue life prediction methods and how the procedure generates and uses a reference distribution of fatigue life for benchmarking.

First, it was tested if the analytical and simulation-based methods contain any modelling errors or simplifications that could cause systematic inaccuracies. For this test, the 'true' distributions for fatigue strength and manoeuvre loads are known with very high accuracy and available to the two prediction methods. This test case is discussed in section 2.6.3 and reveals that the analytical model makes use of modelling simplifications that could lead to systematic prediction errors if the 'true' populations of random fatigue strength and manoeuvre loads would be known with full accuracy. The simulation-based case is demonstrated to improve on the analytical method by being capable to provide fully accurate predictions.

Second, the accuracy and precision of the two methods were tested for cases where the 'true' population of random fatigue strength and manoeuvre loads was not known and had to be estimated from a limited number of samples. The corresponding test results in section 2.6.4 demonstrated that the potential accuracy and precision improvements of the simulation-based method over the analytical method are not always significant in practise. It was demonstrated that the modelling simplifications of the analytical model do not always cause significant errors in comparison to overall estimation uncertainty. The modelling errors can be insignificant in comparison to prediction errors coming from estimation errors for the distributions of fatigue strength and manoeuvre loads.

The test results in section 2.6.4 also compare the difference between quantile predictions made with 95% confidence and without a confidence level. This comparison explicitly and numerically demonstrates a large difference in predicted fatigue life quantiles between these two cases.

2.6.2 Definition of a synthetic reference problem for reliability testing

The synthetic reference problem uses a model to simulate a random distribution of fatigue strength. The definition of this model in the form of an S-N-P curve is given in Figure 2.24. The standard deviation of the strength factor in this definition is set to a realistically low value to maximise the relative influence of variance in loads on fatigue life. This is important as the simulation-based model is meant to improve accuracy by explicitly accounting for the influence of uncertainty in loads on fatigue life.

Random synthetic flight regimes are used to do 'virtual manoeuvre load testing' and to sample in-service manoeuvre loads. Flight manoeuvre loads are composed by a Fourier series that forms a random load signal for the i^{th} synthetic regime of the i^{th} virtual manoeuvre load test:

$$[\text{Load signal}]_i = \sum_{n=1}^{k} a_{i,n} \sin(f_{i,n} \cdot t + \phi_{i,n}) + m_{i,n} \tag{2.24}$$

where t is a synthetic time vector discretizing the domain $[0,2\pi]$ into 150 points and where $\{a,f,\phi,m\}$ are randomly drawn load signal parameters defining an ordinary Fourier series.

For each manoeuvre i, random manoeuvre type parameters set a multivariate distribution from which the load signal parameters are drawn. $K = 5$ signal parameters are randomly drawn from the distributions that these random manoeuvre type parameters define, each time a virtual manoeuvre load test is performed:

$$[a_i, f_i, \phi_i, m_i] = \mathrm{N}\left(\left[\mu_{a,i}, \mu_{f,i}, \mu_{\phi,i}, \mu_{m,i}\right], \left[\sigma_{a,i}, \sigma_{f,i}, \sigma_{\phi,i}, \sigma_{m,i}\right]\right) \tag{2.25}$$

To define the virtual flight manoeuvres, the manoeuvre 'type' parameters for $i = 15$ different manoeuvres are randomly drawn from the following uniform and/or normal distributions:

$$\begin{aligned} \mu_m &= U[-10,10] \cdot 2.7 & \sigma_m &= \mathrm{N}(0,1) \cdot 1.4 \\ \mu_f &= \mathrm{N}(0,1) \cdot 45 & \sigma_f &= \mathrm{N}(0,1) \cdot 1.5 \\ \mu_a &= \mathrm{N}(0,1) \cdot 45 & \sigma_a &= \mathrm{N}(0,1) \cdot 0.4 \\ \mu_\phi &= \mathrm{N}(0,1) \cdot 0.2 \cdot \pi & \sigma_\phi &= \mathrm{N}(0,1) \cdot 1.2 \end{aligned} \tag{2.26}$$

where the scaling factors were set by tuning of the synthetic reference problem such that it is representative and realistic, as determined by manual and heuristic comparison with a dataset from industry. Changing the parameters in (2.26) can be used to change the nature of the load spectra occurring in the synthetic reference problem.

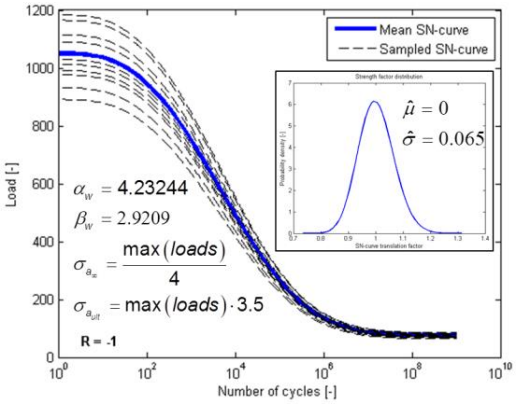


Figure 2.24: Definition the S-N-P curve in the reference problem making use of equations (2.1) and (2.5). *"Loads"* refers to all sampled load signals, as in Figure 2.25.

Some load signals generated by the random flight regime model are shown in Figure 2.25. Corresponding distributions for regime minimum and maximum load are given in Figure 2.26. Figure 2.27 then shows corresponding regime damage distributions, computed with strength factors according to the distribution defined in Figure 2.24. (More examples are presented in Appendix F.2)

The mission profile is randomly defined by drawing a random sequence of 150 flight regimes and setting the regime timeshare proportional to the number of occurrences of the regime in the random sequence. Figure 2.28 shows an example of a drawn sequence of manoeuvre extreme loads.

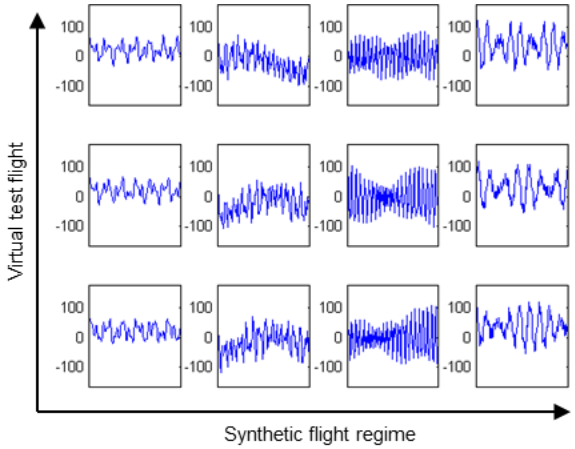


Figure 2.25: Example of artificially generated test flight data. and how there is similarity between samples for the same flight regime and distinction between different flight regimes.

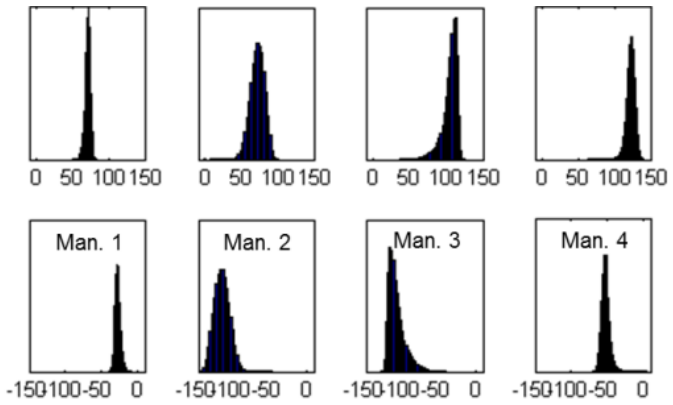


Figure 2.26: Example of marginal distributions for flight regime maximum (above) and minimum (below) loads that are generated to form a reference distribution. ("Man." abbreviates manoeuvre)

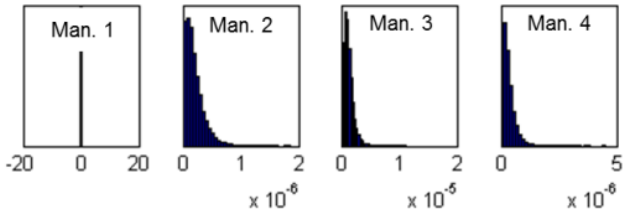


Figure 2.27: Example of marginal distributions for flight regime damage that are generated to form a reference distribution. ("Man." abbreviates manoeuvre)

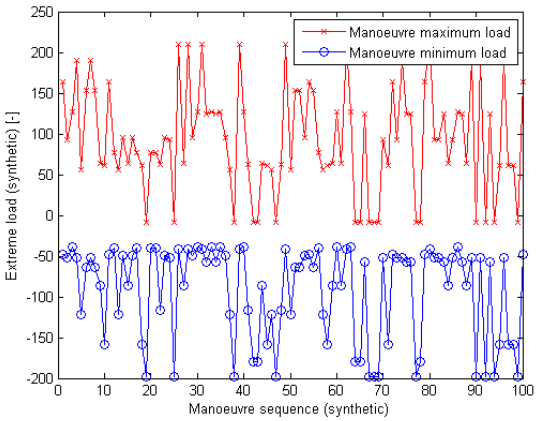


Figure 2.28: Example of sampled GAG extreme manoeuvre loads before extreme load and Peak Valley filtering.

The configuration of the reference problem makes that approximately 5% of the fatigue life is determined by manoeuvre damage. This relatively low percentage is representative of many components in the dynamic system. However, this percentage can be changed by adjusting the random load parameters in equation (2.26) and the parameters of the S-N-P curve, as in Figure 2.24. More influence of manoeuvre damage on fatigue life enables more stringent testing of the manoeuvre damage model in the stochastic load spectrum model, as introduced in Section 2.5.2.2.

All the reference distributions that are used for validation contain 10^5 samples. The CoV of the 'true' P_{fail} of the 'true' 10^{-3} lifetime quantile is then 10%, according to equation (2.18). This means that it is roughly 99.7% certain that the P_{fail} of a 'true' 10^{-3} lifetime quantile is actually between $1.3 \cdot 10^{-3}$ and $0.7 \cdot 10^{-3}$. This imprecision must be considered when regarding observed estimation errors of the models.

The 10^{-3} quantile of the 'true' lifetime distribution can thus be estimated with high precision by BMC simulation and without making any assumptions about the distribution of fatigue life. However, tests for estimating a more realistic 10^{-6} quantile can only be conducted when the distribution of the 'true' reference sample of size 10^5 is extrapolated. To do this, it is assumed that fatigue life follows a GEV distribution. Although GEV distribution models generally matched simulated lifetime distributions very well, few cases have been observed where the fit appeared to model the lower tail too conservatively, potentially leading to the presentation of (slightly) over-conservative test results in present work. The use of dedicated tail modelling may remediate this inaccuracy and is recommended for future work.

2.6.3 Reliability testing under idealised circumstances

First, the ideal performance of the standard analytical (section 2.2) and new simulation-based (section 2.5) fatigue life substantiation models are tested to see if these models are methodologically correct. The distributions for fatigue strength and manoeuvre loads are considered as known from $5 \cdot 10^5$ fatigue tests and 10^4 flight tests[4]. Hence, if a model makes wrong estimates, then this must be due to fundamental shortcomings in the model itself, as there is practically no uncertainty in the fitted strength and load distributions that serve as input to the models.

The standard analytical method is expected to non-conservatively overestimate reliability as this model only computes with the average (extreme) loads and neglects effects of their variance. The simulation-based model, in contrast, aims to fully simulate the effects of random loads and should, therefore, make an accurate reliability estimate.

2.6.3.1 Reliability testing of standard analytical method for fatigue life prediction

The standard method is tested by using the 'true' lifetime distribution to compute the actual P_{fail} of the 10^{-3} lifetime quantile predicted by the standard method. As in Figure 2.29, the actual P_{fail} is about $7 \cdot 10^{-3}$, i.e. the failure probability of the predicted lifetime is about seven times higher than the target of 10^{-3}. A repetition of the test while instead targeting a more realistic and challenging 10^{-6} lifetime quantile also revealed that the estimated quantile actually corresponded to a 'true' fatigue life quantile of about $5.9 \cdot 10^{-5}$, i.e. was also biased non-conservatively. These results thus indicate that the modelling simplifications of the standard reliability substantiation model can cause non-conservative prediction inaccuracies. The cause is that the standard method only computes with the average (extreme) loads and neglects effects of their variance.

[4] These sample sizes followed from limitations in memory capacity of the computational resources used to conduct presented work.

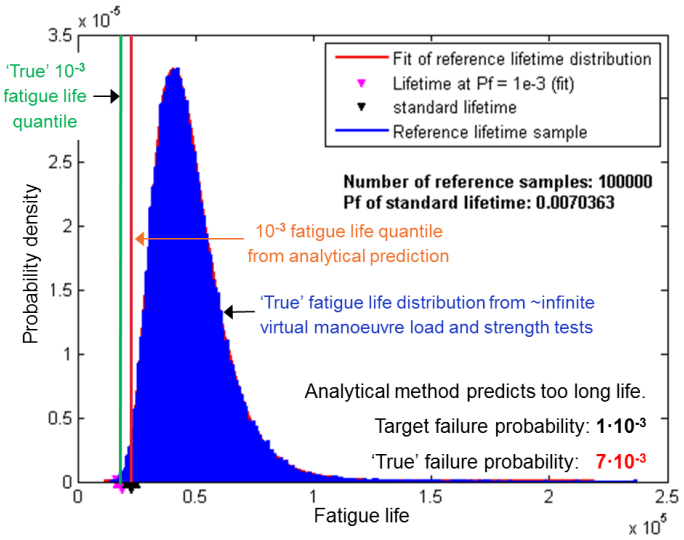


Figure 2.29: Comparison between the (synthetic) 10^{-3} lifetime quantile according to the reference distribution and the standard prediction method. (test ID = 1 in Table 2-1)

2.6.3.2 Reliability testing of simulation-based method method

The new simulation-based fatigue life substantiation model is tested differently as it does not directly predict a lifetime quantile but only predicts the probability of failure of a given SLL. Therefore, it tested if the simulation-based model indeed predicts a 10^{-3} probability of failure for the lifetime that is already known to be the 10^{-3} quantile of the 'true' reference lifetime distribution.

The test result is depicted in Figure 2.30. The circles in the blue line show $P_{fail}\left(\mathrm{SLL}_{ref}\right)$ for the i^{th} strength interval. The probability of having a component in the i^{th} strength interval is displayed by the squared red line. The triangulated black line shows the point-wise multiplication between $P_{fail}(\mathrm{SSL},s_i)$ given strength and the probability of this given strength. The dotted green line finally shows the cumulative probability of failure, which here accumulates to $1.05\cdot10^{-3}$.[5] [6]

The predicted $P_{fail}\left(\mathrm{SLL}_{ref}\right)$ of $1.05\cdot10^{-3}$ is practically a perfect result, as the estimate is well within an approximate 'one sigma' confidence interval of the 'true' reference quantile. Repetition of the test for predicting $P_{fail}\left(\mathrm{SLL}_{ref}\right)$ for a reference SLL corresponding to a more realistic but also more challenging 'true' 10^{-6} lifetime quantile, demonstrated similar results. For this test case, the 'true' 10^{-6} lifetime quantile was only slightly over-optimistically estimated to correspond to the $6.75\cdot10^{-7}$ quantile of fatigue life, i.e. P_{fail} was estimated too low as $6.75\cdot10^{-7}$ instead of 10^{-6}.

Finally, the test as in Figure 2.30, i.e. estimating P_{fail} of a true 10^{-3} lifetime quantile, was repeated for different strength discretization grids to evaluate the influence of strength discretization coarseness. As in Figure 2.31, increasing the coarseness of the grid, hence decreasing computational costs, leads to more conservative

[5] The $P_{fail}(\mathrm{SSL},si)$ estimates are made for sequentially increasing strength intervals, starting at the lower tail. When these probability estimates become very small, and as soon as the product of the i^{th} estimated probability of failure and the probability of a strength value in the i^{th} interval itself no longer provides a significant contribution to the overall reliability integral (2.21), the Pfail(SSL,si) estimates for the remaining strength intervals are conservatively assumed to save computational costs. For the case in Figure 2.30, the failure probability was conservatively assumed for intervals with normalised strength higher than about 0.9.

[6] The computations were executed with 10^3 samples per subset and a strength distribution discretized in 250 intervals. This represents a very accurate but computationally expensive configuration.

reliability estimates. This is as expected due to conservative truncation of the strength by which high-frequency fatigue damage is computed.

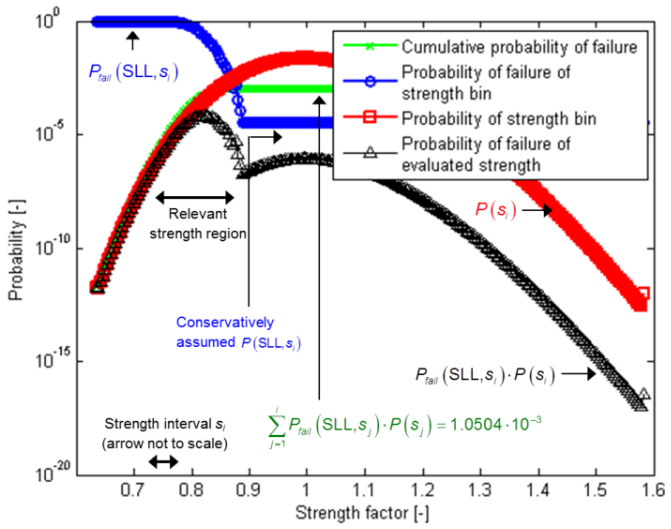


Figure 2.30: Detailed representation of results from Subset Simulation that are obtained under ideal circumstances, i.e. large sample-size conditions. (Test ID = 3 in Table 2-1)

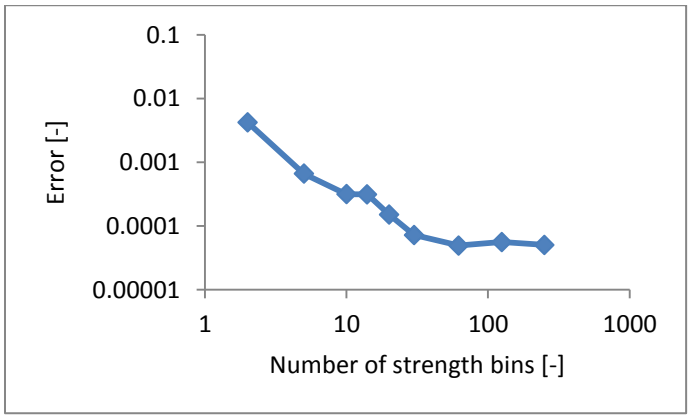


Figure 2.31: Graph demonstrating the effect of increasing the coarseness of the strength discretization grid. A positive estimation error is conservative.

Overall, the test results demonstrate that the newly proposed simulation-based fatigue life substantiation model does not contain methodologic errors. This is an improvement to the standard analytical model, which was demonstrated to be significantly biased in previous section 2.6.3.1.

2.6.4 Reliability testing with realistic small samples

In practice, the number of fatigue and flight tests that can be used to estimate their statistical distribution is small and computational resources are limited. Therefore, the validation tests are repeated but now while assuming that the results of only seven fatigue tests can be used to estimate the distribution of fatigue strength and that every flight regime was only test-flown fifteen times. Computational costs are limited by dividing the strength distribution into wide intervals and by using a low number of samples per subset distribution.

It can now no longer be expected that any of the models perfectly predicts the 10^{-3} fatigue life quantile. Because only few test results are available to estimate the distribution of fatigue strength and manoeuvre loads, fatigue life predictions are inevitably made based on inaccurate distribution estimates. Therefore,

predictions are made while targeting a 95% confidence level and it is tested if the models give a conservative estimate of the 10^{-3} fatigue life quantile in 95% of repeated prediction cases. In addition and for comparison, it is tested what inaccuracies can arise if predictions are made based on estimated distributions of fatigue strength and manoeuvre loads but without targeting a confidence level.

2.6.4.1 Reliability testing of standard analytical method

The standard analytical method was tested first by 250 repeated estimations of the same conservative 10^{-3} lifetime quantile. Each prediction was made based on a new and independent set of simulated fatigue strength and manoeuvre load tests. The distribution of the predicted fatigue life quantiles is shown in Figure 2.32. The predictions were made with and without targeting a 95% confidence level and thus also compare the effect of estimating the 10^{-3} lifetime quantile with and without targeting a 95% level of confidence. Seven virtual fatigue tests and fifteen virtual tests per manoeuvre were newly performed per repetition.

It was demonstrated that if no confidence interval is targeted while predicting the 10^{-3} quantile of fatigue life, only about 40% of the lifetime predictions would actually meet the targeted 0.999 reliability requirement. This automatically means that the lifetime percentile is non-conservatively overestimated in 60% of the cases, as strength dominates the prediction when using the analytical method.

The prediction bias can be understood by noting that the estimator of the variance, most notably of fatigue strength, is biased towards underestimating the variance. Simulations in sections 3.2.3 and 3.3 confirmed that it is 'normal' to underestimate the standard deviation of fatigue strength in roughly 60% of the cases if only seven tests are done. Systematic underestimation of the MLE variance of fatigue strength due to small sample-size effects is also observed in an industry data set shown in Figure 3.8 and discussed in section 3.5.1.

However, the distribution of fatigue life predictions in Figure 2.32 also demonstrated that if the 10^{-3} lifetime quantile is predicted with a single-sided 95% confidence interval, then indeed 241 out of 250 (96.4%) repeated predictions met the targeted 0.999 reliability requirement. This demonstrated that the 95% confidence level target was met accurately and that the use of confidence level analysis can accurately mitigate the effects of biased MLE estimates due to small sample-size effects.

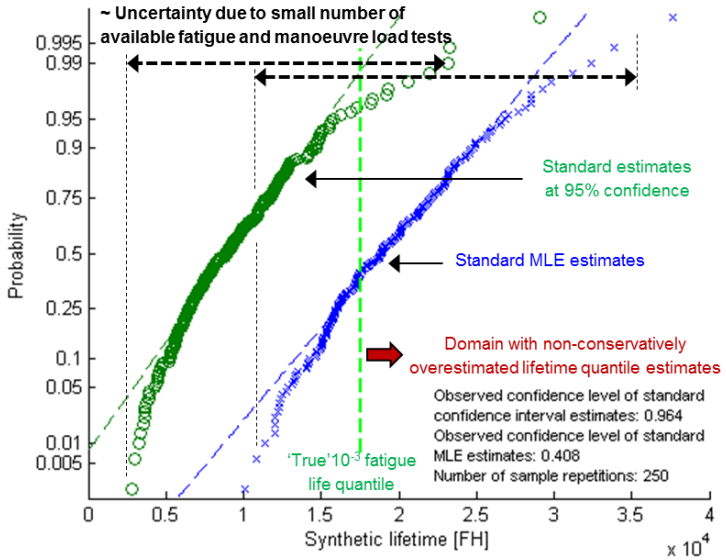


Figure 2.32: Distribution of test results for the standard fatigue life prediction method using realistically small samples as input. Probability plots assume a normal distribution. (Test ID = 5 in Table 2-1)

The test as in Figure 2.32 was repeated 25 times for redrawn synthetic problems. Each redrawing from equation (2.25) generates a slightly different fatigue life prediction problem by modifying the overall behaviour of the distributed flight regime loads. This approach explicitly tests the repeatability of the accurate

behaviour of the standard method. Following the test process in Figure 2.33, Figure 2.34 shows that for all the repeated test cases the targeted 95% confidence level was met with about ±2% accuracy. Using equation (2.18) and noting that only 25 test cases have been simulated, it follows that an approximate 'one-sigma' confidence interval of the realised confidence levels themselves have an approximate width of 2.8% for this test. Therefore, it can be concluded that the standard method yields practically perfect estimates, at least for the tested problem family as defined in section 2.6.2.

To further increase confidence in the accuracy of the standard method, the test as in Figure 2.34 was repeated but now while simulating that 'only' seven, instead of fifteen, manoeuvre load tests were performed per manoeuvre. This makes that the relative uncertainty in estimated manoeuvre loads is increased. The realised confidence levels followed a comparable normal distribution as in Figure 2.34 but with slightly increased variance (imprecision). The observed 'bottom-of-scatter' of the decimal meeting the reliability requirement reduced slightly to 91.2%, instead of 94% before.

Overall, the accurate prediction results demonstrated that the standard analytical method can yield accurate results under small sample size conditions. The methodological simplification that the analytical method makes by neglecting any effects of uncertainty in loads can in practise be small in comparison to the effects of uncertainty in strength, which the analytical method does duly account for.

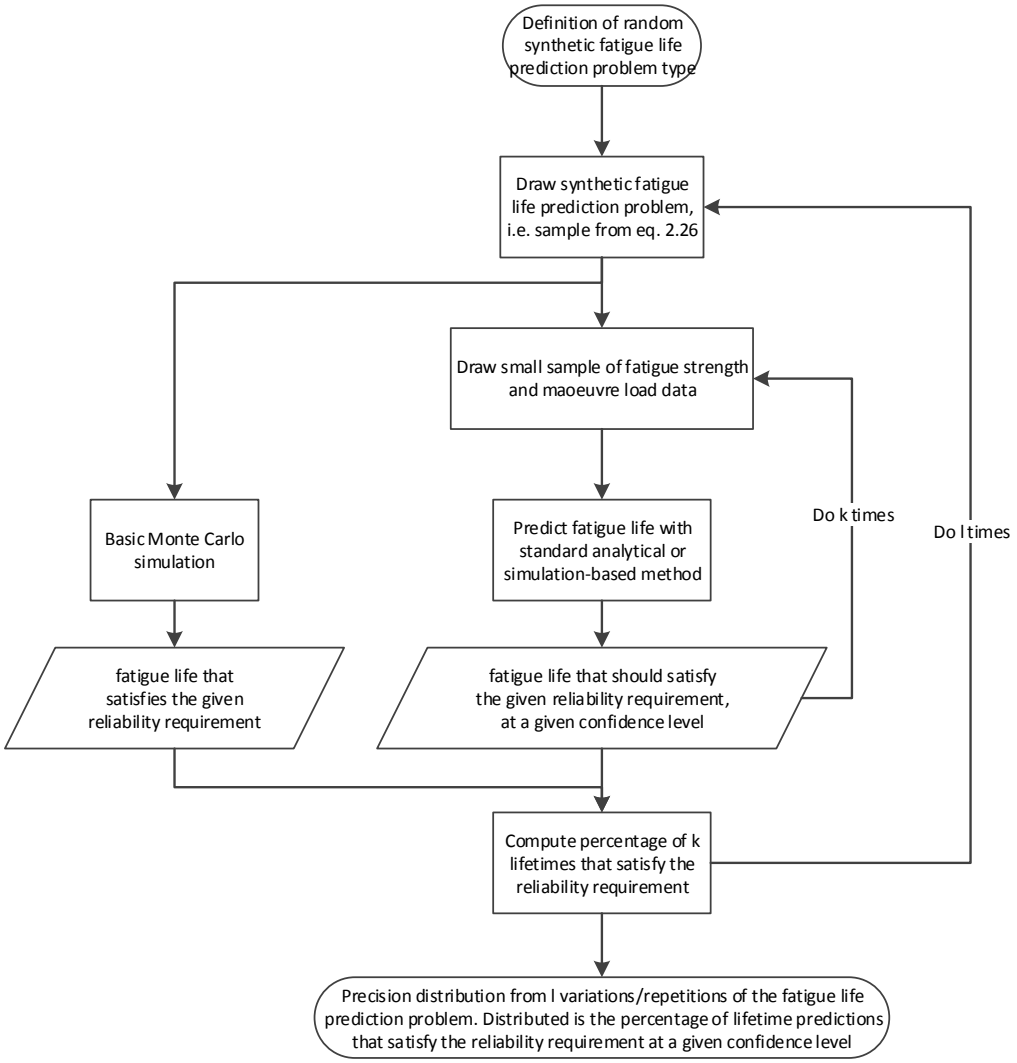


Figure 2.33: Summary of the procedure for repeated precision testing under small sample size conditions

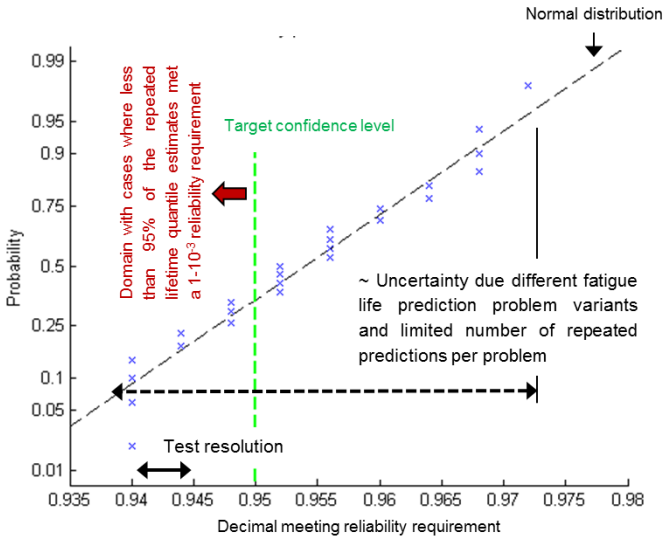


Figure 2.34: Distribution of the verified confidence level of repeated reliability predictions made by the standard fatigue life prediction method when the method makes use of realistically small samples as input. (Probability plot assumes a normal distribution) (Test ID = 6 in Table 2-1)

As a last validation step, the tests as in Figure 2.32 and Figure 2.34 were repeated once more but under even more challenging circumstances. It was again simulated that 'only' seven manoeuvre load tests were available per manoeuvre. However, now the reliability target was more realistic. It was required to estimate the 10^{-6} fatigue life quantile instead of the 10^{-3} quantile previously. The 'true' 10^{-6} fatigue life quantile was estimated by a fitted GEV CDF distribution through 10^{6} samples from BMC simulation.[7]

Figure 2.35 shows fifty repetitions of estimating the same conservative 10^{-6} lifetime quantile with the analytical method. The estimates were made both with and without targeting a 95% level of confidence. Seven virtual fatigue tests and seven virtual tests per manoeuvre were newly available per repetition. The distribution of the predictions demonstrated that when disregarding uncertainty due to the small amount of fatigue strength and manoeuvre load tests, i.e. when not targeting a confidence level for the fatigue life quantile predictions, only about 28% of the lifetime predictions actually met the targeted 1-10^{-6} reliability requirement. However, if the 10^{-6} quantile was estimated with a single-sided 95% confidence interval, then the analytical method was successful in meeting the targeted confidence level requirement. 48 out of 50 (96%) of the repeated predictions indeed met the 1-10^{-6} reliability requirement.

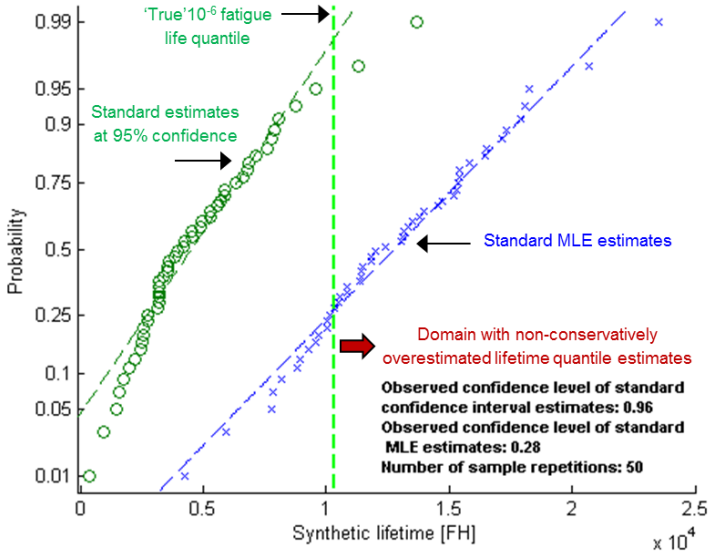


Figure 2.35: Distribution of 10^{-6} fatigue life quantiles estimated by the standard analytical fatigue life prediction method when the method uses realistically small sample sizes as input. (Probability plots assume a normal distribution) (Test ID = 8 in Table 2-1)

The distribution of the predictions of the same 10^{-6} fatigue life quantile in Figure 2.35 shows significant prediction variance in terms of lifetime. Figure 2.36 shows the same test result as in Figure 2.35 but now with the horizontal lifetime axis transformed to the corresponding probability of failure. This gives an estimate about what this variance in lifetime corresponds to in terms of variance in the probability of failure. The test demonstrates that the distribution of fatigue life predictions can span about 10 orders of magnitude of reliability. This span originates from estimation uncertainty about the distributions of fatigue strength and manoeuvre loads.

[7] Such an extrapolation-based quantile estimate can only be regarded as fully accurate and precise when fatigue life can indeed be correctly described by a GEV distribution. BMC simulations with 10^{5} samples justified this assumption, as well as the negligence of confidence bounds around the GEV distribution fit.

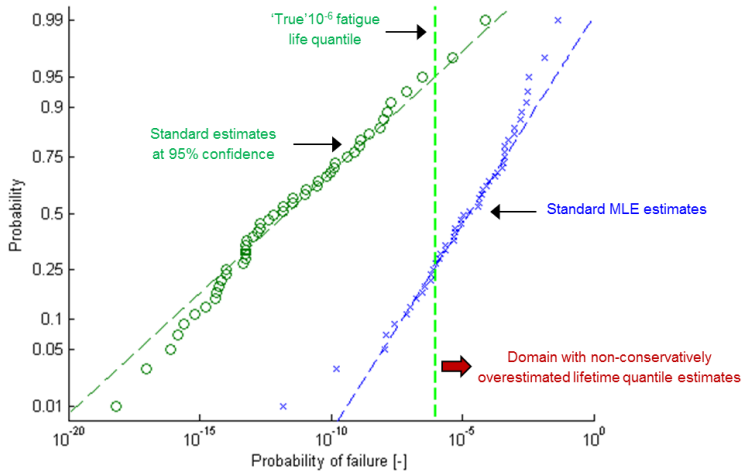


Figure 2.36: Distribution of the 'true' reliability of 10^{-6} fatigue life quantiles estimated by the standard analytical fatigue life prediction method when the method uses realistically small sample sizes as input. (Probability plots assume a normal distribution) (Test ID = 8 in Table 2-1)

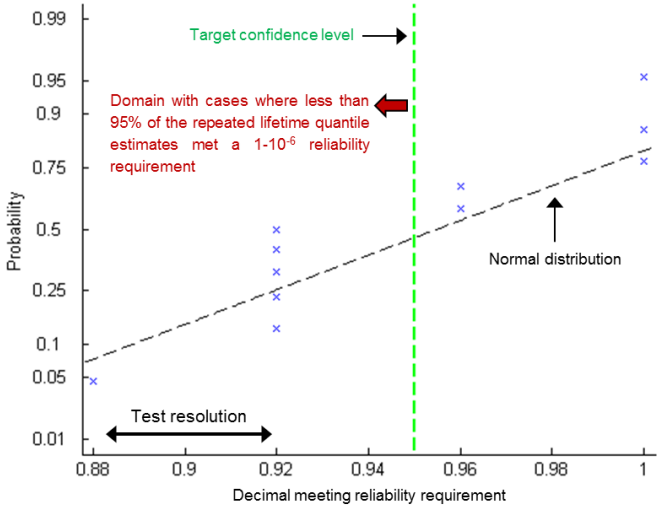


Figure 2.37: Distribution of the result of 12 similar test cases of the standard fatigue life prediction method for realistically small samples and 10^{-6} quantiles. One test case consists of 25 repetitions to estimate the 'true' confidence level. Seven load tests per manoeuvre were available per repetition. (Probability plot assumes a normal distribution) (Test ID = 9 in Table 2-1)

The validation of the analytical method's ability to accurately predict a 10^{-6} quantile of fatigue life, as presented in Figure 2.35, was repeated twelve times for redrawn synthetic problems (i.e. resampling the parameters in equation (2.26)) to further increase confidence in the accurate behaviour of the standard method. The distribution of the decimal meeting the reliability requirement is shown in Figure 2.37. Despite the limited power of the test (i.e. few bootstrap repetitions), it can be concluded that the analytical method again meets the targeted 95% confidence level reasonably well (i.e. -10% +5%).

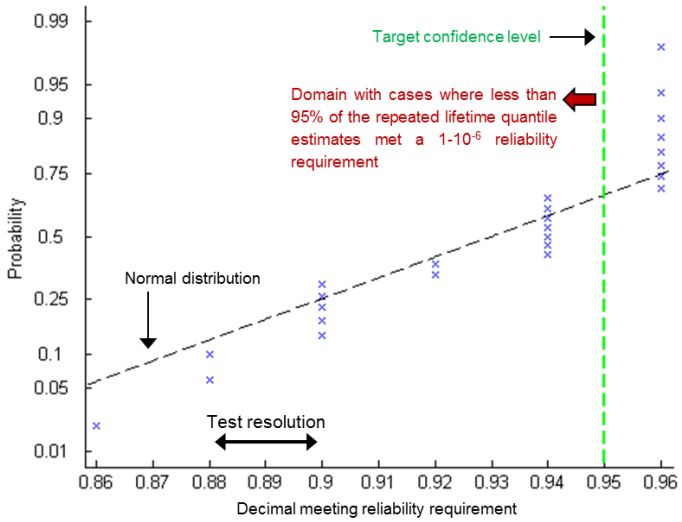


Figure 2.38: Distribution of the result from 25 similar test cases for the standard fatigue life prediction method. One test case consists of 50 repetitions by the standard fatigue life prediction method to estimate 10^{-6} quantile of fatigue life with 95% confidence and using realistically small samples as input. (Probability plot assumes a normal distribution) (Test ID = 10 in Table 2-1)

To demonstrate that the analytical method is not generally applicable though, a final test case has been defined. In this final test case, the relative influence of uncertainty due to manoeuvre loads is increased even more by simulating that seven fatigue tests and only three load tests per manoeuvre were available to estimate a 10^{-6} fatigue life quantile. According to the test result in Figure 2.38, the analytical method only met the targeted 95% confidence level with limited accuracy, i.e. -10% +1%.

Although the analytical method passed most defined test cases successfully, one should thus nevertheless be careful with using the analytical model under circumstances where uncertainty from regime loads is very high with respect to uncertainty from fatigue strength. Due to the modelling assumption of the analytical method that uncertainty from regime loads may be neglected, it is not reasonable to expect that the analytical method can precisely meet the targeted confidence level requirement under all circumstances. For future work, it is recommended to systematically determine the boundaries of applicability of the analytical method with more detail.

2.6.4.2 Reliability testing of simulation-based method

The new simulation-based method was first tested by checking if it indeed predicts a 10^{-3} probability of failure for a lifetime that was already known to be the 10^{-3} quantile of the 'true' reference lifetime distribution. When making the prediction with 95% confidence, the predicted P_{fail} may not be lower than 10^{-3} for 95% of the load and strength sampling repetitions. Figure 2.39 shows that 5/100 of the repeated predictions were too optimistic regarding the probability of failure of the true 10^{-3} lifetime quantile. This is practically 'perfect' performance when considering the precision of this 'true' reference.[8]

The test was also repeated while simulating that 'only' seven, instead of fifteen manoeuvre load tests were performed per manoeuvre. As a result, 89/100 MLE estimates and 99/100 upper confidence level estimates met the actual reliability requirement.[9] This result demonstrated over-conservative predictions is believed to be caused by an over-conservatively designed custom procedure that mitigates implementation problems in

[8] Repetition of this validation test using a strength distribution discretized in only fifteen intervals resulted in a verified confidence level of 0.40 and 0.90 for the MLE and 95% confidence estimates respectively.
[9] This validation test was run twice. The repeated test yielded very similar results: 89/100 MLE estimates and 100/100 upper confidence estimates were observed to meet the actual reliability requirement.

fitting multi-dimensional distributions through few sample points.[10] It is expected though that adjustments of the distribution estimation procedure, possibly in combination with more bootstraps per repeated sample, will yield more accurate results. Appendix G discusses implementation issues and their handling in more detail.

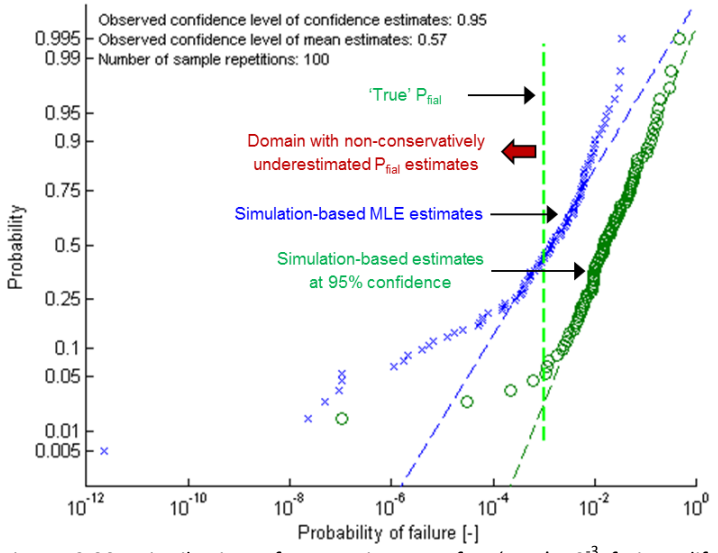


Figure 2.39: Distribution of P_{fail} estimates of a 'true' 10^{-3} fatigue life quantile made by the simulation-based fatigue life substantiation model making use of realistically small samples as input. The simulation used 150 samples per subset, a strength distribution discretized in 25 intervals and 25 bootstraps per repeated sample. This is a computationally 'cheap' configuration. (Probability plots assume a normal distribution) (Test ID = 11 in Table 2-1)

As a test of using the custom RBDO application from section 2.5.6, Figure 2.40 shows 50 repetitions of estimating the same conservative lifetime 10^{-3} lifetime quantile, while having seven fatigue and fifteen manoeuvre load tests available. It shows that none of the repeated lifetime designs fell below the 'true' 10^{-3} lifetime quantile. As a 95% upper single sided confidence level was targeted, this test demonstrated too conservative results. The validation test of simulation-based P_{fial}(SLL) estimates where the same amount of samples were available to estimate distributions, as in Figure 2.39, was passed successfully before. Therefore, it is expected that adjustments of the RBDO application will yield significantly more accurate results.

2.6.4.3 Reliability comparison between analytical and simulation-based prediction models

As a last testing step, predictions from both the analytical and simulation-based prediction methods are compared while they use the same dataset. The distribution of predictions from the two methods in Figure 2.40 demonstrated that lifetime quantiles designed by the simulation-based method are similar to estimates from the standard method, though somewhat over-conservative. In general, though, the small difference between the results from the analytical and simulation-based prediction methods suggests that, at least for the tested problem family and under realistically small sample size conditions, the precision and accuracy for estimating a lifetime quantile is simply governed by the precision and accuracy up to which a quantile of a lognormal strength distribution can be estimated.

[10] As detailed further in Appendix G.

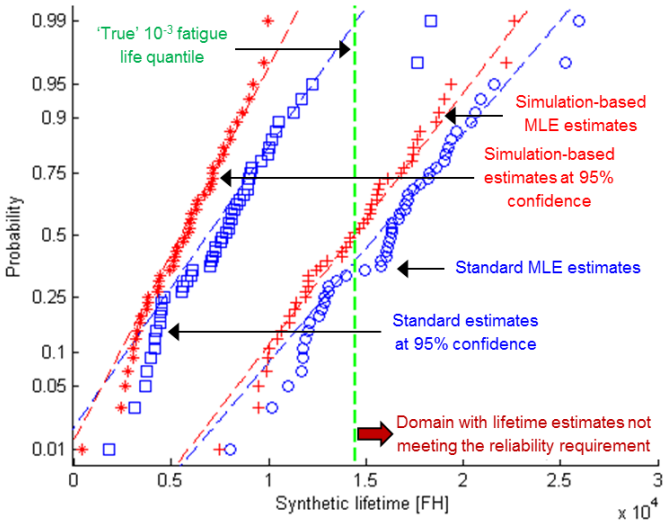


Figure 2.40: Distribution of 10^{-3} quantile estimates of fatigue life made by both the simulation-based and standard fatigue life quantile prediction models and by making use of reliability-based design optimisation for the simulation-based model. The simulation used 150 samples per subset, a strength distribution discretized in 20 intervals and 25 bootstraps per repeated sample. (Probability plots assume a normal distribution) (Test ID = 15 in Table 2-1)

The comparative test as in Figure 2.40 was repeated for estimating a 10^{-6} quantile (instead of 10^{-3}) and while having 'only' seven instead fifteen manoeuvre load tests available. The distribution of predicted lifetime quantiles in Figure 2.41 clearly indicated that the simulation-based quantile estimates were too conservative and underestimated permissible lifetimes. As discussed before, it is expected that practical implementation and design issues concerning the RBDO application are the main cause, in addition to the following two issues:

- Difficulties in small sample size distribution fitting
- limited computational resources, e.g. forcing few samples per subset and a coarse strength discretization grid

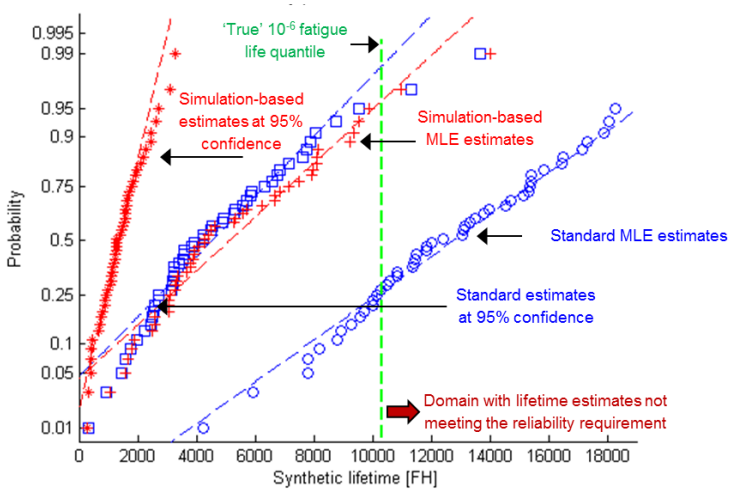


Figure 2.41: Distribution of 10^{-6} quantile estimates of fatigue life made by both the simulation-based and standard fatigue life quantile prediction models and by making use of reliability-based design optimisation for the simulation-based model.. (The simulation used 150 samples per subset, a strength distribution discretized in 16 intervals and 25 bootstraps per repeated sample.) (Probability plots assume a normal distribution) (Test ID = 16 in Table 2-1)

2.6.5 Summary of results from reliability testing

Table 2-1 contains an elaborate summary of all validation test results discussed in sections 2.6.3 and 2.6.4. The results demonstrate that under ideal circumstances, i.e. with large samples available to determine distributions of fatigue strength and flight regime loads, the simulation-based method provides predictions with very high accuracy, whereas the simplified analytical method is prone to non-conservative estimation bias.

Under realistic circumstances, however, where only a few data points are available to estimate distributions of fatigue strength and flight regime loads, the analytical method yields accurate results for most tested cases. Its simplification to not model and mitigate uncertainty from manoeuvre loads mostly only results in small errors which can be neglected in comparison to uncertainty from fatigue strength.

The simulation-based method performs less well under realistic small sample-size conditions. Especially in circumstances where few samples are available to fit distributions of regime loads and damage, the simulation-based method results in significant prediction errors. Nevertheless, biases are often conservative and manageable. It is expected that improvements to the (over-)conservative practical implementation of the simulation-based model can result in better accuracy and are thus recommended for future work.

Also included in the summary of the test results are the reliability levels achieved by using MLE estimates of the distributions of fatigue strength and flight regimes loads. Using this simplified approach where confidence level analysis is not applied, it can be expected that the probability of conservatively estimating a required fatigue life quantile is significantly less than 50%. As discussed earlier in section 2.6.4.1, this non-conservative estimation bias primarily originates from a non-conservative and fundamental estimation bias for the MLE estimator for the standard deviation under small sample size conditions. As demonstrated by the probability plots before in section 2.6, using MLE distribution parameters may lead to large and non-conservative errors resulting in orders of magnitude less reliability than targeted.

Table 2-1: Table summarising all the validation test results used in chapter 2.

Test ID number	Target probability of failure	Target confidence level	Fatigue tests	Load tests per manoeuvre	Reliability substantiation model	Load and strength dataset variations	Load and strength dataset bootstrap variations	Lifetime prediction problem variations	Strength distribution discretization points	Samples per subset	Percentage of MLE estimates satisfying reliability requirement	Percentage of estimated confidence interval satisfying reliability requirement	Bottom-of-Scatter of percentage of estimated confidence interval satisfying reliability requirement	Reliability requirement	Probability of failure before predicted lifetime	Predicted probability of failure
1	1E-03	-	5E+05	1E+04	Analytical	-	-	-	-	-	-	-	-	7E-03	-	-
2	1E-06	-	5E+05	1E+04	Analytical	-	-	-	-	-	-	-	-	1.20E-05		
3	1E-03	-	5E+05	1E+04	Simulation-based	-	-	250	1000	-	-	-	-	-	1.05E-03	
4	1E-06	-	5E+05	1E+04	Simulation-based	-	-	250	1000	-	-	-	-	-	6.75E-07	
5	1E-03	95%	7	15	Analytical	250	-	-	-	-	40.5%	96.4%	-	-	-	-
6	1E-03	95%	7	15	Analytical	250	-	25	-	-	-	-	94%	-	-	-
7	1E-03	95%	7	7	Analytical	250	-	25	-	-	-	-	91.2%	-	-	-
8	1E-06	95%	7	7	Analytical	50	-	-	-	-	28%	96%	-	-	-	-
9	1E-06	95%	7	7	Analytical	25	-	12	-	-	-	-	88%	-	-	-
10	1E-06	95%	7	3	Analytical	50	-	25	-	-	-	-	86%	-	-	-
11	1E-03	95%	7	15	Simulation-based	100	25	-	25	150	57%	95%	-	-	-	-
12	1E-03	95%	7	15	Simulation-based	100	15	-	25	150	40%	90%	-	-	-	-
13	1E-03	95%	7	7	Simulation-based	100	25	-	25	150	89%	99%	-	-	-	-
14	1E-06	95%	7	15	Simulation-based	25	25	-	25	150	28%	92%	-	-	-	-
15	1E-03	95%	7	15	Simulation-based (RBDO)	50	25	-	20	150	52%	100%	-	-	-	-
16	1E-06	95%	7	7	Simulation-based (RBDO)	50	25	-	16	150	96%	100%	-	-	-	-

2.7 Conclusion

This chapter confirmed that, under idealised circumstances, i.e. knowing the 'true' distributions of fatigue strength and manoeuvre loads, a methodological and non-conservative error is made when the reliability of a predicted fatigue life is substantiated using only the distribution of fatigue strength and simplifying the flight manoeuvre load distributions to their mean values. As a solution, a new simulation-based fatigue life prediction method was successfully validated.

However, it was also demonstrated that the simple analytical method did nevertheless yield accurate results under all studied realistic engineering conditions, i.e. where the distributions of fatigue strength and manoeuvre loads have to be estimated from a small number of test results. Direct comparison under these realistic conditions between the analytical and simulation-based method revealed small differences in precision and accuracy, excluding cases for which implementation problems caused the simulation-based method to be over-conservative. Generalising, the accurate results of the analytical method suggest that under small sample size conditions, uncertainties in manoeuvre loads may be fully neglected and the full reliability substantiation may be derived from the fatigue strength distribution only.

Table 2-2 provides a summary of all test results.

Table 2-2: Table synthesizing the results of the verification and validation tests conducted in chapter 2.

Probability of failure	fatigue tests	load tests per manoeuvre	Analytical	Simulation-based
10-3	$> 10^3$	$> 10^3$	NOK	OK
10-6	$> 10^3$	$> 10^3$	NOK	Near OK
10-6	7	15	OK	OK
10-3	7	7	OK	Near OK
10-6	7	7	Near OK	NOK
10-6	7	3	Near OK	N.A.

However, the set of prediction problems for which the analytical method was validated was small and simulations discussed in section 4.2.6 suggest that its accurate performance cannot be generalised in full. Therefore, future work is thus recommended to include expansion of the synthetic test conditions to estimate boundaries for reliable application of the analytical and simulation-based methods. Expanded test conditions may include:

- Increased variation in manoeuvre loads
- A broad range of S-N curve shapes, strength variations and mission profiles

Additionally, future work may also include a detailed study on the numerical efficiency of the presented methods.

Although the simulation-based method was demonstrated to provide methodologic improvements over the simplified analytical method, its practical implementation is complex. The tested implementation did result in over-conservative predictions for some test cases. It is, therefore, recommended to only make use of the new and complex simulation-based method when circumstances are encountered where the simple and easy-to-apply analytical method is clearly not applicable, i.e. when variance and uncertainties from manoeuvre loads are no longer insignificant in comparison to variance and uncertainties from fatigue strength.

Given the work's modelling assumption that only full-scale component fatigue tests can provide relevant data to estimate an S-N-P curve, which was introduced in section 2.2.2 is discussed in more detail in chapter 3, the emphasis is put on the importance of properly mitigating uncertainty coming from inadvertent inaccurate fatigue strength estimates, due to the availability of only a few samples. The reach of such uncertainty was

clearly exemplified by the test results discussed in section 2.6. It is therefore recommended to explicitly determine a confidence interval for any critical fatigue life quantile prediction and clearly state modelling assumptions. This may prevent misconceptions on the reliability that can really be guaranteed by statistical methods.

In addition, for future work it is also recommended to perform sensitivity studies to determine the impact of the design assumptions that underpin both the analytical and simulation-based fatigue life prediction models and which were discussed in sections 2.2.1 - 2.2.3. In particular the effect of alternative S-N curve models, fatigue strength distribution models, or fatigue damage accumulation models could be quantified. Such a sensitivity study may provide more insight on the maximum attainable accuracy and precision that fatigue life prediction models can have in the first place. This could thereby effectively limit accuracy and precision requirements that fatigue life prediction models could reasonably be subjected to.

3 Tolerance interval estimation for fatigue strength

The fatigue life of rotorcraft is typically substantiated by a combination of conservative manoeuvre load assumptions and conservative fatigue strength. For cases where the statistical analysis to determine a conservative value for fatigue strength must be performed based on very few samples or even none at all, the use of common frequentist statistics are demonstrated to yield inaccurate and unpractical results. As an alternative, the value of Bayesian statistics is exemplified using real data and a simplified fatigue life substantiation model. The case study demonstrates how well-established industry practise can be formally substantiated by traceable and explicit statistical analysis. It is thus exemplified how Bayesian statistics can reduce the number of critical engineering assumptions in fatigue life substantiation, or make them more explicit and traceable. In addition, it is also demonstrated how explicit statistical analysis can now be used to yield practical results, even under very small sample-size conditions.

An adapted version of this chapter was published by Dekker *et.al.* in the "International Journal of Fatigue" [56].

3.1 Introduction

Fatigue strength is a random variable. This means that if several specimens are fatigue tested under an identical repetitive loading profile, then each item fails after a different and random number of load cycles. This variation in fatigue strength can be significant, can result in major scatter in cycles-to-failure and can often not be predicted accurately without extensive full-scale component fatigue testing. The costs of full-scale component fatigue tests are generally high and often the number of tests, therefore, remains small. The statistical modelling of fatigue strength is however of high importance since uncertainty about fatigue strength can dominate the reliability of complete service life predictions as, for example, earlier simulations conducted in section 2.6 demonstrate.

A case study using common frequentist statistical methods, i.e. relying on sample data only, to substantiate the reliability of reduced working S-N curves is performed for aerospace-typical small sample size conditions. The example demonstrates reduction factors either not meeting the target reliability requirement or reduction factors too low for practical use. Typically, rotorcraft manufacturers must, therefore, resort to strong engineering assumptions and/or the use of fixed reduction factors [16] when dealing with very small sample sizes.

As an alternative solution, a simple Bayesian model is introduced. Using Bayesian statistics, it is illustrated how prior experience and other related knowledge about the variance in fatigue strength can be taken into account as well. In this example, using a real dataset from industry, Bayesian statistics can substantiate realistic reduction factors for fatigue strength, even if few or no results from component fatigue testing are available for a specific component. The demonstrated case makes use of a simplified model for fatigue strength variance and is simple to apply. The method can be customized easily and provides for straightforward integration in numerical and simulation-based reliability models for fatigue life prediction of the type as earlier introduced in chapter 2. Despite the simple modelling assumptions of the exemplified method, it is compatible with reliability substantiation methodologies from several rotorcraft manufacturers [16, 57, 35] and the authority guideline AC-27-1B MG 11 [58]

3.2 Introduction to statistical fatigue strength modelling

This section first introduces a set of modelling assumptions and a statistical model for fatigue strength, which are the same as the assumptions and model already introduced in section 2.2. For convenience, these are however summarized here again. Second, this section presents details of the way a lognormal distribution is treated in this work to ease understanding of specific implementation details and results. Third, this section briefly revisits the concept of tolerance intervals, which was already introduced in chapter 2.

3.2.1 Fatigue strength modelling by an S-N-P curve

Figure 3.1 exemplifies the basic and simplified model for probabilistic fatigue strength by which the use of Bayesian statistics is demonstrated. The illustration features constant amplitude fatigue tests through which a four-parameter Weibull curve is fitted, which is the expected S-N curve. Scatter in fatigue strength is completely modelled by the distribution of normalized residuals on the load axis. Figure 3.2 shows the corresponding normalized fatigue strength distribution.

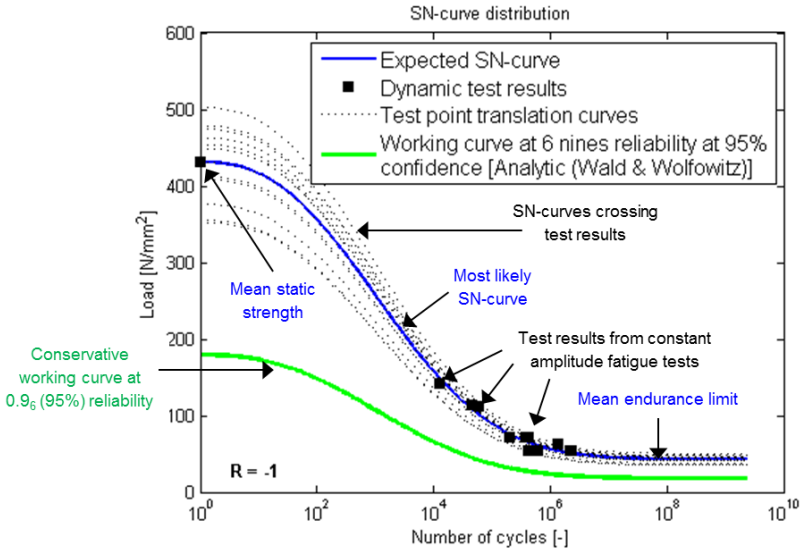


Figure 3.1: Example of results from constant amplitude and full-scale fatigue tests for a component from the dynamic system of a helicopter, the S-N curve fitted through these results and the derived conservative working curve. (Figure replicated from chapter 2)

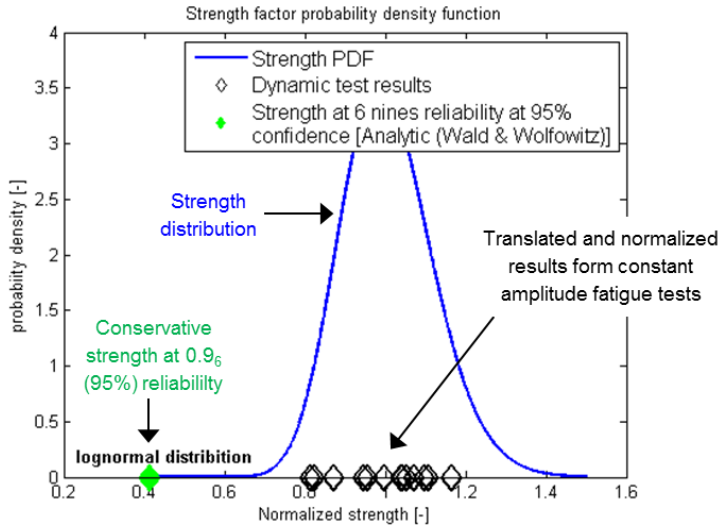


Figure 3.2: Example of one-dimensional fatigue strength distribution fitted through fatigue test data corresponding to Figure 3.1 and normalised by the fitted S-N curve. (Figure replicated from chapter 2)

This one-dimensional distribution to model uncertainty about fatigue strength and to form an S-N-P curve is a highly simplified model and includes the following simplifications and engineering assumptions:

- Definition of the shape parameters of the S-N curve is perfect, i.e. it is appropriate to assume that all estimation uncertainty considering the S-N curve parameters may be captured by a one-dimensional fatigue strength distribution

- Normalized fatigue strength follows a lognormal distribution
    - Data presented by Thompson & Adams based on pooling a large number of full-scale fatigue strength results of metallic rotorcraft components suggest the assumption of lognormal or normal distributed fatigue strength is reasonable, with the latter being more conservative [18]
- Fatigue strength is homoscedastic, i.e. scatter in load direction is independent of the number of cycles to failure
    - Although the assumption of homoscedasticity does not generally hold and can be invalidated by examples where scatter positively correlates with *N*, e.g. Schijve [1], this engineering assumption is acceptable to aviation authorities and general engineering practise in the rotorcraft industry, where scatter is often estimated in the load dimension based on test results falling in the important region around $N = 10^5$.
- The results from full-scale fatigue tests are fully representative:
    - Fatigue testing is done under conditions that are representative of operational conditions
    - Tested components are representative of the population of in-service components, i.e. their manufacturing and quality assurance processes are equivalent
- Right-censored data, i.e. run-outs, can be neglected or considered as an observed failure at the time of test termination

These modelling assumptions correspond to, or are highly similar to, long-standing practice in rotorcraft industry and are in compliance with airworthiness requirements, see for example chapter 4.1 of the AGARD-AG-292 Helicopter Fatigue Design Guide [16], a review of industry practise by Everett [35] or Advisory Circular AC-27-1B MG 11 [58]. Different manufacturers generally make use of different models and design assumptions to comply with airworthiness regulations. Everett [35] observed that fatigue life predictions by different manufacturers for the same component can vary significantly. Many alternative and more elaborate probabilistic models for fatigue strength exist. Examples include the use of an exponential S-N curve, normal or Weibull distributed fatigue strength, multi-variate S-N-P distribution models, or models specialized for composite materials [18, 1, 59, 60, 61].

It can be expected that such differences, in particular concerning one-dimensional lognormal distributed fatigue strength, may result in significantly different tolerance intervals. However, it is expected that the general case presented in this chapter, illustrating the value of Bayesian statistics to substantiate a safe value for fatigue strength, remains unaffected.

3.2.2 Introduction to the lognormal distribution

Implementation details on the lognormal distribution used to model fatigue strength are introduced to ease understanding further on. A transformation of the standard normal distribution $\mathrm{N}(\mu = 0, \sigma = 1)$ can describe the probability density function (PDF) $p(S \mid \mu, \sigma)$ of a quantity S, whose population follows a lognormal distribution with mean μ and standard deviation σ [30]:[11]

[11] As an alternative to using the natural logarithm and exponential function, a logarithm and exponential with base-10 are also used by many computer programs and authors. Present work distinguishes these cases by appending the mean μ and standard deviation σ with an additional subscript: $_{10}$. The mean and standard deviation of both variants are different and cannot be compared directly. However, both formulations result in an identical PDF and there is thus no difference between statistics (e.g. quantiles) from both variants. Multiplication by $\log_e(10)$ transforms a base-10 standard deviation σ_{10} to the corresponding standard deviation σ of a base *e* lognormal distribution.

$$p(s\,|\,\mu,\sigma)=\exp\left(\sigma\cdot \mathrm{N}(\mu=0,\sigma=1)+\mu\right) \quad \text{with} \quad \mathrm{N}(\mu,\sigma)\equiv \mathrm{p}(z\,|\,\mu,\sigma)=\frac{1}{\sigma\sqrt{2\cdot\pi}}\exp\left(\frac{-(z-\mu)^2}{2\cdot\sigma^2}\right) \tag{3.1}$$

A sample with *n* independent and random realizations of *S*, $\{S_1, S_2, \ldots S_n\}$, can be used to estimate the population mean and standard deviation of the associated normal distribution analytically by sample estimates $\hat{\mu}$ and $\hat{\sigma}$ respectively:

$$\hat{\mu}=\frac{1}{n}\cdot\sum_{i=1}^{n}\log_e(S_i) \tag{3.2}$$

$$\hat{\sigma}=\sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}\left[\hat{\mu}-\log_e(S_i)\right]} \tag{3.3}$$

3.2.3 Introduction to confidence and tolerance intervals

Confidence intervals are paramount to the appropriate understanding of small sample size statistics, i.e. as applied in fatigue life prediction. Statistics (e.g. the mean, variance or quantile of a distribution) are often computed based on a limited number of test results. Consequently, estimated statistics themselves are subject to significant uncertainty. However, this imprecision can be accounted for by computing confidence intervals. A confidence interval on an estimate of a quantile is a tolerance interval, where a quantile is the same as a percentile but expressed as a probability instead of a percentage. A confidence level specifies the probability that the confidence interval includes the true value of the statistic if a large number of independent samples of equal size would each estimate that statistic.

For example, consider that the strength of some component is normal distributed and that its value used for design, S_{design}, must be conservative for 999,999 out of every 1,000,000 manufactured products. If many firms, all using identical material batches and manufacturing, testing and quality control processes, would independently carry out a number of strength tests to estimate the distribution of strength and to estimate the 10^{-6} quantile, then several of them would estimate S_{design} such that the true probability of failure is significantly higher than 10^{-6}. This phenomenon has been exemplified and discussed before in section 2.6 where fatigue life predictions were made repeatedly, each time based on a new and independent sample of the input data.

These tests explicitly exemplified that without confidence level analysis predictions can be biased systematically and significantly. To explain this phenomenon in more detail, Figure 3.3 illustrates how the estimator of the standard deviation is only asymptotically biased and how this leads to systematic underestimation of target quantiles. In addition, Figure 3.3 exemplifies the likelihood and severity of possible underestimations of a target quantile if these are made without a confidence level. The smaller the size of the sample to estimate the distribution, the more likely it is to significantly underestimate the required quantile. However, if in contrast a tolerance interval is specified, then it can be compensate that each sample can only characterize the true population with limited precision. For example, with a 10^{-6} (95%) tolerance interval specified, on average only 5/100 firms underestimate the 10^{-6} quantile.

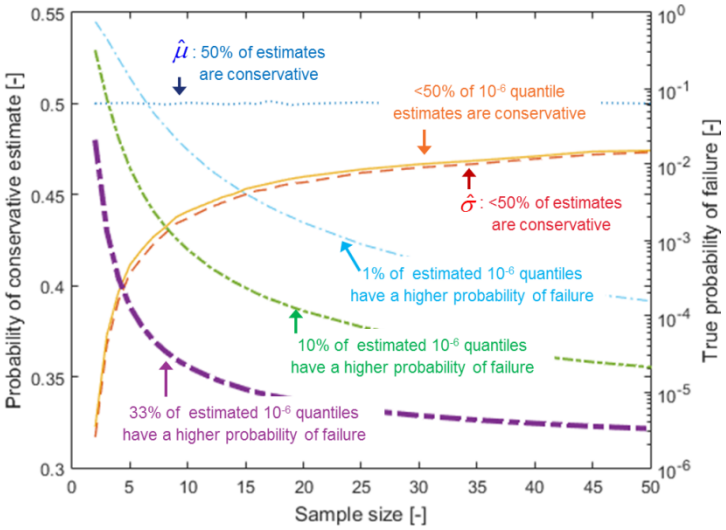


Figure 3.3: Example simulation illustrating how the precision of estimating 10^{-6} quantiles of a standard normal distribution depends on the sample size used to estimate the normal distribution. The example also illustrates how the estimator of the standard deviation becomes asymptotically un-biased with increasing sample size and is significantly biased for small and medium sample sizes.

3.3 Benchmarking and reliability testing of classic methods for tolerance interval estimation for fatigue strength substantiation

Many classic frequentist methods to estimate a tolerance interval are available. Table 3-1 selects some common methods applicable to a lognormal distributed population. Their accuracy is benchmarked in this work under small sample-size conditions relevant for deriving a reduction factor for a working S-N curve.

Table 3-1: Tabulated overview of selected methods to estimate a tolerance interval of a lognormal distributed quantity.

Approximate analytical (Wald & Wolfowitz) [62, 63]	• Classical method to estimate a tolerance interval of a normally distributed quantity
Approximate analytical (ESDU 91041) [64]	• Analytical approximation function to estimate a tolerance interval of a normal distributed quantity
Approximate analytical (AGARD-AG-292) [16]	• A custom analytical approximation from industry for a tolerance interval of a normally distributed quantity • Assumes a sample estimate of the standard deviation perfectly represents the true standard deviation of the population
Observed likelihood [65]	• Semi-analytical approximation of a tolerance interval of a normally distributed quantity • Standard built-in functionality in the MATLAB distribution fitting toolbox
Likelihood profile [66]	• Semi-analytical method using a likelihood distribution of fitted distribution parameters • Can estimate confidence intervals around any statistic that is a function of the distribution parameters • The likelihood distribution can be approximated analytically or by simulation • Can straightforwardly handle run-outs from fatigue testing
Parametric bootstrapping [66]	• Simulation-based method using exact analytical distributions for the uncertainty of fitted distribution parameters • Can estimate confidence intervals around any statistic that is a function of the distribution parameters
Non-parametric bootstrapping [67]	• Generic simulation-based method that can simulate the uncertainty distribution of any statistic

3.3.1 Reliability testing of selected quantile estimation methods

Accuracy verification of the tolerance interval estimators follows a methodology based on Monte Carlo simulation summarized in Figure 3.4. First, many samples are drawn from a known lognormal population. Then, for each sample, the targeted tolerance interval is computed. Finally, it is checked if the proportion of estimates that exceeds or matches the target reliability requirement (i.e. quantile) corresponds with the targeted confidence level, as illustrated in Figure 3.5. Though computationally inefficient, this test procedure is relatively intuitive, easy to implement and highly accurate if the simulation is carried out with sufficient samples. Test accuracy can be estimated by the estimator of the coefficient of variation *CoV* of the basic Monte Carlo estimator of an estimated quantile P_{sim}:

$$CoV_{P_{sim}} = \frac{\hat{\sigma}_{P_{sim}}}{\hat{\mu}_{P_{sim}}} = \sqrt{\frac{1-P_{sim}}{P_{sim} \cdot n}} \tag{3.4}$$

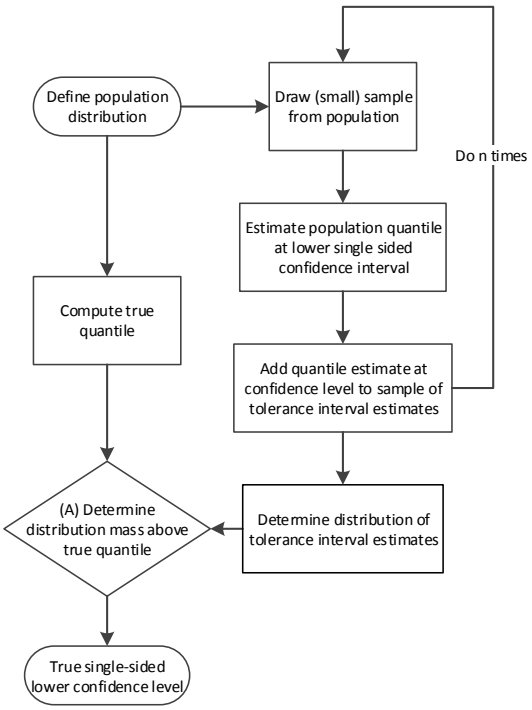


Figure 3.4: Summary of a procedure to test the accuracy of tolerance interval estimators**.** Step (A) is also illustrated in Figure 3.5.

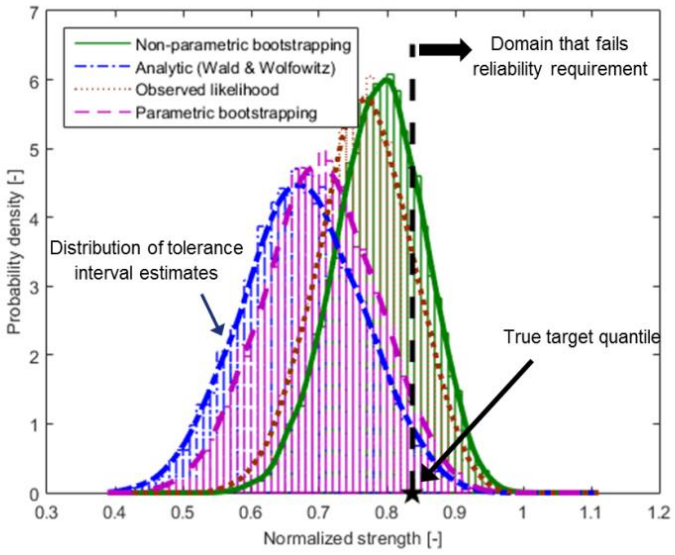


Figure 3.5: Simulation result showing the precision and accuracy of selected estimators to estimate a 10^{-3} (95%) tolerance interval using a sample with size six from a lognormal distribution with $\mu = 0$, $\sigma = 0.058$.

The verification test is performed for several reliability requirements targeting probabilities of failure of 10^{-3} to 10^{-6}, each time with a targeted confidence level of 95% and for a realistic range of σ_{10} from 0.015 to 0.085 (and $\mu = 0$). For all of these test cases, the verified confidence levels are similar to the results in Table 3-2. These results indicate that not all tested methods consistently meet the targeted reliability requirement. Under small sample size conditions, only the analytical method by Wald & Wolfowitz, the analytical method by ESDU 91041 and parametric bootstrapping accurately meet the targeted reliability requirement. As is already

well known in the state-of-the-art, and here again demonstrated, care should thus be taken when selecting a tolerance interval estimator.

Note that the estimator in chapter 4.1 of the NATO AGARD-AG-292 Helicopter Fatigue Design Guide [16] assumes that the true variance is equal to the sample variance and that this assumption leads to inaccuracy in this test. Nevertheless, the inclusion of the AGARD-AG-292 estimator in the verification test implicitly reveals that if there is no prior knowledge about the variance of the population, and given the modelling assumptions in section 3.2.1, then uncertainty about the population variance is the overall main driver for uncertainty about fatigue strength and about the reliability of a working curve.

Table 3-2: Table showing the confidence levels that can be demonstrated for different tolerance interval estimators estimating the same $\gamma = 10^{-6}$ quantile of a lognormal distribution with $\sigma_{10} = 0.015$. Tabulated is the rounded percentile of estimated quantiles that meet the γ-quantile requirement for a target of $\chi = 0.95$. The reliability test uses 10^3 Monte Carlo samples. [12]

Estimator↓ Sample size→	**2**	**3**	**4**	**6**	**8**	**11**	**16**	**100**
Analytic (Wald & Wolfowitz)	96	96	96	95	95	96	97	96
Analytic (ESDU 91041)	4	97	95	94	93	94	95	94
Observed likelihood	71	78	80	83	84	88	89	91
Empirical likelihood profile	69	80	83	86	86	90	91	92
Parametric bootstrap	96	95	95	94	93	94	95	94
Non-parametric bootstrap	n.a.	48	62	75	79	86	89	93
Analytic (AGARD-AG-292)	45	50	53	56	59	59	60	63

3.3.2 Benchmarking of classic tolerance interval estimation methods for fatigue strength

After positive verification of the analytic tolerance interval estimators by Wald & Wolfowitz, ESDU-91041 and by parametric bootstrapping, Figure 3.6 presents tolerance intervals computed using Wald & Wolfowitz for a common 'six-nines' reliability requirement [57], a range of sample sizes and realistic[13] standard deviations for fatigue strength.

It is well-known that for identical components and fatigue test results, values for conservative fatigue strength to obtain airworthiness certification can differ significantly between rotorcraft manufacturers [35]. Nevertheless, there is usually one clear commonality and that is the value of normalized fatigue strength that would be used in the worst-case, in which no component fatigue tests are available at all and when the S-N curve must, for example, be determined based on handbook material data only. For this case, accepted means of compliance to rotorcraft airworthiness requirements CS-27/29 by means of AC-27-1B MG 11 [68] prescribe a generic value of 1/3 for normalized fatigue strength. Lower values, as for example computed using the method of Wald & Wolfowitz in Figure 3.6, can, therefore, be regarded as too strict or even unrealistic. Although the tolerance intervals computed using the method of Wald & Wolfowitz for normalized fatigue strength can be used to substantiate the reliability of a conservative working curve, they can be considered as too strict and unrealistic for very small sample sizes.

Even if somewhat larger sample sizes are available, then the tolerance intervals in Figure 3.6 are still unusually low. According to the standard airworthiness certification guideline AC-27-1B MG 11, common values for

[12] Percentiles presented in Table 3-2 are estimated by 10^3 Monte Carlo samples and rounded to the nearest integer. Before rounding, and under assumption of a normally distributed error distribution and using equation (3.4), results are considered accurate to ±1.4% (with 95% confidence) or ±2.1% (with 99.7% confidence) for a 95% percentile estimation.

[13] Based on analysis of the Airbus-proprietary dataset summarized in Figure 3.8.

reduction factors in the rotorcraft industry to ensure the reliability of a working S-N curve are between 0.5 and 0.75 for metallic alloys, if four or more component fatigue tests are available.

Some reliability substantiation methods prescribe the empirical mean fatigue strength with the bottom-of-scatter, as for example in AGARD-AG-292, and target a lower reliability in return. However, on average, translation of the mean to the bottom-of-scatter negligibly contributes to substantiated reliability. Straightforward simulations, using the same methodology as in section 3.3.1, verified this and illustrated that for small sample sizes, the statistically expected bottom-of-scatter lies close to the mean.

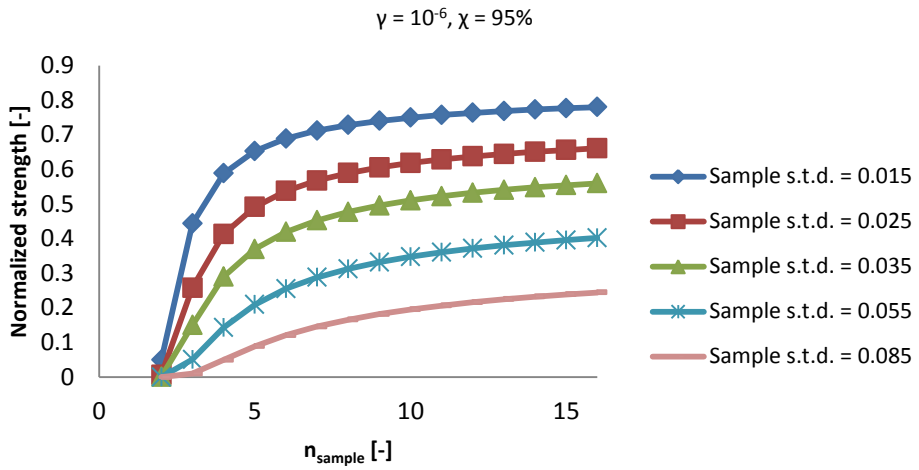


Figure 3.6: Graph showing generic reduction factors for fatigue strength that meet a $\gamma=10^{-6}$ (95%) computed with the analytical method by Wald & Wolfowitz. "Sample s.t.d." denotes the sample estimate of the standard deviation $\hat{\sigma}_{10}$. of normalized fatigue strength.

3.4 Introduction of Bayesian statistical analysis for tolerance interval estimation of fatigue strength

A clear shortcoming of frequentist tolerance interval estimators is their inability to account for information from sources other than results from full-scale component fatigue tests of the specific component in question. Especially if very few test results are available, frequentist methods, therefore, stipulate too low intervals. Such behaviour is fully in line with the mathematical situation: few test results provide very little knowledge about the actual strength, particularly concerning its variance. Thus, to substantiate a strict reliability requirement mathematically, the tolerance interval on strength must be extremely low - to counter that it cannot be ruled-out that variance may be very high.

In practice though, there is often already some well-founded expectation on the fatigue strength of a component. For example, results from material coupon tests or full-scale component fatigue tests from similar components can already give information on what are reasonable values for fatigue strength and its variance, even before the specific component in question has been tested. Bayesian statistical analysis provides an explicit framework that allows taking such prior knowledge into account when computing tolerance intervals.

The work of Edwards & Pacheco [69] already introduced a method to estimate confidence intervals for a most-likely estimate of a log-linear S-N curve by means of Bayesian statistics and a non-informative prior. An advantage of their method is its ability to take results from run-outs into account accurately. The specific work does however not extend to the use of informative priors and only substantiates a very high level of confidence for the most likely S-N curve and does not estimate actual tolerance intervals.

Guida & Penta [61] previously also identified and demonstrated the potential of Bayesian statistical analysis and introduced a Bayesian approach for tolerance interval estimation for S-N curves. The method uses a prior

based on material data from coupon tests and includes explicit modelling of uncertainty in the estimated shape parameters of the S-N curve. Their approach is however not readily applicable if compatibility with chapter 4.1 in AGARD-AG-292 shall be kept. Therefore, the following introduces an alternative and more simplified methodology. This simplified methodology will be used to exemplify the value of Bayesian analysis for rotorcraft fatigue life prediction.

3.4.1 Introduction to modelling approach for Bayesian estimation of fatigue strength quantiles

If parametric bootstrapping as previously introduced in section 2.2.2 or in appendix B.6 is used to estimate a tolerance interval of lognormal distributed fatigue strength, then Bayes' Theorem can be used to impose a prior on the uncertainty distributions of the standard deviation and mean. To develop a prior on the standard deviation, it is assumed that:

- there exists one random population of the standard deviation of normalized fatigue strength
- this distribution can be estimated
- this distribution can serve as an appropriate prior on the uncertainty distribution of the standard deviation of the fatigue strength of a component at hand

Attempting to formulate a prior on the mean can be unattractive for several reasons:

- The uncertainty distribution of the mean is actually dependent on the sampled standard deviation and such coupling would lead to complications in determining its likelihood distribution.
- The influence of uncertainty about the mean on an estimated tolerance interval is actually modest, as implicitly demonstrated in section 3.3.1.
- Defining a prior on the mean fatigue strength likely requires a detailed case-by-case analysis and may be associated with a considerable effort.

The next sections outline the development of a methodology to set a prior on the standard deviation of lognormal distributed fatigue strength.

3.4.2 Introduction of Bayes' Theorem

The uncertainty distribution for a population standard deviation σ, also designated as the posterior distribution, can be estimated using Bayes' Theorem, see also Box & Tiao [70]:

$$p(\sigma|\hat{\sigma},n,\alpha)=\frac{p(\hat{\sigma}|\sigma,n)\cdot p(\sigma|\alpha)}{p(\hat{\sigma}|\alpha)}\propto p(\hat{\sigma}|\sigma,n)\cdot p(\sigma|\alpha) \tag{3.5}$$

The posterior distribution of the population standard deviation σ is here denoted by $p(\sigma|\hat{\sigma},n,\alpha)$ and is conditional on the Maximum Likelihood Estimate (MLE) of the population standard deviation $\hat{\sigma}$, based on a sample of size *n*, as well as a prior expectation α.

The term $p(\hat{\sigma}|\alpha)$ may be considered as normalization constant and is independent of the 'function parameter' σ. It is therefore not necessary to explicitly compute the distribution $p(\hat{\sigma}|\alpha)$ itself as it may be replaced by an integration constant which can be estimated by straightforward numerical quadrature:

$$p(\hat{\sigma}|\alpha)=\int_{-\infty}^{+\infty} p(\hat{\sigma}|\sigma,n)\cdot p(\sigma|\alpha)d\sigma \tag{3.6}$$

The posterior distribution of σ is thus proportional to the product of two probability distributions:

- $p(\hat{\sigma}\,|\,\sigma,n)$, which models the probability of making MLE estimate $\hat{\sigma}$, given population standard deviation σ and drawing a sample of size *n*.
- $p(\sigma\,|\,\alpha)$, which defines the prior expected probability distribution of σ, given assumption α.

3.4.3 Introduction of the likelihood function

It is customary to designate the probability distribution $p(\hat{\sigma}\,|\,\sigma,n)$ as the normalized likelihood function. Despite its notation, this function can be regarded as a function of σ only. For a given trial value of the population standard deviation σ, it can be determined how probable it is that the sample estimate of the standard deviation $\hat{\sigma}$ is observed, either analytically or by Monte Carlo simulation. Computing this probability over an appropriate domain of σ values allows constructing a PDF of the likelihood function, telling how likely each potential value of the population standard deviation is, given the sample estimate at hand.

Basic Monte Carlo (BMC) simulation can be used to estimate the PDF of $\hat{\sigma}$ given a sample size *n* and a single value of σ, simply by drawing a large number of samples of size *n* from a normal distribution set by σ. Evaluating the resulting PDF at the coordinate of the sample estimate $\hat{\sigma}$ at hand yields one coordinate in the likelihood function in relatively intuitive but computationally costly fashion.

Alternatively, a semi-analytical method can estimate the PDF of $\hat{\sigma}$, given a sample size *n* and σ, much more efficiently based on the analytical distribution function of the biased sample estimate of the standard deviation of a normal distribution (Weisstein [71]):

$$p(\hat{\sigma}_B\,|\,\sigma,n) \propto 2\cdot\frac{\left(\frac{n}{2\hat{\sigma}_A^2}\right)^{\frac{n-1}{2}}}{\Gamma\left(\frac{1}{2}(n-1)\right)}\cdot\exp\left(\frac{-n\cdot\hat{\sigma}_B^2}{2\hat{\sigma}_A^2}\right)\cdot\left(\hat{\sigma}_B\right)^{n-2} \quad \text{with} \quad \hat{\sigma}_A^2 \equiv \frac{n\cdot\hat{\sigma}_B^2}{n-1} \tag{3.7}$$

where $\Gamma(z)$ denotes the Gamma function:

$$\Gamma(z) = \int_0^\infty e^{-x}\cdot x^{z-1}dx \tag{3.8}$$

and where $\hat{\sigma}_B$ is the biased estimator of σ:

$$\hat{\sigma}_B = \sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}\left[\hat{\mu}-\log_e(S_i)\right]} \tag{3.9}$$

and where conversion from the biased estimator of σ to the unbiased estimator follows from arithmetic manipulation of equations (3.9) and(3.3).

Figure 3.7 summarizes a procedure making use of equation (3.7) to compute the likelihood function before an additional normalization step estimates $p(\hat{\sigma}\,|\,\sigma,n)$. In its practical implementation, numerical quadrature and empirical distribution functions can be used for normalization and distribution fitting.

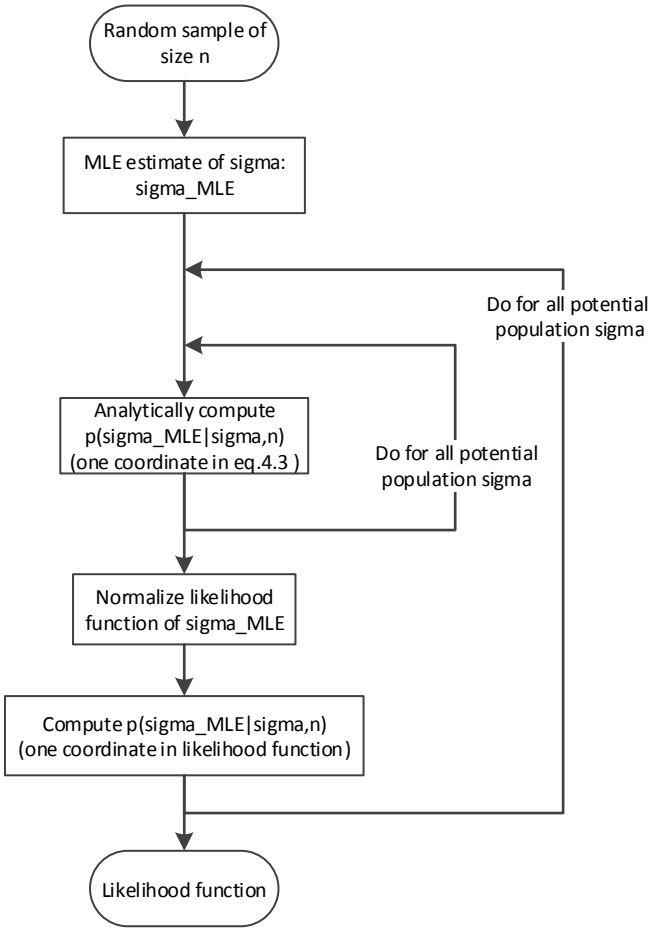


Figure 3.7: Diagram summarizing the implemented process to compute the likelihood function.

3.4.4 Setting a prior on the variance of fatigue strength

The prior should reflect all existing knowledge and expectations on the standard deviation excluding the actual test results. It is, however, difficult to develop a generic method that can take information from a wide variety of sources into account and merge this into a single PDF defining the prior. Prior information can come from engineering judgement, previous experience, tests results from similar designs, basic material data, etc. Due to the very nature of a prior, it is however sufficient to only require that the formulation of the prior is reasonable traceable, and accountable. Then, the prior is merely an explicit vehicle to collect and translate design assumptions until their influence diminishes automatically once applicable test results become available. For the here presented exemplary case study a relatively simple and highly flexible method to formulate a prior is developed. This method can readily be applied once a database of fatigue test results is available, or can easily be adjusted to accommodate other cases.

If results from many comparable full-scale component fatigue test programs are available, then this information can be used to formulate a prior expectation on the standard deviation to be observed in general. To set the prior, present work postulates, based on engineering judgement, that the average of uncertainty distributions for the standard deviation from *k* different but comparable test programs yields an appropriate prior expectation for the standard deviation:

$$p(\sigma|\alpha) \propto \frac{\sum_{i=1}^{k} p(\sigma|\hat{\sigma}_i, n_i)}{k} \tag{3.10}$$

where the uncertainty distribution $p(\sigma|\hat{\sigma},n)$ can be computed according to equation (2.8) or (7.7) in Appendix B.

A major advantage of this method to form a prior is its relative simplicity and ease of application. Another basic feature is automatic 'widening' and 'flattening' the prior if it is mostly based on test programs with very few samples. Also, it is straightforward to manually assign higher weights to prior test programs that are expected to be of particular importance.

The mathematical and statistical appropriateness of using a prior in the form of equation (3.10) was verified under a range of synthetic test circumstances in Appendix C. Additional verification comes from a comparison of results from two alternative methods developed to formulate a prior. Comparisons of Bayesian tolerance intervals estimated with priors from a more sophisticated, complex and computationally expensive approach in Appendix D, as well as from a highly simplified method, in the form of an upper and lower bounded uniform prior, all provided similar results. These similarities provide positive verification and indicate robustness and appropriateness of the exemplified method to formulate a prior.

3.5 Application of Bayesian statistical modelling to estimate fatigue strength quantiles of components of helicopter dynamic systems

A prior in the form of equation (3.10) is implemented using a real dataset from industry and applied to perform statistical substantiation of fatigue strength tolerance intervals for primary mechanical components in the dynamic system of rotorcraft. The application of Bayesian analysis provides major advantages here, especially in comparison to the situation demonstrated in section 3.3.2. There, frequentist methods yielded unrealistically low tolerance intervals on strength when imposing aerospace-typical reliability requirements with high levels of confidence.

3.5.1 Definition of a generic prior for the variance of fatigue strength

To set a prior, data from a large number of full-scale fatigue tests of selected components has been collected. For simplicity, the prior is not specialized according to, for example, type of material (e.g. steel, aluminum or composite), type of component (e.g. shaft, lug or bolt) or manufacturing properties (e.g. machining, forging or casting). In the example application presented here, it is assumed that all primary mechanical components in a helicopter dynamic system are similar and that the variances of the fatigue strength of such components are all comparable.

Despite this crude approach, the results of this example are nevertheless considered to be useful. Particularly because it ensures that the prior stays relatively uninformative. Since the prior 'mixes' component types with high and low variance in fatigue strength, this prevents unintended expectation biases and effectively reduces its role to just enforcing realistic upper and lower limits on fatigue strength variance. Additionally, the simple approach yields only one generic set of tolerance intervals that is realistic, straightforward to apply and corresponds to values already in common use in industry, as is demonstrated later in section 3.5.2. The use of standard reduction factors generically applied to a wide range of components is common and approved practice in the rotorcraft industry; see for example the generic 1/3 reduction factor in AC-27-1B MG11 [58] or the default reduction factors in chapter 4.1 in AGARD-AG-292 [16].

Here, the broad dataset with past test results is thus assumed to be representative for any randomly chosen component of the dynamic system of a rotorcraft. Nevertheless, it can also be envisioned to use an expanded and/or specialized and/or weighted database to define the prior and to meet individual analysis requirements. The current study merely aims to provide an example of the effectivity of simplified Bayesian analyses for statistical fatigue strength substantiation.

Figure 3.8 graphically summarizes the raw dataset by plotting the observed standard deviation from each set of component fatigue tests taken into account. Visual inspection of the fatigue test results does not clearly reveal the presence of data clusters and gives the impression that the data can be considered to roughly originate from one distribution. More importantly, the data also clearly demonstrates a significant spread in the variance of fatigue strength. This last observation underpins the importance of explicitly taking estimation uncertainties in fatigue strength variance into account. It should be noted though, that the dataset may still contain some over-conservative test results due to non-representative but conservative fatigue testing protocols.

It can also be observed that the estimated variance has a rising trend with sample size. This normal statistical behaviour was already observed in the consistent underestimation of population variance previously illustrated in Figure 3.3. The cause is that the estimator of the standard deviation is only an asymptotically unbiased estimator, as $n \to \infty$. For small sample sizes, the estimator of the standard deviation is biased significantly and non-conservatively.

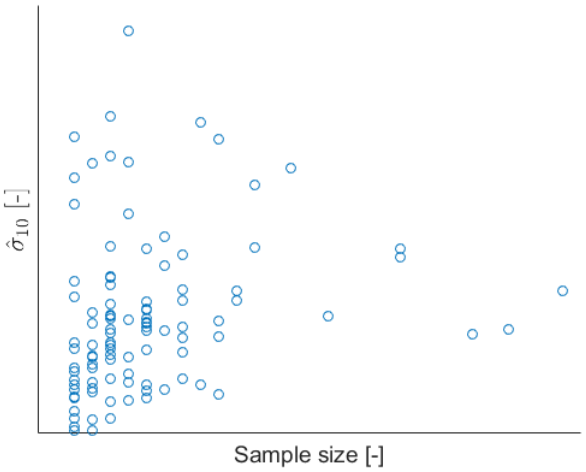


Figure 3.8: Scatterplot displaying the distribution of observed standard deviations from selected full-scale component fatigue tests.

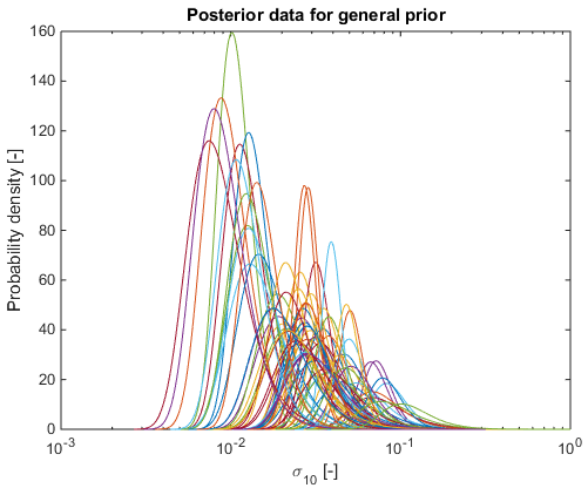


Figure 3.9: Overview showing the scale of the analytical estimation uncertainty distributions for all the estimated values of the standard deviation of fatigue strength shown in Figure 3.8. (Estimates based on less than four samples are not considered)

Estimates of fatigue strength variance based on only two or three fatigue tests have been removed[14] from the data that is used to formulate a prior expectation on σ. There are two main reasons for discarding these data points:

- There is a significant chance of over-fitting an expected S-N curve if only few test points are available, and thereby of underestimating scatter in fatigue strength. The expected S-N curves are defined by a Weibull function, e.g. as in chapter 4.1 in AGARD-AG-292 [16], which has a relatively high number of four fitting parameters. The special statistical analysis in Appendix E indicates that for sample sizes two and three the estimated variances are on average lower than statistically expected, even when correcting for the statistically expected estimation bias. The inclusion of this data can, therefore, introduce a non-conservative expectation bias.

[14] It was verified that nevertheless taking into account data points from two or three fatigue tests still leads to a similar prior distribution as the final prior in Figure 3.10. The major resulting difference lies in the 99.7% and 0.3% percentiles of σ, which would increase by approximately 50%.

- Estimates based on only two or three tests are statistically not very informative, potentially leading to unnecessarily heavy tails of the prior distribution which their data contributes to.

The uncertainty distributions of all remaining σ-estimates that can be computed using equation (2.8), or equation (7.7) in Appendix B, and are shown in Figure 3.9. These uncertainty distributions are subsequently averaged, according to equation (3.10), to form a general prior expectation on the standard deviation σ. The final averaged prior distribution is presented in Figure 3.10.

This final prior distribution is somewhat irregular in its centre domain. This is simply a result of the data at hand. It may be that this irregularity is caused by a 'mixing' of data that in fact come from two separate populations, e.g. two different material or component types. This possibility is however not further investigated and possible consequences are assumed negligible in present work. For future work, it is recommended evaluate the use of specialised priors. If the prior would however indeed result from mixing of two distinct distributions, then this would only increase the generality of the prior and be in line with approved design methods, see for example the generic 1/3 reduction factor in AC-27-1B MG11 [58] or the default reduction factors in chapter 4.1 in AGARD-AG-292 [16].

Practical application of the prior to a range of real components demonstrated smooth and realistic posterior distributions[15] supporting the premise that the effect of slight irregularities in the prior is negligible. Such behaviour is also expected, since the prior should principally only convey an approximate, but explicitly documented, design assumption bounding results to a realistic and commonly accepted domain. Its detailed distribution properties should therefore naturally be of less importance. Otherwise, it may also be expected that addition of more data smooths-out the prior distribution, or that tailored and component individual priors should be used instead.

The final prior is relatively wide and uninformative for base-10 standard deviations between 0.01 and 0.1 and 'drops off' sharply outside this domain. Its 35th lower percentile of 0.027 approximately corresponds to the minimum acceptable value for $\hat{\sigma}_{10}$ of 0.025 implemented in AGARD-AG-292 [16] and indicates its practical value for the industry. It is re-empahsised that these values are the result of reproducible and systematic treatment of a representative and relatively large dataset. And that they have been obtained without applying any implementing any adjustments to data filtering and processing procedures to bring the obtained results in line with common industry practises.

[15] Implementation used the same Airbus dataset used to define the prior and was done for the components listed in section 4.2.3.3. Specific results are Airbus proprietary.

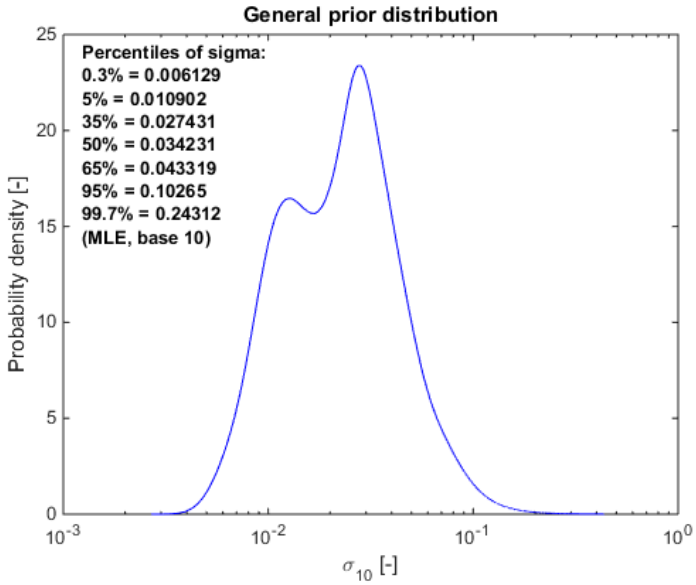


Figure 3.10: Distribution plot showing a generic prior expectation on the standard deviation in fatigue strength.

3.5.2 Computing generic tolerance intervals for normalized fatigue strength

With a prior for the standard deviation of fatigue strength defined, corresponding tolerance intervals for fatigue strength can be computed by a method similar to frequentist bootstrapping in B.6. The implemented variant is summarized in Figure 3.11. All of the results are computed according to the simple and broad prior in Figure 3.10. Basic Monte Carlo simulation during step (A) in Figure 3.11 was executed with 10^5 samples. This leads to negligible estimation errors according to equation (3.4). If only one fatigue test is available, which is a special case not covered by the summary in Figure 3.11, then the uncertainty distribution for the variance in fatigue strength is determined by the prior alone, whereas the uncertainty distribution for μ is computed while assuming that two test results are available. Assuming a sample size of two is here deemed acceptable since the uncertainty distribution of μ has a limited influence on the overall tolerance interval and is even forced in the absence of a prior on μ.

Resulting tolerance intervals for a typical reliability requirement of $\gamma = 10^{-6}$ and $\chi = 0.95$ and various sample sizes are given in Figure 3.12. Comparison with the tolerance intervals according to frequentist methods in Figure 3.6 displays one of the major advantages of using Bayes' Theorem. The introduction of a prior successfully imposes an upper bound on σ and thereby a lower bound on the tolerance intervals. Even if very few full-scale component fatigue test results are available to estimate a tolerance interval, these remain bounded to the feasible domain and do not approach zero.

It is noticeable that in some cases the estimated tolerance interval drops with increasing sample size, even though the observed standard deviation remains constant. This is a normal consequence of the use of a prior expectation. In these rare cases, the observed variance in strength is significantly higher than expected according to the defined prior. Consequently, expectations have to be adjusted for the worse as more test results become available to correct the initially too optimistic prior assumption. Naturally, the probability of initial over-optimistic expectations increases with decreasing confidence levels, as a comparison with the same 10^{-6} quantiles estimated with a less strict confidence level of $\chi = 0.5$ in Figure 3.13 illustrates.

Coincidentally, for sample sizes zero and one, the tolerance interval for a reliability requirement of $\gamma = 10^{-6}$ and $\chi = 0.95$ for normalized fatigue strength approximately coincides with the default reduction factor in AC-27-1B MG 11 [68] of 1/3, as in Figure 3.12.. That the stipulated value of normalized fatigue strength can

subsequently drop below 1/3 does however not necessarily imply that the exemplified Bayesian method to substantiate fatigue strength is incompatible with AC-27-1B MG 11. When using the prior for computing tolerance intervals for a less strict reliability requirement of 10^{-6} (50%) then the results in Figure 3.13 demonstrate that the use of a reduction of 1/3 will virtually always meet this reliability requirement. This is important because the use of such a reduced confidence level requirement is approved practise in industry; see for example Everett [35] or Beale *et.al.* [72]. Currently, neither AC-27-1B [68], AC-29-2C nor ARP 4761 [33] prescribe the use of a specific confidence level.

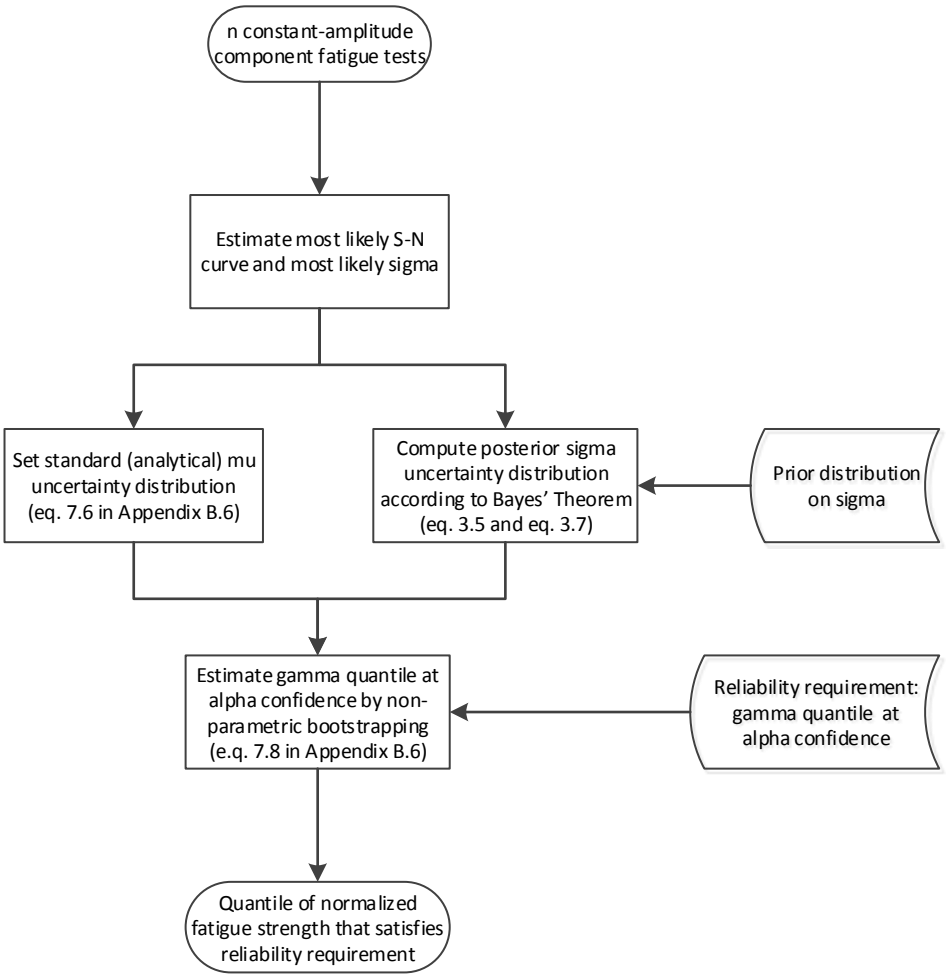


Figure 3.11: Process summary how to compute a conservative value of normalized fatigue strength by Bayesian analysis if the size of the available sample is larger than one (n > 1).

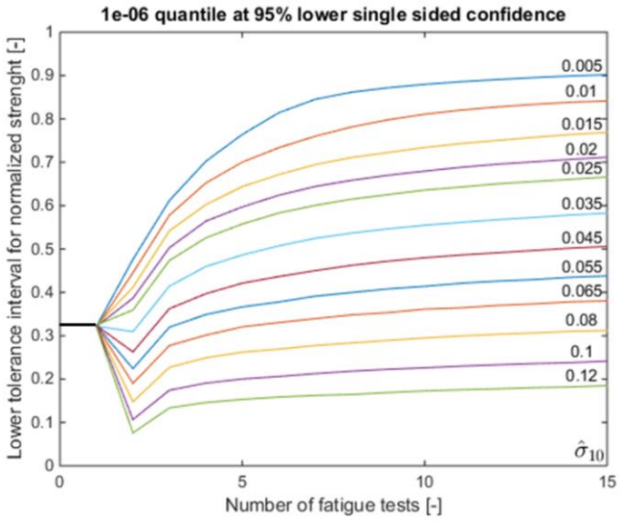


Figure 3.12: Graph showing generic tolerance intervals for normalized fatigue strength according to Bayesian statistical analysis and for a $\gamma=10^{-6}$ (95%) reliability requirement.

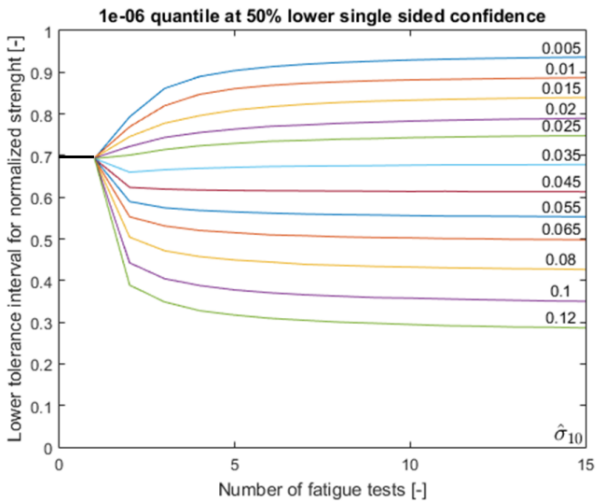


Figure 3.13: Graph showing generic tolerance intervals for normalized fatigue strength according to Bayesian statistical analysis and for a $\gamma=10^{-6}$ (50%) reliability requirement.

3.6 Conclusion

It is exemplified that common frequentist statistical analysis for tolerance interval estimators to substantiate a safe design value for fatigue strength may not yield accurate results when subjected to aerospace-typical small sample size conditions. Moreover, if strict and explicit tolerance interval requirements are enforced, then these non-Bayesian methods stipulate unrealistic strength values approaching zero.

The value of Bayesian statistics as an alternative is demonstrated by means of a case study using simplified modelling assumptions. Based on data from other fatigue tests or even expert judgement and experience, the Bayesian analysis allows defining and taking into account, a traceable and explicit expectation about the fatigue strength of a component. Thereby, even if very few fatigue tests are available for a particular component, its substantiated fatigue strength can still be bounded to the realistic domain and explicit statistical modelling can still substantiate a strict reliability requirement. This can bypass the need for using implicit, difficult-to-trace, and difficult-to-justify engineering assumptions.

The Bayesian method for substantiating fatigue strength is relatively generic, simple to use, and yields results and methodologies compatible with well-established aerospace design practises; such as chapter 4.1 in AGARD-AG-292 or AC-27-1B MG 11. The exemplified approach complements more elaborate approaches, such as by Guida & Penta [61] and its results encourage further research into the use of Bayesian statistics for substantiation of aerospace fatigue strength.

Future work could include further study into the use of different types of priors and modelling assumptions, as well as updated and expanded datasets for priors specialized for different types of components and materials. Moreover, it is recommended to compare results from the simplified model with the more elaborate model of Guida & Penta [61] in a rotorcraft context. Finally, it is recommended to investigate the use of a combination of knowledge from different fatigue life prediction methodologies to formulate a Bayesian prior, for example, alternative S-N-P models, crack growth models, or other material data.

4 Virtual Fatigue Life Monitoring

Using Virtual Fatigue Life Monitoring (VFLM), in-service loads are not measured directly but only estimated. This can introduce estimation errors and this additional source of error must be accounted for when substantiating the reliability of VFLM-based individual SLLs. Based on the modelling framework discussed in chapter 2, two methods for VFLM are introduced. The models are based on a generically applicable load prediction framework that allows for the presence of prediction errors. Direct Load & Damage Modelling (DLDM) is introduced as a simplified implementation where the influence of load prediction errors can be regarded as negligible in comparison to uncertainties about the fatigue strength of the involved part As a more accurate and universally applicable alternative, Probabilistic Load & Damage Modelling (PLDM) is introduced as well. PLDM models and mitigates the influence of combined uncertainty from random fatigue strength and random load prediction errors and is applicable regardless of achievable DLDM prediction accuracy and precision. The accuracy and reliability of their fatigue life estimates was verified using more than one thousand recorded flight hours from two commercially operated helicopters equipped with strain gauges to independently verify in-flight load predictions. Moreover, the potential fatigue life extensions were demonstrated for several components in the dynamic system of three helicopters operated under an Emergency Medical Services (EMS) profile.

4.1 Overview of the State-of-the-Art for Virtual Fatigue Life Monitoring

The development of prediction models and reliability substantiation methods for VFLM is an active field of research, also motivated by some practical results summarized in section 1.5. Examples of comprehensive reviews outlining the classic concepts and challenges of VFLM are published by ADS-79D-HDBK [73] and Wallace *et.al*. [6]. Recent work by Beale & Davis [10] presents a comprehensive treatment of an end-to-end implementation for VFLM by FRR for commercial implementation.

There are two principle implementation strategies for VFLM: by means of Flight Regime Recognition (FRR) or by direct load prediction. An overview of prior work on FRR and direct load prediction for VFLM will be given in sections 4.1.1 and 4.1.2 respectively, whereas section 4.1.3 briefly discusses the potential of physics-based VFLM models. The need for additional measures to guarantee the reliability of VFLM-based reliability has been widely recognized and will be discussed in section 4.1.5.

4.1.1 Review of Flight Regime Recognition for Virtual Fatigue Life Monitoring

An initially intuitive method to implement Virtual Fatigue Life Monitoring is by replacing the sequence and timeshare of flight regimes in the Design Mission Profile by the flight regimes that actually have been flown. This methodology is referred to as Flight Regime Recognition (FRR). Many methodologies have been developed, tested and published to implement FRR. These include linear regression and nearest neighbour interpolation [21], artificial neural networks [74, 75, 76], binary decision trees [77, 78], hybrid methods [23, 10] and hidden Markov Models [79]. Though not covering recent developments, a review of classic results and methodologies has been published by Wallace *et.al.* [6], which covers the traditional implementation approaches.

Isom *et.al.* [80, 81] have presented FRR recognition results making use of a new and dedicated sensor to provide extra information on the state of the main rotor and reported clear benefits in terms of recognition accuracy. The use of custom sensors estimating aircraft weight has been demonstrated as well, e.g. by Beale & Davis [10]

4.1.1.1 Review of statistical Flight Regime Recognition

Commonly, the flight regimes in the DMP are defined heuristically based on, for example, usage surveys, pilot interviews, and expert judgement. This means that there is no strict mathematical mapping between a multi-dimensional flight parameter space and the set of flight regimes defined in the DMP. To address this issue, there has been a wide use of statistical classifiers to ‘learn’ this implicit mapping from Load Classification

Flights, where many elements of the DMP are repeatedly flown, by means of non-linear statistical data modelling. This approach results in easy-to-interpret results, which should directly allow creating a customized DMP for an individual helicopter.

Although some publications report consistent recognition accuracies in the range of 95%-100% for this type of FRR, these may not be reproducible due to inadvertent overspecialization of the models involved. As an example, the method published by Hoffmann *et.al.* [23] was tested with data previously unseen by the models, resulting in significantly lower recognition rates than originally published. The replication efforts included rigorous and manually verified database cleaning and independent application and testing of the original models as well as newly developed ones. Newly developed models included improved ANN-based models, Binary Decision Trees, Deep Learning, advanced feature generation, and probabilistic sequence recognition models similar to modern speech recognition algorithms. Tests have been performed across two different helicopter types with similar results.[16]

Amongst others, successful FRR-based VFLM implementation by statistical learning can be difficult due to the following limitations:

- Flight regime recognition algorithms often only predict flight manoeuvres and usually cannot accurately determine an aircraft's weight and centre-of-gravity during a flight. These parameters have a strong correlation with flight loads. Dedicated algorithms, manual pilot estimations or dedicated sensors are thus required for actual flight regime recognition.
- The low frequency or Ground-Air-Ground (GAG) load cycle determination procedure is very sensitive to the number of recognized flight regimes. This phenomenon was replicated by unpublished research work carried out at Airbus Helicopters Germany in the framework of present work. It was observed that the number of GAG cycles may be severely overestimated due to regime misrecognition, recognition results from poorly defined flight regimes, regime transitions [82], or regime toggling effects. The development of special post-processing algorithms [83, 22, 10, 84] or advanced sequence recognition algorithms (unpublished) can provide a solution to such undesired regime toggling.
- Flight regimes may not be defined as precise and unique mathematical events but rather as qualitative and approximate events open to interpretation. This introduces inaccuracies and subjective data and complicates or hinders the establishment of objective and reliable 'truth-data' by which maneuverer recognition can be tested and validated.
- Some organisations make use of manually and heuristically defined flight regimes that are not guaranteed to uniquely cover the entire flight envelope. Mathematically defined regimes, e.g. which divide the entire flight parameter space into hyper cubes, generally do not suffer from such problem though. Such an approach can be considered as a look-up table variant of machine-learning based FRR. Although the accuracy of mapping from flight data to the appropriate hyper cube can be guaranteed, the accurate mapping to actual loads cannot since it can be expected that significant load scatter exists within a hyper cube. If a large number of hyper cubes would be defined, and load scatter would thus be less, then method can be considered as a look-up table variant of direct load prediction.[17] It is however unlikely that enough flight data can be gathered to make such a mapping

[16] It was concluded that the likely cause of overspecialization in Hoffmann *et.al.* [23] lies in the use of binary decision trees post-processing initial classifications by an ANN. These trees are likely developed without ensuring statistical independence between the data by which the models have been created and the data by which their performance has been tested.

[17] Also note that simple rule-based conservative substitution may not be possible without additional and numerically expensive analysis of which load case would yield more conservative results. Due to the non-linearity of a fatigue surface, it cannot be readily known if an increase in the mean or the amplitude of a load cycle causes more damage. In addition, if rainflow counting is used for cycle counting, then it cannot be readily known to which cycle counted load cycle a particular load sample will belong, as this depends on the load case substitution of the load sample itself but also others.

accurately and for all defined hypercubes, and if so, that it would perform better than direct load prediction.

- Manoeuvre loads, and also the associated in-flight characteristic features, vary with amongst others: differences in manoeuvre execution, atmospheric conditions and mechanical differences between aircraft of the same type. For example, manoeuvre loads can vary due to pilot technique and machine-to-machine variablity can be caused by assembly tolerances, ageing effects, and control system rigging.
- FRR can be seen as an aggregation step after the recognition of a sequence of relatively short duration flight states, for example as implemented by Bates *et.al* [82]. Machine learning classification of flight states is difficult since features of short flight elements may be assigned to multiple manoeuvres. Aggregation methods, e.g. by hidden Markov Modelling, may be able to handle uncertain classification but are generally hampered by an inherently small dataset of examples. Since FRR classification is a manual and expensive process and often requires a test pilot to interpret and execute prescribed manoeuvres, FRR algorithms are usually only trained and validated using a relatively low number of examples per manoeuvres, i.e. 1-10. The application of complex machine learning models, preferably in combination with error modelling and test requirements involving an independent dataset generally demand large datasets though.

4.1.1.2 Review of definition-based Flight Regime Recognition

Alternatively, it is possible to define a mapping between the flight parameter space and flight regimes by heuristic rule-based classification. Usually, this is done by simple rule-based methodologies and may effectively result in dividing the parameter space into hyper-cubes and assigning a flight regime to each hypercube. If it is required that the mapping shall point to manually and heuristically described flight regimes, then rule based classification may however not be able to cover all regimes. Recognition of complex and highly dynamic manoeuvres can be challenging by simple rule-based methods [10, 22]. However, if covered regimes are simple, or if the mapping does not need to correspond to intuitive manoeuvre descriptions, then recognition accuracy is guaranteed since the regimes are simply defined according to their recognition criteria.

4.1.1.3 Introduction to mission profile classification

Using the fatigue life prediction and substantiation methods analysed in chapter 2, a SLL can be expressed as a maximum number of flight hours (FH). As mentioned before, it is possible to define several DMP's, each covering a different type of mission and to let operators assign a DMP to each flight. Such a mixing of DMPs is not analysed in this work. As a benefit though, such an approach would be relatively simple to implement and operate. However, due to the coarseness of the model the benefits would be less than could be attained with VFLM. In addition, the approach would likely still require the use of conservative sub-DMPs and include unnecessary conservatism or complexity to mitigate that DMPs are usually only valid over longer periods of time (>> 10FH).

4.1.2 Review of Direct Load Prediction for Virtual Fatigue Life Monitoring

4.1.2.1 Introduction to Direct Load Prediction

Flight regime recognition essentially takes a two-step approach to reconstructing in-flight loads: first from flight parameters to flight regimes and then from flight regimes to loads. First, each section of flight is classified to a flight regime. Then, the corresponding load spectra for these regimes can be assigned by the same methods used to define load spectra for classic service life limits, as for example implemented in chapter 2. Since the classifiable regimes are often the same as the regimes that make out the DMP, the second step can generally be done using tools and methods already developed.

Alternatively, flight parameters can be mapped to in-flight loads directly, if regression is applied to loads instead of classification to regimes. This approach is referred to as Direct Load Prediction (DLP). The primary

conceptual advantage of DLP over discrete approaches such as classic FRR lies in the ability to use a continuous scale. Whereas FRR is limited to classifying loads by the number of recognizable regimes, DLP should allow modelling of even small variations in loads. The resolution of predicted loads can be important since differences in loads can have highly non-linear effects on overall fatigue life. DLM may even be able to capture load variations due to for example pilot technique, atmospheric conditions, and aircraft weight. Naturally, the only fundamental limitation is the actual correlation between recorded flight parameters and in-flight loads. A practical advantage of DLP is that prediction generally takes place in short timeframes, e.g. with a one-second duration. There is thus automatically a large database available with examples to 'learn' the correlation between flight parameters and loads. The number of available examples can even be increased by oversampling without running into overspecialization issues, as demonstrated by Dekker [21]. The effective increase of unique events that can be used to build a regression model can significantly improve modelling accuracy and precision.

4.1.2.2 Review of current methods for Direct Load Prediction

Despite major industry-focus on FRR-based VFLM developments, an early comparative review by Wallace *et.al.* [6] already suggested that DLP features many advantages over FRR and would be most attractive for future development. An early example of DLP was presented by Haas *et.al.* [85]. Artificial neural networks (see Appendix H for an introduction to machine learning and artificial neural networks) were used to predict in-flight loads. Loading was modelled by a constant amplitude loading block with a single and component-generic characteristic frequency and with the duration of a single rotor revolution. The prediction involved the amplitude and mean load for each loading block. Typical results showed correlation coefficients in the range of 84% to 97% for high-speed manoeuvres and selected components; a reference model using linear regression demonstrated correlation coefficients of 79% to 95%. Follow-up work involved improvements to the regression methodologies and expanded test conditions to low-speed manoeuvres for which correlation coefficients as low as 49% were obtained [86, 87]. Similar results have been presented by Cabell *et.al.* [88], by David *et.al.* [89, 90] for airplane tail loads and by Allen *et.al.* for fighter jet wing loads [91]. Polanco presented a comprehensive overview of early DLP methods for helicopters [92], followed by Wallace *et.al.* [6], amongst others identifying that none of the work on DLP so far included sufficient reliability substantiation to let it be used for virtual fatigue life monitoring.

More recently, Bendisch *et.al.* [93] introduced a discretized variant of the classic approach by Haas *et.al.* [85]. Bendisch used high-frequency loading blocks with a characteristic frequency and bounded by predicted and discrete minimum and maximum load classes. Another advanced and recent implementation of direct load monitoring has been presented by Isom *et.al.* [80] and Beale [11] where high-frequency load signals are reconstructed by regression to principle components of load signals (see Appendix H for an introduction to Principle Component Analysis). Cheung *et.al.* [94, 95, 96] have also presented several recent examples of DLP demonstrating good regression results with artificial neural networks. Recent work on aircraft digital twins and associated virtual crack growth monitoring solutions, such as presented by Rudd [97], are also closely related to VFLM.

A form of DLP is used in the implementation of Direct Load & Damage Modelling, a new method for VFLM introduced in section 4.2.

4.1.3 Review of physics-based in-flight load reconstruction for Virtual Fatigue Life Prediction

Physical modelling could be considered as an alternative to machine learning for predicting in-flight loads. Recorded flight parameters could then be used to constrain the model and a solution for the corresponding load could be found, as for example proposed by Prendergast [98] using a surrogate interpolation model with pre-computed solution points. Although such an approach might at first seem attractive as it would make use

of well-established principles of physical modelling, this approach has not been pursued in present work. Based on analysis and interviews with experts [99], the following argumentation supports this decision:

- The current State-of-the-Art in physical modelling does not provide high accuracy predictions for transient and high-load flight conditions. Especially these conditions are likely to have a major contribution to the accumulated load spectrum.
- The effort of developing appropriate physical models is likely to be high; especially in comparison to machine learning models where standard methods can be used.
- The computational costs of running a statistical regression model are low once the model has been pre-generated. For a physical model, it is expected that a computationally expensive problem needs to be solved for each time element under consideration.
- A physics-based model is not necessarily a statistics-free model. It is likely that essential modelling parameters still need to be tuned based on test data so that the physics-based model can also be interpreted as an expert- or knowledge guided regression model. In addition, a physics-based model will likely also yield estimation errors that need to be accounted for by statistical methods.
- It can be argued that the complexity of a physics-based model is relatively high. The practical difference in terms of model understanding between 'black-box' machine learning models and physics-based models may therefore disappear.

4.1.4 Review of direct load measurement for Fatigue Life Monitoring

Virtual fatigue life monitoring solutions making use of custom and integrated sensor networks such as presented by Hajek *et.al.* [100] can significantly improve prediction accuracy, especially for components on or near rotor controls, as also demonstrated by Isom *et.al.* [80, 81]. However, they may add significant cost, weight and sensor maintenance.

4.1.5 Overview of existing methods for reliability substantiation of Virtual Fatigue Life Monitoring

Practical implementation of VFLM requires an analysis of the entire end-to-end process, as for example reviewed by Augustin [101], required by FAA AC-27 MG-15 [58], and recently illustrated by Beale & Davis [10], or exemplified by Larder *et.al.* [102] for vibration health monitoring. This chapter is limited in scope and only addresses reliability substantiation requirements in the framework of AC-27 MG-11, as chapters 2 and 3 did. This means that modelling or prediction uncertainties concerning in-flight loads, usage, and fatigue strength are analysed and mitigated to demonstrate a sufficient reliability level for VFLM-based SLLs, e.g. $\gamma = 10^{-6}$ (95%), with reliability defined as 1-γ with a single-sided lower confidence level of (α) . End-to-end process analysis and operational system safety analysis are mostly neglected in this work and it is assumed that their effects on VFLM-based service life limits can be mitigated entirely by means of operational, monitoring or watchdog processes, in addition to system safety assurances, e.g. by DO-178C, DO-254, DO-232 or Design Organisation Approval quality assurances.

There is currently one case in which extensions of service life limits have been granted based on analysis of recorded in-service flight data. An S-92 main rotor hub component whose service life limit depends on the time spent in a single, steady-state, flight regime, easily recognizable by definition-based flight regime recognition, has received VFLM-based SLL extensions. In this case, Beale & Davis [10] and Adams & Zhao [103] have presented a methodology to substantiate the reliability of-VFLM-based SLL extensions for a fleet using definition-based flight regime recognition. Based on simulation results from Thompson & Adams they estimated that substitution of the conservative design mission profile would reduce the reliably of service life limits by at least one order of magnitude.

The reliability of their classic service life limits relies on a reliability contribution based on the conservatism of the design mission profile, despite the uncertain effects this can be expected to have on helicopter individual reliability, as discussed earlier in section 1.3.2. In any case, substitution of the conservative design mission profile by actual individual usage thus results in a general reduction of SLL reliability. As a solution, they introduced multiplication factors to be added to usage recognitions such that reliability would still meet a $\gamma = 10^{-6}$ reliability requirement. The multiplication factors are based on the observed fleet-wide distribution of usage and are independent of individual recognition results.

Since this method relies on the assumption of perfect recognition from definition-based FRR, it is challenging to modify this reliability framework to apply to statistics-based VFLM methods which inherently feature estimation inaccuracies. The fatigue life model used by Beale & Davis [10] and Adams & Zhao [103] practically relied on constant amplitude load block modelling and a random model similar to the common model summarized in Figure 4.1 (repeated from chapter 2) and is thus not compatible with the load spectrum model from chapter 4.1 in AGARD-292 [16] and used in chapter 2.

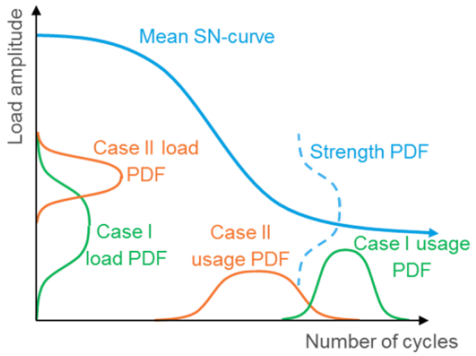


Figure 4.1: Schematic graphs illustrating how the probabilistic modelling framework used in many recent SLL reliability models depends on the statistical definition of load cases.

For statistical FRR-based VFLM application for military applications, ADS-79D-HDBK [73] provides some guidelines on reliability requirements. Herein, it is recommended to demonstrate that any misrecognized or unidentified flight regime should correspond to, or result in, a conservative regime classification. In addition, it is suggested that a classification accuracy of 97% should suffice to achieve a 99.5% reliability of VFLM-based SLLs. The guideline does however not specify how the recognition rate should be computed or demonstrated.

Simulation-based estimations by Hong [104] defined a correlation measure working with the difference between FRR-classified loading blocks and point-by-point load signal predictions based on linear regression. Hong indicated that under this measure, a regime recognition requirement of 95% may actually result in 10% underestimation of predicted fatigue life. Hong also re-iterated the practical difficulties of performing independent verification of flight regime recognition statistics due to the commonly subjective nature of their definitions, and that regime recognition accuracy over 95% may not be achievable.

4.1.6 Identification of requirements to improve Virtual Fatigue Life Monitoring

The accuracy of statistics-based load spectrum reconstructions, either based on FRR or DLP, is inherently limited. The cost of additional data acquisition to improve regression models is high due to the need of dedicated flight test campaigns or in-flight load and data recording equipment. Therefore, it is generally necessary to account for modelling and prediction inaccuracies in VFLM-based service life limits.

Based on the state-of-the-art in VFLM, several challenges have been identified that present work addresses:

- There is a need to develop a reliability substantiation method for SLLs determined by DLP

- There is a need to develop a VFLM substantiation method that can mitigates uncertainties from prediction inaccuracies, i.e. all prior methods assert perfect recognition or prediction
- There is a need to develop a VFLM substantiation method that can provide aircraft-individual mitigation of prediction errors. I.e. all prior methods assume that prediction errors, on average, have the same influence on each aircraft, independent of its usage profile.
- There is a need to develop a VFLM substantiation method that can analyse, compare or mitigate uncertainties due to limited sample sizes or data availability, i.e. include confidence levels. This need is especially pronounced by the demonstration in chapters 2 and 3 of significant uncertainties that can arise from a limited amount of fatigue and flight test results available to estimate uncertainty distributions and generate prediction models.
- There is a need to validate VFLM methods using long-term and independent reference load measurements, for example with strain gauges installed on commercially operated helicopters.

Present work is however limited in its scope and the methods introduced and verified in present work alone are not expected to be sufficient for commercial implementation of VFLM. In particular, present work, assumes an idealized data acquisition and processing chain. I.e. it is assumed that all flights of a particular component have been recorded and that none of the recordings contains errors. Also, it is assumed that no data transfer and processing errors exist and that the administrative process to register parts, their flights, and to guarantee in-time maintenance is error-free. It is thus not expected that the present methods to substantiate the reliability of VFLM-adjusted SLLs alone are sufficient to enable safe industrial implementation, as also indicated by certification guidance material, e.g. FAA AC-27-1B MG-15 [17].

4.2 Direct Load & Damage Modelling for Virtual Fatigue Life Monitoring

A new VFLM method based on direct load prediction is introduced in section 4.2.1. This method is fully compatible with the fatigue life modelling framework already introduced and validated in chapters 2 and 3 and is summarized in section 4.2.2. The flight data by which the DLDM models are generated and tested is specified briefly in section 4.2.3. After that, the procedure to generate the models is exemplified in section 4.2.4. The accuracy and reliability of the models is subsequently tested in sections 4.2.5 and 4.2.6. Finally, the potential for in-service application and SLL extensions for several components is presented in section 4.2.7.

4.2.1 Definition of modelling approach

Direct Load & Damage Modelling (DLDM) is a DLP-based VFLM variant that has been introduced by Dekker *et.al.* [105, 20]. Various properties and configurations of the model have already been analysed and tested previously in [21]. DLDM is based on the same simplified and analytical fatigue life and reliability modelling method successfully verified in chapter 2, and is fully compatible with chapter 4.1 of AGARD-292 [16].

Several alternative VFLM models, other than DLDM, have been considered, developed and tested in the preparation of this work, including FRR and new modelling strategies introduced in Appendix J. The two methods introduced in Appendix J are specifically developed to provide modelling simplifications and reduced computational costs. However, testing of prototype implementations did not result in sufficiently accurate results to encourage further development and suggested applicability to a very limited class of problems only. Comparative testing between DLDM and alternative VFLM methods, including methods introduced in the state-of-the-art in section 4.1, may be subject to future work.

4.2.1.1 Definition of load spectrum modelling

The load spectrum model employed by DLDM is equivalent to the load spectrum model introduced and validated in chapter 2, but the timescale on which it is applied is significantly different. Although the load spectrum is still represented by a sequence of loading blocks, whose transitions are conservatively modelled

by an extreme load curve, each loading block no longer represents an entire manoeuvre but rather the loads occurring within a timeframe with duration of approximately one second, as shown in Figure 4.2.

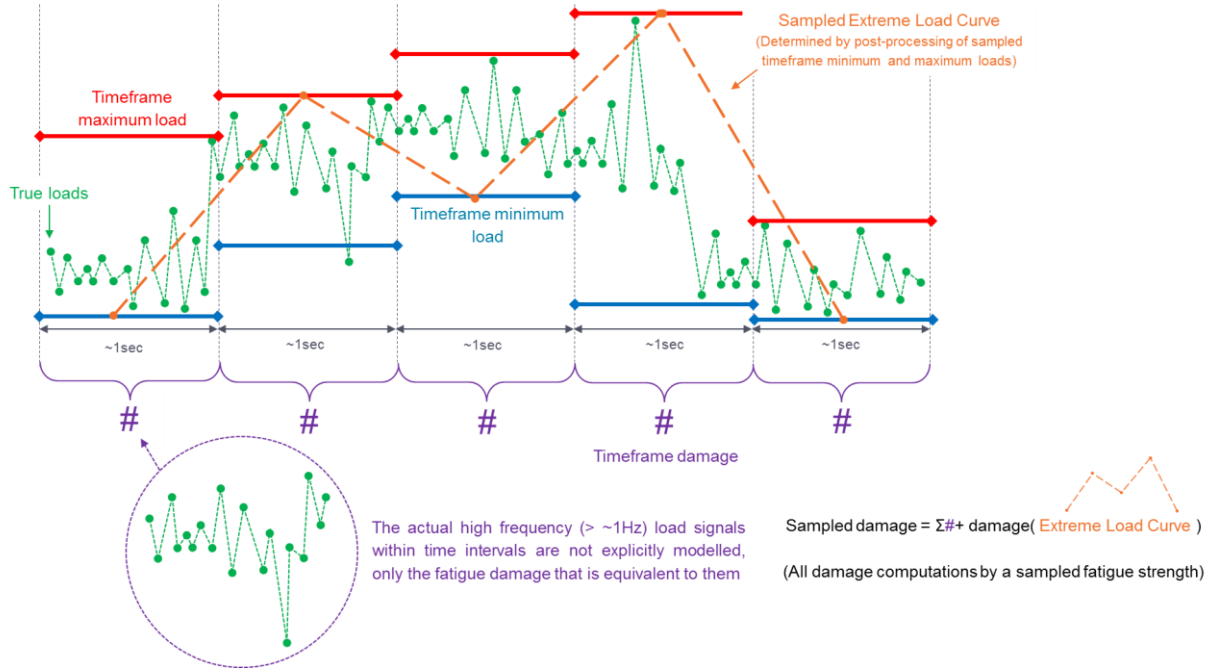


Figure 4.2: Schematic summarising how DLDM models accumulated fatigue damage as a function of predicted timeframe extreme loads and a summation of predicted timeframe damage. The example contains five timeframes/time intervals

DLDM models the fatigue damage of a flight by dividing it into consecutive timeframes, each with a specified duration, for example one second. The total accumulated fatigue damage during the flight consists of two parts: a low-frequency part and a high-frequency part.

High-frequency damage consists of the fatigue damage that is caused by the load spectrum within each individual timeframe and is also referred to as timeframe damage. For example, if a flight would last 100 seconds, then the accumulated high-frequency damage would be made out of the fatigue damage caused by the addition of 100 individual load spectra. Total accumulated fatigue damage D_{total} is thus composed of a low-frequency part D_{LF} and a high-frequency part D_{HF}:

$$D_{total} = D_{LF} + D_{HF} \tag{4.1}$$

If a component`s time history consists of N timeframes, then the high-frequency part is made up by the sum of the high-frequency damage of all individual timeframes $D_{HFT,j}$:

$$D_{HFT} = \sum_{j=1}^{N} D_{HFT,j} \tag{4.2}$$

The high-frequency damage of a timeframe is computed according to the Miner linear damage accumulation hypothesis, and is determined by the load spectrum $L_{spectrum,HF,j}$ that results from the loads within the timeframe and the working S-N curve S_{Work}:

$$D_{HFT,j} = f_{Miner}\left(L_{Spectrum,HF,j}, SN_{Work}\right) \tag{4.3}$$

The load spectrum $L_{spectrum,HF,j}$ is based on the continuous recorded loads ω_j of the j^{th} timeframe determined by a function $f_{ELC,RFC,LBC}$ performing peak-valley filtering, rainflow cycle counting and class binning, as also previously introduced in 2.2.1.3.

The low-frequency damage is computed equivalently to the high-frequency damage, with the distinction of how the load spectrum is composed. The low-frequency load spectrum $L_{spectrum,LF,j}$ is based on the sequence of maximum and minimum loads that occur within each timeframe:

$$L_{Spectrum,LF} = f_{ELC,RFC,LBC}\left(\left[\max(\omega_1),\min(\omega_1)\right]_1,\ldots,\left[\max(\omega_j),\min(\omega_j)\right]_j,\ldots,\left[\max(\omega_N),\min(\omega_N)\right]_N\right) \quad (4.4)$$

Crucially, DLDM does not attempt to reconstruct the actual load spectrum or load signal ω_j that occurred during a timeframe, but rather estimates $D_{HF,j}$ and $\left[\max(\omega_j),\min(\omega_j)\right]$ directly.[18] DLDM assumes that it is easier and more accurate to directly recognize the fatigue damage that is caused by the load signal within the timeframe, rather than the load signal itself.

The load spectrum model of DLDM makes use of rainflow cycle counting as defined in section 2.2.1.3. This implies that counted load cycles can span multiple timeframes and it is thus necessary to account for the load signal transitions between timeframes as well. This is achieved by recognizing the minimum and maximum load occurring within each timeframe. An Airbus proprietary method is then used to determine the most conservative load signal through the resulting sequence of extreme loads. This is called the Extreme Load Curve (ELC) and the fatigue damage corresponding to its cycle-counted load spectrum is referred to as low-frequency fatigue damage.

The entire DLDM fatigue damage recognition process is summarized in Figure 4.3. After the quality and integrity of the recorded flight data is ensured, it is divided into subsequent timeframes. Then, for each timeframe, high-frequency timeframe damage is estimated and summed over all timeframes to estimate the total accumulated high-frequency damage, given the working S-N curve of the component under consideration.

In parallel, the maximum and minimum load occurring during each timeframe is predicted for each timeframe. Then, an extreme load curve is fitted through the sequence of timeframe extreme loads, and accumulated low-frequency D_{LF} damage is computed. Summing low- and high-frequency damage finally results into the total accumulated fatigue damage, given the conservative S-N working curve.

[18] Given the results from Haas *et.al.* [82] and the good timeframe extreme load recognition results that are demonstrated in section 4.2.5.2, it would instead also be feasible to bound a constant amplitude loading block and follow classic DLP approaches. However, this would restrict the application to components whose primary loading is composed of time-invariant, few and very narrow frequency bands only, or lead to overly conservative results. Accurately choosing the high-frequency load content of individual timeframes is difficult unless the loading blocks are composed of known and simple waveforms, such as for time-invariant, constant amplitude and single-frequency loading. Nevertheless, high-frequency load signal reconstruction results using principle component decomposition presented by Beale *et.al.* [10] do however indicate the feasibility of PCA-based load signal reconstruction. Although the results by Beale *et.al.* [10] are not meant to be used for DLP, further research in the area of high-frequency load signal or spectrum recognition is therefore recommended. Especially as it could simplify reliability substantiation as in section 4.3 and significantly reduce computational costs.

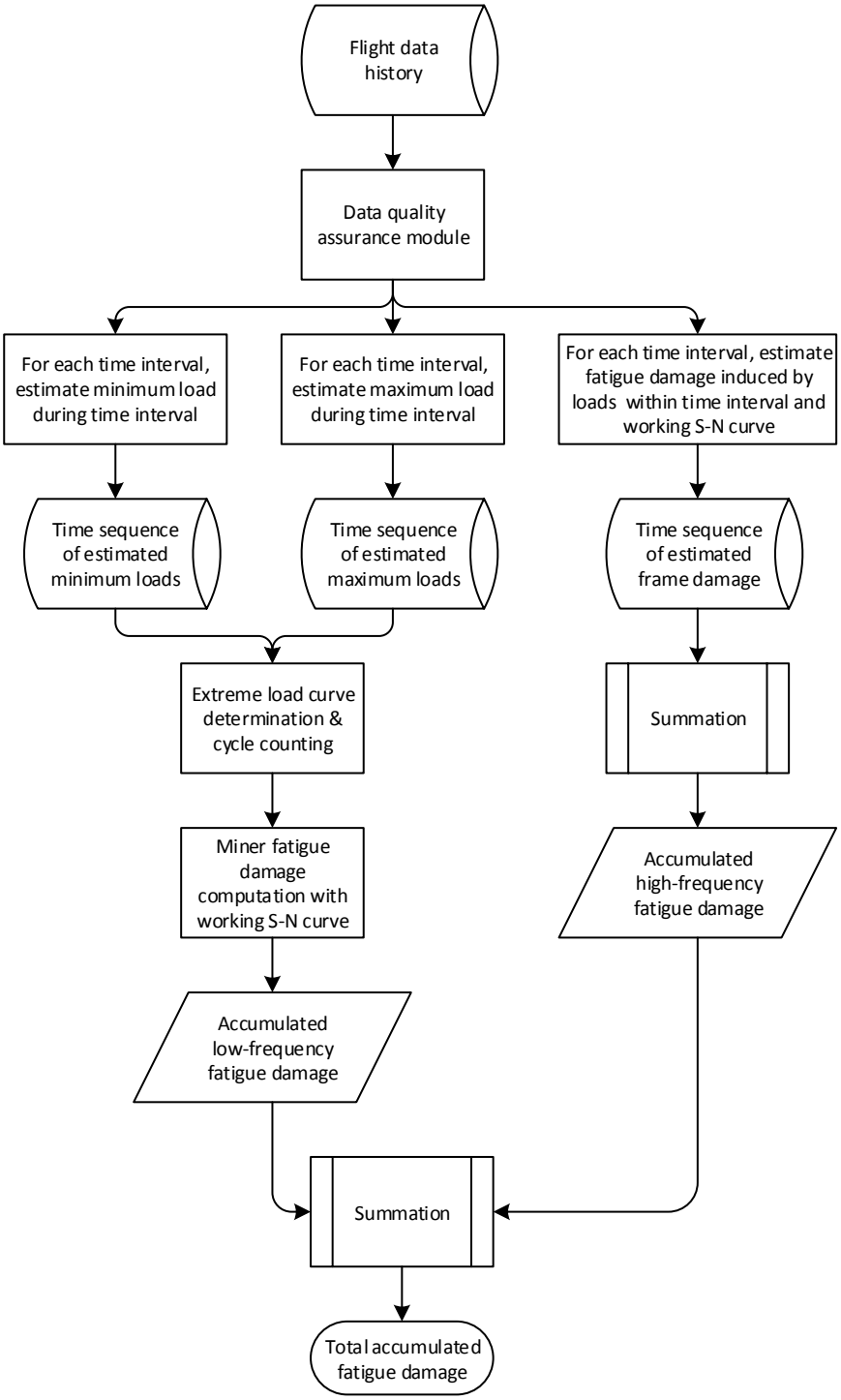


Figure 4.3: Process summary of how Direct Load & Damage Modelling makes usage-based estimations of accumulated fatigue damage and implements Virtual Fatigue Life Monitoring

4.2.1.2 Reliability substantiation concept

DLDM makes use of a simplified reliability substantiation concept that is based on prior experience from Dekker [21] and the simulation results from chapter 2, and is based on the following notions:

- The negligibility of load estimation errors: Simulation results from chapter 2 demonstrate that small errors in the distribution of flight regime loads are negligible in comparison to uncertainties about fatigue strength when substantiating the reliability of SLLs.

- Small model prediction errors: Prior test results for DLDM [21] and results for DLP by, amongst others, Haas *et.al.* [85] and Wallace *et.al.* [6] indicate that DLP can obtain good recognition results and that prediction errors may remain small, i.e. without leading to significant errors in predicted fatigue life
- Predicted fatigue life must be conservative, not individual timeframe predictions: It is not strictly necessary to substantiate the reliability and accuracy of timeframe-individual DLDM recognitions themselves, it is only required to ensure that resulting estimates of total accumulated fatigue damage meet a reliability appropriate for the component in question.

Therefore, it is reasonable to assume that small prediction errors from DLDM can be allowed and that their effect on overall fatigue damage accumulation estimates may be neglected. It is thus hypothesized that the effects of DLDM estimation errors on timeframe extreme load and damage have a negligible effect on the reliability of estimates for accumulated fatigue damage, if all damage is computed using a conservative working S-N curve. Crucially, it is postulated that all reliability may be substantiated by the working curve, i.e. a DLDM-based service life computed while making use of a working S-N curve with a reliability of $\gamma = 10^{-6}$ (α=95%), should imply that the DLDM-based estimate of accumulated fatigue damage also has a reliability of $\gamma = 10^{-6}$ with a confidence level α of 95%. This hypothesis is tested in section 4.2.6.

The reliability substantiation model of DLDM is thus equivalent to the reliability substantiation model of the analytical fatigue life prediction and substantiation method discussed and tested before in chapter 2. In addition, DLDM makes use of all of the assumptions concerning the modelling of fatigue damage, fatigue strength and associated uncertainty models as already used and discussed before in chapters 2 and 3. DLDM's reliability model continues to assume the full validity of these modelling assumptions.

4.2.1.3 Definition of two-step regression method for timeframe damage

Since DLDM assumes that all reliability can be substantiated by the reliability of the working curve, timeframe fatigue damage must also be computed using the working S-N curve, for example with a reliability of $\gamma=10^{-6}$ (95%). Fatigue damage is a highly non-linear quantity with a discontinuity caused by the fatigue limit, i.e. it can be that for many timeframes damage is zero, since all the corresponding load cycles lay below the fatigue limit. To model this discontinuity, timeframe damage is predicted using the two-step process introduced earlier for the simulation-based stochastic load spectrum model in chapter 2 and summarized in Figure 4.4. First, it is predicted if timeframe damage is expected to be equal to zero, or more than zero. Second, in the case of the latter, timeframe damage is further detailed by a regression model.

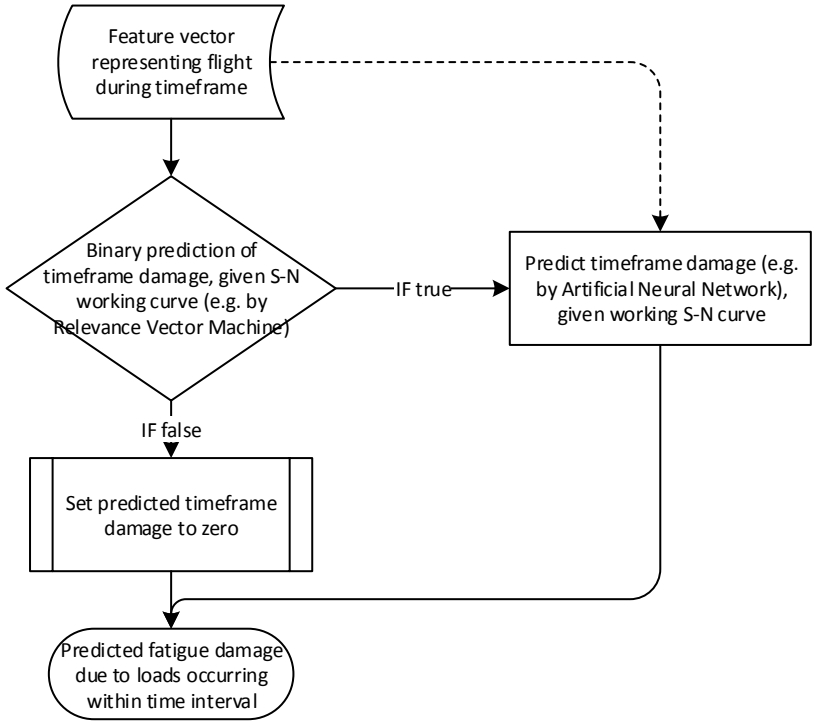


Figure 4.4: Process overview detailing how DLDM predicts timeframe damage by a conditional and two-staged regression process. The method uses a Relevance Vector Machine, which is a binary classifier whose details are elaborated in Appendix H.

Expressed mathematically, timeframe damage is thus predicted as follows:

$$D_{HFT,j}\left(\omega_j\right) = \delta_j\left(\omega_j\right) \cdot D_{HFT,j}\left(\omega_j\right) \quad \text{with} \quad \delta_j\left(\omega_j\right) \in \{0,1\} \tag{4.5}$$

Where $\delta_j\left(\omega_j\right)$ denotes a binary delta function, whose value is a function of the feature space and for example determined by a non-linear statistical data model.

4.2.2 Overview of fatigue damage modelling for Direct Load & Damage Modelling

All fatigue damage modelling employed in this chapter is performed as defined earlier in section 2.2, except fatigue strength distribution modelling, which uses additional methods from chapter 3. It is, therefore, compatible with chapter 4.1 in AGARD-AG-292 [16] and corresponds to approved and common practice in the rotorcraft industry. Modelling of fatigue strength uncertainty is augmented by the Bayesian modelling technique introduced in chapter 3. The Bayesian prior distribution introduced earlier in section 3.5 is used, as it allows performing realistic fatigue strength distribution modelling for all selected components that will be used to test VFLM.

4.2.3 Introduction of data for testing and generating prediction models for Virtual Fatigue Life Monitoring

The recognition of timeframe extreme loads and damage is implemented by means of non-linear data modelling or machine learning. The following sections summarize which data has been used to generate and test VFLM models.

4.2.3.1 Definition of data sources

VFLM is implemented and tested for several components of the dynamic system of a light, twin-engine, and multi-role helicopter. Flight test data from Load Classification Flights (LCF) for this helicopter was used to generate VFLM models and to provide initial test statistics. This data contains manoeuvres covering the entire

flight envelope and manoeuvre execution is often considered more aggressive than expected in operational practise. This increases the likelihood that the data covers all cases that can be encountered during practical operations. However, the following considerations are made on the validity and representativeness of this LCF data:

- Load classification flights are generally performed on only one machine, therefore model overspecialization may occur and no data is generally available on the variability across machines. LCF data set aside for testing purposes will, therefore, be referred to as quasi-independent and not as fully independent – which would be the case for data from helicopters and operations that were not part of the LCF campaign. Data analysis in Appendix L indeed indicates significant differences in the properties and distribution of data collected during load classification flights and commercially operated helicopters.
- Since load classification flights do not represent a realistic operational flight profile:
    - it is challenging to estimate the actual benefits of VFLM application. Apart from the conservative and composite design mission profile, helicopter OEMS are often unable to forecast realistic in-service load spectra.
    - it is also challenging to estimate the effect of modelling errors during actual operations

The availability of flight data collected during actual commercial operations on different helicopters is, therefore, crucial in obtaining representative statistics on the potential benefits of VFLM. Three helicopters have consequently been equipped with a data recorder and all their flight data has been collected during more than one year. These three helicopters are all commercially operated under a Helicopter Emergency Medical Service (HEMS) mission profile. In total, almost 3,000 flight hours have been recorded during approximately 20,000 flights.

Two of these helicopters have additionally been equipped with a custom strain gauge installation to continuously measure Fenestron tail driveshaft torque. This load data can be used to test VFLM predictions using continuous and real flight data. The load data recorded on these two helicopters is used exclusively to validate the accuracy and reliability of VFLM-estimated loads and fatigue damage. It has never been presented to any of the VFLM prediction models to avoid any chance of overspecialization or data contamination of test results. Recorded data has been anonymized and these three helicopters are subsequently referred to as helicopters 1-3, where helicopters 1-2 have been equipped with the additional load sensor for reference and validation purposes.

In general, the following assumptions are made about the data quality, validity, and representativeness:

- The effect of modelling errors due to limitations of the statistical data model on the reliability substantiation of accumulated fatigue damage may be neglected. The following sources of inaccuracy are especially concerned:
    - model overspecialization
    - non-representativeness of the load classification flights for serial helicopters (e.g. centre-of-gravity, external equipment, atmospheric conditions, manoeuvre execution, sensor bias, sensor calibration, and sensor precision)
    - regression extrapolation due to operational conditions outside of the tested flight envelope during load classification flights[19]
- The integrity of the recorded data is guaranteed and the presence of invalid data can be detected
- The removal or filtering of data does not influence the representativeness or accuracy of derived statistics

[19] Non-linear statistical data models are used to predict timeframe damage and extreme loads. Such models can become highly unstable if they have to perform extrapolation, i.e. are exposed to a feature vector that is not spanned by feature vectors used during model generation.

The validity of these assumptions is partially ensured by the implementation measures listed in the next section 4.2.3.2. Otherwise, associated errors should be incorporated in the accuracy and reliability test results in sections 4.2.5, 4.2.6, 4.3.3, and 4.3.4. As long as the test results using data from helicopters 1-3 can indeed be considered as independent and representative, then these test results should include errors due to limited data quality, validity and representativeness as well.

4.2.3.2 Summary of methods for data pre-processing

The flight data used to predict timeframe extreme load and damage is listed in Table 4-1 and corresponds to flight data that can be recorded on this helicopter without significant modification of the avionics.[20] Although different VFLM publications make use of different flight parameters, these differences are often simply the result of what data is available on the helicopter type under consideration and may not be the result of an elaborate feature selection process.

It is assumed that the available database for model generation is sufficiently large and representative and that any redundant or irrelevant features in the data do not cause adverse overspecialization of the regression models.

Studies have been performed in the framework of present work on automatic feature selection in order to evaluate if the number of recorded flight parameters could be reduced without significant loss of regression accuracy. However, results were inconsistent and the resulting feature vectors could not be used to reliably generate accurate regression models. Further research may be recommended in this area if computational costs need to be reduced or if specific over-specialization issues are encountered.

All flight parameters have been uniformly sampled at 10Hz. However, sensors, or data buses, themselves may operate at different frequencies. The chosen recording frequency is compatible with recommendations in ADS-79D-HDBK [73]. In addition, prior work by Dekker [21] demonstrated good DLDM regression accuracy using a sampling rate as low as 2Hz. A timeframe length of one second has been chosen based on earlier work by Dekker [21] where the influence of different timeframe lengths on regression accuracy was tested. Timeframe data is concatenated and transformed by principle component analysis for noise reduction, decorrelation, and data compression before actual statistical regression takes place.

Table 4-1: Table defining the recorded flight parameters that are used for VFLM

Designator	Description	Unit
DALPHA	**Angle main rotor lateral cyclic pitch**	Deg
DBETA	**Angle main rotor longitudinal cyclic pitch**	Deg
DDELTA	**Angle main rotor collective pitch**	Deg
DTHETA	**Angle fenestron collective pitch**	Deg
NZ	**Load factor**	[-]
NRO	**Rotor RPM**	%
P	**Pitch rate**	Deg/s
Q	**Roll rate**	Deg/s
R	**Yaw rate**	Deg/s
PHI	**Roll attitude**	Deg
THETA	**Pitch attitude**	Deg
IAS	**Indicated Air Speed**	Knot
ZP	**Pressure altitude**	Feet

[20] The use of the features listed in Table 4-1 has been tested extensively in prior and unpublished work performed at Airbus Helicopters Germany and its applicability has been verified through expert judgement.

MQTW1	**Torque engine #1**	Nm
MQTW2	**Torque engine #2**	Nm
MQTW12	**MQTW1 + MQTW2**	Nm
MMO	**Absolute moment on main rotor mast**	%

Data quality and validity have been ensured by a range of methods and processes:

- All load classification data has been reviewed manually with a special focus on the validity of the load signals by Airbus Helicopters specialists. Where applicable, data has been filtered and repaired before use by custom and proprietary software. In general, all LCF data is subject to procedures approved by aviation authorities.
- In addition, a new proprietary software has been developed by Airbus Helicopters in the framework of present work to semi-automatically check LCF data for errors and inconsistencies, including: missing data, discontinuities, outliers, invalid acquisitions, high noise, offset errors, data out-of-range and sign errors. After field-trial and expert validation, this tool has been used for detailed analysis of both LCF data as well the data recorded on commercially operated HEMS helicopters 1-3. No anomalies related to standard avionics equipment were detected on the data from the HEMS helicopters. However, a significant portion of HEMS data has been removed due to inconsistencies or recording errors caused by the flight test equipment specially integrated in these helicopters, reducing the useful size of the database to approximately 1,800 flight hours.
- All the recorded flight data from the commercially operated HEMS helicopters has been pre-processed by an additional specialized and proprietary software tool that was developed in the framework of present work to check for the presence of time synchronization errors related to the flight data recorder. This tool revealed an occasional error in the timestamp of recorded data. Although sometimes repairable by post-processing, any data with a timestamp error has been removed from the database employed in this work to guarantee data quality. It is assumed that the occurrence of timestamp errors occurs randomly and does not influence the representativeness of any presented statistics.
- Data recorded during ground operations of the three HEMS helicopters has been filtered out and has not been considered. Since ground operations have scarcely been included in LCF data, accurate regression during ground operations was not achievable. This is acceptable since ground operations are also not considered by the classic fatigue life predictions for the helicopter type under consideration.

In addition, limitations in the avionics and data recording equipment, and due to the rushed nature of HEMS operations, made that flights often commenced before the valid recording of the reference strain gauge on the tail driveshaft could start. Hence, it was not possible to record a reference load signal during many take-off procedures. In most cases, recording of regular flight parameters could commence in-time. Where appropriate, corresponding limitations and cautions are included to presented statistics.

Additional details on the practical implementation details of how data is formatted and preprocessed are provided in Appendix I.

4.2.3.3 Overview of selected components to test and benchmark methods for Virtual Fatigue Life Monitoring

The results in this work primarily focus on predicting torque on the Fenestron tail driveshaft and the fatigue damage that the lower housing of the main gearbox experiences due to this torque. A Fenestron is a ducted tail rotor. Fenestron torque depends on how engine torque is distributed between the main rotor and the Fenestron. This division depends on the detailed state and loading of the main rotor blades and the Fenestron blades, as well as atmospheric conditions. A complete list of all the components for which VFLM models have

been generated and applied in the framework of this work is included in Table 4-2. These components have been selected to provide a wide coverage of relevant service life limit parts and for which a reasonable amount of load classification test flight data was available for model generation.

Table 4-2: Table defining the selection of components for which VFLM is applied in present work. Only for the lower gearbox casing is independent and continuous reference data available from strain gauge measurements on two commercially operated helicopters. The primary focus in present work lies on component 6, whose row is highlighted in bold.

Component ID	Designator	Fatigued component	Loading determined by	Load type
1	FBTHETA	Casing of hydraulic actuator	Main rotor collective control booster	axial
2	FBTHETAP	Control rod	Main rotor collective control booster	axial, tension only
3	FKAR	Gimbal	Combined loading by main rotor collective, longitudinal and lateral control boosters	axial
4	FSTA	Forked lever	Combined loading by collective and lateral control boosters	axial
5	FSTY	Upper gearbox casing	Main gearbox anti-torque strut in lateral direction	axial
6	**MQF or MF**	**Lower gearbox casing**	**Torque in Fenestron tail driveshaft**	**torque**
7	MMO	Main rotor mast	Effective main rotor torque	torque

All load signals have been sampled at 100Hz to cover high-frequency loading content with sufficient detail. This sampling rate is approved by aviation authorities for the helicopter under consideration.

4.2.4 Generation of prediction models for Direct Load & Damage Modelling

For model generation and initial testing, the available data from load classification flights (LCF) has been divided into a training set, a validation set and a semi-independent test set by means of interleaving. The test set is regarded as semi-independent, only due to the considerations outlined before in section 4.2.3.1. Timeframes have been cut-out of the continuous flight data in an overlapping fashion as it was found during earlier work that this effectively up-samples the available data, increases accuracy, and reduces overspecialization [21]. Timeframe minimum and maximum loads are each predicted by a shallow feed-forward Artificial Neural Network (ANN) [106]. Further implementation details are specified in Appendix I. The use of deep learning and deep artificial neural networks [107] has been tested but did not lead to significant accuracy gains. Their use has been discontinued to avoid unnecessary model complexity and computational costs.

The two resulting regression models for predicting the minimum and maximum load during a timeframe are tested on the portion of LCF test data set aside for semi-independent testing. The resulting regression plots in Figure 4.5 and Figure 4.6 demonstrate promising regression accuracy and precision comparable with state-of-the-art results, e.g. load prediction models discussed by Haas [85] and Wallace [6].

Initial classification of timeframes with zero timeframe damage and more-than-zero high frequency timeframe damage is done by a Relevance Vector Machine (RVM) developed by Tipping *et.al.* [108, 109, 110, 111], whose implementation details are further specified in Appendix I and whose theoretical background is further clarified in Appendix H. The use of an RVM was considered after ANNs and Support Vector Machines (SVM) [106] were tested unsuccessfully, as already outlined in earlier work by Dekker [21]. In contrast to regular classifiers, an RVM does not provide a binary classification but rather an actual probabilistic estimate of binary class membership. Predictions on the probability of timeframe damage being more-than-zero for timeframes in the LCF test set are presented in Figure 4.7 and demonstrate good classification accuracy. Turning these

probabilistic predictions into binary classifications by means of ordinary rounding enables classification performance to be analysed by a confusion matrix as shown in Figure 4.8. In general, it can be observed that only a small portion of the timeframes causes high-frequency fatigue damage, given the conservative S-N working curve, and that classification accuracy is reasonably accurate.

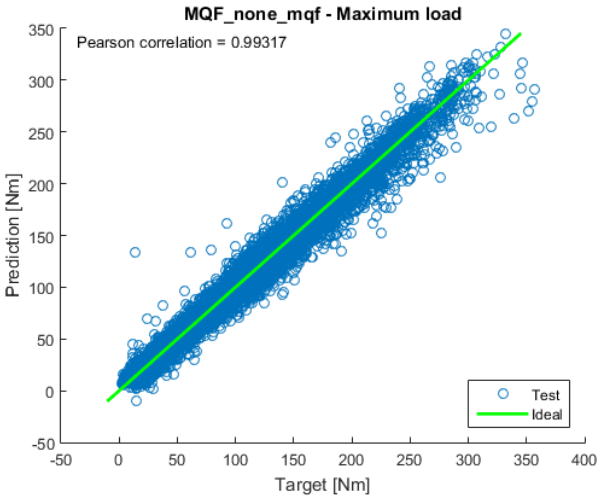


Figure 4.5: Regression plot showing the correlation between predicted maximum torque load on the Fenestron driveshaft and the actually measured maximum torque load during timeframes in the portion of LCF flight test data set aside for model testing. The predictions are made by a shallow Artificial Neural Network.

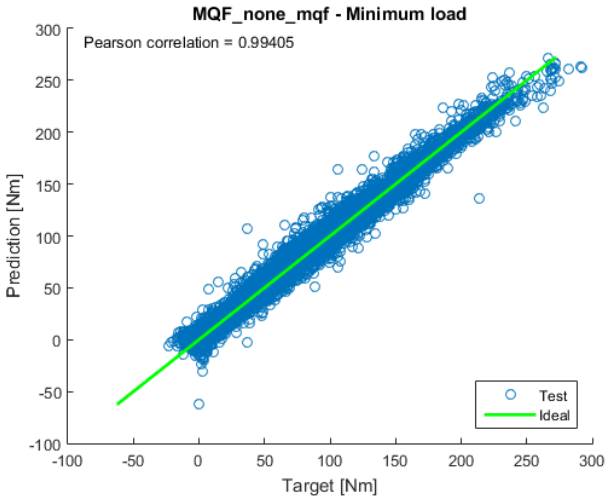


Figure 4.6: Regression plot showing the correlation between predicted minimum torque load on the Fenestron driveshaft and the actually measured minimum torque load during timeframes in the portion of LCF flight test data set aside for model testing. The predictions are made by a shallow Artificial Neural Network.

After timeframes have been classified as damaging, their actual damage value is predicted by a shallow ANN. Regression of timeframes previously correctly classified as damaging is tested on the LCF test data in Figure 4.9-left and for all timeframes classified as damaging in Figure 4.9-right. The regression plot in Figure 4.9 demonstrates that the order of magnitude of timeframe fatigue damage can be predicted with reasonable accuracy and that confusion in timeframe classification generally only occurs for timeframes causing little damage. This is further clarified by the marginalised distributions in Figure 4.10. Importantly, the regression plots also demonstrate the robustness of the timeframe damage regression. Timeframes incorrectly classified as damaging do not induce an unstable response by the ANN. Despite that, the ANN has been trained on relatively few samples and has not been trained with timeframes not causing timeframe damage.

Overall, the test results presented in this section demonstrate that DLDM can be implemented with reasonable accuracy for component 6 in Table 4-2. Tests for six other components are included in Appendix K and demonstrate similar results, suggesting a broad applicability for DLDM.

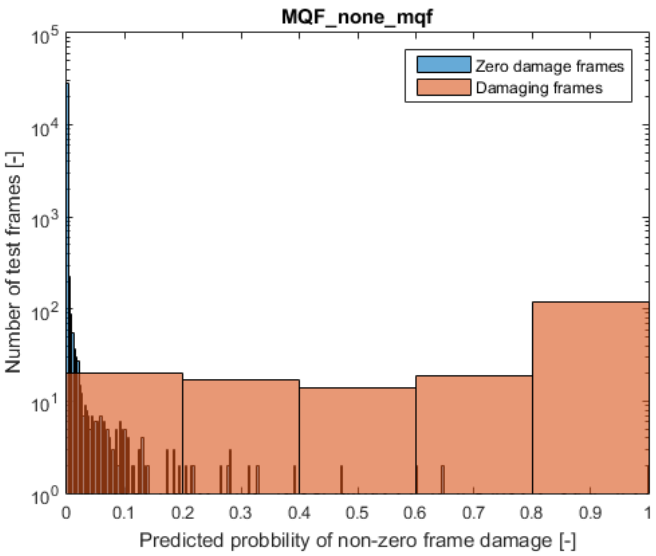


Figure 4.7: Chart showing the distribution of predicted timeframe damage probabilities and a comparison with the actual occurrence of timeframe damage. The predictions are made with an RVM classifier for LCF data set aside for model testing. The chart shows how most zero-damage timeframes are correctly predicted to have a low probability of causing frame damage. Whereas the chart also shows that most of the timeframes with positive timeframe damage are indeed predicted to have a high probability of causing timeframe damage. (Blue bins showing frames with zero damage are partially overlaid by read bins with non-zero timeframe damage. These bins are displayed as dark/grey-red)

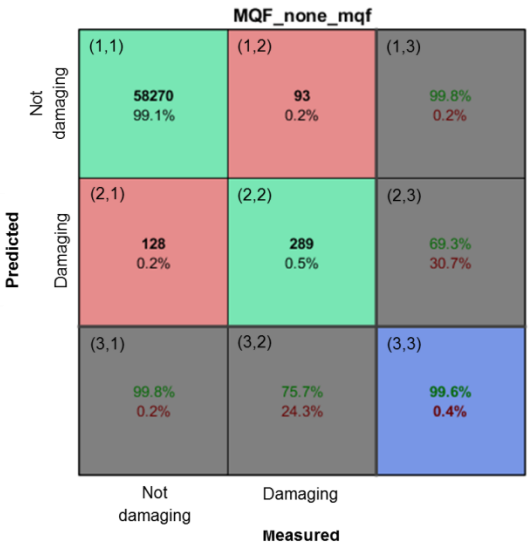


Figure 4.8: Confusion matrix[21] showing the accuracy of timeframe damage classification using an RVM classifier and for LCF data set aside for testing.

[21] (3,3): Overall classification accuracy of the test set, i.e. 99.6% of all timeframes in the set were correctly classified
(1,1): 99.1%, or 58,270 timeframes, of all tested timeframes were correctly classified as non-damaging
(2,2): 0.5%, or 289 timeframes, of all tested timeframes were correctly classified as damaging
(2,1): 0.2%, or 128 timeframes, of all tested timeframes were incorrectly classified as damaging. This error is conservative.
(1,2): 0.2%, or 93 timeframes, of all tested timeframes were incorrectly classified as not damaging. This error is not conservative.
(3,1): Given that a timeframe is not damaging, it is correctly classified as such with a probability of 99.8%
(3,2): Given that a timeframe is damaging, it is correctly classified as such with a probability of 75.7%
(1,3): Given that a timeframe is classified as not damaging, this classification is correct with a probability of 99.8%
(2,3): Given that a timeframe is classified as damaging, this classification is correct with a probability of 69.3%

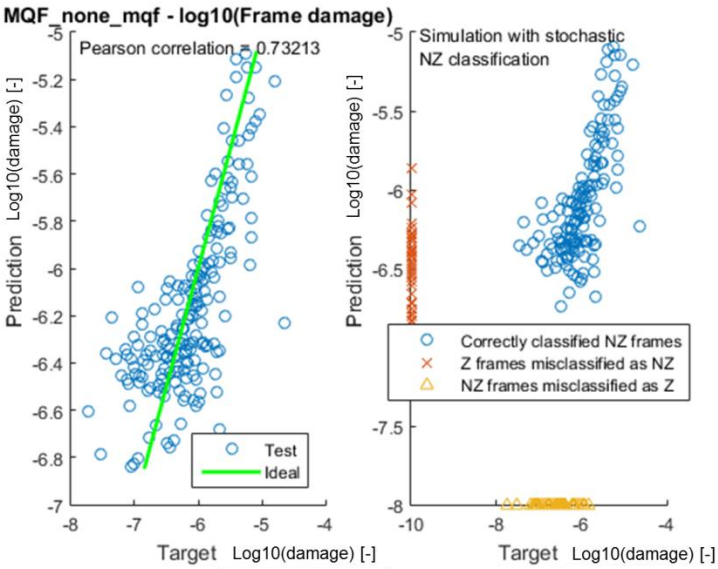


Figure 4.9: Regression plot showing the correlation between predicted and actually measured timeframe damage for correctly classified damaging timeframes in the LCF test set. (left)
Regression plot showing the correlation between predicted timeframe damage and actually measured timeframe damage for all timeframes classified as damaging in the LCF test set. (including incorrectly classified timeframes)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes). (right)

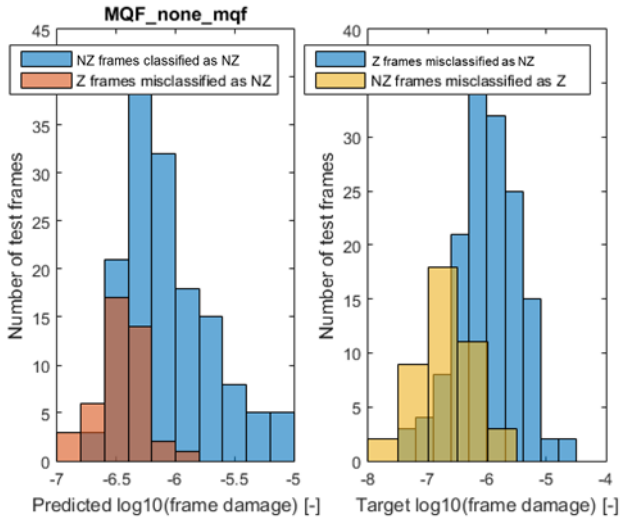


Figure 4.10: Regression plot showing the distribution of predicted timeframe damage for frames misclassified as damaging or correctly classified as not damaging in LCF test set (left)
Regression plot showing the distribution of the true timeframe damage of non-damaging timeframes that are misclassified as damaging and damaging timeframes that are misclassified as non-damaging in the LCF test set (right)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes)

4.2.5 Accuracy and precision testing for Direct Load & Damage Modelling

4.2.5.1 Definition of testing strategy

The regression test results discussed in the previous section 4.2.4 demonstrated highly accurate DLDM regression models. However, these test results can be considered to have been generated with semi-

independent LCF data only and may thus not be fully representative for prediction errors during commercial operation. As outlined previously in section 4.2.3, estimates of predictive accuracy and precision based on test data originating from the same load classification flight test campaign that was used to generate the predictive DLDM models, and thus from the same helicopter, may be subject to errors. Therefore, the performance of the DLDM regression and classification quality is tested again, but now using fully independent data recorded on two commercially operated helicopters.

The accuracy verification tests are performed as standard regression tests. The predictive DLDM models are given flight data from the two in-service reference helicopters 1 and two. Subsequently comparing model predictions with actually measured loads on these two helicopters provides realistic and fully independent test statistics.

4.2.5.2 Testing regression accuracy for timeframe extreme loads with independent data

Comparing ANN predictions for timeframe extreme loads with actually measured loads for helicopters 1 and 2, as shown in Figure 4.11, demonstrates significant differences with the previous and semi-independent test results in section 4.2.4. Using fully independent data from helicopters 1 and 2 instead of using semi-independent LCF data demonstrates significantly less accurate and precise regression. For helicopter-1, it can be seen that regression errors are more heteroskedastic than observed previously based on LCF test data in Figure 4.5 and Figure 4.6, i.e. errors increase significantly in magnitude with increasing load. And for both helicopters, there is now a slight proportional bias that increases with load, especially pronounced for helicopter 2. Results for minimum load recognition are similar but not presented.

4.2.5.3 Testing the consistency of prediction results

The predictive models for timeframe extreme loads are generated using machine learning. Therefore, if the relationship between flight parameters and loads differs between helicopters, then the application of a model generated by data from one helicopter on data from another helicopter should cause significant regression errors. Therefore, the difference in regression performance between using LCF data or data from helicopters 1 and 2 may have been caused by a systematic measurement fault or inaccuracy. To rule out this possibility, the data from helicopters 1 and 2 have each separately been used to generate an ANN regression model to predict timeframe extreme load, i.e. a model is generated based data from helicopter-1 and another based on data from helicopter-2.

Testing the model from helicopter-1 on data from helicopter-2, and vice-versa, as shown in Figure 4.12, brings confidence that there is no systematic measurement error in the data from the two helicopters. If the data collected on one of the helicopters contains a measurement error, either from the recorded flight data or the specially installed strain gauge, then this should result in the observation of significant regression errors when the predictive models is tested on data from a different helicopter. Such errors are not observed in Figure 4.12. The regression plots actually demonstrate a reversal of the bias in comparison to the regression results obtained with LCF-based models in in Figure 4.11.

The incidental and gross misrecognitions that were observed on helicopter-2 in Figure 4.12 are suspected to originate from unstable ANN response in an area of the feature space where helicopter-1 may barely have spent any time, thus not generating training examples for this area of the feature space. This explanation is supported by comparative coverage analysis presented in Appendix L.4, where significant differences in the domains of the feature spaces spanned by data from helicopters 1-2 are demonstrated.

In general, the cross-comparison of the ANN prediction models also suggests that a prediction bias resulting from a regression model generated by data from another helicopter is common and may be the result of different typical configurations, e.g. typical weight or centre-of-gravity, or limitations in strain gauge calibration accuracy and precision. The load data from helicopters 1-2 can thus be used as is and the observed bias with LCF-based regression results will have a conservative influence on observed reliability levels in

sections 4.2.6 and 4.3.4. Prediction reliability is thus tested under realistic and non-ideal circumstances and under the presence of significant prediction errors on timeframe damage and extreme loads.

It is re-iterated that, except for this single consistency test in Figure 2.20, the recorded load data from helicopters 1-2 has never been used to generate, modify, or optimize any regression model and has been used exclusively for independent accuracy and reliability testing.

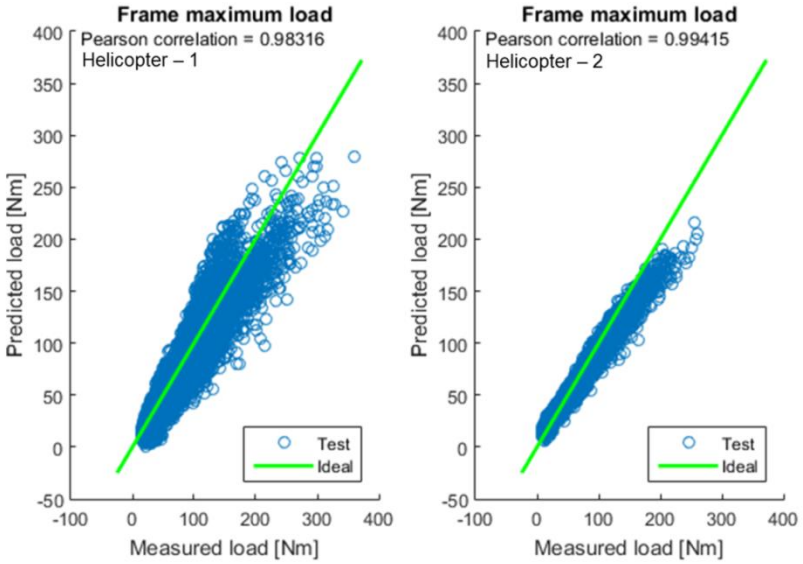


Figure 4.11: Regression plots showing the correlation between predicted values of timeframe maximum torque on the Fenestron driveshaft and independently measured torque maximums on helicopters 1-2. The predictions are made by an ANN generated by LCF data.

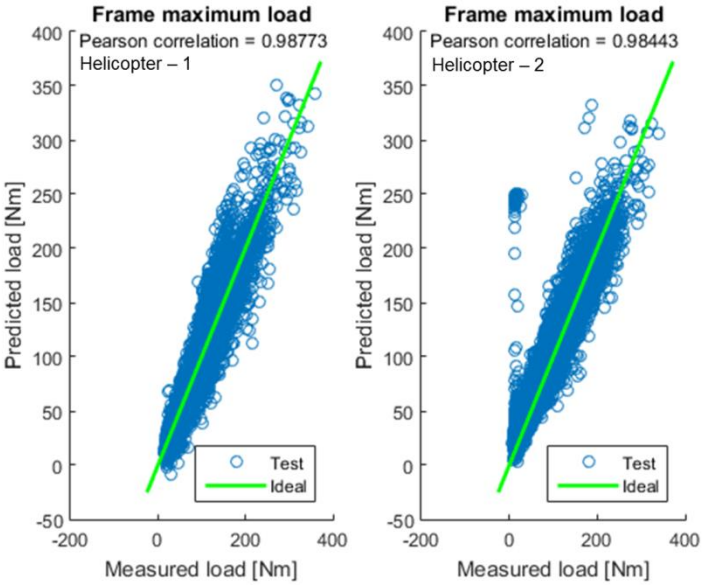


Figure 4.12: Regression plots testing the prediction of timeframe maximum load on the Fenestron driveshaft by an ANN generated by data from helicopter-2 and tested on helicopter-1 (left), and vice-versa (right).

4.2.5.4 Testing regression accuracy for timeframe damage regression with independent data

Prediction of timeframe damage is tested with less detail than in the case previously shown in Figure 4.8. For simplification and to reduce computational costs, it is only tested if the binary presence of timeframe damage

in entire flights is correctly classified, instead of testing the classification accuracy of individual timeframes. The result in Figure 4.13 demonstrates reasonable accuracy given that the probability of misrecognition is vastly increased because misrecognition of a single timeframe makes that the entire flight is counted as misrecognized.

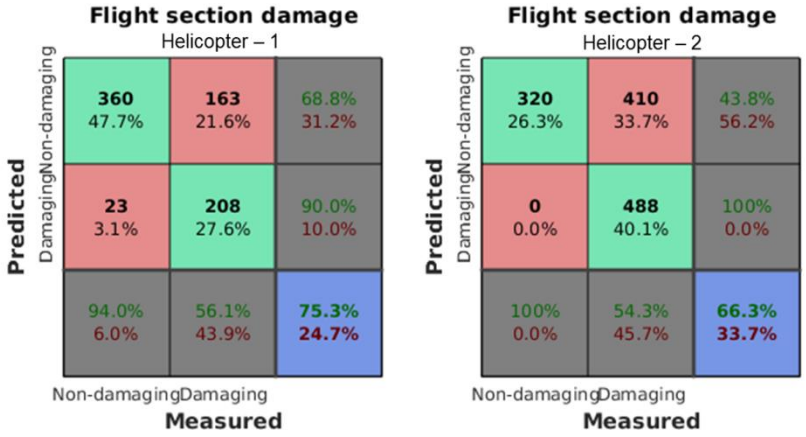


Figure 4.13: Confusion matrices showing the accuracy by which DLDM can correctly classify the occurrence of high-frequency damage during a flight. The predictions are made by an RVM classifier generated from LCF training data. Verification data comes from independently measured loads in helicopters 1 and 2.

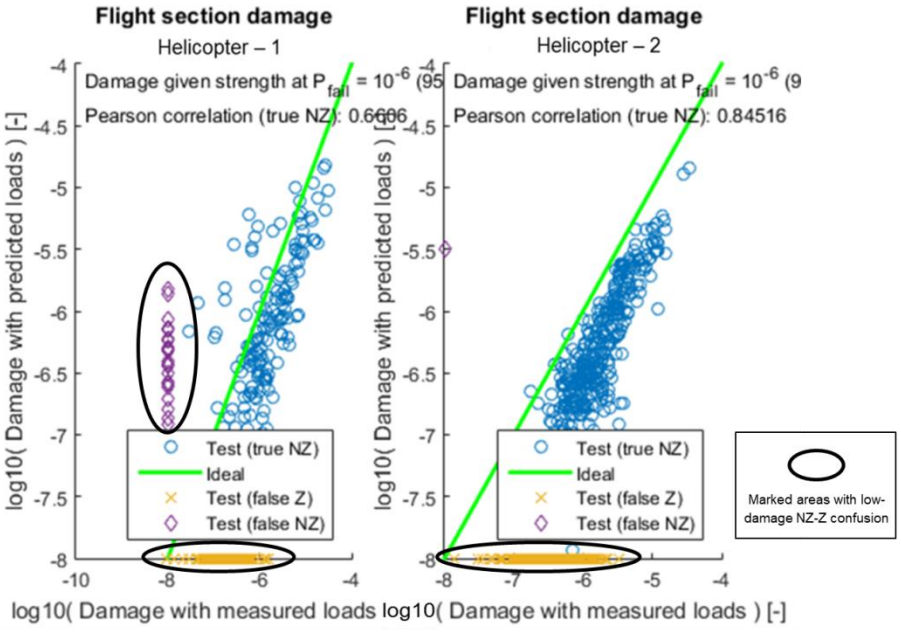


Figure 4.14: Regression plots showing the correlation between predicted and independently measured values of accumulated high-frequency timeframe damage during a flight. (Z denotes zero high-frequency damage and NZ more than zero timeframe damage)

To test the regression of timeframe damage, prediction test results of high-frequency damage accumulated during an entire flight are displayed in Figure 4.14. The regression plots demonstrate that accumulated high-frequency damage is consistently underestimated and that confusion between flights during which timeframe damage was accumulated, or not, generally only occurs at low damage values, limiting the importance of such errors. There regions where such confusions occur are marked by continuous ovals in Figure 4.14.

4.2.6 Reliability testing of estimates of accumulated fatigue damage by Direct Load & Damage Modelling

The previous sections 4.2.5.2 and 4.2.5.4 verified that DLDM regression and classification results are reasonably accurate. Test results demonstrated regression coefficients of 0.98-0.99 for extreme load prediction and 0.66-0.84 for aggregated timeframe damage predictions Nevertheless, the primary test objective is to determine if the effect of DLDM regression errors can be neglected and if the reliability of DLDM-based estimates of accumulated fatigue damage can be substantiated by the reliability of the S-N working curve only.

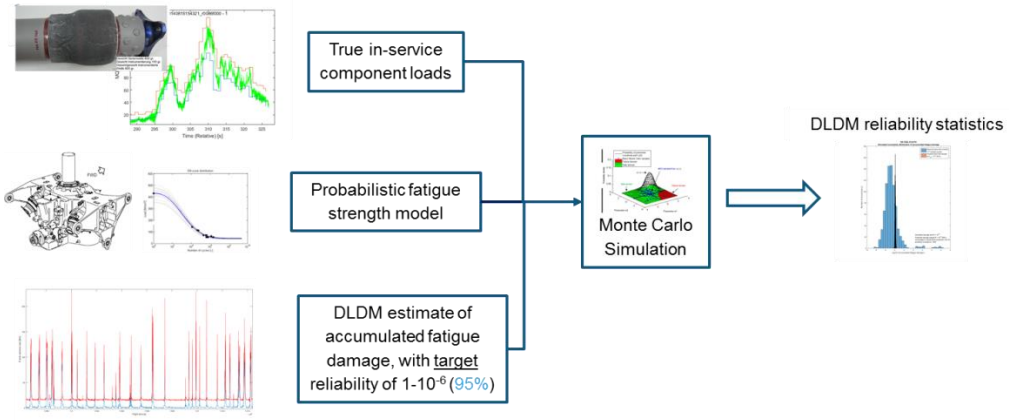


Figure 4.15: Schematic summarising how reliability testing is performed for DLDM-based estimates of accumulated fatigue damage.

4.2.6.1 Definition of method for reliability testing

The methodology to test the reliability of DLDM predictions of accumulated fatigue damage uses simulations based on Monte Carlo simulation and the reliability testing methods detailed earlier in chapters 2 and 3. The testing methodology is conceptually sketched in Figure 4.15 and clarified in more detail in Figure 4.16. In accordance with chapters 2 and 3, the reliability target is set to $\gamma = 10^{-6}$ (95%), i.e. the probability that actually accumulated fatigue damage is higher than predicted shall not exceed 10^{-6} with 95% single-sided confidence.

The test was conducted under several assumptions and conditions:

- All the assumptions on fatigue damage accumulation made in chapter 2 are valid and do not introduce modelling errors, for example the Miner linear damage accumulation hypothesis, rainflow cycle counting [1], a four-parameter Weibull function for S-N curve modelling, and the load spectrum model from chapter 4.1 in AGARD-AG-292 [16] are all assumed to be perfect modelling assumptions. All of these assumptions are accepted by aviation authorities and are common practise in the rotorcraft industry.
- All the assumptions on the modelling of the uncertainty of fatigue strength made in chapter 3, including the simplified Bayesian framework, are valid and do not introduce a modelling error. As already discussed in chapter 3, these assumptions should be reasonable to aviation authorities or are already common and approved practise in the rotorcraft industry.
- The generic prior on the variance of the fatigue strength distribution developed in chapter 3 is valid and applicable
- The predicted uncertainty distribution for fatigue strength is perfect, i.e. the reliability test focuses on the relative influence of DLDM regression errors in comparison to uncertainties about fatigue

strength. This is a reasonable approach because the modelling and mitigation of uncertainty from fatigue strength have been validated before in chapters 2 and 3.

- The recorded loads on helicopters 1 and 2 are the true loads

The reliability test thus does not make use of an artificial reference case as employed in chapter 2. Although this would allow increasing the scope of the reliability test, it would come with a large additional computational cost as the full testing scope would require the generation of artificial flight data and corresponding flight loads. Instead, preference is given to testing reliability using real flight data and true loads, but with an asserted uncertainty model for fatigue strength. This test case is as real and directly applicable as possible. However, future work could include the generation of artificial test cases to do additional reliability testing.

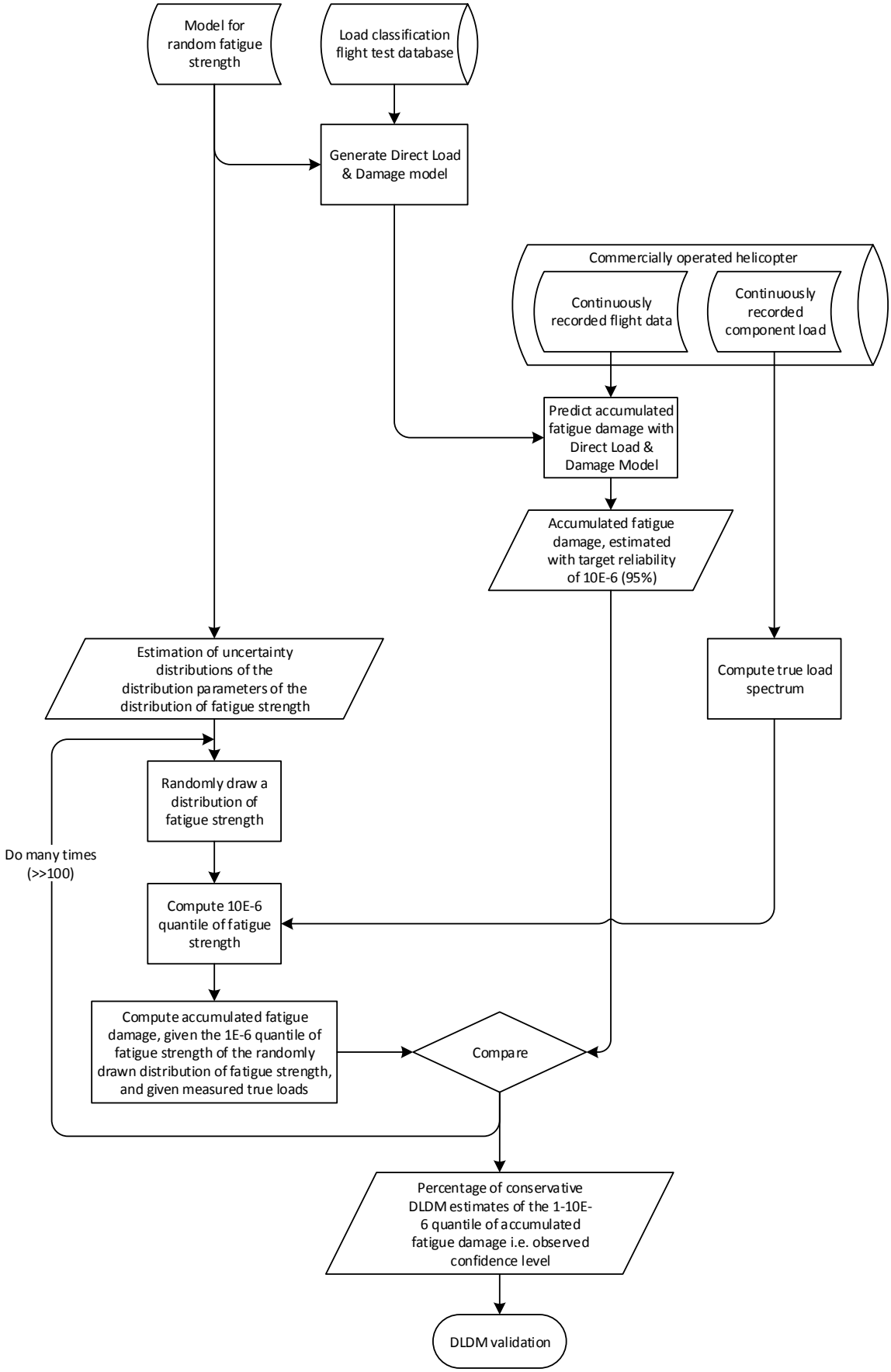


Figure 4.16: Process overview defining how reliability of DLDM estimates of accumulated fatigue damage is tested and benchmarked

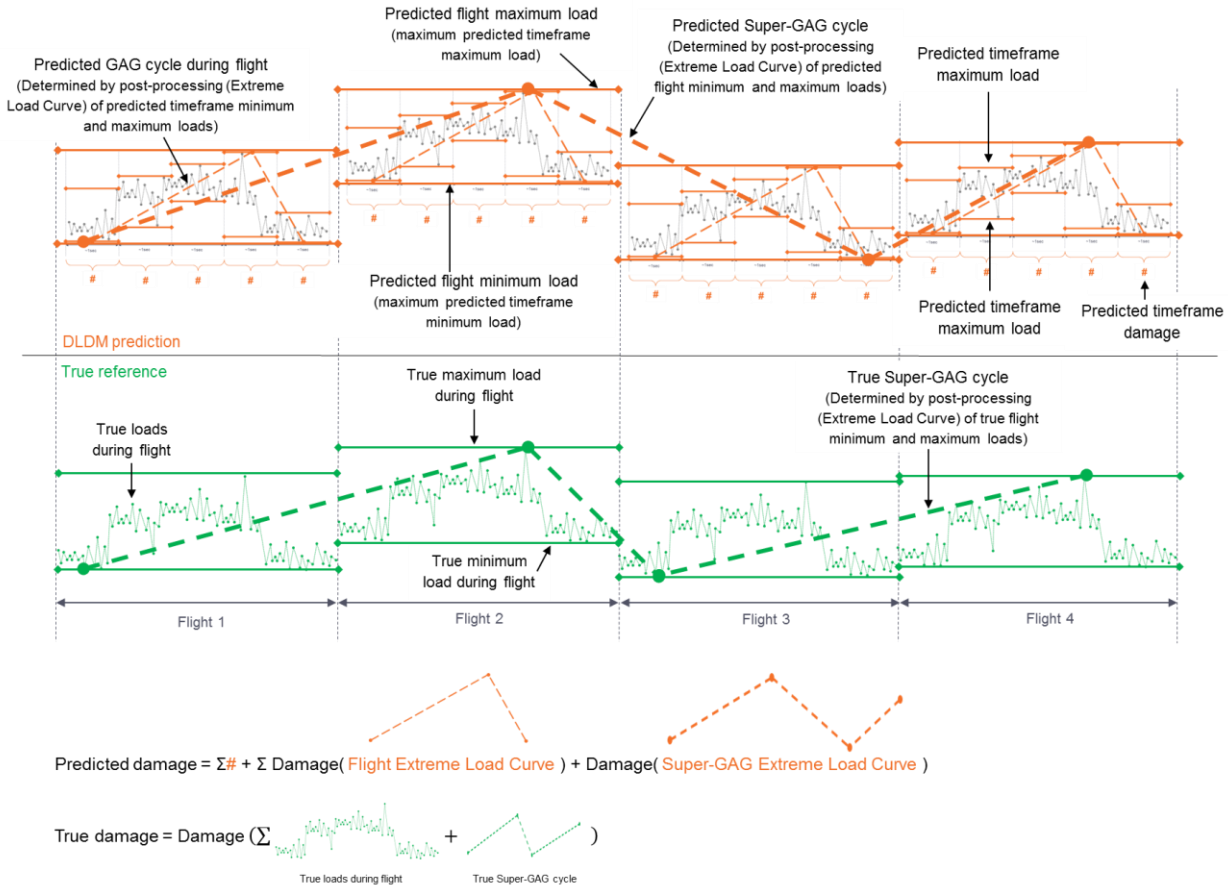


Figure 4.17: Schematic explaining the difference between the load spectrum accumulation model employed by DLDM and the 'true' reference load spectrum created from recorded loads from helicopters one and two that is used as a `true` reference during reliability testing.

The way in which load spectra are modelled by DLDM and for the true reference loads from helicopters 1-2 differs significantly, as outlined in Figure 4.17. In order to save computational costs and prevent loading of the entire flight history in computer memory when artificially repeatedly reshuffling the sequence of flights, for the true reference load spectrum, the load cycle contribution due to transitions between flights is modelled by an additional 'super-GAG' cycle. This super-GAG cycle is determined by the most conservative load signal that can be fitted through the sequence of maximum and minimum loads that occur during each flight and causes that true damage is overestimated in comparison to a simple concatenation of all flight loads. DLDM models low-frequency load cycles by a GAG cycle within each flight and additionally also makes use of a super-GAG cycle. By definition, DLDM should thus overestimate fatigue damage in comparison to the defined true reference as its theoretical load spectrum model is biased more conservatively.

The test procedure essentially consists of generating the uncertainty distribution of accumulated fatigue damage over the test period and comparing this distribution with DLDM's prediction of accumulated fatigue damage. Since the loads on the tail driveshaft have been measured and recorded, the true load spectrum is known; whereas the fatigue strength of the lower housing of the main gearbox is unknown and a randomly distributed variable. For a given distribution of fatigue strength, its 10^{-6} quantile can be computed analytically as well as the correspondingly accumulated fatigue damage. The distribution of fatigue strength itself is however also subject to significant uncertainty, as it has been determined by a small number of full-scale fatigue tests. Using Bayesian statistical modelling from chapter 3, Monte Carlo simulation based on parametric

bootstrapping is employed to generate a large[22] and random set of variations of the fatigue strength distribution, and their corresponding 10^{-6} quantiles. Computing the accumulated fatigue damage according to each sampled 10^{-6} fatigue strength quantile results in the uncertainty distribution of accumulated fatigue strength with a reliability of $\gamma = 10^{-6}$. The upper 95th percentile of this distribution corresponds to the 'true' 95% confidence level and should equal DLDM's prediction of accumulated fatigue life.

4.2.6.2 Results of reliability testing of fatigue damage estimates by Direct Load & Damage Modelling

The simulated uncertainty distributions of $\gamma=10^{-6}$ quantiles of accumulated fatigue damage are displayed in Figure 4.18 for component number 6 (see Table 4-2), i.e. MQF, of helicopters 1 and 2. Comparison with DLDM estimates of accumulated fatigue strength using a working S-N curve with a reliability of $\gamma = 10^{-6}$ (95%) demonstrates that the predictions actually satisfy a $\gamma = 10^{-6}$ reliability requirement with more than 92% confidence.

Some of the simulated fatigue damage values in Figure 4.18 are greater than 1. This corresponds to cases in which randomly sampled fatigue strength was so low that the SLL should have been set to a duration shorter than the number of flight hours accumulated during the test period to guarantee a reliability of $\gamma = 10^{-6}$. All cases for which accumulated fatigue damage exceeds unity should thus have been rejected by the fact that the main gearbox housing actually did not fail on helicopters 1 and 2 during the test period. However, this has been neglected and it is assumed that this modelling inaccuracy does not significantly influence the accuracy of the simulation. This is a conservative assumption since removal or resampling of the data points above unity would reduce the weight of the upper tail and increase the demonstrable confidence level.

Close inspection of Figure 4.18 reveals that the upper 10^{-6} quantile of accumulated fatigue damage estimated by DLDM is also modelled as a distribution. This is the result of an additional test element verifying the size of the database with flights from helicopters 1 and 2. If the period over which data has been collected on these helicopters would have been too small, then non-parametric bootstrapping, i.e. randomly recombining their flights into new and random sequences, would have resulted in a significant variance of accumulated fatigue damage and demonstrable confidence level. However, as this variance was observed to be small, it is concluded that the size of the database is sufficiently large. To reduce complexity and computational costs, all subsequent reliability simulations presented in section 4.3 have therefore been carried out using only the actual sequence of flights.

Repetition of the reliability test for DLDM predictions with a constant target reliability of $\gamma = 10^{-6}$ (95%) for other quantiles than 10^{-6} in Figure 4.19 illustrates the interrelation between the demonstrable probability of failure and confidence level. Higher confidence levels can be achieved, but only for lower reliabilities, and the other way around, higher reliability levels can be substantiated too, but only for lower levels of confidence.

It can thus be substantiated that the influence of DLDM regression errors as previously observed in section 4.2.5 is negligible and can be ignored while substantiating the reliability of the overall estimate of accumulated fatigue damage by the reliability of the working curve only. However, this conclusion may not be generally valid. Since the fatigue strength of component 6, MQF, has been determined by very few full-scale fatigue tests, the uncertainty in fatigue strength is large. It is, therefore, more likely that uncertainty due to DLDM regression errors has a small effect on overall prediction reliability. To test the generality of the negligibility of DLDM regression errors, their relative influence is artificially increased by reducing the standard deviation of the distribution of normalised fatigue strength. In practise, this has been realized by introducing a fixed σ-

[22] The simulation contains 10^4 samples and uncertainty due to limited Monte Carlo sample size is negligible for quantile estimates of approximately 0.95 according to the standard estimator of Monte Carlo simulation precision)

multiplication factor adjusting all sampled values of the standard deviation of normalised fatigue strength σ_{SF} by a strength factor ϕ :

$$\sigma_{SF} = \phi \cdot \sigma_{SF} \tag{4.6}$$

Simulation results for σ-factors of 0.75 and 0.5 are presented in Figure 4.20 and Figure 4.21 and in the comparative summary overview in Table 4-3. It demonstrates that with decreasing uncertainty in fatigue strength, the significance of DLDM regression errors increases and the confidence by which a reliability of $\gamma=10^{-6}$ can be substantiated drops by more than 10% below the intended target.

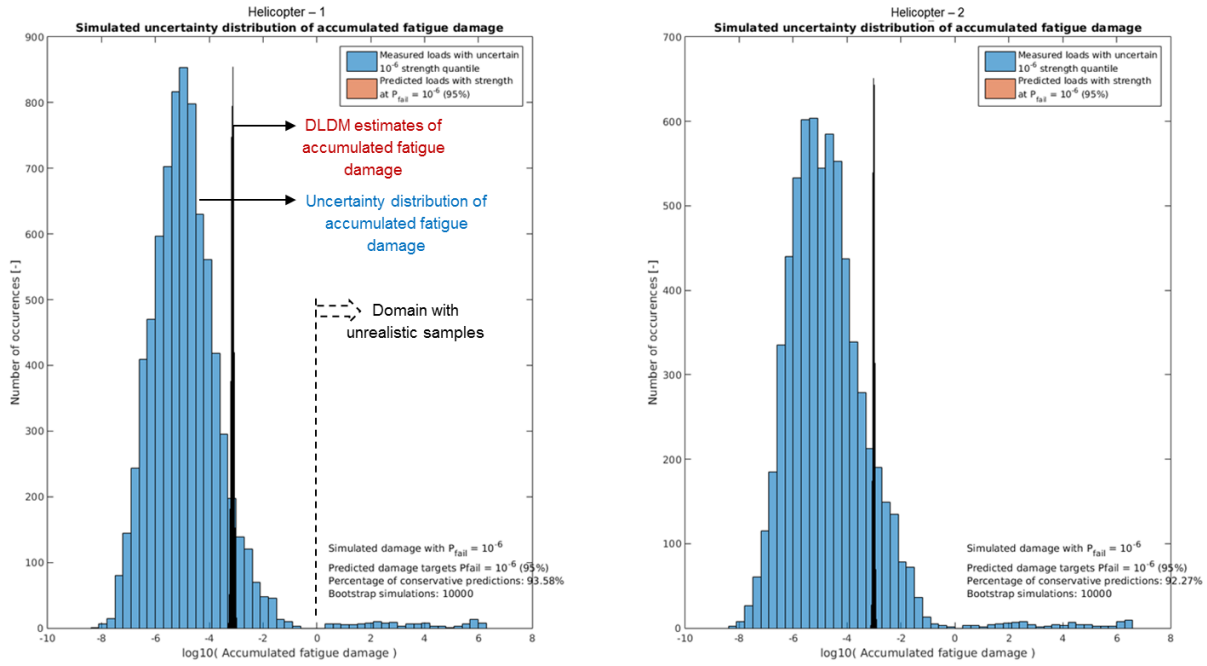


Figure 4.18: Charts comparing the distribution of the `true` uncertainty distribution of the upper 10^{-6} quantile of accumulated fatigue damage given the `true` distribution of fatigue strength (blue) and the distribution of DLDM estimates caused by bootstrapping of the dataset containing `true` reference loads (red) for helicopters 1 (left) and 2 (right). (Due to figure scaling, the red bars contain the distribution of DLDM predictions are not visible and visually aggregated to thick black lines.)

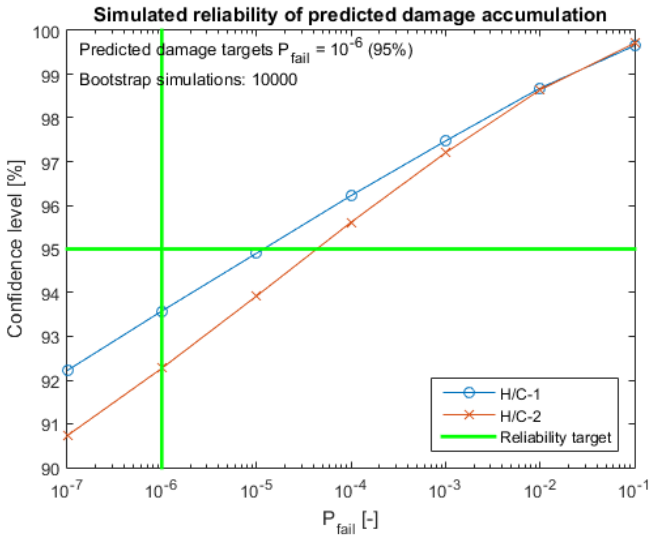


Figure 4.19: Graph showing the demonstrable reliability level of DLDM predictions that have a target reliability of $\gamma=10^{-6}$ (95%). The graph also shows how the demonstrable reliability level can be varied as a function of the demonstrable reliability quantile and confidence level (i.e. not as function of DLDM reliability target, which is constant).

This means that DLDM's reliability substantiation methodology is only valid for situations where DLDM regression is accurate and where there is a relatively high uncertainty in fatigue strength. The simulation result also indicates that a similar limitation applies to the analytical substantiation method verified in chapter 2 and highlights the recommendation to better determine the conditions under which this simplified method can be used.

It can also be observed from Figure 4.18 and Figure 4.20 that the bootstrap uncertainty due reshuffling of the flights increases with decreasing uncertainty in fatigue strength. Since the lower quantile of fatigue strength increases with a decreasing standard deviation of its distribution, less load cycles become damaging. As the occurrence of damaging conditions becomes rarer, the effect of reshuffling the flights becomes more pronounced. Nevertheless, it is hence further still assumed that any uncertainty due to the limited number of tested flights may be neglected.

Table 4-3: Table showing how the confidence level with which a $\gamma=10^{-6}$ reliability level can be demonstrated for estimates of the accumulated fatigue damage of the lower gearbox casing made by DLDM for helicopters one and two reduces with synthetically lowering the variance of fatigue strength (decreasing σ-factor). The DLDM predictions are made with a target reliability of $\gamma=10^{-6}$ (95%).

σ-factor	H/C-1	H/C-2
1	93.6	92.3
0.75	85.4	84.4
0.5	62.7	61.0

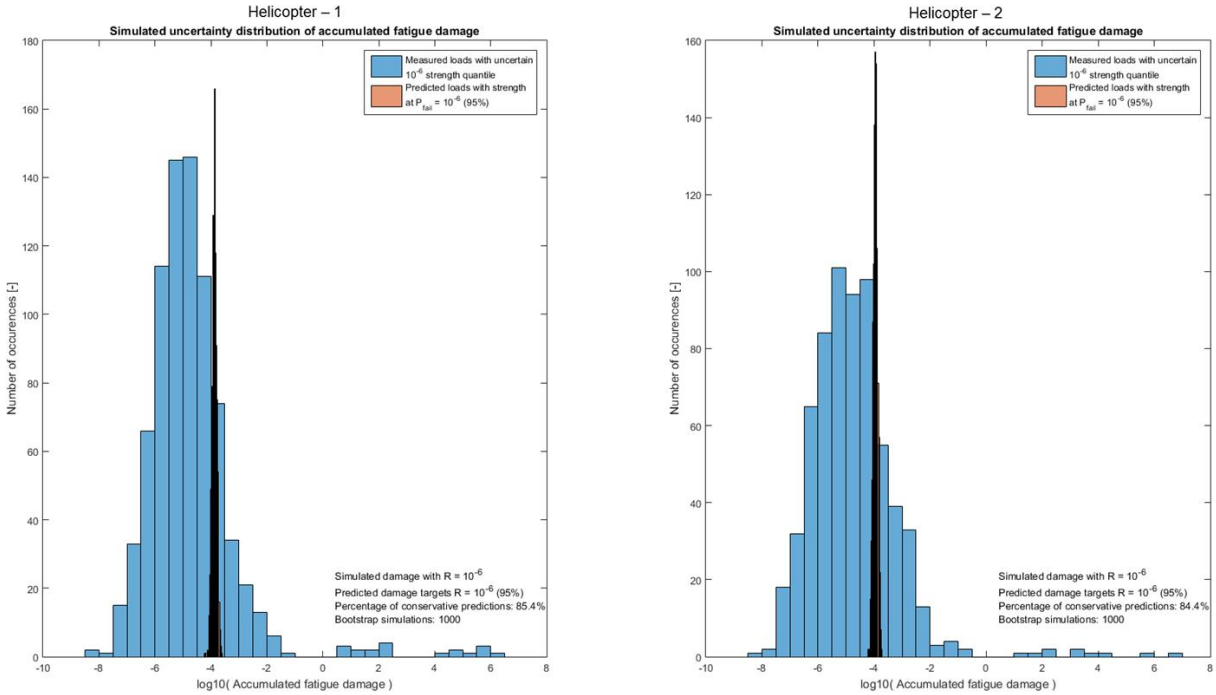


Figure 4.20: Charts comparing the distribution of the `true` uncertainty distribution of the upper 10^{-6} quantile of accumulated fatigue damage given the `true` distribution of fatigue strength (blue) and the distribution of DLDM estimates caused by bootstrapping of the dataset containing `true` reference loads (red) for helicopters 1 (left) and 2 (right). (Due to figure scaling, the red bars contain the distribution of DLDM predictions are not visible and visually aggregated to thick black lines.) The simulation uses an artificially reduced value for the standard deviation of the fatigue strength of the lower gearbox casing by a σ-multiplication factor of 0.75.

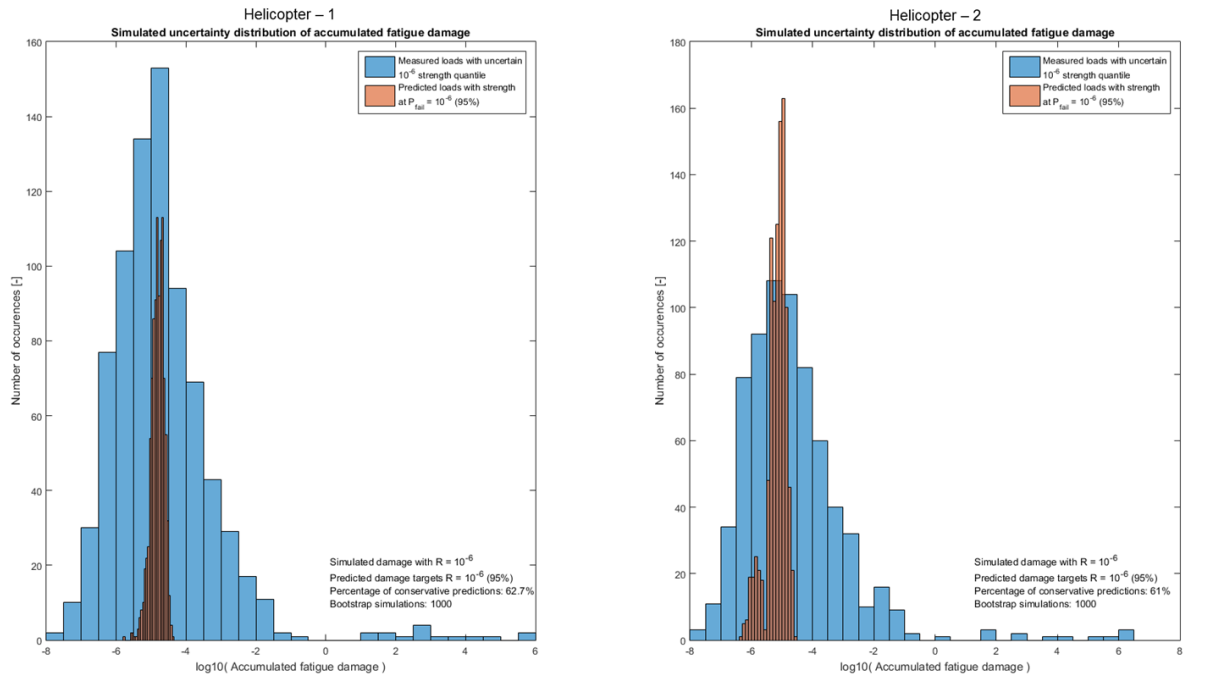


Figure 4.21: Charts comparing the distribution of the `true` uncertainty distribution of the upper 10^{-6} quantile of accumulated fatigue damage given the `true` distribution of fatigue strength (blue) and the distribution of DLDM estimates caused by bootstrapping of the dataset containing `true` reference loads (red) for helicopters 1 (left) and 2 (right). (Due to figure scaling, the red bars contain the distribution of DLDM predictions are not visible and visually aggregated to thick black lines.) The simulation uses an artificially reduced value for the standard deviation of the fatigue strength of the lower gearbox casing by a σ-multiplication factor of 0.5.

4.2.7 Benchmarking of in-service application of Direct Load & Damage Modelling

Since DLDM has been successfully verified for component 6, i.e. MQF or the lower main gearbox housing loaded in torque, it is possible to evaluate the benefit of VFLM for this component. In Figure 4.22, DLDM estimates have been normalized with respect to fatigue damage accumulation as expected by the Design Mission Profile (DMP). The fatigue damage accumulation rates for helicopters 1 and 2 demonstrate that fatigue life extensions by a factor of five are feasible. SLL extensions of this magnitude can effectively turn component 6 into a component with a practically unlimited life, eliminating the need to replace the component during the commercial life of the helicopter it is installed on. Although the estimated fatigue damage accumulation for helicopter 3 remains well below the conservative design assumption, this helicopter also features an accumulation rate steeper than the DMP expects for a prolonged period of time. This underlines that the DMP accumulation rate is only valid and conservative over very long periods of time, e.g. >>100FH, and that usage severity of helicopters from the same operator under a similar mission profile can still differ significantly. For the other six tested components, DLDM results indicate that fatigue life extensions in excess of a factor of ten are feasible for all helicopters. However, these results should be interpreted with caution as DLDM's validity for these components has not been fully verified.

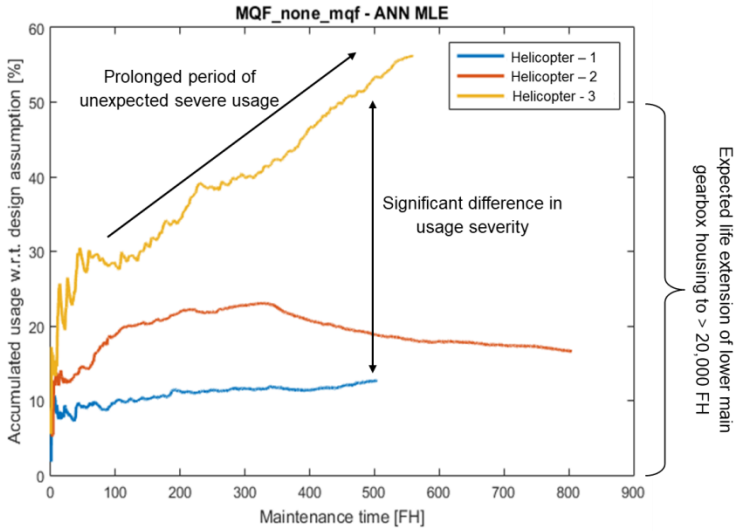


Figure 4.22: Chart show how the rate of fatigue damage accumulation that is predicted by DLDM for the lower gearbox casing differs between in-service helicopters. CAUTION: Service life limits underlying this graph are computed for academic purposes only and are not approved by any OEM or airworthiness authority.

Finally, the source of damage accumulation as modelled by DLDM is differentiated in Table 4-4. It follows that the primary contributor to accumulated fatigue damage can differ significantly between components and that accurate modelling of both timeframe damage and timeframe extreme loads is important. The result also shows a limitation in the scope of the performed test of DLDM. Since there is virtually no damage caused by timeframe damage for component 6, the accuracy of recognition of timeframe damage is thus not tested thoroughly. For future work, it is recommended to define an artificial S-N curve such that the proportion of high-frequency damage is increased and to repeat the reliability test for such an artificial test in order to more accurately validate the accuracy of DLDM's load spectrum model under real circumstances.

Table 4-4: Table identifying how DLDM models the accumulated fatigue damage on average for helicopters 1-3 as a sum comprising high frequency timeframe damage, low-frequency damage caused by timeframe extreme loads, and a super-GAG cycle caused by flight-by-flight extreme loads.

		DLDM damage source [%][23]		
Component ID	**Designation**	**HF**	**LF**	**Super-GAG**
1	FBTHETA	2	18	80
2	FBTHETAP	20	40	40
3	FKAR	54	5	41
4	FSTA	33	19	48
5	FSTY	74	0	26
6	MQF	0	68	31
7	MTM	0	39	60

4.3 Probabilistic Load & Damage Modelling for Virtual Fatigue Life Monitoring

As an alternative to DLDM, Probabilistic Load and Damage Modelling (PLDM) is introduced to mitigate the shortcomings of DLDM. Whereas DLDM assumes that the reliability influence of regression errors may be neglected, PLDM attempts to fully model and mitigate the influence of regression errors based on the simulation-based reliability substantiation concept successfully demonstrated in chapter 2. The modelling and reliability substantiation method of PLDM will be introduced in sections 4.3.1 and 4.3.2 and its validation by real in-service flight data is presented in section 4.3.4, followed by an assessment of the achievable service life limit extensions in section 4.3.5. A dedicated analysis to the validity of some of the modelling assumptions employed by PLDM for its prediction error models is presented in section 4.3.3.

4.3.1 Definition of modelling approach

A typical load profile of the load on the Fenestron driveshaft, i.e. the loading of component 6, or MQF, during a flight is displayed in Figure 4.23, along with DLDM-predicted timeframe maximum and minimum loads. The flight consists of some manoeuvring during take-off during which the Fenestron is loaded intensively, followed by long periods of steady forward flight during which only one manoeuvre is executed. At the end of the flight, some aggressive manoeuvring takes place before landing.[24]

[23] Percentages are rounded to the nearest integer and may not sum to 100%

[24] It can be seen that the true load signal has not been recorded during the initial take-off phase. Because the boot-up and internal calibration procedure of the recorder can take significant time, recording could not allways start before flight commenced. The recording of flight data begins as soon as the helicopter takes off and is cut immediately upon touch-down. Take-off and touchdown have been determined by an algorithm verified by experts and extensive manual flight data analysis.

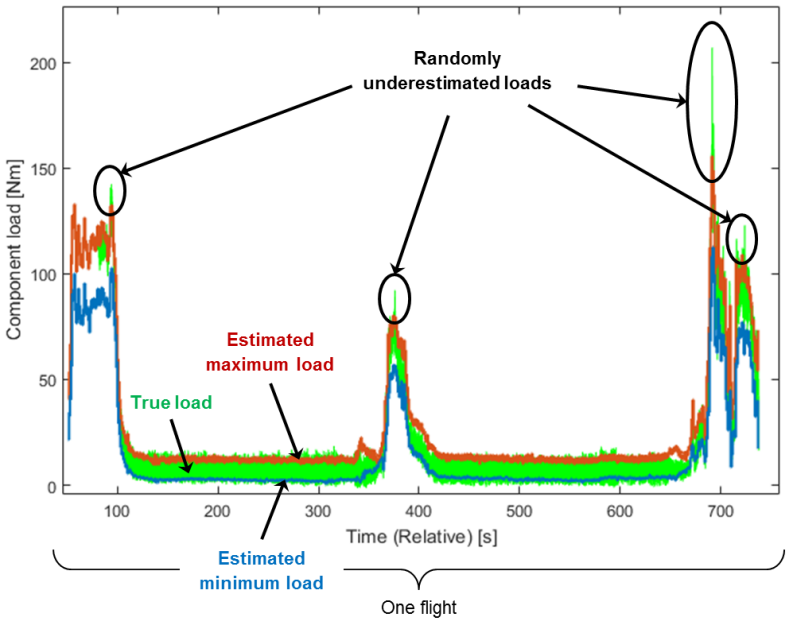


Figure 4.23: Chart comparing DLDM predictions for timeframe extreme loads with independently recorded `true` loads during a flight of helicopter one or two. The chart also illustrates a case where `true` loads could not be recorded during the beginning of the flight.

Although DLDM predictions are generally accurate, DLDM only makes a prediction of the most likely extreme loads occurring within a timeframe. In Figure 4.23 it is exemplified how DLDM, therefore, fails to capture the occurrence of brief but intense load excursions. The occurrence of such spikes cannot be predicted accurately since the recorded flight and vehicle data cannot be correlated with these spikes accurately. It could be imagined to install additional sensors, collect more data, and vastly expand the size of the load classification flight database in order to improve the achievable accuracy by non-linear statistical data modelling. However, such an approach would be very costly and may still not lead to the desired level of accuracy. Instead, it is proposed to accept that in-flight loads can only be predicted with limited accuracy and that as long as the occurrence of regression errors is predictable; their effects can be mitigated statistically.

Hence, PLDM no longer makes a point-prediction of timeframe damage, minimum load, and maximum load, but rather a probabilistic estimate thereof. For each timeframe, the high-frequency fatigue damage, minimum load, and maximum load is predicted as a statistical distribution that specifies how likely the occurrence of a certain value has been. This concept is schematically illustrated in Figure 4.24, where PLDM's probabilistic load spectrum model is summarized in a simplified fashion. Taking a sample from all the timeframe distribution estimates of maximum and minimum load allows computing an extreme load curve, as DLDM did before, and adding the corresponding low-frequency damage to the sum of a sample from all timeframe damage distribution estimates, to yield a sample of accumulated fatigue damage. Sampling many times results in a distribution of accumulated fatigue damage and allows computing the desired lower quantile of accumulated fatigue damage in order to meet a specified reliability requirement. PLDM thus does not make use of fleet-generic multiplication factors to ensure reliability but instead mitigates prediction uncertainties for each helicopter separately by probabilistic simulation applied to individual flight records.

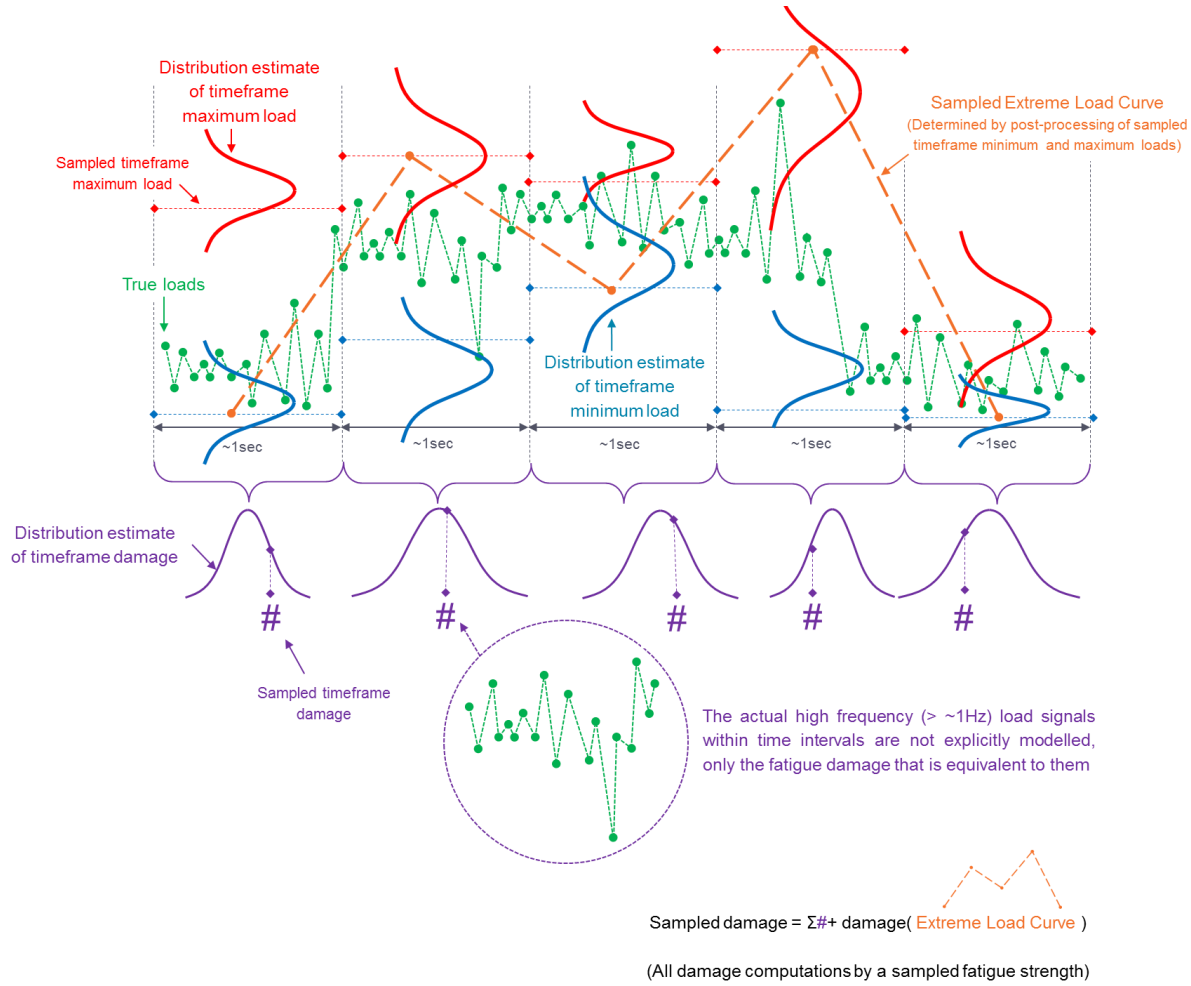


Figure 4.24: Schematic summarising how PLDM models accumulated fatigue damage in the same way as DLDM but with probabilistic estimations of the determining parameters, i.e timeframe extreme loads and timeframe damage. The schematic also illustrates how accumulated fatigue damage is computed from extreme load and timeframe damage samples from five subsequent timeframes. Distributions are not drawn to scale.

Actual results from PLDM for extreme load recognition in Figure 4.25 illustrate some of the benefits and features. The highest load, occurring approximately 692s after flight commences, is underestimated by the Maximum Likelihood Estimation (MLE) prediction. However, the prediction of the high load event acknowledges that the prediction comes with relatively high uncertainty and the actual load falls well within the upper 95% probability range. During the prediction period, it can be seen that the uncertainty ranges are adapted realistically and that actual loads are bounded reasonably well by the 5% and 95% lower and upper predicted quantiles.

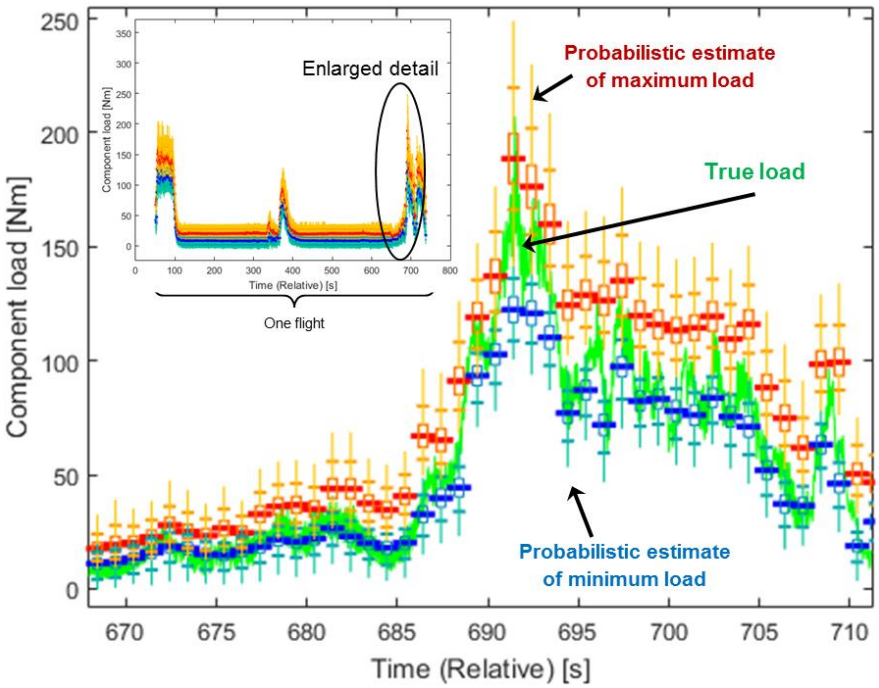


Figure 4.25: Chart showing a comparison between probabilistic estimates and independent 'true' recordings of the minimum and maximum torque on the Fenestron driveshaft during a flight of helicopter one or two.[25]

4.3.2 Method to estimate accumulated fatigue damage by Probabilistic Load & Damage Modelling

The basic fatigue damage model that PLDM uses is introduced in section 4.3.2.1 and equivalent to the model used by DLDM. In contrast to DLDM, PLDM however uses explicit statistical simulation to estimate conservative quantiles of accumulated fatigue damage. The basic modelling processes that PLDM uses for its reliability prediction simulations is introduced in section 4.3.2.2 and is similar to the simulation-based reliability modelling process already introduced and validated in sections 2.5 and 2.6 respectively. The practical implementation of PLDM`s reliability model by means of Subset Simulation is also similar to the practical implementation of the simulation-based SLL substantiation model from section 2.5 and is illustrated in 4.3.2.3. PLDM`s specific implementation for confidence level analysis is introduced in section 4.3.2.4, which also similar to the implementation that was introduced for the simulation-based SLL substantiation model in section 2.5. The core load spectrum prediction model that PLDM uses is equivalent to the implementation that DLDM uses and which was introduced in section 4.2.1.1. However, the implementation and generation of the specific probabilistic prediction models that are unique to PLDM is introduced in 4.3.2.5.

A general overview of the process to generate a PLDM prediction model and to apply it to a new flight record of another helicopter is shown in Figure 4.26. The flight test data base and the flight data mentioned in elements [A] and {B] respectively are the same as introduced for DLDM in section 4.2.3. The method generate a probabilistic fatigue strength model mentioned in element [C] is the Bayesian model for fatigue strength introduced and tested in sections 3.4 and 3.5. The process to generate the probabilistic prediction models for timeframe damage and extreme loads mentioned in element [D] is introduced in section 4.3.2.5 and in Figure 4.36. The data quality and assurance module mentioned in element [E] is not introduced in detail in present work but was generally introduced in section 4.2.3.2. The basic Monte Carlo process contained in element [H] to estimate a quantile of accumulated fatigue damage based on a series of predicted timeframe load and damage distributions, as well as a random fatigue strength model, is detailed further in Figure 4.27. This process underlies the more efficient implementation for quantile estimation by Subset Sampling introduced in

[25] Boxplot describing probabilistic extreme load prediction features markings for the {0.005 0.05 0.25 0.5, 0.75 0.95 0.995} quantiles, from bottom to top.

4.3.2.3. The implemented additional process to estimate confidence levels, and thus to be able to make conservative tolerance interval predictions for accumulated fatigue damage, which is mentioned in element [J] is elaborated in section 4.3.2.4 and in Figure 4.35.

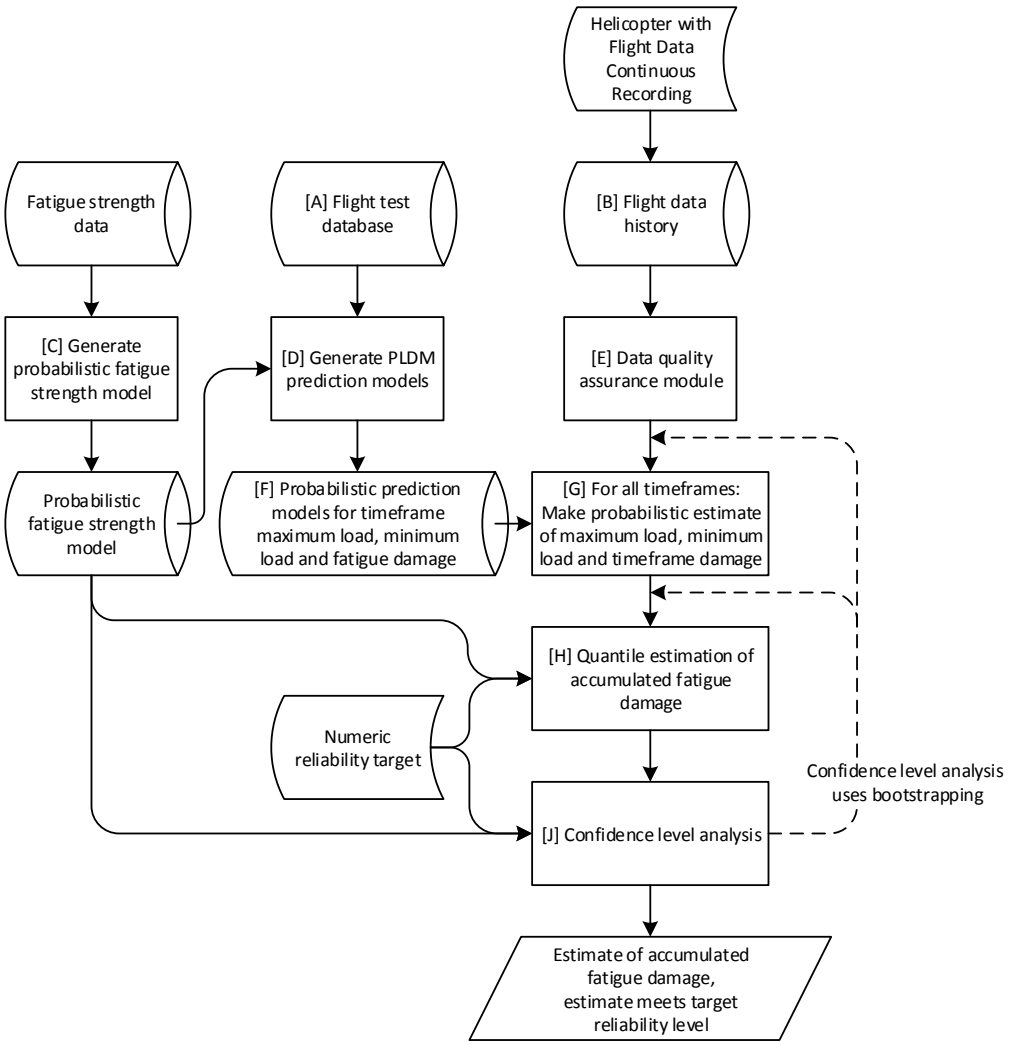


Figure 4.26: Process overview how PLDM model are generated and used to make usage-based estimations of accumulated fatigue damage.

4.3.2.1 Definition of modelling assumptions for fatigue damage modelling

The probabilistic framework of PLDM makes use of the modelling assumptions which are listed below:

- Prediction uncertainty is time-independent, i.e. the prediction error between two subsequent timeframes is uncorrelated
- Prediction errors between timeframe damage, maximum load, and minimum load are independent
    - Some feasibility checks on predictions and samples from timeframe uncertainty distributions are however applied by custom filters, as further detailed in Appendix I.
- The effect of uncertainty due to modelling limitations or limited sampling is negligible, i.e. uncertainty or limitations due to limited sample sizes during subset simulation and a small number of bootstrap iterations may be neglected while substantiating the target reliability requirement. This limitation can be lifted though by increasing computational costs or by explicit inclusion of uncertainty simulation for subset simulation estimation uncertainty and bootstrap estimation uncertainty. However, the inclusion of these uncertainties in chapter 2 did not lead to significant improvements in accuracy.

- The prediction models are reasonably accurate, i.e. within a bootstrapped set of prediction models there exist a reasonably large number of prediction models that feature a reasonably small prediction bias
- All assumptions listed for DLDM before in section 4.2 are valid, except for the negligibility of prediction errors
- The example implementation of PLDM presented in this chapter assumes that uncertainty distributions for fatigue strength and timeframe damage and extreme loads can be modelled by continuous and unbounded distributions. This assumption is however not strictly required and can be lifted straightforwardly by modifying the uncertainty distribution models, or other measures with the same effect.
- Prediction errors between timeframe- damage, minimum load, and maximum load are uncorrelated, i.e. it is assumed that underestimation of maximum load does not increase the probability to underestimate minimum as well.

In section 4.3.3 it is shown that some of these assumptions do not hold in practise. However, these assumptions are important to reduce the complexity and computational costs of the PLDM framework. Validation tests in section 4.3.4 actually demonstrate that PLDM's reliability substantiation does consistently meet a stringent $\gamma=10^{-6}$ (95%) and that its probabilistic modelling framework is sufficiently accurate, in contrast to the limited performance of DLDM discussed earlier.

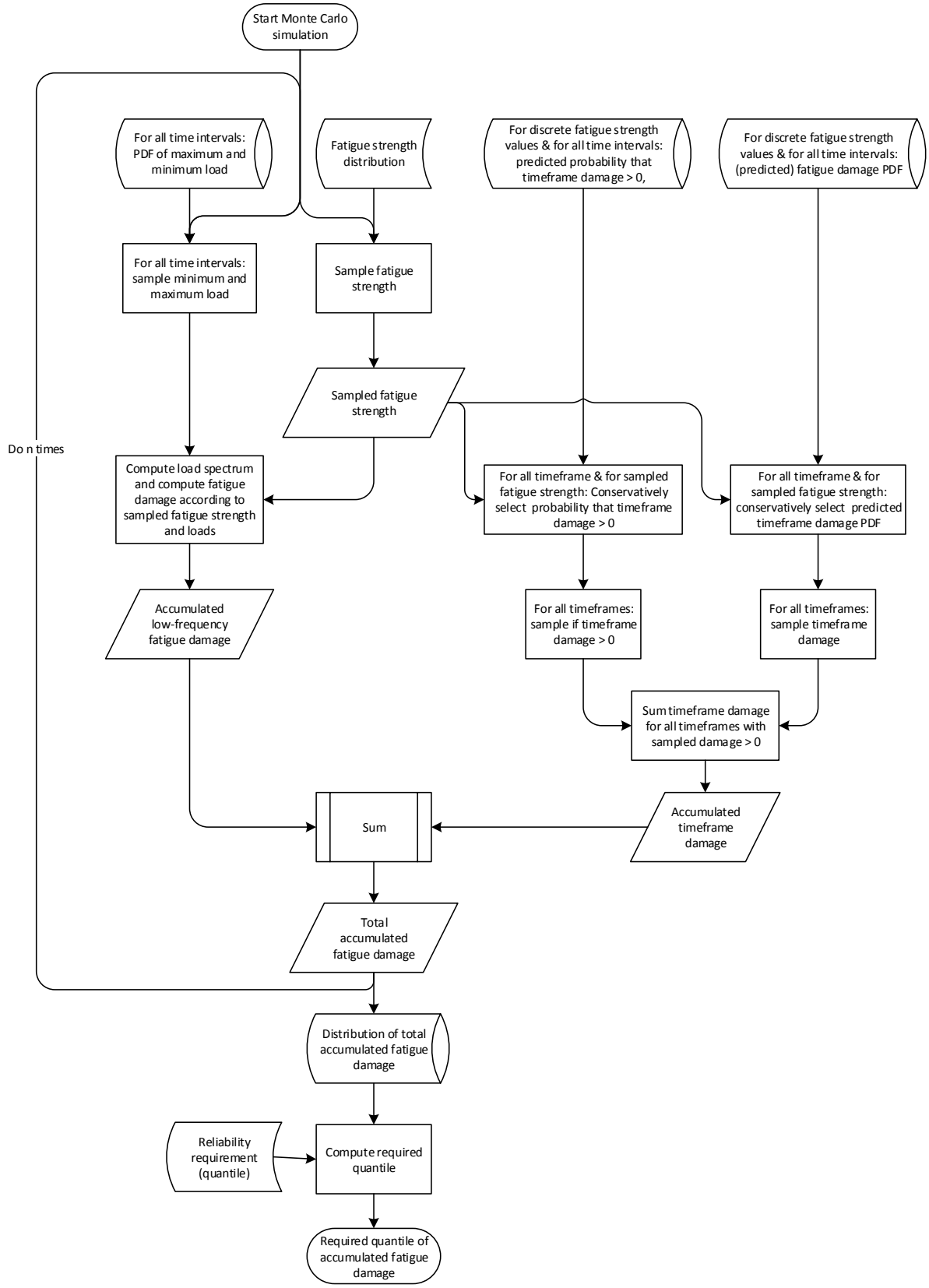
4.3.2.2 Definition of Monte Carlo simulation method to estimate a quantile of accumulated fatigue damage

As outlined before in chapter 2, it is not feasible to estimate a quantile of accumulated fatigue strength analytically while using PLDM's complex and non-linear probabilistic modelling framework. Quantiles are therefore estimated by means of stochastic simulation. There are many simulation-based methods to estimate a statistical quantile, some of which are summarized in Appendix A, but the simplest process by which PLDM's quantile estimation method can be clarified is by means of a Monte Carlo simulation, an implementation of which is in shown in Figure 4.27. Having defined the way in which a Monte Carlo simulation can be carried out, implementations using other methods, such as Subset Sampling shown in the next section 4.3.2.3, can be derived.

In summary, a single Monte Carlo sample can be computed by sampling a load sequence from PLDM's probabilistic load estimations for a sequence of timeframes, and by sampling a value for fatigue strength, and then using this fatigue strength to compute a low-frequency damage value. Accumulated high frequency damage can then be computed similarly for the sampled value of fatigue strength. Summing high and low frequency damage finally gives a sample for total accumulated fatigue damage given the flight history under evaluation. Repeating this sampling procedure many times with randomly sampled values for loads, fatigue strength and timeframe damage results in a simulated uncertainty distribution for accumulated fatigue damage, from which finally a required conservative quantile can be computed.

PLDM's Monte Carlo simulation process in Figure 4.27 starts with pre-computing a stochastic prediction of timeframe damage for all timeframes and relevant discrete intervals of fatigue strength. The introduction of these discrete intervals improves computational efficiency as they reduce the number of fatigue strength values for which timeframe damage must be predicted. Since timeframe damage is predicted directly by PLDM and as timeframe damage is a function of random fatigue strength, this implementation prevents PLDM from requiring the generation of a very large number of prediction models, one for each possible value of fatigue strength.[26] A more detailed introduction of these discrete intervals is provided in Appendix sections I.4 and I.5.

[26] As an alternative implementation, it can be envisioned to include fatigue strength as a predictive feature for .timeframe damage prediction. In practise, this would however cause practical problems and significant additional



complexity in the implementation of subset simulation as this would introduce a correlation between the proposal distributions and sampled values for fatigue strength. A more detailed study to the feasibility and computational efficiency of this alternative implementation vaiant is recommended for future work.

Figure 4.27: Process flow defining a Monte Carlo simulation that can be used by PLDM to estimate a required quantile of accumulated fatigue damage. The process can be considered as equivalent to the high-level process element [H] in Figure 4.26.

4.3.2.3 Introduction to Subset Simulation for Probabilistic Load & Damage Modelling

Following the implementation of the simulation-based substantiation model in chapter 2, which uses a fatigue damage accumulation model very similar to DLDM and PLDM, Subset Simulation [112, 54, 113] is used to estimate a quantile of accumulated fatigue damage instead of brute-force Monte Carlo simulation. Subset simulation is presented in more detail in Appendix A with specific implementation details in Appendix I and was introduced in section 2.5.3.2. This section presents some simulation results to illustrate the specific implementation for PLDM.

The subset distributions in Figure 4.28 illustrate how Subset Simulation generates progressively more severe distributions of PLDM-predicted fatigue damage. In the illustrated case, PLDM would estimate that the probability that the example component has accumulated a normalized fatigue damage value of more than about 10^{-3} is at most 10^{-6}.

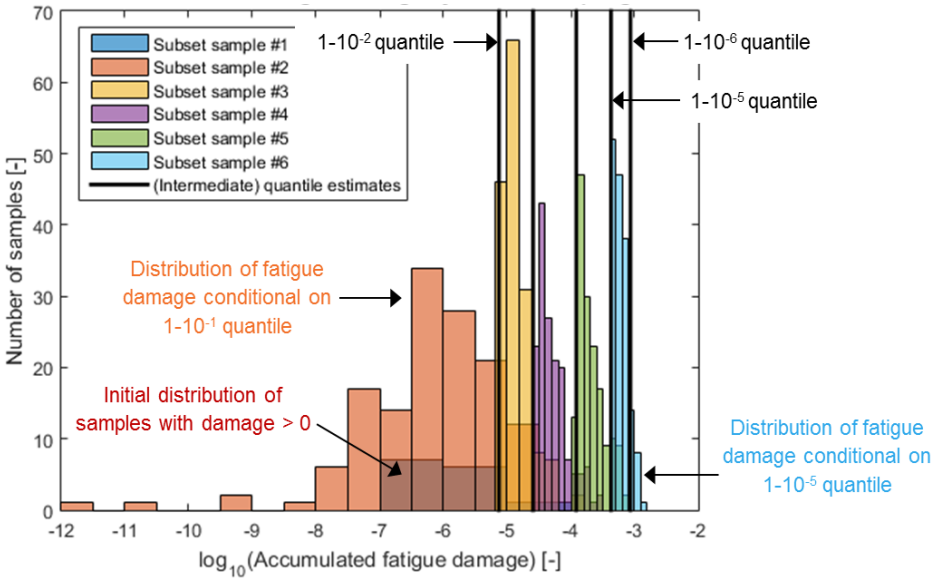


Figure 4.28: Chart illustrating how PLDM uses Subset Simulation to estimate a conservative $\gamma=10^{-6}$ quantile of accumulated fatigue damage by a lower gearbox casing.

Simultaneously, as subsets with sampled fatigue damage become more severe and unlikely, Figure 4.29 illustrates how normalized fatigue strength reduces correspondingly. Transforming these samples of fatigue strength by the inverse of the cumulative distribution of normalized fatigue strength, as in Figure 4.30, reveals that the 10^{-6} quantile of accumulated fatigue damage is effectively substantiated using a fatigue strength quantile of approximately 10^{-4}. This means that the remaining two orders of magnitude of reliability are mostly coming from conservative timeframe extreme loads. This clearly illustrates that PLDM indeed substantiates its reliability by a combination of conservative strength and load estimates, in contrast to DLDM which uses conservative strength only.

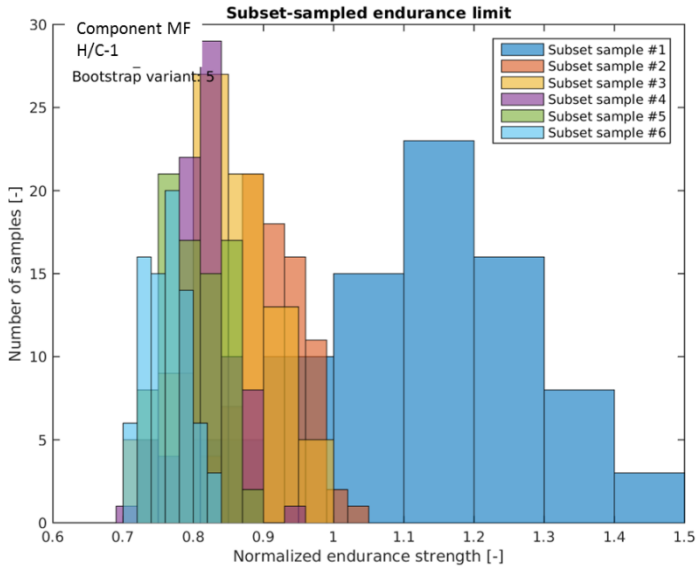


Figure 4.29: Chart illustrating how sampled values for the fatigue strength of a lower gearbox casing reduce as the subsets during Subset Simulation become more severe and correspond to increasingly unlikely events.

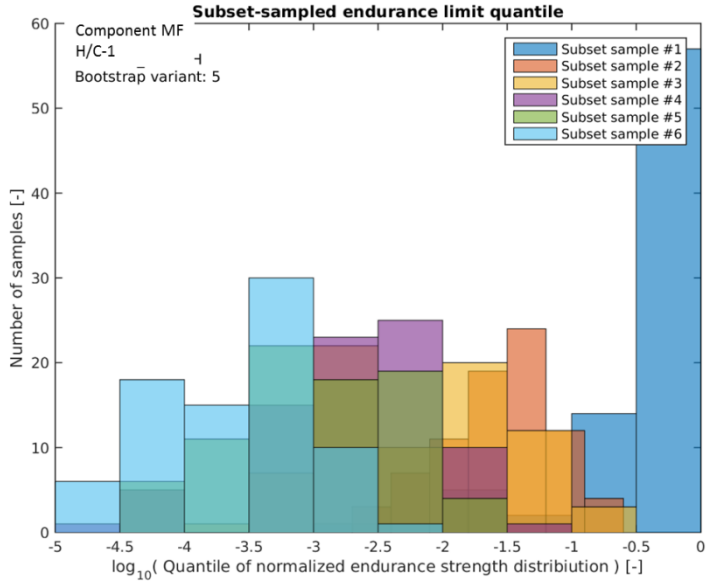


Figure 4.30: Chart illustrating how increasingly severe and unlikely values for fatigue strength are being sampled as the Subset Simulation process progresses towards more unlikely and severe cases of accumulated fatigue damage. The example also demonstrates that the sixth subset sample contains fatigue strength values approximately corresponding to the 10^{-5} to 10^{-3} quantiles of the distribution of fatigue damage.

Subset samples in Figure 4.31 of maximum load, summarized as the maximum load sampled over all timeframes, illustrate that overall maximum load is initially roughly invariant and remains distributed approximately in the region of maximum observed loads during load classification flights. During the last two to three iterations, maximum load, however, increases sharply to load levels well beyond anything observed during load classification flights.

Looking at the time domain representation of a sample of timeframe extreme loads from the sixth and last iteration of subset simulation in Figure 4.32 identifies that during this example the extremely high load is only sampled for a single timeframe. It is likely that the single occurrence of this unusually high load has a significant influence on overall accumulated fatigue damage. Since timeframe extreme loads are sampled completely independently from each other, and since the number of one-second timeframes necessary to describe the entire service history of a component easily reaches an order of magnitude of 10^7, PLDM statistically expects that some extremely rare and high load events will occur. During subset simulation, these load samples are pushed even further into the tails of the predicted extreme load distributions as the cases of accumulated fatigue damage become ever more severe and remote. Naturally, though, exceedance of the static strength of the component cannot have occurred, as it is known that the component did not fail so far, and loads sampled exceeding the sampled static strength are thus rejected and re-sampled. As fatigue strength reduces during subset simulation, this upper limit becomes tighter and more restrictive. Precise definitions of the applied filtering and rule-based re-sampling processes are given in Appendix I.

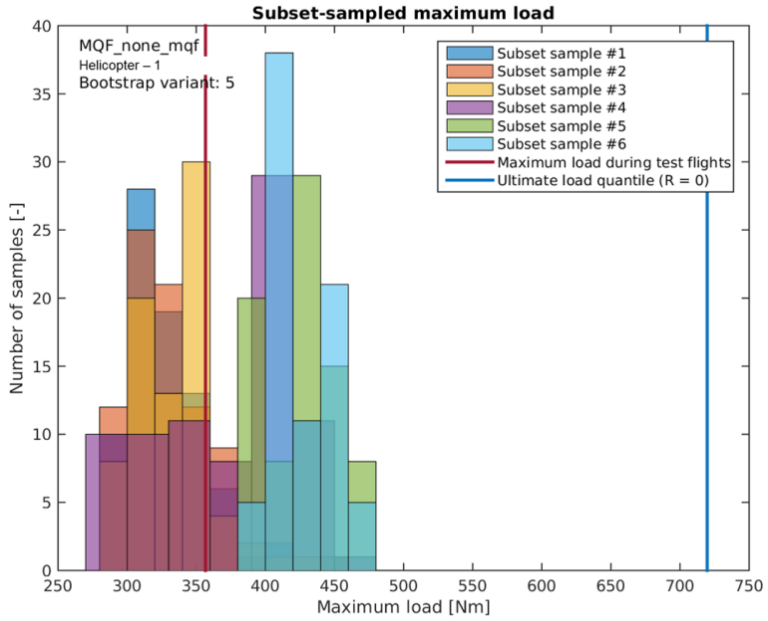


Figure 4.31: Chart illustrating how samples of maximum load increase as Subset Simulation moves towards subsets with ever more unlikely events and more severe cases of accumulated fatigue damage for the lower gearbox casing of helicopter-1.

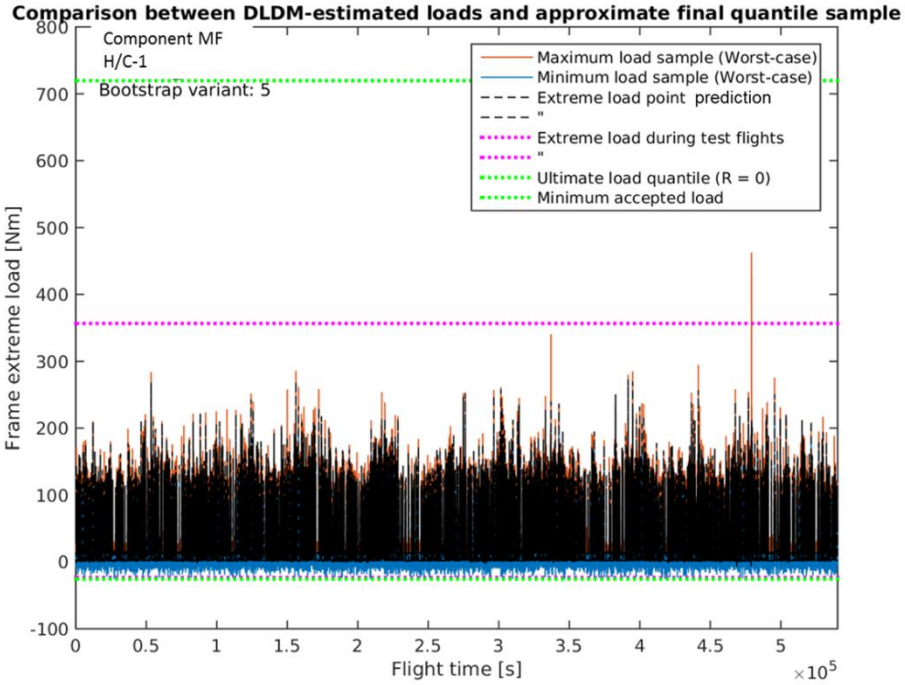


Figure 4.32: Graph comparing the initial stage-1 MLE point estimations of the maximum and minimum Fenestron torque during timeframes with the endurance limit and sampled extreme loads that are determined by Subset Simulation to correspond to a case of accumulated fatigue damage with $\gamma=10^{-6}$ reliability. The illustrative case also shows how loads sampled during Subset Simulation can be significantly higher than the initial Maximum Likelihood point estimates and how sampled values for maximum load are allowed to incidentally exceed the maximum torque value ever observed during LCF flights.

4.3.2.4 Definition of method to for confidence level analysis

To mitigate uncertainty about the true distribution of fatigue strength, PLDM makes use of confidence level analysis. If distribution parameters for fatigue strength have been estimated by a small number of test results, as is generally the case for rotorcraft, then these fitted distribution parameters may randomly deviate from their true values. Chapters 2 and 3 previously introduced a method to mitigate this uncertainty using Bayesian statistics and parametric bootstrapping and demonstrated that this uncertainty can have a significant influence on the reliability of estimated fatigue damage.

The effectiveness of parametric bootstrapping to mitigate uncertainty from a limited number of tests to determine the distribution of fatigue strength has already been demonstrated by the simulation-based substantiation of regular SLLs introduced in chapter 2. For PLDM, the method consists of repeating the quantile estimation method summarized previously in section 4.3.2.2 and in Figure 4.27 for many randomly selected variants of the distribution of fatigue strength, as outlined in Figure 4.33. Variants of the distribution of fatigue strength are sampled according to the Bayesian method introduced earlier in chapter 3. A result from bootstrapping by PLDM shown in Figure 4.34 illustrates how bootstrapping results in a distribution of estimates of accumulated fatigue damage with $\gamma=10^{-6}$ reliability. The example also illustrates the relevance of bootstrapping and that this process indeed mitigates a significant source of uncertainty.

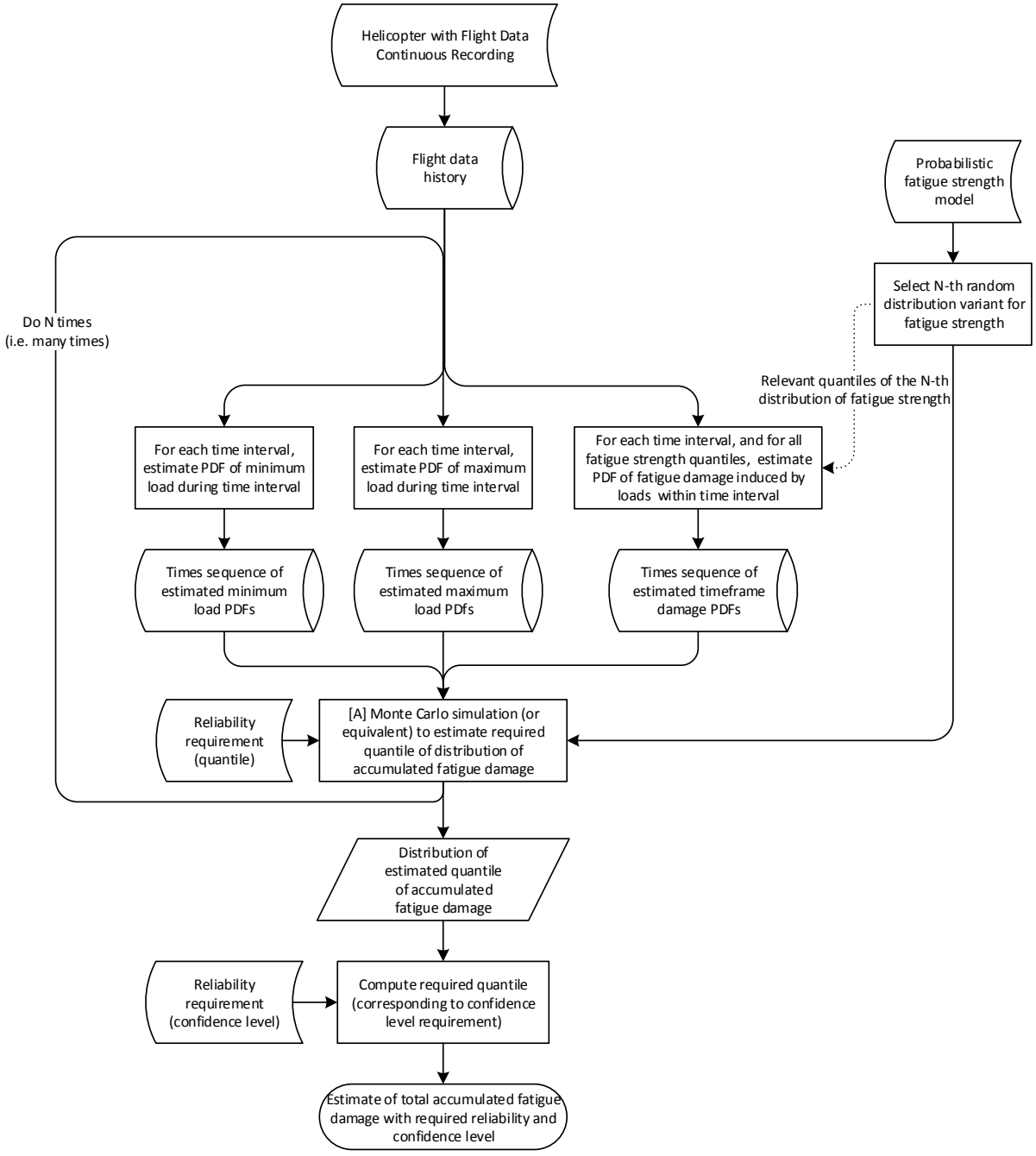


Figure 4.33: Process chart summarising how Probabilistic Load & Damage Modelling uses bootstrapping of a fatigue strength distribution to perform confidence level analysis for its estimations of a reliability quantile of accumulated fatigue damage. Process element [A] is detailed in Figure 4.27 and sections 4.3.2.2 and 4.3.2.3.

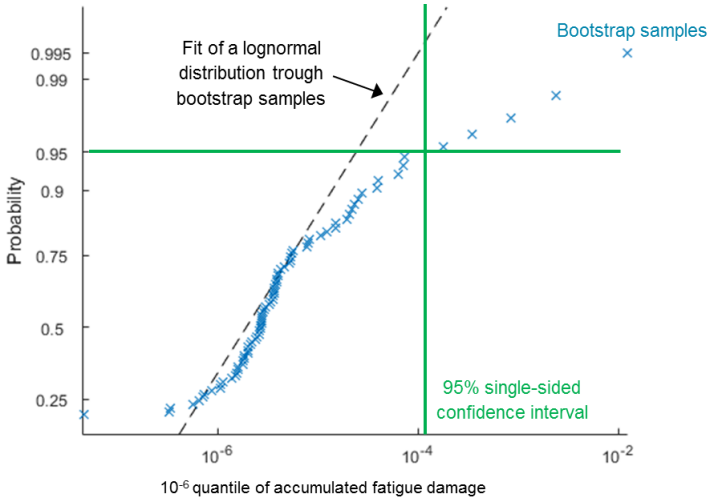


Figure 4.34: Graph showing how the result of bootstrap simulation implemented by PLDM is used to estimate a single-sided 95% upper confidence level for the γ=1-10^{-6} reliability quantile of accumulated fatigue damage.

PLDM's prediction models for timeframe damage and extreme loads are generated by non-linear statistical data models. Although these models are fully deterministic once they have been generated, the state to which they converge during their training process is subject to random variation. In particular, the training process of the artificial neural networks used for the prediction of extreme loads and timeframe damage starts with a random initialization of its parameters.[27] However, this practise also makes that the final state that the network converges to is subject to slight random variation.

Another source of uncertainty comes from the limited amount of data that can be used to generate the regression models for timeframe damage and extreme load. Just as with distribution fitting, or any other function fitting procedure, the use of a slightly different set of data will produce a different variant of the regression model. This variation becomes smaller as more data can be used to generate the regression model. In the case of PLDM, the available amount of flight data from load classification flights is relatively small in comparison to the size of the feature space and the complexity of the required regression model. To account for this uncertainty, a database with regression models is generated by non-parametric bootstrapping. I.e., each model in this database is generated according to a random selection of the LCF database and, if applicable, has been trained starting from a random initial state. Although other methods to estimate confidence intervals for artificial neural networks exist [114, 115, 116] and may be more accurate, non-parametric bootstrapping is simple to implement and assumed to capture uncertainties with sufficient accuracy.

In the example application of PLDM presented in this chapter, the accuracy of the recorded flight data is considered as a source of uncertainty as well as it could be considered that sensor data is recorded with limited precision and that this causes uncertainty. However, although it is acknowledged that small recording errors occur, it is also assumed that these change randomly throughout time and that this uncertainty is already represented by modelled prediction uncertainty by the regression models themselves and that this uncertainty is thus already accounted for by the quantile estimation method introduced in section 4.3.2.2. Furthermore, it is assumed that all sensors are calibrated and that there exist no measurement biases in flight data from LCF or commercial operations.

Due to the inclusion of the position of the rotor controls as predictive features for the regression models, as listed in Table 4-1 (DALPHA - DTHETA), the rigging of these controls introduces another source of uncertainty.

[27] This done in order to break certain symmetry conditions and to promote convergence to a global optimum

The actual rotor rigging for helicopters 1-3 is not known for the period of recorded operations and this is thus considered as a uniformly distributed sensor bias bounded by the allowable rigging configurations, as detailed in Appendix I. In contrast to load classification flights, the position of the rotor controls has not been recorded for helicopters 1-3. For these helicopters, the position of the controls must be estimated by means of the recorded position of the pilot controls, the state of the intermediate autopilot and stability augmentation system and a proprietary conversion model. Randomly choosing the rigging this conversion model assumes results in a random variation of the recorded flight data from helicopter 1-3. Each bootstrap thus makes use of a random variation of the recorded flight data by assuming a randomly drawn rigging configuration.

The complete bootstrapping process to estimate confidence levels for PLDM is summarized in Figure 4.35.

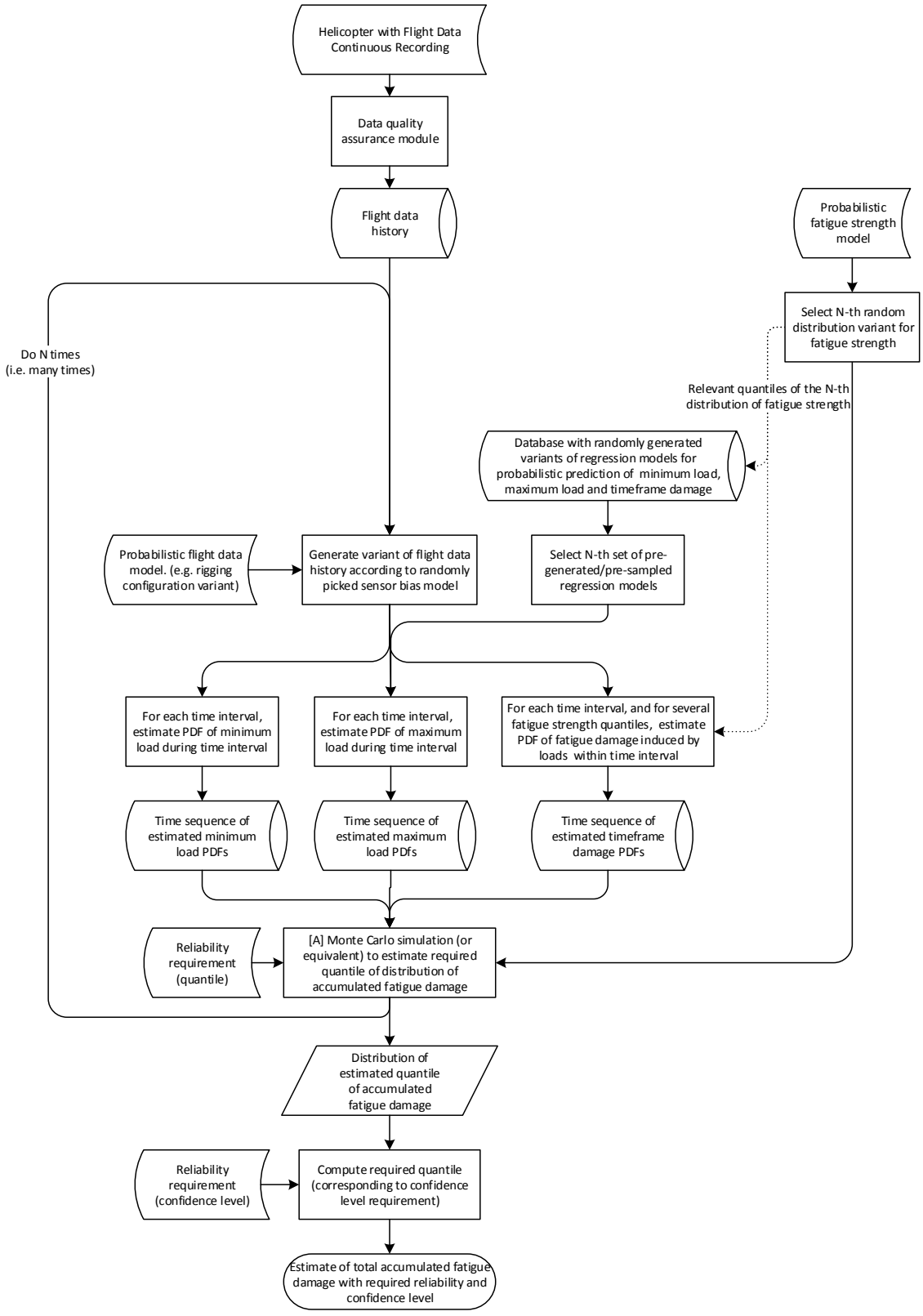


Figure 4.35: Process chart summarising the complete implementation of Probabilistic Load & Damage Modelling that present work uses to estimate confidence levels of predicted quantiles of accumulated fatigue damage. The chart specifically defines how bootstrapping of an estimated distribution of fatigue strength and bootstrapping of the model generation process to create regression models for the probabilistic estimation of timeframe damage and minimum and maximum load is carried out to estimate confidence levels. Process element [A] was elaborated in Figure 4.27 and sections 4.3.2.2 and 4.3.2.3.

4.3.2.5 Generation of probabilistic prediction models

Ideally, prediction intervals for timeframe damage and extreme loads are directly dependent on the timeframe's coordinate in feature space, i.e. prediction intervals are different for different manoeuvres, even though the most likely load to occur may be the same. Although the load that occurs during a timeframe in a high-speed turn may be the same as one during an aggressive push-over, the uncertainty of corresponding load predictions may be significantly different. Given the successful application of artificial neural networks for DLDM, it is preferred to also use ANNs for PLDM. Several methods to estimate prediction intervals exist for ANN regression models [117, 118, 119, 120, 121, 122, 123].

A relatively straightforward method is to have the ANN regression model predict distribution parameters instead of making a scalar point prediction of the regression target itself. For example, instead of directly predicting the load, a distribution mean and variance can be predicted. The predicted distribution parameters then describe the predicted probability distribution of the load. This approach could however not be implemented successfully and was abandoned.[28]

Most other regression error models, including relevance vector machines, assume homoscedastic prediction errors and are often also limited to a Gaussian prediction error model. Regression tests presented later in this section and in section 4.3.3 and Appendix K clearly demonstrate that prediction errors for PLDM's prediction of timeframe damage and extreme loads are neither Gaussian nor homoscedastic. The application of these two modelling assumptions may thus lead to significant modelling errors. Modification of some of the aforementioned probabilistic modelling methods, e.g. relevance vector machines, to feature non-Gaussian and homoscedastic prediction error models would require considerable mathematical development and is beyond the scope of present work.

Instead, a simplified model for prediction uncertainty is generated which fits a probability density function through observed errors from test data. The model for prediction uncertainty makes the following assumptions:

- Prediction uncertainty is only a function of predicted load; i.e. if the same load is predicted for timeframes from different manoeuvres, the prediction uncertainty is considered equal
- The predictive error distributions or prediction uncertainties, derived from load classification flights are assumed to be representative, i.e. the distribution of prediction errors observed from load classification flights is representative for prediction errors on other helicopters even though these are likely to have a different manoeuvre distribution and profile
- Uncertainty from the fitting of the error distributions is assumed to be fully covered by the bootstrapping process of the probabilistic regression models, i.e. by the process outlined in section 4.3.2.4

The practical implementation of the model divides the prediction domain into intervals such that each interval contains a sufficient number of test points to fit an error distribution. The distribution fitting processes uses a composite performance function, which is a function of the Bayesian and Akaike Information Criterion [106, 124], to select the best fit through the error distribution while deselecting models with too many distribution parameters. Considered distribution models are a Gaussian, the Generalized Extreme Value Distribution, a Gaussian kernel density function and a custom distribution model consisting of a Gaussian with its tails replaced by Pareto distributions. It is assumed that these distribution models cover all real error distributions and are thus perfect. The model requires the load classification data to be distributed into three parts: one for maximum likelihood regression, one for the fitting of the prediction error distributions, and one for semi-

[28] An implementation using a custom combination of artificial neural networks, trained using a Particle Swarm Optimization algorithm and while assuming that prediction errors are distributed according to a Generalized Extreme Value distribution, led to uncertainty distributions with too conservative variance, i.e. width or imprecision. This approach was abandoned.

independent testing. Where necessary, databases for model generation are further split into training and validation sets in order to prevent overfitting. Further details about the prediction error model are given in Appendix I.

A high-level process summary for the generation of the prediction models for timeframe damage and extreme loads is given in Figure 4.36.

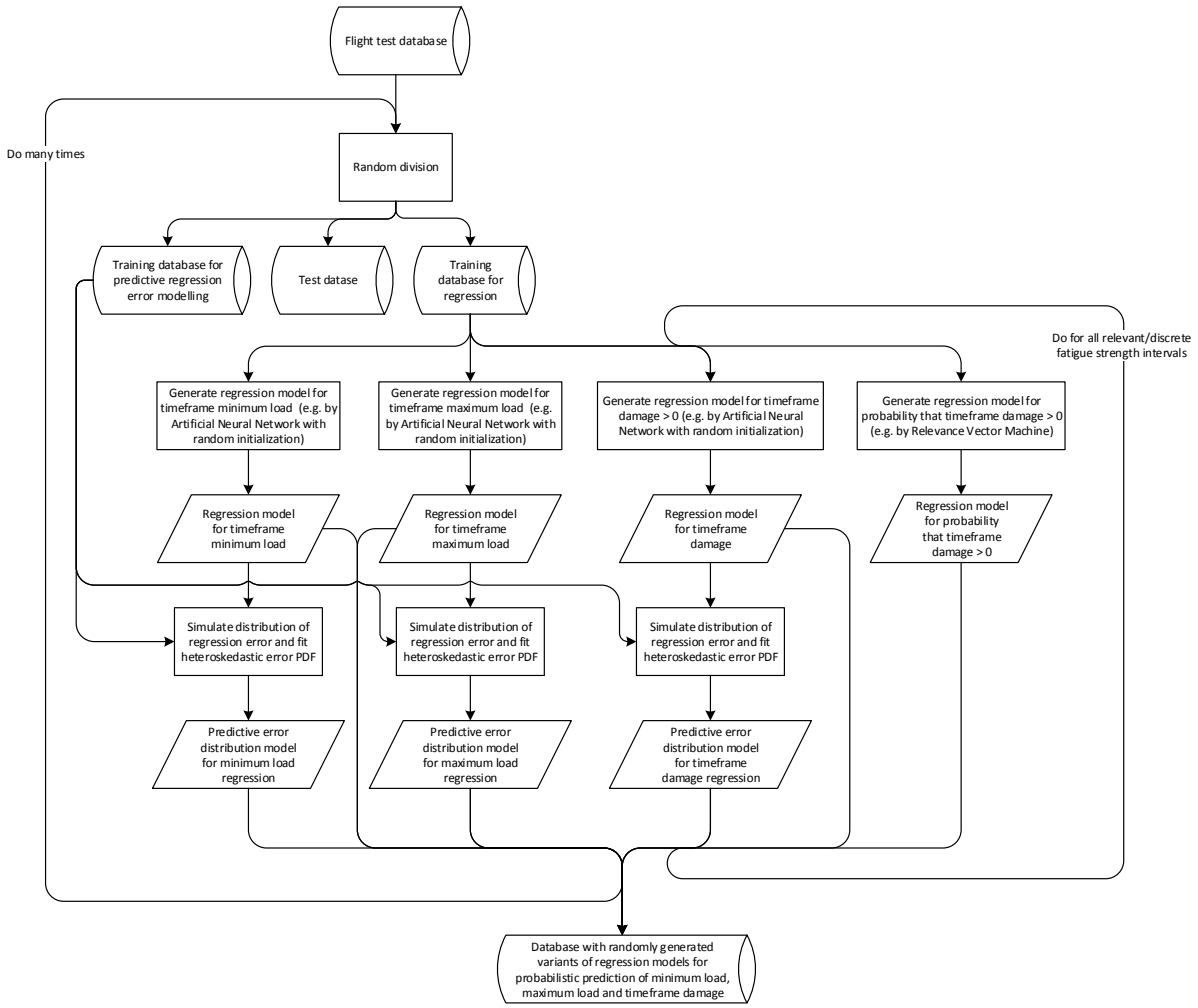


Figure 4.36: Process chart summarising how a database is created that contains randomly generated variants of regression models for probabilistic prediction of minimum load, maximum load and timeframe damage. The entire process elaborates process element [D] in Figure 4.27.

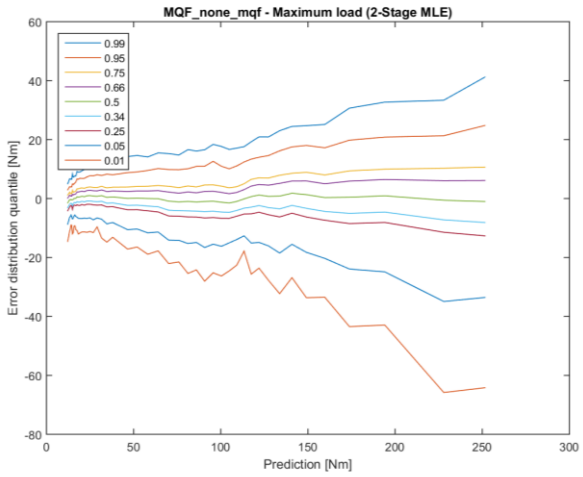


Figure 4.37: Graph showing how selected quantiles of a predicted distribution of regression errors for the timeframe maximum torque in the Fenestron driveshaft varies with the MLE point prediction.

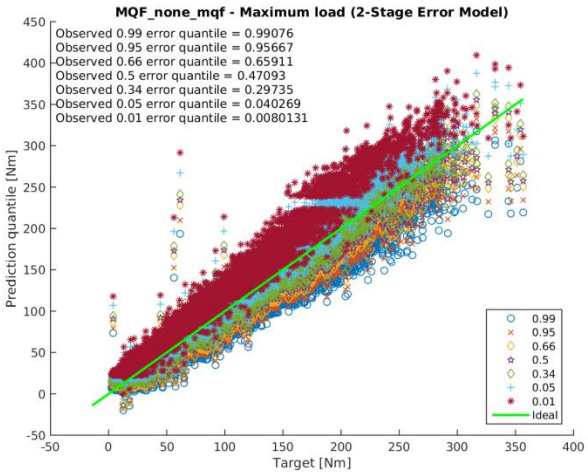


Figure 4.38: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe maximum load predictions for the Fenestron driveshaft. Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

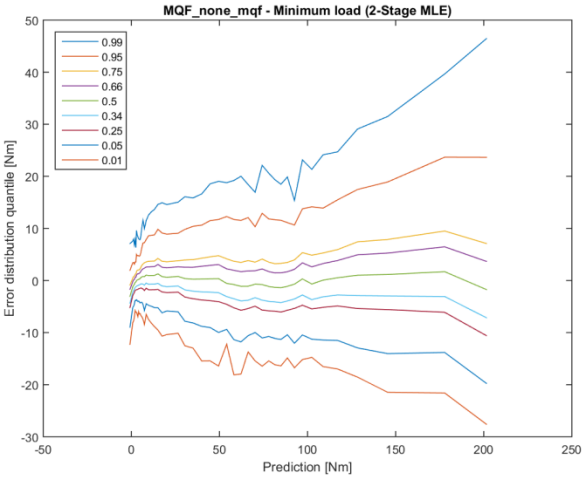


Figure 4.39: Graph showing how selected quantiles of a predicted distribution of regression errors for the timeframe minimum torque in the Fenestron driveshaft varies with the MLE point prediction.

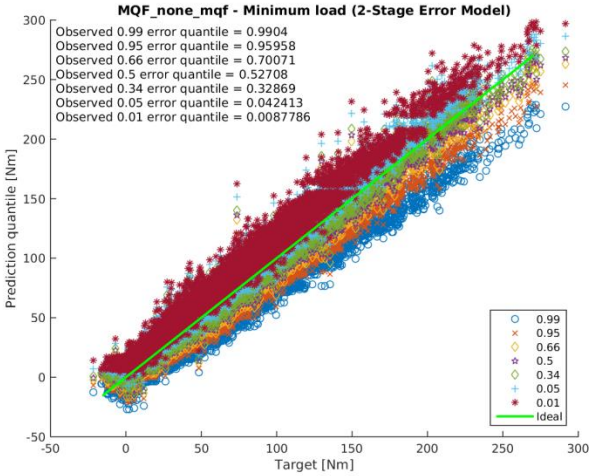


Figure 4.40: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe minimum load predictions for the Fenestron driveshaft. Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

Examples of fitted prediction error distributions for timeframe maximum and minimum loads in Figure 4.37 and Figure 4.39 illustrate results for component number six, or MQF. It can be seen that prediction errors are heteroscedastic and increase with load magnitude. The variance, or effective width, of the distributions with prediction errors is reasonable. Adding error quantiles to maximum likelihood point-estimates in Figure 4.38 and Figure 4.40 demonstrates that, at least on load classification flight test data, moderate error quantiles do not result in unacceptable load spikes and that the error distributions are of similar magnitude as tested prediction errors. This is confirmed by comparing predicted error quantiles with observed quantiles from LCF test data, tabulated in the upper left corner of Figure 4.40.

Initial binary classification for timeframe damage can be carried out using the same Relevance Vector Machine (RVM) as used before for DLDM since an RVM already makes a probabilistic estimate about the occurrence of timeframe damage. Where necessary, subsequent probabilistic estimation of the non-zero value of timeframe damage is performed in the same way as for timeframe maximum and minimum loads. Additional test results, also for other components, are included in Appendix K.

4.3.3 Testing regression accuracy and the validity of associated modelling assumptions

PLDM makes many modelling assumptions and simplifications, as discussed earlier in sections 4.2, 4.3.1, and 4.3.2. The validity of some of these is tested using independent test data from helicopters 1 and 2 and quasi-independent test data from load classification flights. For brevity, results for helicopter 2 are omitted in the presentation, as they are equivalent to the results for helicopter 1.

4.3.3.1 Testing the accuracy of predicted regression error distributions

Testing the predicted error distribution using quasi-independent load classification data shown in Figure 4.38 and Figure 4.40 demonstrates good regression performance of the ANNs predicting timeframe extreme loads. More detailed repetition of the test using independent data from helicopter 1 does, however, reveal significant deviation between the predicted and observed regression error distribution for timeframe maximum loads in Figure 4.41 to Figure 4.43. For low loads, there is a small estimation bias and the weight of distribution tails is overestimated, as illustrated in Figure 4.41. For medium loads, there is also a small bias but more importantly also a significant underestimation of the tails, as illustrated in Figure 4.42. And the predicted error distribution for high loads is clearly overoptimistic and severely overestimates the precision of predictions, as illustrated in Figure 4.43. The observed irregularities of the predicted uncertainty distributions are the result of an imperfect implementation of a custom distribution model using a Gaussian with Pareto tails and may be removed by some technical improvements in distribution fitting and definition.

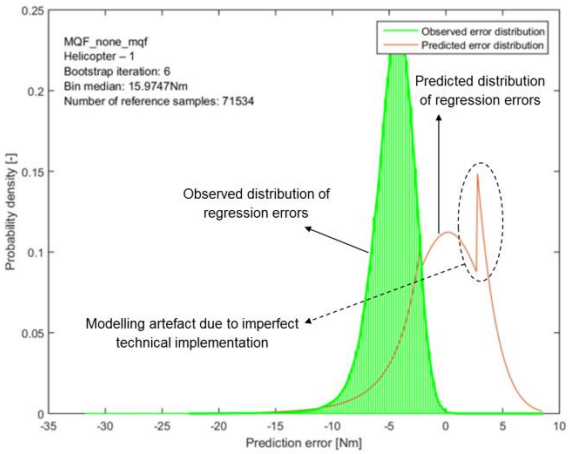


Figure 4.41: Graph showing a comparison between the distributions of predicted and actually measured prediction errors for relatively low values of timeframe maximum Fenestron torque loading on helicopter 1. A positive error denotes that torque is over-estimated.

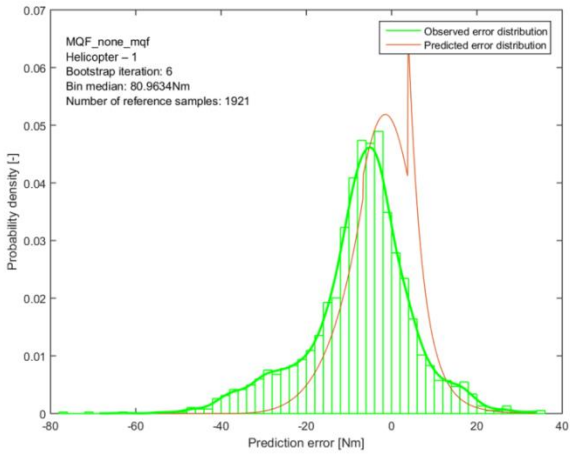


Figure 4.42 Graph showing a comparison between the distributions of predicted and actually measured prediction errors for medium values of timeframe maximum Fenestron torque loading on helicopter 1. A positive error denotes that torque is over-estimated.

An even more detailed comparison between predicted and observed error distribution quantiles in Figure 4.44 to Figure 4.46 again demonstrates that:

- significant differences between predicted and observed quantiles of prediction uncertainty distributions occur throughout the prediction range
- the bootstrapped distribution of predicted error quantiles is non-conservatively biased
- few or none of the bootstrapped cases include true errors

Especially Figure 4.44 and Figure 4.45 demonstrate that important error distribution quantiles are consistently underestimated and that bootstrapping cannot be used to account for these consistent prediction biases.

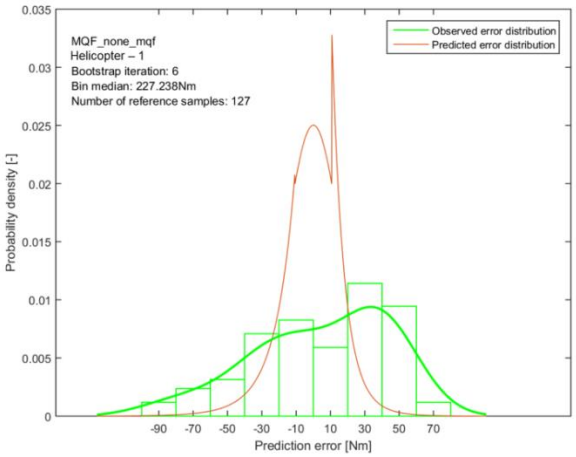


Figure 4.43: Graph showing a comparison between the distributions of predicted and actually measured prediction errors for relatively high values of timeframe maximum Fenestron torque loading on helicopter 1. A positive error denotes that torque is over-estimated.

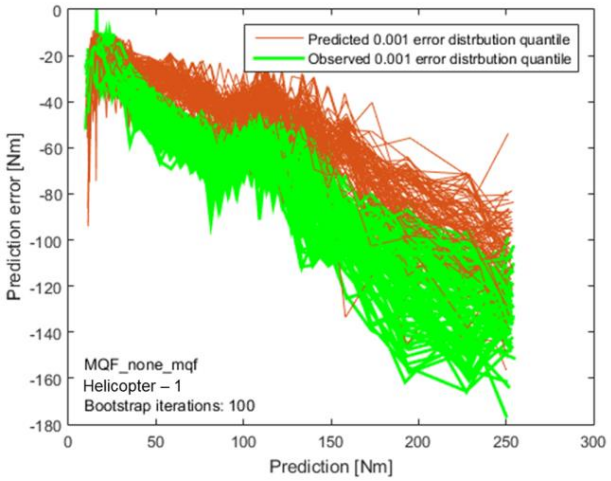


Figure 4.44: Graph showing a bootstrapped comparison between the predicted and actually measured $\gamma=10^{-3}$ error distribution quantiles for the timeframe maximum of the Fenestron torque loading on helicopter 1. The bootstrap distributions result from PLDM prediction bootstrapping as well bootstrapping of the dataset from helicopter 1. A positive error denotes that torque is over-estimated.

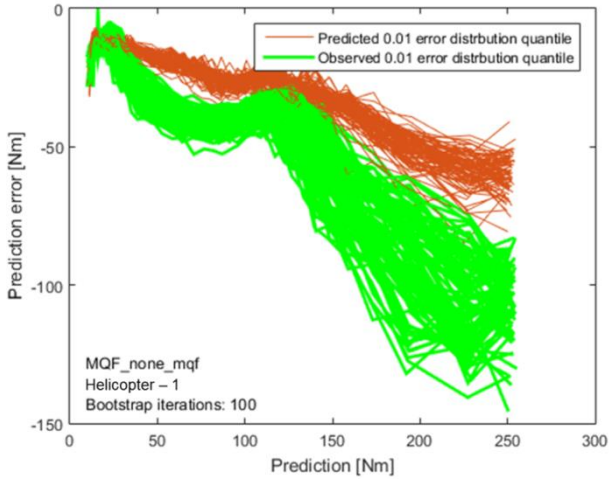


Figure 4.45: Graph showing a bootstrapped comparison between the predicted and actually measured $\gamma=10^{-2}$ error distribution quantiles for the timeframe maximum of the Fenestron torque loading on helicopter 1. The bootstrap distributions result from PLDM prediction bootstrapping as well bootstrapping of the dataset from helicopter 1. A positive error denotes that torque is over-estimated.

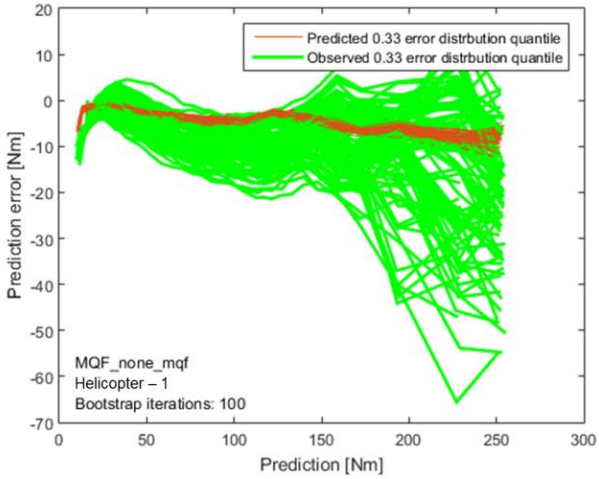


Figure 4.46: Graph showing a bootstrapped comparison between the predicted and actually measured 1/3 error distribution quantiles for the timeframe maximum of the Fenestron torque loading on helicopter 1. The bootstrap distributions result from PLDM prediction bootstrapping as well bootstrapping of the dataset from helicopter 1. A positive error denotes that torque is over-estimated.

4.3.3.2 Testing the accuracy of selected regression modelling assumptions

Comparison of true timeframe maximum loads with maximum likelihood predictions from 100 bootstrapped ANN prediction models clearly illustrates that bootstrapping is a mostly ineffective method to model and mitigate prediction biases, as displayed in Figure 4.47. If significant prediction errors are observed, then these

biases remain present in most bootstrap repetitions. Only for low loads does bootstrapping appear to reflect true prediction uncertainties without considerable biases.

Since there are fewer learning examples for the prediction of higher maximum loads, due to their rare presence in the LCF example database, it could be expected that bootstrap variance would increase with increasing loads. However, the systematic analysis in Figure 4.48 demonstrates that this is not the case and that the normalized bootstrap variance remains approximately constant for medium to high loads. The sharp rise in the bootstrap variance of ANN maximum likelihood predictions, i.e. DLDM point predictions, for low loads is not well understood but may be caused by inherent numerical instability of the performance indicator - the coefficient of variation – around zero.

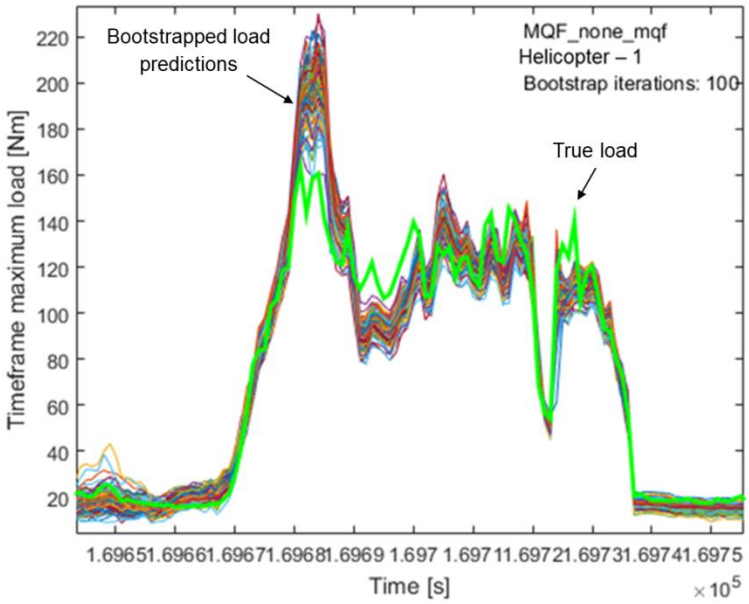


Figure 4.47: Graph showing a detailed comparison between bootstrapped predictions and actually measured values of the timeframe maximum torque load during a flight of helicopter 1. The illustrated variation of timeframe maximum loads is due to bootstrapping of ANN prediction models and associated training database. (The horizontal axis displays maintenance time in seconds)

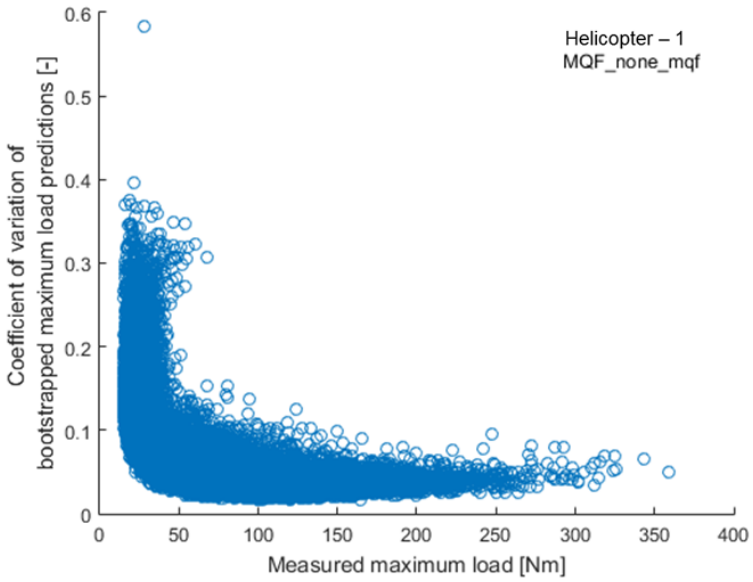


Figure 4.48: Scatterplot showing how the coefficient of variation of bootstrapped MLE point predictions varies with the actually measured timeframe maximum torque for the Fenestron in helicopter 1. The bootstrap variation is the result of bootstrapping of ANN prediction models and associated training database.

A major modelling simplification that PLDM makes in comparison to the simulation-based random load model introduced in chapter 2 is independence between prediction errors for timeframe damage, minimum load and maximum load. This is reasonable since regression takes place independently and should result in maximum-likelihood predictions with random prediction errors. Error correlation of timeframe maximum and minimum load predictions for quasi-independent load classification flight data in Figure 4.49 validates this assumption. However, independent test results from helicopter 1 in Figure 4.50 and Figure 4.51 demonstrate that the assumption does not hold in practise and that the prediction errors for timeframe minimum and maximum loads can exhibit 30%-70% linear correlation.

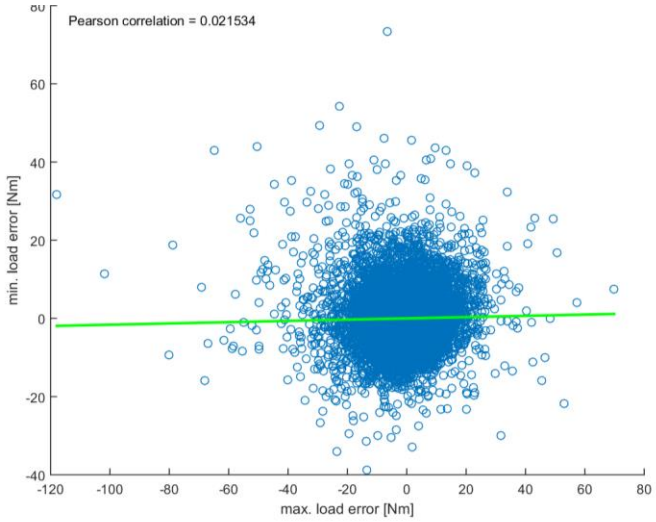


Figure 4.49: Regression plot showing how prediction errors for the timeframe minimum and maximum torque on the Fenestron are correlated for the quasi-independent LCF data.

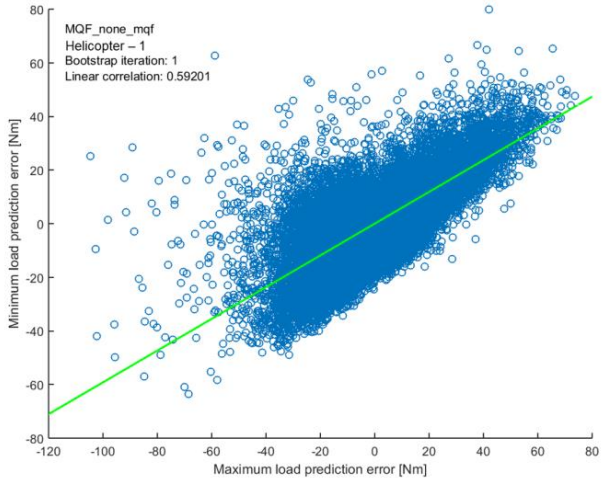


Figure 4.50: Regression plot showing how prediction errors for the timeframe minimum and maximum torque on the Fenestron are correlated for independent data recorded on helicopter 1.

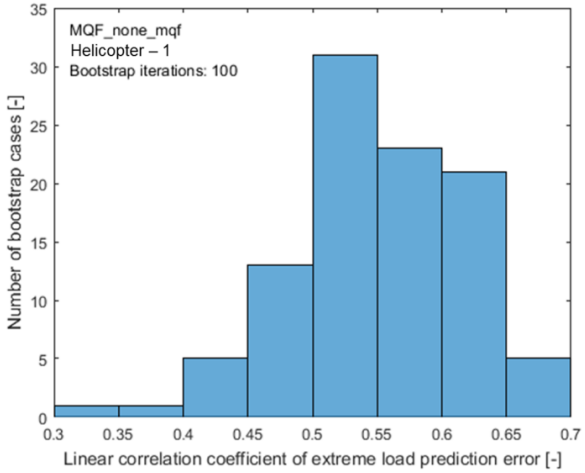


Figure 4.51: Chart showing the bootstrap distribution of the correlation of prediction errors for the timeframe minimum and maximum torque on the Fenestron of helicopter 1. The bootstrap variation is the result of bootstrapping of ANN prediction models and associated training database.

Another assumption carried over from the simulation-based probabilistic load model in chapter 2 is independence between prediction errors from subsequent timeframes. This assumption allows considerable modelling simplifications but does not reflect reality well, as demonstrated by test results in Figure 4.52 to Figure 4.54 where correlation coefficients in excess of 0.75 are observed. However, more detailed analysis in Figure 4.55 and Figure 4.56 reveals a significant drop in correlations for important high load events. These are more similar to random load spikes and less influenced by persistent estimation biases, but even for these events temporal correlation generally does not drop below 0.5.

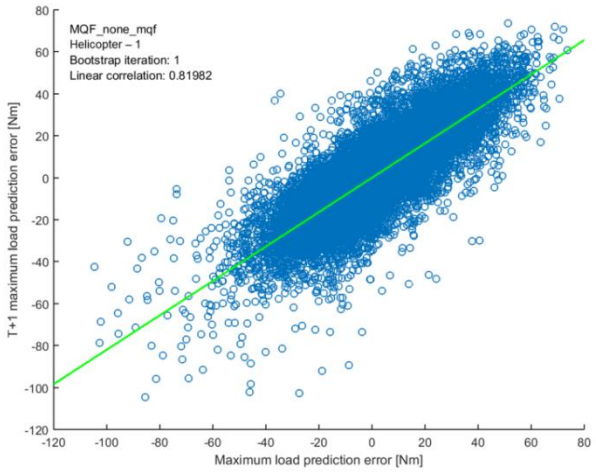


Figure 4.52: Regression plot showing how prediction errors for the maximum torque are correlated between subsequent timeframes recorded on helicopter 1.

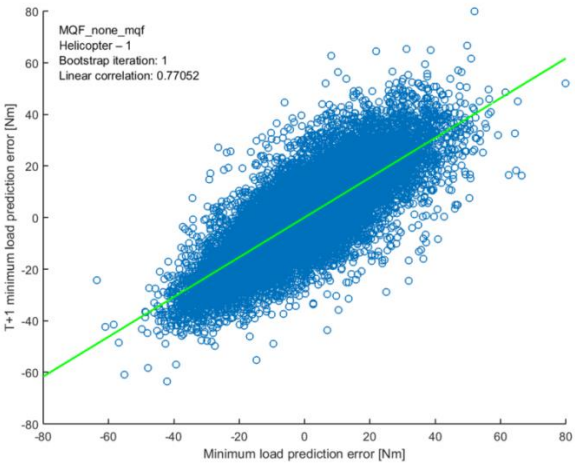


Figure 4.53: Regression plot showing how prediction errors for the minimum torque are correlated between subsequent timeframes recorded on helicopter 1.

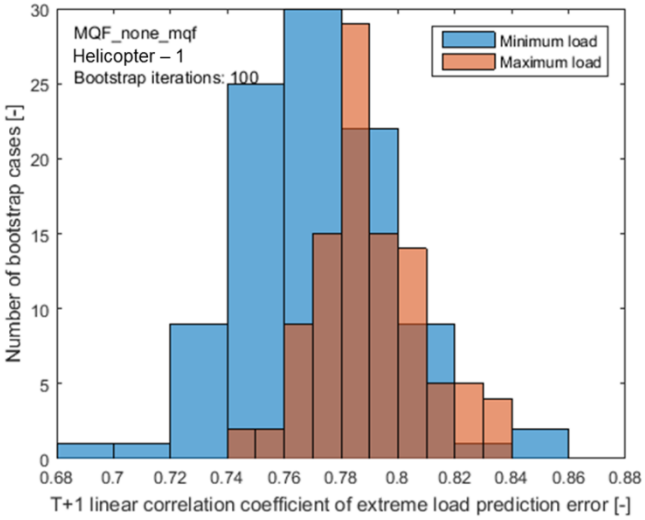


Figure 4.54: Chart showing the bootstrap distribution of the correlation of prediction errors for the extreme loads of subsequent timeframes. The bootstrap variation is the result of bootstrapping of ANN prediction models and associated training database. Where the distributions for maximum (red) and minimum (blue) load overlap, the bars may appear as grey.

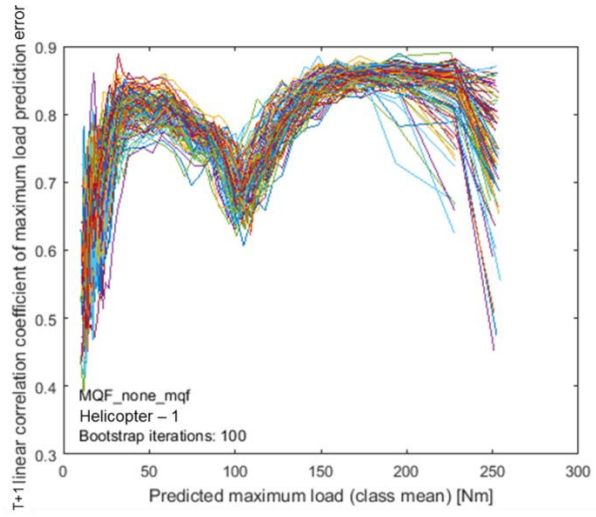


Figure 4.55: Chart showing how the correlation between subsequent timeframes for the prediction error for the maximum torque on the Fenestron of helicopter 1 varies with the MLE prediction of the timeframe maximum load. Bootstrapped estimations of magnitude dependence of correlation between maximum load prediction errors of subsequent timeframes.

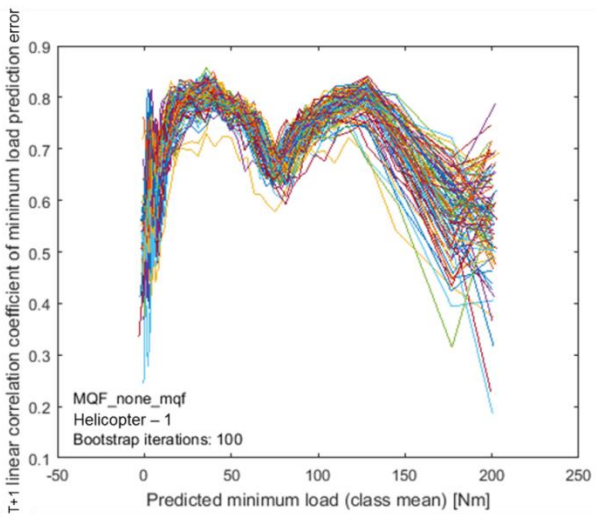


Figure 4.56: Chart showing how the correlation between subsequent timeframes for the prediction error for the minimum torque on the Fenestron of helicopter 1 varies with the MLE prediction of the timeframe minimum load. Bootstrapped estimations of magnitude dependence of correlation between minimum load prediction errors of subsequent timeframes.

4.3.4 Reliability testing of estimates of accumulated fatigue damage made by Probabilistic Load & Damage Modelling

The actual reliability of PLDM estimates of accumulated fatigue damage is tested using the same test procedure as introduced earlier for DLDM in section 4.2.6 and summarized for PLDM in Figure 4.57. For helicopters 1 and 2, accumulated fatigue damage is predicted using PLDM with a target reliability of $\gamma=10^{-6}$ (95%). The prediction is then compared with a distribution of accumulated fatigue damage according to the recorded load spectrum and randomly sampled 10^{-6} quantiles of fatigue strength.

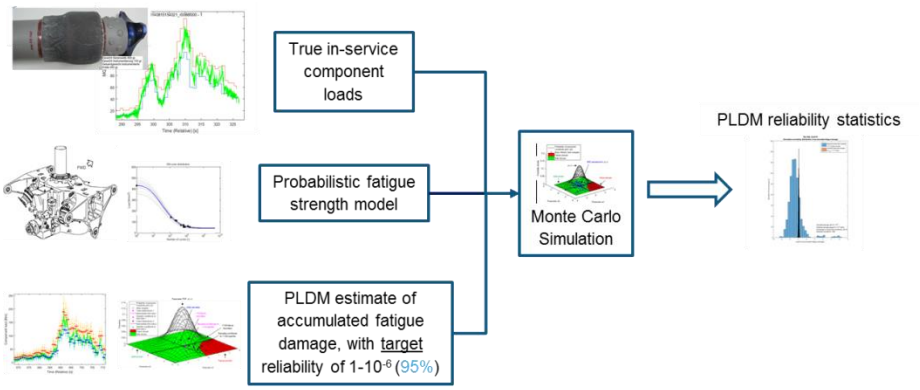


Figure 4.57: Schematic introducing the reliability testing procedure for PLDM estimates of accumulated fatigue damage.

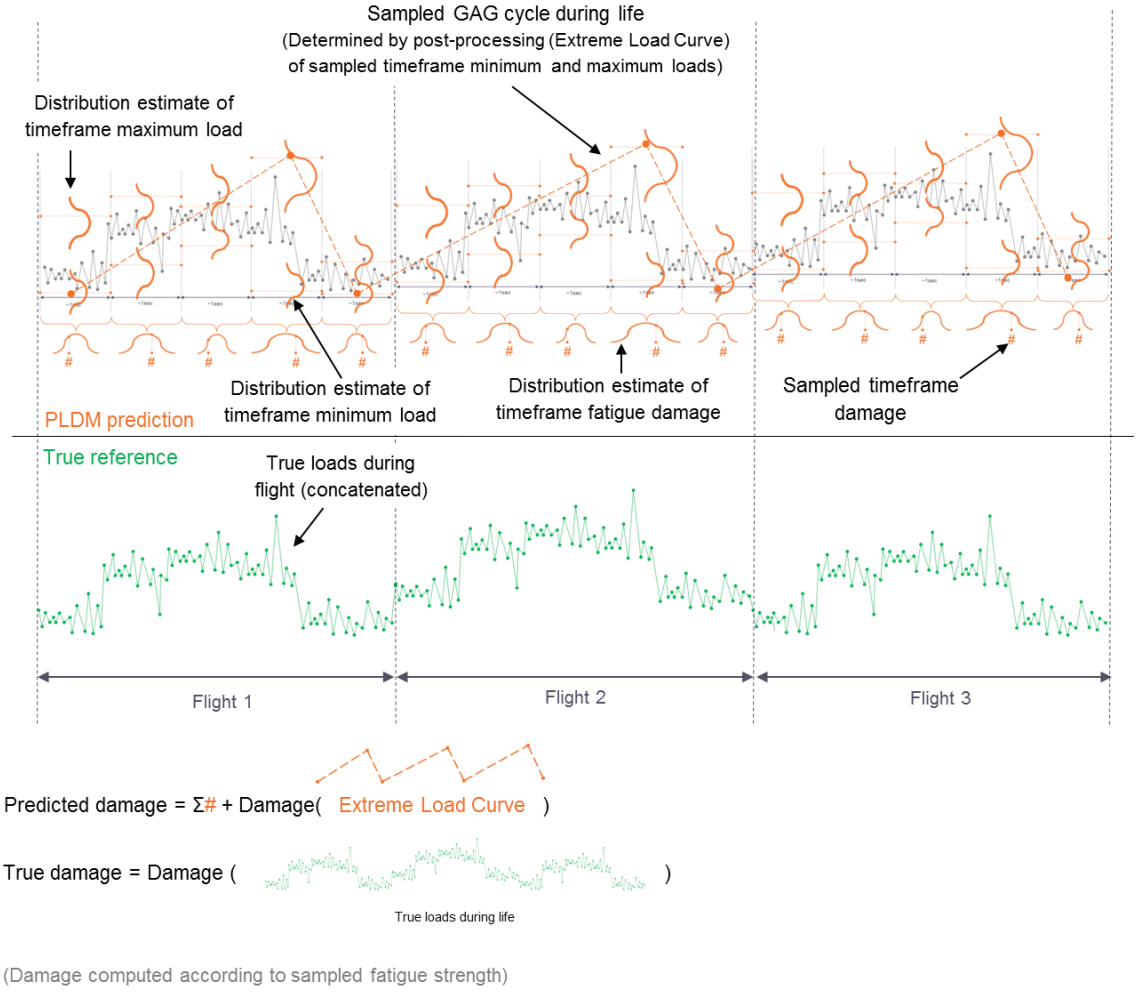


Figure 4.58: Schematic explaining the difference between the load spectrum accumulation model employed by PLDM and the 'true' reference load spectrum created from recorded loads from helicopters one and two that is used as a `true` reference during reliability testing.

In contrast to the reliability test for DLDM, PLDM is tested using a load spectrum model without the use of a super-GAG cycle. Although this leads to less conservative fatigue lives, it does add computational costs as the entire sequence of flights need to be considered at once, as shown in Figure 4.58.

The results shown in Figure 4.59 and Table 4-5 of the reliability testing procedure for PLDM, as defined in Figure 4.60, demonstrate that PLDM meets its reliability target for all tested prediction cases. In contrast to DLDM, PLDM predictions meet their reliability target even if uncertainty about fatigue strength is artificially reduced in order to increase the significance of prediction errors on timeframe damage and extreme loads. This result demonstrates that PLDM's reliability substantiation model can successfully mitigate the effect of combined uncertainties from both fatigue strength as well as in-service load spectrum prediction. Also, the results demonstrate that the effects of the inaccurate PLDM modelling assumptions discussed in sections 4.3.3.1 and 4.3.3.2 can be neglected and do not adversely influence demonstrable reliability of predicted fatigue damage.

Table 4-5: Table showing the confidence level with which a $\gamma=10^{-6}$ reliability level can be demonstrated for estimates of the accumulated fatigue damage of the lower gearbox casing made by PLDM for helicopters one and two and how this demonstrable confidence is not significantly reduced with synthetically lowering the variance of fatigue strength (decreasing σ-factor). The PLDM predictions of accumulated fatigue damage are made with a target reliability of $\gamma=10^{-6}$ (95%).

σ-factor	H/C-1	H/C-2
1	96.7	96.3
0.75	95.6	94.9
0.5	93.6	93.3

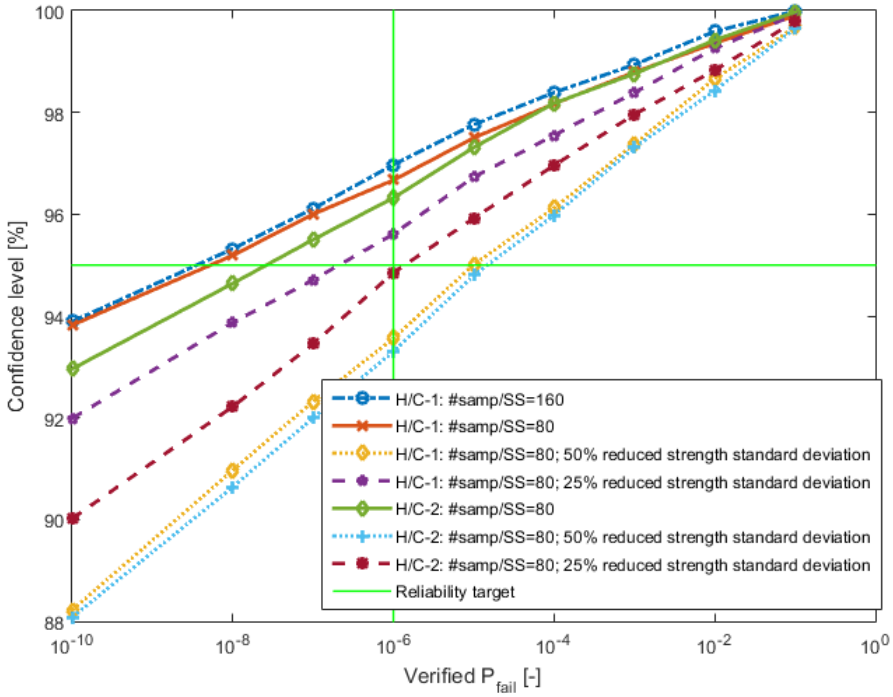


Figure 4.59: Graph showing the demonstrable reliability level of PLDM predictions that have a target reliability of $\gamma=10^{-6}$ (95%) and are made with a varying amount of bootstrap samples and for different synthetically generated cases for the variation of fatigue strength. The single test case using 160 bootstrap samples verifies the convergence and stability of the predictions and yields similar results to the other predictions made with a computationally `cheaper` configuration using 80 bootstrap samples. The graph also shows how the demonstrable reliability level can be varied as a function of the demonstrable reliability quantile and confidence level (i.e. not as function of PLDM reliability target, which is constant).

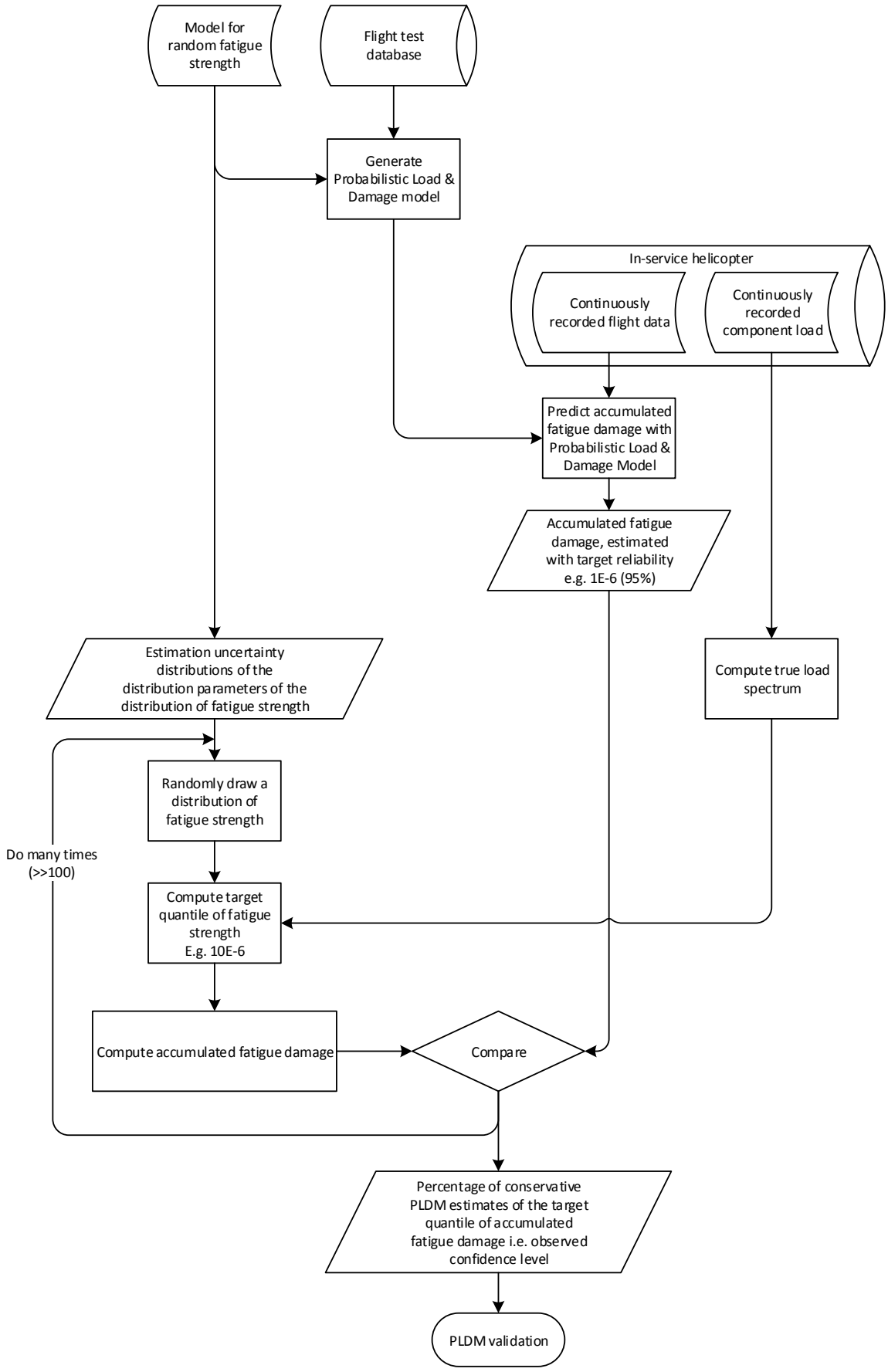


Figure 4.60: Process overview defining how reliability of PLDM estimates of accumulated fatigue damage is tested and benchmarked.

4.3.5 Benchmarking of in-service application of Probabilistic Load & Damage Modelling

After the successful verification of PLDM's reliability model, recorded flight data from helicopters 1-3 is analysed to estimate the magnitude of helicopter-individual SLL extensions that can be enabled by PLDM. The results are summarized in Table 4-6. The first four columns indicate how many flight hours of data have been analysed by VFLM, for which components, for which helicopters, and what the classic Service Life Limit[29] is for these components. PLDM results have been computed for different numbers of PLDM bootstrap samples N_b. The number of components, helicopters, and bootstrap samples for which results have been computed is limited due to the high computational costs involved, limited availability of computing resources, and time constraints in finishing present work. For future work, it recommended to compute results for more selected components and helicopters and to systematically test the effect of the number of PLDM bootstrap samples N_b on the convergence of PLDM estimates of accumulated fatigue damage.

For the lower gearbox housing, the results in Table 4-6 indicate that PLDM can substantiate SLL extensions in excess of five times current SLLs for flown mission profiles. Comparison with its currently published SLL shows that, given the observed usage trend, the SLL can be extended to beyond 20,000 flight hours. This is a regular economic lifetime of a multi-purpose helicopter. The projected SLL extension thus implies that by recording the entire flight history of this component, the lower gearbox casing can effectively lose its operational life limit and would not need to be replaced during the life of the helicopter.

By design, PLDM automatically compensates for cases where regression is inaccurate and automatically mitigate the primary sources of prediction uncertainty. The reliability of PLDM usage-based estimates of accumulated fatigue damage is verified for component 6 (MQF) and by derived synthetic test conditions in section 4.3.4. However, the scope of these synthetic test cases is limited and PLDM is not specifically validated for other components than the lower gearbox housing. Nevertheless, due to PLDM's generic and adaptive design and the successful passing of the more challenging synthetic test cases, it is expected that PLDM also yields accurate results for the other components listed in Table 4-6. Another indicator that PLDM has comparable accuracy comes from comparing the model generation results for the other components presented in Appendix K with the model generation results for the lower gearbox housing presented in sections 4.2.5.2 and 4.3.2.5. It follows from such a comparison that the expected prediction accuracy is similar for all components. It is thus expected that the verification results for component 6 (MQF) also apply to the other components listed in Table 4-6 and that the listed predictions of accumulated fatigue damage are accurate enough to estimate the economical potential of VFLM. However, due to practical constraints, the number of bootstrap samples that are used to compute results for the other components is limited to 25 samples. This means that the expected accuracy of the targeted 95% confidence is low and that it is expected that the addition of more bootstrap samples will increase the predicted amount of accumulated fatigue damage. Nevertheless, the results generally indicate that PLDM can be expected to enable fatigue life extensions of at least around a factor five for all tested components.

In more detail, the difference between estimates of accumulated fatigue damage by DLDM and PLDM in Table 4-6 is significant for component 6. Although the removal of the super-GAG load cycle should reduce PLDM's conservatism, it should only account for an improvement of about 30%, as indicated before in Table 4-4. Nevertheless, it is argued that this result does not challenge the validity of PLDM's reliability tests. But rather that the difference is caused by the large and highly non-linear effect that small changes in the substantiated confidence level of high reliability levels can have on the predicted value of accumulated fatigue damage. This is for example illustrated before by many of the simulation results in section 2.6.4 and in Figure 4.34 and is especially prominent for the lower gearbox housing since its fatigue strength is based on very few full-scale test results and thus features large confidence bounds. This explanation is further supported by analysis in Appendix M, where it is indicated that the substantiated confidence level of the fatigue damage prediction has

[29] The classic Service Life Limits have been computed with the DMP from the OEM but while making use of the new statistical fatigue strength model introduced in chapter 3.

a larger effect on predicted fatigue damage accumulation that the reliability quantile itself. The difference between DLDM and PLDM predictions of accumulated fatigue damage for the other components listed in Table 4-6 instead follows the more general expectation that PLDM predictions are more conservative than DLDM predictions. This expectation follows from the more conservative design of PLDM which accounts for prediction errors for timeframe load and fatigue damage, whereas DLDM does not and assumes perfect predictions.

Table 4-6: Table comparing accumulated fatigue damage computed by Virtual Fatigue Life Monitoring with planned damage accumulation according to the conservative Design Mission Profile. PLDM results have been computed using 25, 56 and 100 bootstrap repetitions.. PLDM and DLDM results are presented as a percentage of damage accumulated according to the conservative design mission profile. All predictions have a target reliability of $\gamma=10^{-6}$ (95%). Cells marked with "-" correspond to test cases whose computations could not be finished within the scope of present work. CAUTION: Service life limits presented in this table are computed for academic purposes only and are not approved by any OEM or airworthiness authority.

Component	**Component ID**	**SLL** [FH]	**Helicopter**	**FH**	**DLDM** [% DMP]	**PLDM** [% DMP]		
						$N_b = 25$	$N_b = 56$	$N_b = 100$
Casing of hydraulic actuator	1	11515	1	504.5	1	5	-	-
Control rod	2	2406			1	11	-	-
Gimbal	3	7231			0	17	-	-
Forked lever	4	2489			0	1	-	-
Lower gearbox casing	6	11116	1	504.5	10	0	-	2
		11116	2	804.9	13	0	-	1
		11116	3	559.6	51	0	17	-

4.4 Conclusion

Two new methods to estimate and substantiate usage-based extensions of permissible fatigue life for selected parts by means of recording and analysing helicopter individual flight data are introduced: Direct Load and Damage Modelling (DLDM) and Probabilistic Load and Damage Modelling (PLDM). Their performance and reliability substantiation methods are verified with the use of over one-thousand hours of recorded data from actually in-service HEMS helicopters. The recorded data includes continuous high-frequency strain gauge data signals to obtain a complete, highly accurate and independent reference to test the VFLM methods.

DLDM is introduced as a simplified Direct Load Prediction based method which assumes that prediction errors for in-flight loads can be neglected and that the reliability of resulting estimates of accumulated fatigue damage can be substantiated by a reliability model for fatigue strength only. DLDM thus numerically substantiates the reliability of its predictions by the reliability of the conservative S-N working curve only. Simulation-based reliability testing of DLDM demonstrates that there are indeed conditions under which its simplified reliability substantiation framework is valid and can be used to safely and significantly extend the service life of individual parts. However, artificial broadening of the scope of reliability testing by artificially increasing the relative influence of prediction errors of in-service loads demonstrates that DLDM`s reliability model is not generally valid and that there is a need to explicitly model and mitigate the influence from uncertain and error-prone in-service load predictions.

PLDM is introduced to meet this requirement. PLDM uniquely makes probability distribution estimates of the parameters which determine accumulated fatigue damage. PLDM then simultaneously simulates and mitigates the influence of prediction uncertainties by means of numerical reliability modelling. PLDM's modelling framework is designed to be generic. The framework can be applied to any component with any loading profile, and regardless of achievable load prediction accuracy. The consistent validity and accuracy of PLDM's

reliability substantiation model is demonstrated for the lower gearbox housing of two commercially operated helicopters, and for a reliability requirement of 0.999999 with 95% single-sided confidence.

The performed reliability tests are passed successfully using a real industrial case with real data. The scope of testing is however expanded by the introduction of artificially generated test cases. The value of the generation of additional synthetic test cases is highlighted by their identification of the limitations of DLDM. It is therefore recommended to further expand the scope of simulated reliability test cases in present work in order to explore the boundaries of applicability of the introduced VFLM models more accurately and thoroughly.

Although the accuracy and validity of the PLDM model is demonstrated, the analysis introduced in present work is not sufficient to enable the in-service deployment of PLDM. Further research is required to enable the practical implementation of PLDM. The reliability and safety effects of data processing errors or limited availibity of sensors, recorder and computers needs to be further analysed and mitigated to comply with airworthiness regulations. Encouragingly however, analysis in Appendix M demonstrates that the reliability requirements that must be imposed on VFLM implementations can be limited. Using numerical reliability simulation it is demonstrated that even if VFLM fails entirely to recognize any fatigue damage, then the effects on the demonstrable reliability of predicted fatigue life are still limited for most of the components analysed in present work. Futher work is recommended to further develop this argumentation and to test its acceptance by airworthiness authorities.

Comparison between test results from quasi-independent Load Classification Flights and from actual in-service load measurements demonstrate significant differences between obtained error statistics. This underlines the need for independent verification data for statistics-based VFLM methods. However, the results also identify that despite the presence of significant modelling errors, the reliability of VFLM predictions can still be substantiated and verified. In particular, the verification results for DLDM in sections 4.2.5 and 4.2.6, and for PLDM in sections 4.3.3 and 4.3.4, demonstrate that required modelling accuracy for DLDM and PLFM is limited and that the presence of significant modelling inaccuracies can be allowed without affecting demonstrable prediction reliability. PLDM is a computationally expensive and simulation-based method to implement VFLM and to substantiate reliability. It is therefore recommend for future work to investigate the introduction design simplifications for PLDM that can significantly reduce its computational costs but which do not significantly reduce demonstrable prediction reliability. For example, it is demonstrated that bootstrapping of the prediction models does not accurately capture modelling uncertainties in section 4.3.3.2. But in section 4.3.4 it is demonstrated that despite considerable errors in predicted error distributions; overall prediction reliability for accumulated fatigue damage can still be substantiated. Bootstrapping of the prediction errors directly affects the costs to generate prediction models, and more importantly, the costs to process flight data. Therefore, it is recommended to benchmark up to what extent these bootstrap repetitions can be omitted without affecting overall substantiated prediction reliability.

In general, the reliability test results of DLDM demonstrate that under certain circumstances the influence of prediction errors of in-flight loads on the reliability of predicted fatigue damage accumulation can be neglected. Otherwise, the reliability test results demonstrate the effects of prediction errors can be mitigated by statistical simulation and by approximate and simplified prediction error models, for example as implemented by PLDM. It is therefore expected that similar conclusions can be made for alternative VFLM methods, e.g. FRR. It is expected that PLDM's prediction and reliability substantiation model can be modified to apply to FRR as well. It is expected that the associated effort is relatively low due to PLDM's generic design, and since FRR can be seen as a discretized version of DLDM, and since the core fatigue damage model of PLDM and DLDM is the same. For future work, the development of new and further simplified methods for VFLM is therefore recommended, in combination with tests to benchmark achievable SLL extensions with a common level of substantiated reliability.

5 Conclusions and recommendations

Based on the work presented in chapters 2-4, the conclusions to the research questions introduced in section 1.7 are discussed.

1. Which additional uncertainties about predicted fatigue life are introduced if it is not assumed that the variance of a component's fatigue strength can be estimated without uncertainty?

The fatigue strength of a component is a random variable and can thus never be known exactly. It is common to estimate the uncertainty distribution of fatigue strength based on fatigue tests. Following AGARD-AG-292 [16], present work models this distribution as a lognormal distribution defined by a mean and a variance. In aerospace, these two distribution parameters usually need to be estimated based on few sampled test values for fatigue strength only. Under such small sample-size conditions, the estimated values for the distribution mean and especially variance can be inaccurate. Many aerospace applications, including AGARD-AG-292, model and mitigate estimation uncertainty for the mean of a fatigue strength distribution but not for its variance.

Present work includes simulations that quantify the attainable reliability of computed quantiles of distributed fatigue strength. These simulations test the effects of not mitigating the effects of small sample-size induced estimations errors for the distribution variance by confidence level analysis. In summary, the simulations evaluate the probability that a target quantile of a lognormal fatigue strength distribution was non-conservatively estimated based on a small sample. The results of these simulations are discussed in section 3.3.1 and are summarized in Table 3-2. In this table, the particularly applicable test case is referenced to as the "analytical AGARD-AG-292" method.

The test results demonstrate that whereas a 95% level of confidence was targeted, the actual confidence level that could be achieved was approximately 50%-60%. I.e. there is an approximate 40%-50% chance that a conservative quantile of fatigue strength is non-conservatively overestimated. This is demonstrated for different quantile targets and for different cases of true variance. Another simulation result presented in Figure 3.3 demonstrates that the effective use of a confidence level as low as 50% results in approximately 10%-30% of simulated cases where a targeted 0.999999 reliability level for estimated fatigue strength is overestimated by at least one order of magnitude.

In addition, simulation results discussed in sections 2.6 and 4.2.6 demonstrate that in many cases the uncertainty distribution of random fatigue strength determines most of the uncertainty about random fatigue life. Predicting fatigue life using a working S-N curve with a low level of confidence should thus be expected to result in a Service Life Limit with an equally low confidence level. In particular, the simulation result of test-8 discussed in section 2.6, and presented in Figure 2.36, demonstrates the effect of predicting a 10^{-6} quantile of uncertain fatigue life without targeting a confidence level. The simulation demonstrates that the proportion of simulated cases in which the resulting reliability level is at least 1 order of magnitude less than targeted is approximately 50%-70%.

In conclusion, present work demonstrates that the effect of small sample-size induced estimation errors on the attainable precision of fatigue strength quantile estimation is significant. These estimation errors can reduce the effective confidence level of estimated quantiles of estimated fatigue strength from 95% to about 50%-60%. The work also demonstrates that classically predicting a 10^{-6} quantile of fatigue life without confidence level analysis, i.e. effectively targeting a confidence level of at most 50%, should be expected to result in reduced reliability levels of at least one order of magnitude in about 50%-70% of simulated cases.

<u>Recommendations:</u>

All the simulation results that underlie the conclusions above assume that no prior knowledge or other justified expectation about the variance of a component's fatigue strength exists. Such an assumption may however not be realistic and can lead to over-pessimistic conclusions. This is argued in more detail in sections 3.4 and 3.5. It is therefore recommended for future to work to repeat the simulation tests while making use of the new Bayesian method and Bayesian prior introduced in sections 3.4 and 3.5. In particular the simulations discussed in section 2.6, and section 3.3.1, and the simulation summarized in Figure 3.3, could be repeated. Taking into account the generic Bayesian prior introduced in 3.5.1 should effectively bound the possible variances of fatigue strength distributions to a realistic domain. Therefore, the simulated negative consequences of inadvertent estimation errors of fatigue strength variance should be reduced.

As an extension, it is also recommended for future work to investigate how informative, i.e. limiting, a prior must be to effectively allow the assumption that all sample estimates of the variance of distributed fatigue strength are accurate. In future work it can be simulated how the use of increasingly informative Bayesian priors on the variance of distributed fatigue strength reduces resulting uncertainty about the distribution of fatigue strength.

2. Can the accuracy of fatigue life predictions be improved by accounting for the effects of combined randomness of fatigue strength and flight regime loads?

Present work postulates that fatigue life prediction can be considered as a non-linear function in which a non-linear 'mixing' of the reliabilities of fatigue strength and regime loads occurs. Therefore, it cannot be expected that the use of a working S-N curve with a high reliability translates to the same high-reliability value for predicted fatigue life when computing with average flight regime loads at the same time. Nevertheless, the analytical fatigue life prediction and substantiation method defined in section 2.2, and based on AGARD-AG-292 [16], assumes that the numerical reliability of a fatigue life prediction can be substantiated by the reliability of the conservative S-N working curve only.

Present work comprises simulations which test the reliability that classic fatigue life predictions can ideally attain when considering the reliability of fatigue strength and flight regime loads independently. A test result presented in Figure 2.29 in section 2.6.4.1 demonstrates that even though a conservative working S-N curve with a reliability of 0.999 was used to predict fatigue life, the resulting probability of failure was about 7 times higher. Since the prediction was made without significant estimation errors for the distribution of fatigue strength or flight regime loads, such an error is systematic. A new simulation-based fatigue life prediction and substantiation method introduced in section 2.5 demonstrates full accuracy under the same test circumstances. This simulation-based method models and mitigates the simultaneous and combined uncertainty effects of random fatigue strength and flight regime loads.

Present work also considers that, in practise, there generally is significant estimation uncertainty about the distribution of fatigue strength and flight regime loads. Under such conditions, the improved accuracy of the new simulation-based prediction model may not be significant. Simulations discussed in section 2.6.4.1 indeed demonstrate that fatigue life prediction problems exist for which the modelling errors of the standard analytical method are not significant in comparison to inherent estimation errors due to the availability of a limited amount of fatigue strength and flight tests.

Extended simulations that test the effect of neglecting the influence of in-flight load prediction errors made by Direct Load & Damage Modelling (DLDM) demonstrate that another class of fatigue life prediction problems also exists. Section 4.2.6.2 includes test results on the accuracy of Virtual Fatigue Life Monitoring (VFLM) by DLDM. In the tests it was assumed that the reliability of an employed working S-N curve can be used in full to substantiate the numerical reliability of a VFLM fatigue life prediction. As summarized in Table 4-3, it was empirically found that with increasing certainty of fatigue strength, the relative influence of random prediction errors for in-flight loads also increases. It was observed that synthetically reducing scatter in fatigue strength

by up to 50%, results in demonstrable confidence levels of predicted 10^{-6} fatigue life quantiles of little more than 60%, even though 95% was targeted by the use a working S-N curve with 0.999999 (95%) reliability.

As a solution, present work introduces Probabilistic Load & Damage Modelling (PLDM) for VFLM in sections 4.3.1 and 4.3.2. Since PLDM accounts for both random estimation errors for fatigue strength as well as in-flight loads, increasing the relative influence of random load prediction errors does not significantly reduce demonstrable prediction reliability in comparison to targeted prediction reliability. Present work demonstrates in section 4.3.4 that PLDM's demonstrable prediction reliability for a 10^{-6} fatigue life quantile remains constant between 93% and 97% confidence for the same fatigue life prediction problems that DLDM was tested for.

In conclusion, present work demonstrates that for certain fatigue life prediction problems, the use of extended reliability substantiation models that are able to assess and mitigate the combined influence of random fatigue strength and in-service loads can significantly improve prediction accuracy and precision. However, present work also explicitly demonstrates that these improvements are not significant for all fatigue life prediction problems.

Recommendations:

Although present work has introduced new models that demonstrate improved prediction reliability, these models are comparatively demanding, complex, and computationally expensive. Present work demonstrates that there are fatigue life prediction problems for which simpler models can be used as well. Simulations demonstrate that their systematic modelling errors do not always lead to significant reductions in the reliability of predicted fatigue life. It is therefore recommended to for future work to develop straightforward decision criteria that can be used to decide if the use of simulation-based models can significantly improve accuracy in comparison to the use of simplified reliability substantiation models.

The development of such a decision criteria can be considered even more important when taking into account that it is currently common industry practise to assert that no uncertainty exists about the variance of distributed fatigue strength. In effect, this approach makes that the relative influence of random loads increases significantly in comparison to uncertainties caused by random fatigue strength. The likelihood that a simulation-based numerical model significantly improves prediction reliability thereby increases as well.

The extended numerical reliability substantiation models introduced and validated in present work do still incorporate multiple modelling assumptions, as mainly summarized in sections 2.2 and 2.3. Each of these assumptions could by themselves represent considerable modelling and prediction uncertainty which has been neglected so far. It is thus also recommended for future work to assess and prioritize the potential sources of uncertainty on predicted fatigue life. Such extended studies should confirm that when numerically substantiating the reliability of fatigue life predictions, these quantifications address primary significant sources of uncertainty. Future work could explicitly confirm that the accuracy and precision of all statistical models to substantiate the reliability of fatigue life predictions are really significant in comparison to the uncertainty effects from all other implicitly incorporated modelling assumptions.

3. What is the importance of confidence level analysis for fatigue life prediction?

The reliability substantiation method for fatigue life predictions makes use of estimated distributions of fatigue strength and flight regime loads. In aerospace applications these distributions are generally estimated based on a small number of test results. Estimation errors due to small sample-size effects can thus have a significant influence on the precision of fatigue life predictions and their predicted reliability. Simulations discussed in section 2.6.4 provide uncertainty distributions for predicted quantiles of classically predicted fatigue life due to small sample-size induced estimation errors of random fatigue strength and flight regime loads.

In particular, the simulation result of test #8 presented in Figure 2.36, demonstrates the effect of predicting a 10^{-6} quantile of random fatigue life without targeting a confidence level. It is observed that the expected probability of failure of the fatigue life prediction is significantly, and potentially several orders of magnitude, higher than the targeted 10^{-6} probability. The cause of the non-conservative estimation bias is illustrated in Figure 3.3, where it is explicitly exemplified that the estimator of a standard deviation is only asymptotically unbiased. For small sample sizes, i.e. when few fatigue strength tests have been performed, the estimator is only expected to yield an accurate or conservative result in about 40%-45% of cases.

Examples presented in Table 3-2 demonstrate that if confidence level analysis is conducted, then it is important that accurate and applicable methods are used. The use of methods that are only asymptotically accurate as sample size increases, i.e. as the relevance of confidence level analysis decreases, may reduce demonstrable confidence levels by as much as 20%-30% if a confidence level of 95% is targeted.

The simulation-based fatigue life prediction method introduced in section 2.5 makes use of Subset Simulation to estimate quantiles of randomly distributed fatigue life. The precision of Subset Simulation is however limited due to constraints on computational resources. Present work includes a comparison of uncertainties caused by small sample-size induced estimation errors for the distribution of fatigue strength and flight regime loads with uncertainties caused by the use of Subset Simulation to estimate reliability quantiles of random fatigue life. The simulation result in Figure 2.19 in section 2.5.5 demonstrates that that modelling uncertainty due to the use of Subset Simulation is comparatively small.

In addition to precision errors from Subset Sampling, present work also demonstrates that VFLM prediction models themselves do cause significant load prediction uncertainty. These imprecisions are for example caused by random convergence behaviour during the training phase of the prediction models and the availability of a limited number of training examples. Examples of these uncertainties are discussed for PLDM in section 4.3.3. The significance of imprecisions caused by regression errors is demonstrated in section 4.2.6.2 for DLDM by Figure 4.18, Figure 4.20, and Figure 4.21. These results demonstrate that the relative influence of prediction model imprecisions is generally small but increases with increasing certainty about the distribution of fatigue strength.

In conclusion, present work demonstrates that the primary source of uncertainty that confidence level can mitigate is estimation uncertainty about the distribution of fatigue strength. Small sample-size effects can cause systematic under-estimation of the variance of fatigue strength. If these effects are not mitigated, then the reliability of predicted fatigue lives can be several orders of magnitude less than expected.

Recommendations:

It should be noted that all simulations in chapter 2 do not limit the uncertainty range for estimated variances of fatigue strength by the use of Bayesian statistics and the generic prior introduced in in section 3.5.1. The use of such a prior decreases the observed variance of predicted fatigue life observed during bootstrapping and thus decrease the impact of small sample-size induced estimation errors. For future work, it is thus recommended to repeat the simulations in section 2.6.4 with the use of the generic Bayesian prior.

Present work has numerically demonstrated the potential and significant non-conservative effects of omitting a confidence level analysis when predicting a fatigue life quantile. Therefore, it is recommended for future work to test the significance of confidence level analysis on a wider set of fatigue life prediction problems. Such work should result in decision logic or guidelines for engineers to decide when the inclusion of confidence level analysis shall be required in order to guarantee that a numerical substantiation of the probability of failure of a predicted fatigue life is precise.

4. **When substantiating the reliability of a fatigue life limit under the consideration that the variation of fatigue strength is a random variable, can the use of Bayesian statistics prevent over-conservative fatigue life predictions and enable more economical test requirements?**

Many aerospace applications, and in particular applications following AGARD-AG-292 [16], do not consider that estimated values for the variation of fatigue strength are subject to random estimation error. Present work does however take a different modelling approach and covers this uncertainty by explicit statistical analysis. Present work computes which tolerance intervals for fatigue strength are necessary with this different modelling approach to substantiate 0.999999 reliability of a working S-N curve with a confidence level of 95%. Figure 3.6 indicates that strength reduction factors in the neighbourhood of 0.3-0.6 are necessary if more than about 5 fatigue tests are available. However, for fewer available test results, a lack of data forces regular explicit numerical and mathematical analysis to yield highly conservative results. For example, the availability of less than 4 test results makes it necessary to use reduction factors less than 0.1 if a reliability level of 0.999999 (95%) shall be guaranteed by explicit statistical analysis. The use of such low reduction factors is not common in industry and would result in high weight or cost penalties.

To incorporate a means to acknowledge the presence of alternative engineering knowledge, experience and data, Bayesian statistical analysis is introduced. This enables the use of significantly higher strength values without compromising numerically and explicitly demonstrable reliability. Present work used a large dataset encompassing fatigue test results from many components from dynamic systems of multiple helicopters to generate a generic prior expectation about which distributions of fatigue strength should be expected for typical components. The resulting prior distribution is summarized in Figure 3.10. This prior distribution effectively defines a statistical expectation that the probability that the standard deviation of normalized base-10 lognormally distributed fatigue strength is more than about 0.1 is approximately less than 5%. This prior thus effectively puts an upper bound to the standard deviation that may be expected.

The use of the prior typically enables numerical justification of single-sided tolerance intervals for normalized fatigue strength in a range of about 0.3-0.8, depending on sample size and actual test results for the particular component under consideration. The prior thus allows using explicit, traceable and numerical methods to justify the use of fatigue strength values above 0.3 for working S-N curves, even if less than 5 fatigue strength tests have been performed for the particular component under consideration. The detailed range of typical fatigue strength values that the use of the prior enables is illustrated in Figure 3.12.

In conclusion, the use of Bayesian statistics can enable the use of typical fatigue strength reduction factors for conservative working S-N curves but while employing considerably less restrictive modelling assumptions to numerically substantiate their reliability. In addition, the use of Bayesian statistics provides a framework to make modelling assumptions that are typically implicit or hidden more traceable, transparent, and flexible.

<u>Recommendation:</u>

The Bayesian method introduced in present work is relatively simple. Thereby, it can be used readily in industry. However, for future work, it is recommended to investigate the use of more elaborate models addressing more sources of uncertainty, i.e. all S-N curve parameters, and to broaden the set of supported scatter distribution types.

5. **What are suitable and generic reduction factors for S-N working curves for classical SLL substantiations when these are based on few or no results from directly applicable full-scale fatigue tests?**

It is shown in the frame of research question 4 that the use of Bayesian statistics and the developed generic prior can numerically justify that the use of 1/3 of normalized fatigue strength will lead to a working S-N curve with a reliability of 0.999999 (95%) without having any knowledge about the variance of the fatigue strength of a particular component.

However, the availability of fatigue test results for the specific component under consideration can still stipulate the use of normalized fatigue strength values less than 1/3, as presented in Figure 3.12. These cases occur when fatigue tests reveal standard deviations significantly higher than the mean of the prior distribution. It should, therefore, be acknowledged that if a numerical reliability level of 0.999999 (95%) shall be substantiated, and if scatter for the specific component under consideration is significantly higher than expected, then normalized fatigue strength values more conservative than 1/3 are applicable.

6. Can the reliability of Virtual Fatigue Life Monitoring by Direct Load & Damage Monitoring be substantiated without accounting for the influence of regression or recognition errors?

In the framework of research question 2, simulations discussed in section 2.6.4.1 demonstrate that the reliability of a classic fatigue life prediction with assumed and known usage can under certain circumstances be numerically substantiated by the reliability of a conservative working S-N curve only. DLDM uses an identical model to numerically substantiate the reliability of its usage-based and component-individual fatigue life predictions. Using DLDM for an actual helicopter component demonstrates that its predictions of in-flight loads and induced fatigue damage are accurate. Present work demonstrates in section 4.2.6 that regression coefficients of more than 99% are obtainable for timeframe extreme loads. In the example summarized in Figure 4.39, about 90% of load estimates can be expected to fall within ±10% of the true load. With such accurate and precise load prediction capability, DLDM can indeed numerically substantiate a 10^{-6} probability of failure with 95% confidence by just computing accumulated fatigue damage using an S-N working curve with 0.999999 (95%) reliability. This is demonstrated by simulation results presented in Figure 4.18 and Figure 4.19 in section 4.2.6.2.

However, the component studied in the example discussed in section 4.2.6.2 featured relatively high uncertainty about its fatigue strength since only a single fatigue strength test was performed for it. Extended synthetic test cases, as summarized in Table 4-3, demonstrate that reducing the relative uncertainty about fatigue strength increases the relative influence of load prediction errors. Although justifiable fatigue life increased with decreasing fatigue strength uncertainty, the confidence level with which accumulated fatigue damage was predicted by DLDM dropped by up to 35% in the tested example. Thus, instead of being able to demonstrate a target reliability of 0.999999 with a 95% level of confidence, the actual demonstrable level of confidence reduced to approximately 60%.

In conclusion, the reliability of usage-based fatigue damage predictions by DLDM can be substantiated for cases in which the relative uncertainty about fatigue strength is high. This is however not always the case. There are also cases for which the influence of random load prediction errors is not negligible and for which the actual reliability of fatigue damage predictions made by DLDM can be significantly lower than targeted.

Recommendation:

Although future work could include the development of more accurate regression methods, DLDM's application to a wider range of components in Appendix K indicates that the development of highly accurate regression models for in-flight load prediction may not be realistic. Significant improvements in prediction accuracy are expected to require the addition of dedicated and additional sensor equipment and a considerable increase in the number of test flights that can be used build the statistical regression models. But even if significant improvements in the accuracy of load predictions can be achieved, then DLDM's safe application may still not be assured if a component's fatigue strength features relatively low scatter and is known with relatively high precision through many fatigue test results. DLDM deployment requires that the relative influence of load prediction errors is small. To determine if this condition holds, requires case-by-case synthetic simulation analysis or future work to develop decision criteria that stipulate under which conditions DLDM`s simplified reliability substantiation model can be employed. If the reliability substantiation model of

DLDM cannot be expected to yield reliable results, then the more elaborate PLDM method shall be used instead, as outlined in the frame of research question 7.

7. **When using uncertain estimates of in-service loading, and resulting fatigue damage accumulation, can the reliability of derived fatigue life limitations still be predicted accurately?**

In the frame of research question 2, present work introduces a simulation-based model in section 2.5 that enables the modelling and mitigation of simultaneous randomness of fatigue strength and flight regime loads. Simulation-based and synthetic validation work in section 2.6.3.2 demonstrates that this simulation-based model can be considered as unbiased and accurate. Having available the actual distributions of fatigue strength and flight regime loads, the simulation-based model predicted the 0.999 quantile of distributed fatigue life with high accuracy. A test result presented in Figure 2.30 demonstrates that the introduced simulation-based method predicted the corresponding probability of failure within an approximate accuracy of 5%.

Building on these results, present work introduces an extension to the DLDM modelling framework, referred to as Probabilistic Load & Damage Modelling (PDLM) in sections 4.3.1 and 4.3.2. In addition to DLDM, PLDM also predicts how the errors of its predictions are distributed. By simulating how the random distribution of fatigue strength and load prediction errors results in an uncertainty distribution of accumulated fatigue damage, PLDM can accurately predict and mitigate the effects of both these uncertainties.

Simulations in present work indeed demonstrate that PLDM does not justify all its reliability by the conservativeness of a working S-N curve. For example, in Figure 4.30 in section 4.3.2.3 it is illustrated that only a conservative 10^{-4} quantile of fatigue strength is used by PLDM to substantiate a probability of failure of 10^{-6}. PLDM thus closes the remaining reliability 'gap' by using conservative load values.

Based on more than one thousand hours of recorded flight data from two commercially operated helicopters present work demonstrates that PLDM estimates of accumulated fatigue life are accurate in practise. The accuracy of the reliability substantiation of PLDM is independent of the relative magnitude of uncertainties coming from fatigue strength and predicted loads. In Table 4-5 in section 4.3.4 it is summarized that the demonstrable confidence level of PDLM predictions of the 0.999999 quantile of accumulated fatigue damage is within ±2% of the targeted 95%. A reduction of the standard deviation of fatigue strength of up to 50%, and thus a major increase in the relative importance of load prediction errors, does not cause significant changes in demonstrable prediction accuracy.

In conclusion, present work demonstrates that VFLM predictions of in-service loads or usage do not need to be fully accurate or precise. The use of probabilistic methods can enable the accurate modelling and mitigation of the effects of prediction errors. As a prerequisite, it is only required that the distribution of estimation errors is consistent and predictable.

<u>Recommendations:</u>

The reliability substantiation model for PLDM uses prediction error distributions to simulate and mitigate the influence from expected load prediction errors. PLDM assumes that its predicted prediction error distributions are on average unbiased. Although this is true for the error distributions observed from semi-independent flight test data, fully independent reference data obtained from commercially operated helicopters indicates that this assumption does not hold in practise, as illustrated in sections 4.3.3.1 and 4.3.3.2. Although the tested application for PLDM does not reveal that these biases cause significant errors in the predictions of quantiles of accumulated fatigue life, it is recommended for future work to assess if this also holds for a wider range of cases. If not, then it is recommended to also develop decision criteria and procedures to assess up to which extent prediction biases can be neglected. If there are cases for which the prediction bias of prediction error distributions is not negligible, then it is recommended to research if a temporary in-service test can be

used calibrate the bias for each individual machine, or if calibration can be achieved by some other analytical or statistical means.

The practical implementation of PLDM that present work introduces and outlines in detail in Appendix I comprises several sampling filters and dynamic re-sampling conditions. PLDM uses these while simulating the effect that expected prediction errors can have on the reliability of predicted fatigue damage. These filters and conditional resampling measures prevent the tails of fitted prediction error distributions from yielding unrealistic results. For example, in the practical implementation of PLDM, the load prediction error distributions are fitted by unbounded distributions. Since the entire flight history of a component can consist of more than 50 million subsequent timeframes, some of the initial load samples during reliability simulation can thus be expected to come from 10^{-6} upper prediction error quantiles, and become progressively more conservative as the practical implementation of the simulation in the form of Subset Simulation progresses into ever more conservative load case scenarios. Without filtering, this simulation process can lead to the consideration of in-flight loads that are unrealistic or that even imply in-flight static failure, which is known not to have occurred. It is thus recommended for future work to develop dedicated tail models for the generation of PLDM prediction error distributions, or make use of Bayesian statistics in order to bound samples of in-flight loads to the feasible domain during PLDM reliability simulation. Present work already introduces similar Bayesian statistics in chapter 3 to bound the variance of fatigue strength to a realistic domain.

Another recommendation for future work is the development of alternative VFLM methods. PLDM not only predicts the extreme loads during a timeframe but also predicts timeframe fatigue damage as a function of fatigue strength. This coupling between fatigue strength and predicted timeframe damage results in high computational costs and implementation complexity. The development of an alternative load modelling framework in which loads and fatigue damage can be considered as fully independent is thus recommended. And although efforts summarized in Appendix J were not successful, more work on the development of more simplified and non-probabilistic methods for VLFM, in general, is recommended as well.

8. Can Probabilistic Load & Damage Monitoring accurately and usefully predict and substantiate component-individual and usage-based fatigue damage accumulation?

Practical application of PLDM on 500-800 recorded flight hours each flown by three commercially operated helicopters demonstrates the potential economic benefits of VFLM. The example results summarized in Table 4-6 illustrate that VFLM by PLDM can lead to fatigue life extensions in the range of 200% to more than 10,000%, depending on the application. For many components, such fatigue life extensions imply that the component no longer needs to be replaced during the economic life of the helicopter it is installed on.

Recommendations:

Present work only includes a limited assessment of the practical benefits of PLDM. The limited amount of application tests is for a significant part due to the high computational costs of the current practical implementation of PLDM. For future work, it is recommended to extend the number of cases for which PLDM's economic potential is tested, as well as to increase the computational efficiency of the effort involved.

Furthermore, despite that present work demonstrates and verifies PLDM's capability to predict individually accumulated fatigue life adequately for a wide range of different components and circumstances, additional work is required for actual commercial implementation. To comply with airworthiness guidelines outlined in AC-27 MG-15, safety analysis of the end-to-end application chain is necessary. Such an end-to-end analysis can result in extra hardware and software certification requirements [10][30] which have not been considered so far. For future work, it is recommended, though, to study if relatively strict hardware, software, and process

[30] RTCA DO-178 DAL-C/B requirements for in-flight data acquisition and equivalent requirements for on-ground processing software can be proposed.

control safety requirements can be reduced by using extended statistical analysis to numerically evaluate the potential effects of failures in the end-to-end application chain. For example, analysis discussed in Appendix M demonstrates that even if a VFLM application chain fails entirely, and a component is flown for 20.000 flight hours regardless of its actual SLL, then the probability of a fatigue failure to occur can still be substantiated to be at most 10^{-6}, albeit at a reduced confidence level, e.g. 50% instead of 95%.

References

[1] J. Schijve, Fatigue of Structures and Materials, 2nd red., Springer Science+Business Media, 2009.

[2] Federal Aviation Administration, „Federal Aviation Regulations Part 27 - Airworthiness Standards: Normal Category Rotorcraft," 2014.

[3] P. Edwards en J. Darts, „Part 1: Standardized Fatigue Load Sequences for Helicopter Rotors (HELIX & FELIX); Part 2: Final Definition of HELIX and FELIX," Procurement Executive, Ministry of Defence, Farnborough, Hants, 1984.

[4] SWIFT GmbH, „Fatigue No MATCH for Swift Data loggers and Analysers," *Technology News International,* nr. 78, 2005.

[5] A. Vergoesen, P. Hoek, F. Carati, J. Dominicus, A. ten Have en D. Schütz, „An Automatic In-Flight Data Acquisition System for the RNLN Lynx Helicopter," Amsterdam, 1998.

[6] M. Wallace, H. Azzam en S. Newman, „Indirect approaches to individual aircraft structural monitoring," *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering,* p. 218:329, 2004.

[7] A. Oldersma en M. Bos, „Airframe loads & usage monitoring of the CH-47D "Chinook" helicopter of the Royal Netherlands Air Force," Amsterdam, 2011.

[8] J. Cronkhite, B. Dickson, M. Martin en G. Collingwood, „Operatinoal Evaluation of a Health and Usage Monitoring System (HUMS)," NASA, 1998.

[9] D. Adams, S. Kershner en J. Thielges, „Economical and Reliable Methods of processing HUMS Data for Maintenance Credits," in *American Helicopter Society 55th Annual Forum*, Montreal, 1999.

[10] R. Beale en M. Davis, „Application of Rotorcraft Structural Usage and Loads Monitoring Methods for Determining Usage Credits," Sikorsky Aircraft Corporation, Stratford, Conneticut, USA, 2016.

[11] R. Beale, „Application of virtual monitoring of loads to engineering decision making," in *AHS 75th Annual Forum*, Montreal, Canada, 2014.

[12] S. Kelly, J. Hughes en J. R. D. Kleb, „V-22 Osprey Structural Analysis of Fatigue Effects (SAFE): Assigning Fatigue Life Expended by Means of Recorded Flight Parameters," in *American Helicopter Society 69th Annual Forum*, Phoenix, Arizona, USA, 2013.

[13] J. Tomblin en W. Seneviratne, „Determining the fatigue life of composite aircraft structures using life and load-enhancement factors," 2011.

[14] R. Whiteshead, H. Kan, R. Cordero en E. Saether, „Certification testing methodology for composite structure," 1986.

[15] J. Rouchon, *Certification of rotorcraft composite structures,* 2008.

[16] Advisory Group for Aerospace Research & Development, „Helicopter Fatigue Design Guide," 1984.

[17] Federal Aviation Administration, „AC 27-1B - Certification of Normal Category Rotorcraft," 2014.

[18] A. Thompson en D. Adams, „A computational method for the determination of structural reliability of helicopter dynamic components," in *American Helicopter Society Annual Forum*, Washington D.C., 1990.

[19] Y. Tong, R. Antoniou en C. Wang, „Probabilistic Fatigue Life Assessment For Helicopter Dynamic Components," *Structural Integrity and Fracture,* 2004.

[20] S. Dekker, S. Bendisch en F. Hoffmann, „Fatigue management system and method of operating such a fatigue management system". Patent EP275337 A1, 24 October 2012.

[21] S. Dekker, „Helicopter Fatigue Monitoring with Direct Load & Damage Modelling," 2012.

[22] R. Beale Jr., M. Davis en P. Swindell, „Achieving Usage Based Maintenance with HUMS Regime Recognition," in *AHS 71st Annual Forum*, Virginia Beach, Virginia, USA, 2015.

[23] F. Hoffman, A. Rabourdin, G. Wurzel, S. Bendisch en P. Konstanzer, „Usage Based Maintenance for Dynamic Components of Rotorcrafts," in *American Helicopter Society 67th Annual Forum*, Virginia Beach, VA, 2011.

[24] S. Dekker, G. Wurzel en R. Alderliesten, „Reliability modelling for rotorcraft component fatigue life prediction with assumed usage," *The Aeronautical Journal,* vol. 120, nr. 1232, pp. 1658-1692, 2016.

[25] Federal Aviation Administration,, „Federal Aviation Regulations Part 29 - Airworthiness Standards: Transport Category Rotorcraft," 2014.

[26] European Aviation Safety Agency, „Certification Specifications for Small Rotorcraft - Amendment 3," 2012.

[27] European Aviation Safety Agency, „Certification Specifications for Large Rotorcraft - Amendment 3," 2012.

[28] US Department of Defence, Metallic materials and elements for aerospace vehicle structures, US Departmet of Defence, 2003.

[29] ASTM International, „Standard Practices for Cycle Counting in Fatigue Analysis," 2011.

[30] G. Hahn en W. Meeker, Statistical Intervals: A Guide for Practitioners, John Wiley & Sons Inc., 1991.

[31] A. Wald en J. Wolfowitz, „Tolerance Limits for a Normal Distribution," *The Annals of Mathematical Statistics,* vol. 2, pp. 208-215, June 1946.

[32] J. Darts en D. Schütz, „Development of standard fatigue tests load histories for helicopter rotors - basic considerations and definition of HELIX and FELIX," *AGARD Conference Proceedings,* nr. 297, 1981.

[33] SAE International, „Guidelines and Methods for Conducting the Safety Assessment Process on Civil Airborne Systems and Equipment," 1996.

[34] D. Adams en J. Zhao, „Searching for the Usage Monitor Reliability Factor Using an Advanced Fatigue Reliability Assessment Model," in *Proceedings of the American Helicopter Society 65th Annual Forum*, Grapevine, Texas, USA, 2009.

[35] R. Everett, „A comparison of fatigue life prediction methodologies for rotorcraft," Hampton, Virginia, 1990.

[36] D. Lombardo en K. Fraser, „Importance of Reliability Assessment to Helicopter Structural Component Fatigue Life Prediction," Fishermans Bend, Victoria, Australia, 2002.

[37] J. Zhao, „Development and Demonstration of Advanced Strcuctural Reliability Methodologies for Probabilistic fatigue Damage Accumulation of Aerospace Components," 2008.

[38] J. Zhao en D. Adams, „Achieving Six-Nine's Reliability Using an Advanced Fatigue Relibability Assesment Model," in *American Helicopter Society 66th Annual Forum*, Phoenix, Arizona, 2012.

[39] R. Benton, „Double-Linear Cumulative-Damage Reliability Method," in *American Helicopter Society 67th Annual Forum*, Virginia Beach, VA, 2011.

[40] J. Tang en J. Zhao, „A practical approach for predicting fatigue reliability under random cyclic loading," *Reliability Engineering and System Safety,* nr. 50, pp. 7-15, June 1995.

[41] C. Smith en J.-H. Chang, „Fatigue Reliability Analysis of Dynamic Components with Variable Loadings without Monte Carlo Simulation," in *American Helicopter Society 63rd Annual Forum*, Virginia Beach, Virginia, 2007.

[42] S. Moon en N. Phan, „Component Fatigue Life Reliability with Usage Monitor," in *American Helicopter Society 63rd Annual Forum*, Virginia Beach, Virginia, 2007.

[43] M. A. Brown en J.-H. Chang, „Analytical Techniques for Helicopter Component Reliability," in *American Helicopter Society 64th Annual Forum*, Montreal, Canada, 2008.

[44] J. McFarland en D. Riha, „Uncertainty Quantification Methods for Helicopter Fatigue Reliability Analysis," in *AHS 65th Annual Forum*, Grapevine, Texas, USA, 2009.

[45] D. Adams, „Statistical Analysis of Structural Flight Test Data," in *43rd Annual American Helicopter Society Forum and Technology Display*, St. Louis, Missouri, USA, 1987.

[46] A. Khibnik, M. Atalla, A. Finn, M. Davis, J. Cycon, P. Horbury en H. Bernhard, „Virtual Load Monitoring System and Method". USA Patent US 7,532,988 B2, 12 May 2009.

[47] S. Dekker, „Helicopter Fatigue Monitoring with Direct Load & Damage Modelling," 2012.

[48] C. Genest en A.-C. Favre, „Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask," *Journal of Hydrologic Engineering,* vol. 12, pp. 347-368, July 2007.

[49] MathWorks, „Statistics Toolbox User's Guide - Matlab R2014b," 2014.

[50] J. Hurtado, Structural Reliability - Statistical Learning Perspectives, 1st red., P. F. Pfeiffer en P. P.

Wriggers, Red., Berlin Heidelberg: Springer-Verlag, 2004.

[51] R. Lebrun en A. Dutfoy, „An innovating analysis of the Nataf transformation from the copula viewpoint," *Probabilistic Engineering Mechanics,* vol. 24, August 2009.

[52] B. Echard, N. Gayton en M. Lemaire, „AK-MCS: An active learning reliability method combining Kriging and Monte Carlo Simulation," *Structural Safety,* nr. 33, pp. 145-154, 2011.

[53] V. Caron, A. Guyader, M. M. Zuniga en B. Tuffin, „Some recent results in rare event estimation," *ESAIM Proceedings,* nr. 44, pp. 239-259, January 2014.

[54] S.-K. Au en J. L. Beck, „Estimation of small failure probabilities in high dimensions by subset simulation," *Probabilistic Engineering Mechanics,* nr. 16, pp. 263-277, 2001.

[55] T. Hesterberg, D. Moore, S. Monaghan, A. Clipson en R. Epstein, in *Introduction to the Practise of Statistics - 6th edition*, W.H. Freeman, 2009.

[56] S. Dekker, G. Wurzel en R. Alderliesten, „A Bayesian tolerance interval estimation method for fatigue strength substantiation of rotorcraft dynamic components," *International Journal of Fatigue,* nr. 92, pp. 333-344, 2016.

[57] R. Everett, F. Barlett en W. Elber, „Probabilistic fatigue methodology for six nines reliability," Hamption, Virginia, 1990.

[58] Federal Aviation Administration, „Advisory Circular 27-1B - Cerification of normal cetgory rotorcraft - Fatigue evaluation of rotorcraft structure," 2008.

[59] C. Kassapoglou, „Predicting the Structural Performance of Composite Structures Under Cycliic Loading," Delft, the Netherlands, 2012.

[60] N. Post, J. Bausano, S. Case en J. Lesko, „Modelling the remaining strength of structural composite matrials subjected to fatigue," *International Journal of Fatigue,* nr. 28, pp. 1100-1108, 2006.

[61] M. Guida en F. Penta, „A Bayesian analysis of fatigue data," *Structural Safety,* vol. 32, pp. 64-76, 2010.

[62] A. Wald en J. Wolfowitz, „Tolerance Limits for a Normal Distribution," *The Annals of Mathematical Statistics,* vol. 2, pp. 208-215, June 1946.

[63] A. Weissberg en G. Beatty, „Tables of Tolerance-Limit Factors for Normal Distributions," *Technometrics,* vol. 2, nr. 4, pp. 483-500, 1960.

[64] IHS ESDU, „The statistical analysis of data from normal distributions, with particular reference to small samples," 1993.

[65] B. Efron en D. Hinkley, „Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information," *Biometrika,* vol. 65, nr. 3, pp. 457-482, December 1978.

[66] W. Meeker en L. Escobar, Statistical Methods for Reliability Data, John Wiley & Sons, Inc, 1998.

[67] T. DiCiccio en B. Efron, „Bootstrap Confidence Intervals," *Statisitcal Science,* vol. 11, nr. 3, pp. 189-228,

1996.

[68] Federal Aviation Administration, „Advisory," 2014.

[69] G. Edwards en L. Pacheco, „A Bayesian Method for Establishing Fatigue Design Curves," *Structural Safety,* vol. 2, pp. 27-38, 1984.

[70] G. Box en G. Tiao, Bayesian Inference in Statistical Analysis, Wiley Classics Library Edition Published 1992 red., Addison-Wesley Pub, 1973.

[71] E. Weisstein, December 2014. [Online]. Available: http://mathworld.wolfram.com/StandardDeviationDistribution.html.

[72] R. Beale, M. Davis en P. Swindell, „Achieving Usage Based Maintenance with HUMS Regime Recognition," in *AHS 71st Annual Forum*, Virginia, USA, 2015.

[73] D. Stephan, „Aeronautical Design Standard Handbook Condition Based Maintenance System for US Army Aircraft," 2013.

[74] Brand, „Verfahren zur Ermittlung der individuellen Lebensdauer eines Fluggerätes". Germany Patent DE4336588 C2, 27 October 1993.

[75] R. Bromann, „Theoretical Analyses of the Process of Flight Regime Recognition on EC 145," Eurocopter Deutschland GmbH, Intitut für Angewandete Informatik / Automatiseriungstechnik, Munich, 2010.

[76] A. Rabourdin, „Onboard flight state recognition on the basis of neural networks," Eurocopter Deutschland GmbH, Technische Universität München, Munich, 2008.

[77] E. Laillet, P. Maisonneuve en G. Zuber, „Usage Analysis, SLL Dynamic Assessment Methods, Benefits and Constraints," in *38th European Rotorcraft Forum*, Amsterdam, 2012.

[78] G. Bogendörfer, „Flight Regime Recognition Based on Statistical Approaches," Eurocopter Deutschland GmbH, Technische Universität München, Munich, 2011.

[79] S. Wu, E. Bechhoefer en D. He, „A Practical Regime Prediction Approach for HUMS Applications," in *AHS 63rd Annual Forum*, Virginia Beach, Virginia, USA, 2007.

[80] J. Isom, J. Cycon, E. Eller en M. Davis, „Virtual monitoring of aurcraft fleet loads". United States of America Patent US 2011/0112878 A1, 12 May 2011.

[81] J. Isom, M. Davis, J. Cycon, J. Rozak en J. Fletcher, „Flight Test of Technology for Virtual Monitoring of Loads," in *AHS 69th Annual Form*, Phoenix, USA, 2013.

[82] P. Bates, M. Davis, J. Cycon en P. Sadegh, „Method of determining a maneuver performed by an aircraft". United States of America Patent US 2011/0264310 A1, 27 October 2011.

[83] P. Bates, M. Davis en P. Sadegh, „Rotorcraft Dynamic Component Usage Based Maintenance Process," in *American Helicopter Society 66th Annual Forum*, Phoenix, Arizona, 2010.

[84] P. Bates, M. Davis en P. Sadegh, „Rotorcraft Dynamic Component Usage Based Maintence Process," in

American Helicopter Society 66th Annual Forum, Phoenix, Arizona, USA, 2010.

[85] D. Haas, J. Milano en L. Flitter, „Prediction of Helicopter Component Loads Using Neural Networks," 1993.

[86] D. Haas, L. Flitter en J. Milano, „Helicopter Flight Data Feature Extraction or Component Load Monitoring," vol. 33, nr. 1, 1996.

[87] K. McCool, L. Flitter en D. Haas, „Development and Flight Test Evaluation of a Rotor System Load Monitoring Technology," Washington D.C., 1999.

[88] R. Cabell, C. Fuller en W. O'Brien, „Neural Network Modelling of Oscillatory Loads and Fatigue Damage Estimation of Helicopter Components," vol. 209, nr. 2, pp. 329-342, 1998.

[89] D. Kim en M. Marciniak, „Horizonal Tail Maneuver Load Prediction Using Backpropagation Neural Networks," vol. 37, nr. 2, 2001.

[90] D. Kim en M. Marciniak, „Prediction of Vertical Tail Maneuver Loads Using Backpropagation Neural Networks," vol. 37, nr. 3, 2000.

[91] M. Allen en R. Dibley, „Modelling Aircraft Wing Loads from Flight Data Using Neural Networks," Natial Aeronautics and Space Administration, Edwards, 2003.

[92] F. Polanco, „Estimation of Structural Component Loads in Helicopters: A Review of Current Methodologies," DSTO Aeronautical and Maritime Research Laboratory, Melbourne, 1999.

[93] S. Bendisch, F. Hoffmann en A. Rabourdin, „Fatigue Management System". Patent ECD internal: 2010e50-D0314, 2012.

[94] C. Cheung, B. Rocha, V. J. en J. Puthuparampil, „Improved load estimation and fatigue life tracking demonstrated on multiple platforms using the Signal Approximation Method," in *AHS 72nd Annual Forum*, Palm Beach, Florida, USA, 2016.

[95] J. Valdes, C. Cheung en M. Li, „Towards conservative helicopter loads prediction using computational intelligence techniques," in *WCCI 2012 IEEE World Congress on Computational Intelligence*, Brisbane, Australia, 2012.

[96] C. Cheung, J. Valdes en M. Li, „Use of evolutionary computation techniques for exploration and prediction of helicopter loads," in *WCCI 2012 IEEE World Congress on Computational Intelligence*, Brisbane, Australia, 2012.

[97] J. Rudd, „Airframe Digital Twin," in *The 27th Symposium of the International Committee on Aeronautical Fatigue*, Jerusalem, Israel, 2013.

[98] T. Prendergast, „Fatigue Monitoring Systems and Methods Incorporating Neural Networks". United States of America Patent US 6,480,792 B1, 12 November 2002.

[99] G. Wurzel, „HELIFAM Final Report," 2015.

[100] M. Hajek, S. Manner en S. Sören, „Blade Root Integrated Optical Fiber Bragg Grating Sensors - A Highly Redundant Data Source For Future HUMS," in *AHS 71st Annual Forum*, Virginia Beach, Virginia, USA, 2015.

[101] M. Agustin, „Hazard assessment for usage credits on helicopters using health and usage monitoring system," FAA, Virginia, USA, 2004.

[102] B. Lader, M. Damin, M. Davis en M. Kingsley, „Development, Validation and Demonstration of Health and Usage Monitoring System Technology to Detect Rotorcraft Mechanical Faults," FAA, Virginia, USA, 2013.

[103] D. Adams en J. Zhao Dr., „Searching for the usage monitoring reliability factor using an advanced fatigue reliability assesment model," in *American Helicopter Society 65th Annual Forum*, Grapevine, Texas, 2009.

[104] C. Hong, „Accuracy Assessment of Health and Usage Monitoring System Regime Recignition Algorithms," US Federal Aviation Authority, 2016.

[105] S. Dekker, S. Bendisch, F. Hoffmann en G. Wurzel, „Helicopter Fatigue Monitoring with Direct Load & Damage Modelling," in *5th International HELI World Conference*, Frankfurt, 2013.

[106] T. Hastie, R. Tibshirani en J. Friedman, The Elements of Statistical Learning - Data Mining, Inference and prediction, 2nd red., New York: Springer, 2009.

[107] A. Ng, J. Ngiam, C. Yu, Y. Mai, C. Suen, A. Coates, A. Maas, A. Hannun, B. Huval, T. Wang en S. Tandon, „UFLDL Tutorial," Stanford University, 2016. [Online]. Available: http://ufldl.stanford.edu/tutorial/. [Geopend 12 October 2016].

[108] M. Tipping, „Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learning Research,* nr. 1, pp. 211-244, 2001.

[109] M. Tipping, „Relevance Vector Machine". US Patent US 6,633,857 B1, 14 October 2003.

[110] M. Tipping en A. Faul, „Fast Marginal Likelihood Maximisation for Sparse Bayesian Models," in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Key West, Florida, USA, 2003.

[111] M. Tipping, „miketipping[.com]," Vector Anomaly, March 2009. [Online]. Available: http://www.miketipping.com/downloads.htm. [Geopend 12 October 2016].

[112] S.-K. Au, „On the Solution of First Excursion Problems by Simulation with Applications to Probabilistic Seismic Performance Assessment," Pasadena, California, 2001.

[113] S. Au, J. Ching en J. Beck, „Application of subset simulation methods to reliability benchmark problems," *Structral Safety,* nr. 29, pp. 183-193, 2007.

[114] G. Papadopoulos, P. Edwards en A. Muray, „Confidence Estimation Methods for Neural Networks: A Practical Comparison," *IEEE Transactions on Neiral Networks,* vol. 6, nr. 12, pp. 1278-1287, 2001.

[115] „Confidence Interval Prediction for Neural Network Models," *Chryssolouris, G.; Lee, M.; Ramsey, A.,* vol.

7, nr. 1, pp. 229-232, 1996.

[116] R. Dybowski en S. Roberts, „Confidence Intervals and Prediction Intervals for Feed-Forward Neural Networks," *Clinical Applications of Artificial Neural Networks,* pp. 298-326, 2001.

[117] L. Ungar, R. Veaux en R. Rosengarten, „Estimating Prediction Intervals for Artificial Neural Networks," in *Proceedings of the 9th Yale Workshop on Adaptive Learning Systems*, 1996.

[118] A. Khosravi, S. Nahavandi, D. Creighton en A. Atiya, „A Comprehensive Review Of Neural Network-based Prediction Intervals and New Advances," *Transactions on Neural Networks,* vol. 22, nr. 9, pp. 1341-1356, 2011.

[119] H. Quan, D. Srinivasan en A. Khosravi, „Construction of Neural Network-based Prediction Intervals using Particle Swarm Optimization," in *WCCI 2012 IEEE World Congress on Computational Intelligence*, Brisbane, Australia, 2012.

[120] A. Khosravi, S. Nahavandi, D. Creighton en A. Atiya, „Lower Upper Bound Estimation Method for Construction of Neural Network-Based Prediction Intervals," *IEEE Transations on Neural Networks,* vol. 22, nr. 3, pp. 337-336, 2011.

[121] Z. Zapranis en L. Efstratios, „Prediction intervals for neural network models," in *Proceedings of the 9th WSEAS International Conference on Computers*, 2005.

[122] R. Veaux, J. Schumi, J. Schweinsberg en L. Ungar, „Prediction Intervals for Neural Networks via Nonlinear Regression," *Technometrics,* vol. 40, nr. 4, pp. 273-282, 1998.

[123] G. Hwang en A. Ding, „Prediction Intervals for Artificial Neural Networks," *Journal of the American Statistical Association,* vol. 92, nr. 438, pp. 748-757, 1997.

[124] G. Box en G. Tiao, Bayesian Inference in Statistical Analysis, Wiley Classics Library Edition Published 1992 red., Addison-Wesley Pub, 1973.

[125] J. Hammersley en D. Handscomb, Monte Carlo Methods, Wiley, 1964.

[126] J. Hurtado en A. Barbat, „Monte Carlo Techniques in Computatinal Stochastic Mechanics," *Archives of Computational Methods in Engineering,* vol. 5, nr. 1, pp. 3-30, 1998.

[127] J. Hurtado, Structural Reliability - Statistical Learning Perspectives, 1st red., P. F. Pfeiffer en P. P. Wriggers, Red., Berlin Heidelberg: Springer-Verlag, 2004.

[128] Y.-G. Zhao en T. Ono, „A general procedure for first/second-order reliability method (FORM/SORM)," *Structural Safety,* vol. 21, pp. 95-112, 1999.

[129] Y.-G. Zhao en T. Ono, „New approximations for SORM: part 1 & part 2," *Journal of Engineering Mechanics,* vol. 125, nr. 1, January 1999.

[130] A. Kiureghian en T. Dakessian, „Multiple design points in first and second-order reliability," *Structural Safety,* vol. 20, pp. 37-49, 1998.

[131] J.-M. Bourinet, F. Deheeger en M. Lemaire, „Assessing small failure probabilities by combined subset simulation and Support Vector Machines," *Structural Safety,* nr. 33, pp. 343-353, 2011.

[132] W. Fauriat en N. Gayton, „AK-SYS: An adaptation of the AK-MCS method for system reliability," *Reliability Engineering and System Safety,* nr. 123, pp. 137-144, 2014.

[133] M. Balesdent, J. Morio en J. Marzat, „Kriging-based adaptive Importance Sampling algorithms for rare event estimation," *Structral Safety,* nr. 44, pp. 1-10, 2013.

[134] D. Moore, G. McCabe en B. Craig, Introduction to the Practise of Statistics, 6th red., W.H. Freeman, 2006.

[135] R. Shanmugam, „Correlation between the Sample Mean and Sample Variance," *Journal of Modern Applied Statistical Methods,* vol. 7, nr. 3, 2008.

[136] Q. Wang, „Approximate goodness-of-fit tests of fitted generalized extreme value distributions using LH moments," *Water Resources Research ,* vol. 34, nr. 12, pp. 3497-3502, December 1998.

[137] J. Stedinger en L.-H. Lu, „Goodness-of-Fit Tests for Regional Generalized Extreme Vaue Flood Distributions," *Water Recources Reseach,* vol. 27, nr. 7, pp. 1765-1776, July 1997.

[138] V. Vapnik, E. Levin en Y. Le Cun, „Measuring the VC-Dimension of a learning Machine," *Neural Computation,* vol. 5, nr. 6, pp. 851-876, 1994.

[139] J. Shlens, „A tutorial on principle component analysis," 25 March 2003. [Online]. Available: https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf. [Geopend Tuesday October 2016].

[140] A. Ng, „Stanford CS229 Lecture notes," 2016. [Online]. Available: http://cs229.stanford.edu/notes/cs229-notes10.pdf. [Geopend 11 10 2016].

[141] A. Faul en M. Tipping, „Analysis of sparse Bayesian learning," *Advanced in neural information processing systems,* nr. 14, pp. 383-389, 2002.

[142] K. Trigui, „Optimisation of component lifetimes using fatigue monitoring through flight regime recognition," 2015.

[143] RTCA SC-169; EUROCAE WG-12, „Software Considerations in Airborne Systems and Equipment Certification," 1992.

[144] M. Matsumoto en T. Nishimura, „Mesenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator," *ACM Transactions on Modeling and Computer Simulation ,* vol. 8, nr. 1, pp. 3-30, January 1998.

[145] A. Nieslony, *Rainflow,* 2003.

[146] US Army Research, „Aeronautical Design Standard Handbook for Condition Based maintenance Systems for US Army Aircraft," 2012.

[147] G. Wurzel, „Development of a Condition-Based Maintenance Concept for Helicopter Drive Systems,"

Vienna, 2011.

[148] R. E. J. Benton, „Double-Linear Cumulative-Damage Reliability Method," in *American Helicopter Society 67th Annual Forum*, Virginia Beach, VA, 2011.

[149] R. E. J. P. Benton, „Further Advances in a Recently Developed Cumulative-Damage Reliability Method," in *American Helicopter Society 66th Annual Forum*, Phoenix, Arizona, 2010.

[150] R. E. P. Benton Jr., „Cumulative-Damage Reliability for Random-Independent (Normal-or Weibul-distributed) Fatigue Stress, Random-Fixed Strength, and Deterministic Usage," in *American Helicopter Society 65th Annual Forum*, Grapevine, Texas, USA, 2009.

[151] D. O. Adams en J. Z. Zhao, „Usage Monitor Reliability Factor Using an Advanced Fatigue Reliability Assessment Model". United States of America Patent US 8,200,442 B2, 12 June 2012.

[152] P. Shanthakumaran, T. Larchuk, R. Christ en D. H. E. Mittleider, „Usage Based Fatigue Damage Calculation for AH-64 Apache Dynamic Components," in *American Helicopter Society 66th Annual Forum*, Phoenix, Arizona, 2010.

[153] W. Harris, T. Larchuck, E. Zanoni en L. Zion, „Application of probabilistic methodology in the development of retirement lives of critical dynamic components in rotorcraft," in *American Helicopter Society 55th Annual Forum*, Montreal, Canada, 1999.

[154] S. Maley, P. J. en N. Phan, „US Navy Roadmap to Structural Health and Usage Monitoring - The Present and Future," in *American Helicopter Society 63rd Annual Forum*, Virginia Beach, Virginia, USA, 2007.

[155] S. Bendisch, F. Hoffmann en A. Rabourdin, „Fatigue Management System". USA Patent US8478457 B2, 17 June 2011.

[156] R. Romero, H. Summers en J. Cronkhite, „Feasibility Study of a Rotorcraft Health and Usage Monitoring System (HUMS): Results of Operator's Evaluations," 1996.

[157] J. P. N. P. Scott Maley, „US Navy Roadmap to Structural Health and Usage Monitoring - The Present and Future," Patuxent River, 2007.

[158] J. Cycon, „Sikorksy Aircraft's commitment to Health and Usage Management Systems," in *2011 CHC Safety & Quality Summit*, Vancouver Canada, 2011.

[159] K. McCool en G. Barndt, „Assessment of helicopter structural usage monitoring system requirements," FAA, Virginia, 2004.

Appendix A. Reliability modelling

Six well-established reliability modelling methods are introduced. These methods are introduced such as to provide understanding in their basic working principle and to judge the merits of these approaches. The Subset Simulation method in appendix A.6 is presented in greater detail as it was selected as the method of choice. Appendix A.7 briefly references more recent and complex methods that are not discussed in further detail.

A.1 Analytical reliability

Reliability is mathematically defined as one minus the probability of failure:

$$R = 1 - P_{fail} \quad \text{with} \quad P_{fail} \in [0,1] \tag{6.1}$$

In typical applications, the probability of failure, $P_{fail,}$ is a number much smaller than one, for example, 1/1000, which can be denoted equivalently by 10^{-3} or 0.001. Such a one-in-a-thousand probability of failure thus corresponds to a reliability of 0.999, which is also denoted by 0.9_3 or by referring to 'three-nines' of reliability.

What constitutes a 'failure' depends on the particular application. However, generalizing, failure can be defined as the event that a critical 'demand' parameter *Y* crosses a critical threshold value of a 'capacity' parameter *C*. A well-known example of a failure event is where a load *L* exceeds the static strength *S*, causing static failure:

$$\text{failure} := L > S \tag{6.2}$$

In the case that both component static strength and the applied load are random parameters, the situation as sketched in Figure A.1 applies.

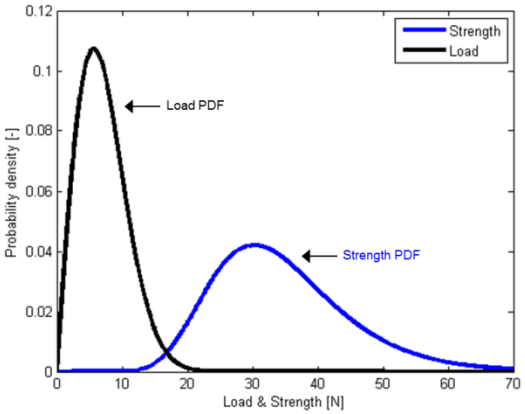


Figure A.1: Example of random load and strength.

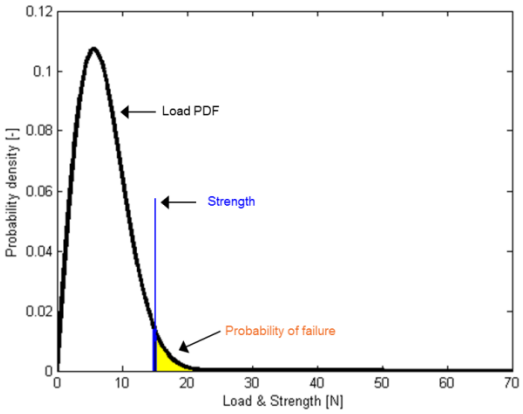


Figure A.2: Example of a probability of failure with random load and known strength.

The probability of static failure can then be computed as follows:

$$P_{fail} = \int_{-\infty}^{\infty} p(S) \cdot \int_{S}^{\infty} p(L) dL \cdot dS = \int_{-\infty}^{\infty} p(S) \cdot P_L(S) dS \tag{6.3}$$

Where:

- $p(S)$ is the probability density function (PDF) of strength
- $p(L)$ is the PDF of load
- $P_L(Z)$ is the cumulative distribution function (CDF) of load, evaluated for value Z.

The problem that is addressed in the context of fatigue life prediction is only a sub-problem of the class of problems that (6.3) belongs to. The problem addressed here can be stated as: what is the probability of failure, given capacity *C* and the random critical parameter *Y*? Remaining in the nomenclature of the example of static failure, this corresponds to the following:

$$P_{fail}(S) = \int_{S}^{\infty} p(L)dL = 1 - P_L(S) \tag{6.4}$$

which is no more than one minus the CDF of the load evaluated at the set strength value, see also Figure A.2.

Consider now that critical parameter *Y* is a function of several random parameters, i.e. *L* is a (complex) function of the random parameter vector ω :

$$L = f(\omega) \tag{6.5}$$

In this case, the probability of failure can no longer be evaluated by a simple one-dimensional integral such as (6.4), especially not if the function *f* is complex or when the domain of values of *L* that lead to failure is complex. In order to solve the problem at hand, it is customary to recast the failure definition (6.2) in the form of an indicator function $\mathrm{I}(\ldots)$:

$$\mathrm{I}\left[L(\omega) \mid S\right] = \begin{cases} 1 \text{ if } S < L(\omega) \\ 0 \text{ otherwise} \end{cases} \tag{6.6}$$

The probability of failure, given strength *S*, now follows from integrating over the parameter space Ω :

$$P_{fail}(S) = \int_{\Omega} \mathrm{I}\left[L(\omega) \mid S\right] \cdot p(\omega) \cdot d\omega \tag{6.7}$$

where $p(\omega)$ is the (multivariate) PDF of the parameter (vector) ω . To aid understanding of equation (6.7), Figure A.3 is included. A multivariate PDF over the two-dimensional parameter space Ω is shown here as a wireframe model. The set of parameter coordinates ω_{fail} for which $S < L(\omega)$ make out the failure domain. This failure domain in Ω_{fail} is coloured red. The probability of failure now corresponds to the probability mass 'above' the red failure domain. Or, recasting (6.4) into a more general form:

$$P_{fail}(S) = \int_{\Omega_{fail}} p(\omega)d\omega \tag{6.8}$$

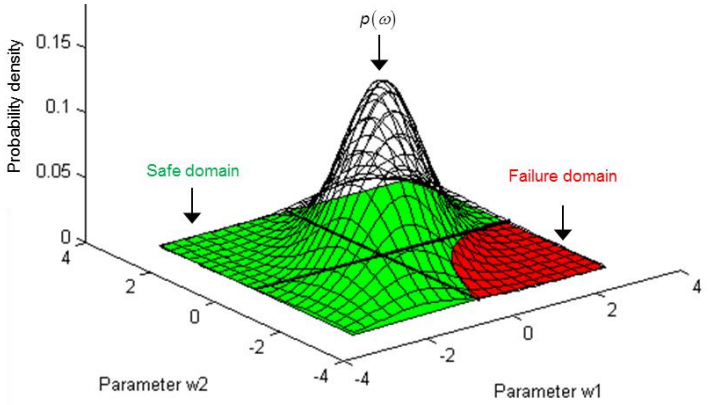


Figure A.3: Illustrative reliability integration problem showing a two-dimensional parameter space over which a multivariate PDF is defined. The parameter region which leads to failure is marked red.

In practise however, the integral (6.8) is virtually never analytically solvable. The exact failure domain in the parameter space is often unknown and sometimes highly complex. Moreover, the (multivariate) PDF of the parameters may also be complex and not be readily evaluated. As such, it is usually necessary to approximate the integral (6.7) or (6.8) by specialized (numerical) methods.

A.2 Basic Monte Carlo

A well-known and intuitive method to solve the reliability integral (6.7) is by means of a Basic Monte Carlo (BMC) simulation. Using the generalized nomenclature from section A.1, the BMC estimate of the failure probability, given a capacity *C,* is as follows:

$$P_{fail}(C) = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \left\{ \mathrm{I}\left[f(\omega_i) \mid C \right] \right\} \quad \text{as} \quad n_{sim} \to \infty \tag{6.9}$$

Where:

- n_{sim} denotes the number of Monte Carlo simulations
- ω_i is the ith parameter sample that is drawn from the Probability Density Function (PDF) $p(\omega)$. (These samples must be independent and identically distributed (i.i.d.))

Essentially, the BMC estimator draws a large number of samples from the sample space, then evaluates for every sample if it causes failure and finally estimates the probability of failure by the fraction of the number of failed samples over the total number of samples. Referring to Figure A.4, the probability of failure can be seen as the fraction of parameter samples that lie within the failure domain.

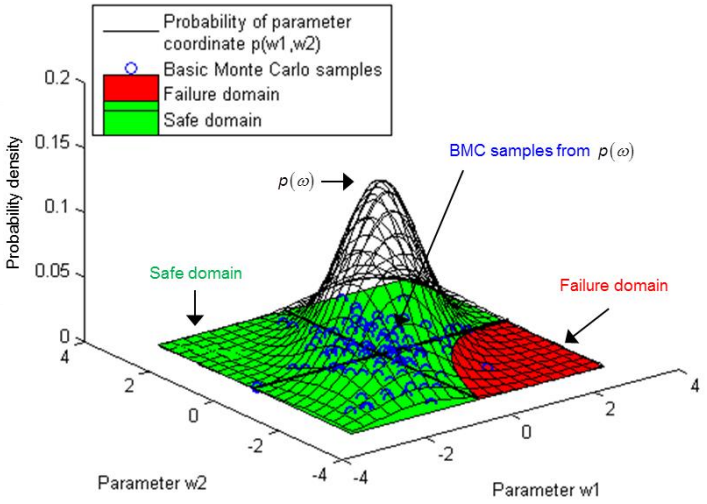


Figure A.4: Illustrative example how BMC can solve a reliability integration problem. Close inspection of the figure will reveal that one out of the in total one hundred drawn samples lies in the failure domain, indicating $P_{fail} = 10^{-2}$.

Considering the strong law of large numbers and that the indicator function $\mathrm{I}\left[f(\omega)|C\right]$ follows a binomial distribution, the coefficient of variation of the BMC estimator for a probability of failure P_{fail} is as follows: (See for example also Hammersley & Handscomb [125])

$$\frac{\sigma_{P_{fail}}}{\mu_{P_{fail}}} = \sqrt{\frac{1-P_{fail}}{P_{fail}\cdot n_{sim}}} \tag{6.10}$$

This indicates that the estimation error is proportional to $1/\sqrt{n_{sim}}$, hence independent of the dimension of the parameter space Ω. This is important as it means that the accuracy of the estimate of failure probability, $\hat{P}_{fail}$, is independent of the number of random variables that determine *f*.

equation (6.10) however also indicates that efficiency is roughly proportional to only $10/P_{fail}$ (corresponding to an upper bound of the CoV of 30%.) [38]. This means that to estimate a probability of failure of 10^{-6} within a reasonable degree of accuracy, it is necessary to perform at least ten million simulations. As such, the BMC estimator is unattractive for reliability problems where a very low probability of failure needs to be shown.

Any search in open literature will provide an abundance of references and further elaboration on the BMC estimator.

A.3 Importance Sampling

To improve the efficiency of the Monte Carlo estimator it is possible to sample from a distribution that 'focuses' on the regions of interest and does not necessarily coincide with $p(\omega)$, the PDF of the parameter vector ω. The Monte Carlo estimator with Importance Sampling (IS) can be expressed as follows:

$$P_{fail}(C) = \frac{1}{n_{sim}}\cdot\sum_{i=1}^{n_{sim}}\frac{\mathrm{I}\left[f(\omega_i)|C\right]}{q(\omega_i)}\cdot p(\omega_i) \quad \text{as} \quad n_{sim}\to\infty \tag{6.11}$$

Where:

- $q(\omega)$ is the PDF from which (i.i.d.) samples ω_i are drawn.

Graphically, this technique is illustrated by Figure A.5. The distribution from which samples are drawn is displayed by a light blue wireframe. This distribution puts most of its probability mass in the direct neighbourhood of the border between the safe and failure domain. Close inspection also reveals that the sampling PDF $q(\omega)$ has its probability mass concentrated on the area where samples drawn from the parameter PDF $p(\omega)$ are most likely to enter the failure domain. Without further elaboration, it is intuitive to see that the IS estimator is much more efficient than the BMC estimator.

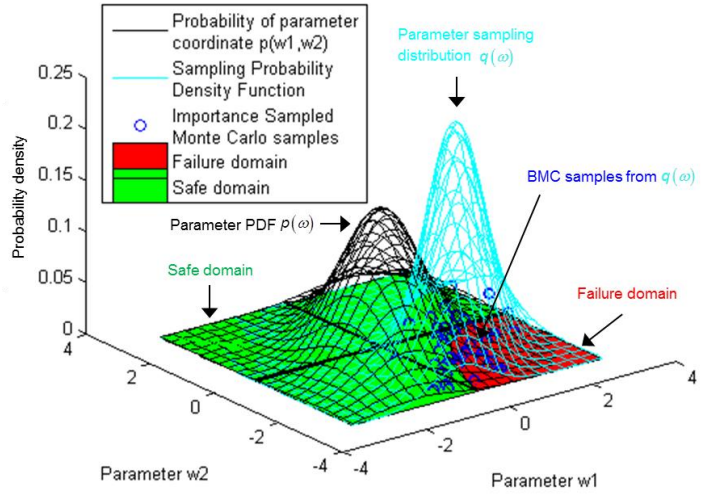


Figure A.5: Illustrative example of Importance Sampling. Samples are generated from a PDF that 'focuses' on the border between the safe and failure domain. This improves the efficiency of the Monte Carlo estimator.

However, the efficiency and trustworthiness of the IS estimator greatly depend on the chosen sampling PDF $q(\omega)$. If $q(\omega)$ is chosen inappropriately, the IS estimator can become highly inefficient and yield high errors. There are no general procedures that guarantee the appropriateness of the chosen sampling PDF $q(\omega)$ and experience and expert judgement must often be relied on to ensure a trustworthy outcome of the IS estimator.

Any search in open literature will provide an abundance of references and further elaboration on the IS estimator. See for example Hurtado & Barbat or Hurtado [126, 127] for a general review.

A.4 First & Second Order Reliability Methods

The limit state function is another way to distinguish the safe and failure domain and can be defined as follows:

$$\Lambda(\omega) = C - f(\omega) \tag{6.12}$$

Where C is the 'capacity' and $f(\omega)$ the critical parameter as a function of the parameter vector ω. The limit state function is thus negative in the failure domain and positive in the safe domain. In terms of the example in section A.1, one can consider the limit state function as computing the strength margin, where a negative margin indicates that the load exceeds static strength and causes failure.

The first step of applying First or Second Order Reliability Methods (FORM or SORM) is to transform the parameter space Ω such that the parameter PDF $p(\omega)$ becomes a (multivariate) standard normal distribution. This can be done for example by a NATAF or Rosenblatt transformation, as detailed by Hurtado [127].

It can now be intuitively understood that the shortest distance between the origin of the multivariate standard normal distribution and the failure domain is proportional to the probability of failure. (Imagine that sampling from the normalized PDF will, on average, result in a symmetrically growing hyper-sphere encompassing the samples. The time that the growing hyper-sphere first intersects the failure domain is directly proportional to the probability of failure.) The point of the failure domain that is closest to the origin of the (multivariate) standard normal parameter PDF is called the Most Probable Point (MPP), whose coordinate is herein designated by β . The MPP is the solution of the following optimization problem:

$$\beta = \underset{\omega}{\arg\min}\left\{\|\omega\|\right\} \quad \text{such that} \quad \Lambda(\omega) = 0 \tag{6.13}$$

This optimization problem can normally be solved by standard gradient-based minimization algorithms.

In the case of FORM, the probability of failure is now estimated by:

$$P_{fail} = \mathrm{N}\left(-\|\beta\|\right) \tag{6.14}$$

Where $\mathrm{N}(z)$ denotes the Cumulative Distribution Function (CDF) of the standard normal distribution. The FORM estimator is only exact if the limit state function is linear and if the parameters ω are Gaussian.

When SORM is applied, an additional correction factor is added. This correction factor takes into account the curvature κ of a second-order approximation of the failure hyper-surface $\Lambda(\omega) = 0$ at the MPP or design-point. There exist multiple methods to approximate the failure surface and the precise approximation of the probability of failure varies accordingly. But in general, the SORM estimate of the probability of failure has the following form:

$$P_{fail} \approx \mathrm{N}\left\{a(\kappa)\cdot\|\beta\| + b(\kappa)\right\} \tag{6.15}$$

Where *a* and *b* are some function of the local curvature κ .

A general illustration of FORM and SORM is given in Figure A.6. In this example, the parameter PDF $p(\omega)$ is transformed to a multivariate standard normal distribution, displayed as a thick black wireframe, and all coordinates in the parameter space Ω are transformed correspondingly. The shortest Euclidian distance from the origin of the transformed parameter space to the also transformed failure domain is shown by a thick purple line. The point belonging to the failure domain and closest to the origin is the Most Probable Point (MPP) and is displayed as a thick open blue circle. The first order approximation to the border between the safe and failure domains is shown as a thick blue dashed line, whereas the second order approximation is displayed with a thick dashed yellow line. The FORM approximation of the probability of failure is the probability mass shown in light blue relative to the one-dimensional standard normal distribution that follows the direction [0,0; MPP], as shown by a thick black continuous line over the thin black wireframe.

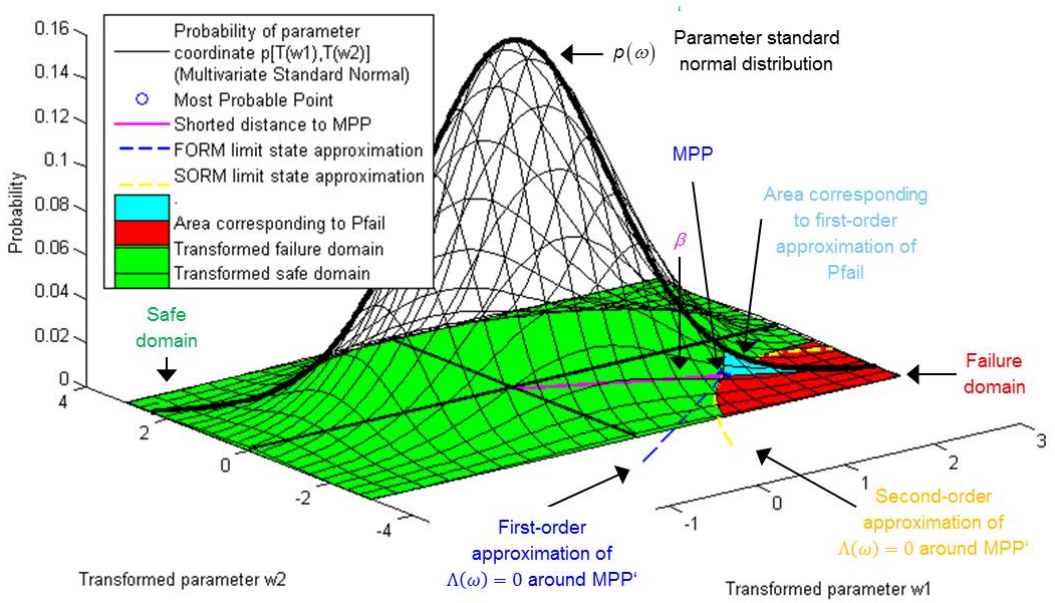


Figure A.6: Example application of the First and Second Order Reliability Method.

High accuracy and trustworthiness of the FORM and SORM are conditional on stringent conditions; see Zhao & Ono or Hurtado [128, 127] for more elaboration on these conditions. These conditions are however not a substitute for quantified tolerance intervals on P_{fail}. These are generally not available for FORM or SORM estimates. In general, it can be said that SORM provides reasonably accurate estimates as the distance $\|\beta\|$ goes up, i.e. with decreasing failure probabilities, and is often able to also handle non-linear limit state functions and non-Gaussian random parameters.

Finding the MPP usually requires a substantial number of evaluations of the local gradient of the limit state function (6.12). If the parameter space Ω is high-dimensional, then the number of evaluations of the performance function $f(\omega)$ can become impractical. The number of required evaluations of the performance function rises even more when the local curvature of the failure hyper-surface, i.e. the border between the safe and failure domain, at the MPP needs to be computed additionally in the case of SORM.

FORM and SORM are well known and widely applied methods and an abundance of open literature is available. Hurtado [127], Zhao & Ono [128, 129] or Kiureghian & Dakessian [130] are examples.

A.5 Basic Monte Carlo Simulation with Surrogate Modelling

Evaluation of the performance function $f(\omega)$ can be computationally expensive, especially if it requires running complex simulations. BMC simulation usually requires a large number of evaluations of the performance function in order to collect a large enough sample size. The computationally expensive performance function can, however, be approximated by a computationally cheap to evaluate surrogate model, e.g. a response surface, Artificial Neural Network or Support Vector Machine. The procedure is to first create a surrogate model of the performance function by means of a small number of examples computed by the real performance function. Then, the surrogate model approximates the performance of a large number of sampling points from Monte Carlo simulation. Figure A.7 and Figure A.8 give a graphical overview of the procedure.

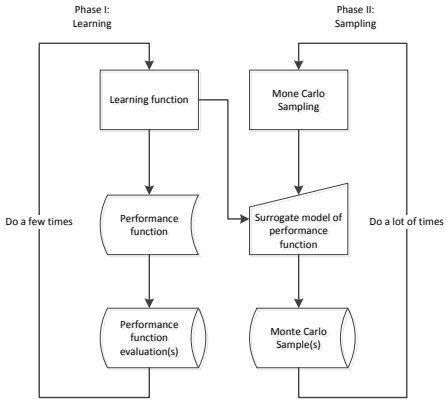


Figure A.7: Overview of surrogate modelling for BMC simulation

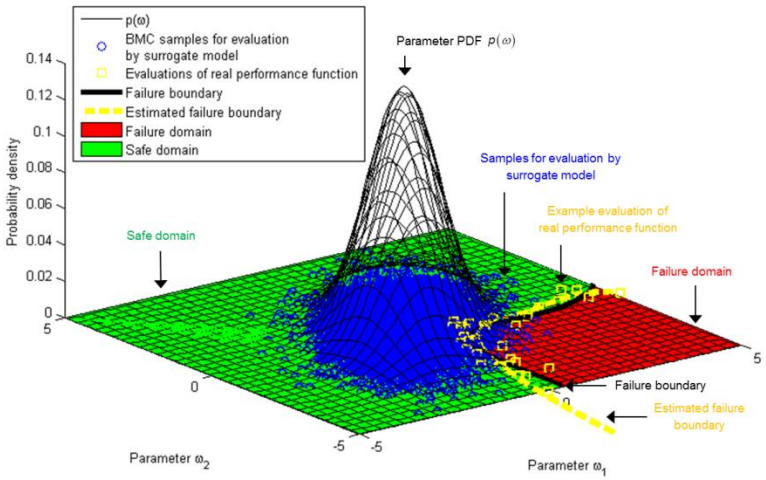


Figure A.8: Illustration of Basic Monte Carlo reliability estimation with Surrogate Modelling.

Hurtado [127] introduced the use of a Support Vector Machine (SVM) (see for example Hastie *et.al.* [106]) to distinguish between BMC samples that lie in the failure domain or in the safe domain. Hurtado [127] provides an in-depth review of surrogate modelling for reliability analysis. Open literature provides many application examples where surrogate modelling has been used for reliability analysis.

However, the use of surrogate models may involve several complications:

- The accuracy of a BMC estimator while using a surrogate model depends on how accurately the surrogate model approximates the real performance function. To establish an accurate surrogate model, it is necessary to generate a proper set of examples with evaluations of the performance function. I.e. if an SVM is used to model the boundary between the safe and failure domain, then it is necessary to generate example evaluations of the real performance function around the length of this boundary.
- Properly generating these examples may require expert- or a-priori knowledge of the reliability problem at hand and may be challenging if the failure surface is complex. Additionally, with increasing dimension of the parameter space, the number of examples necessary to generate an accurate surrogate model increases exponentially.
- Often, the difficulty of generating appropriate examples increases as reliabilities that need to be estimated become larger.
- Strictly, it cannot be assumed that the surrogate model is a perfect approximation of the performance function. When computing confidence intervals for an estimated reliability value, imprecision due to the use of a non-perfect surrogate model may thus need to be taken into account. For example, a confidence interval can be estimated from the distribution of BMC estimates for different alternative surrogate models that are likely, given the (few) available evaluations of the true performance function.

Surrogate modelling may thus be unattractive for high dimensional reliability problems. Application of methods to reduce the problem's dimension may, therefore, be considered. Mapping of the parameter space to a lower dimensional space by using data compression techniques, e.g. Principle Component Analysis, may be used to reduce the dimension of the reliability problem. Sensitivity analysis of the parameters may also reduce the dimension of the parameter space by removing parameters whose distribution does not significantly influence the performance function.

A.6 Subset Simulation

Subset Simulation can be considered as an established methodology which has been applied to a considerable amount of engineering problems. Subset Simulation was developed by Au & Beck [112, 54] and a set of validation results was presented by Au *et.al.* in [113].

The core concept of Subset Simulation (SS) is to divide a difficult problem of estimating a total probability of failure into multiple sub-problems that are by themselves easy to solve. Considering the BMC estimator (6.9), then (6.10) shows that estimating a 1/10 probability of failure is a relatively easy problem to solve, which can be done with a reasonable accuracy with 'only' one hundred evaluations of the performance function, independent of the dimension of the parameter space. Subset Sampling exploits this benefit by estimating the total probability of failure by multiplication of a sequence of conditional failure probabilities.

A set of intermediate failure events can be defined such that:

$$F_1 \supset F_2 \supset \ldots \supset F_m = F \tag{6.16}$$

This means that the failure event $F := Y > C$ (i.e. random 'critical parameter' Y exceeds the 'capacity' C) is a subset of the more probable intermediate failure event $F_{m-1} := Y > C_{m-1}$, which is, in turn, a subset of the even more probable intermediate failure event $F_{m-2} := Y > C_{m-2}$, and so forth.

The total probability of failure can now be computed as:

$$P_{fail} = P_{fail,1} \cdot \prod_{j=2}^{m} P_{fail,j}\Big|_{F_{j-1}} \tag{6.17}$$

Here, $P_{fail,1}$ is the probability of the first intermediate failure event F_1. And $P_{fail,j}\Big|_{F_{j-1}}$ is the probability of failure event F_j, given that the more probable failure event F_{j-1} occurs.

Figure A.10 illustrates the Subset Simulation process.

Computation of $P_{fail,1}$ can be done straightforwardly by a BMC estimator, especially when the first intermediate failure event F_1 is set such that $P_{fail,1}$ equals an easy to compute probability γ, i.e. 1/10. Now, a limited number of samples are drawn, e.g. one hundred, and the random 'capacity' Y is computed for each of these samples. The intermediate failure event F_1 is then defined such that $P(Y_1 > C_1) = \gamma$. I.e. the first intermediate limit state, or intermediate failure boundary (usually a hyper-surface), is set such that ten out of one hundred of the initial samples lie in the first intermediate failure domain.

A similar procedure can be followed for the subsequent (intermediate) failure events. Again making use of a simple BMC estimator, it is now necessary to generate samples that are part of the intermediate failure domain F_{j-1}. Generation of a random sample that is conditional on the domain F_{j-1} can be done with Metropolis-Hastings Markov Chain Monte Carlo Simulation (MH-MCMCS), as detailed by algorithm A-1.

New intermediate sets of samples Y_j are then added until the samples in Y_j have reached the capacity C for which the probability of failure needs to be known, such that the last intermediate probability of failure can be computed as $P_{fail,m} = P(Y_m > C \mid Y_m \in F_{m-1})$.

A-1: Metropolis-Hastings Markov Chain Monte Carlo Simulation for Subset Simulation, Au [112].

For $i = 1,2,3...l_{chain}$ do:

1) Generate a new sample 'candidate' $\tilde{\omega}_{i+1}$:

   a. Draw a sample 'pre-candidate' $\tilde{\tilde{\omega}}_{i+1}$ from the proposal PDF $\tilde{p}\left(\tilde{\tilde{\omega}}_{i+1} \mid \omega_{i|F_{j-1}}\right)$ (the mean of the proposal distribution is the current sample $\omega_{i|F_{j-1}}$)

   b. Compute acceptance ratio of 'pre-candidate' $\tilde{\tilde{\omega}}_{i+1}$:

$$r_{i+1} = \frac{p\left(\tilde{\tilde{\omega}}_{i+1}\right) \cdot \tilde{p}\left(\omega_{i|F_{j-1}} \mid \tilde{\tilde{\omega}}_{i+1}\right)}{p\left(\omega_{i|F_{j-1}}\right) \cdot \tilde{p}\left(\tilde{\tilde{\omega}}_{i+1} \mid \omega_{i|F_{j-1}}\right)} \tag{6.18}$$

   c. With probability $\min\{1, r_{i+1}\}$:

   - Accept 'pre-candidate' $\tilde{\tilde{\omega}}_{i+1}$ as 'candidate' $\tilde{\omega}_{i+1}$ (i.e. $\tilde{\omega}_{i+1} = \tilde{\tilde{\omega}}_{i+1}$).

   With remaining probability $1 - \min\{1, r_{i+1}\}$:

   - Set current sample $\omega_{i|F_{j-1}}$ as 'candidate' $\tilde{\omega}_{i+1}$ (i.e. $\tilde{\omega}_{i+1} = \omega_{i|F_{j-1}}$)

2) Accept sample 'candidate' $\tilde{\omega}_{i+1}$ according to the presence in failure domain F_{j-1}:

- If $\tilde{\omega}_{i+1} = \omega_{i|F_{j-1}}$
  - then $\omega_{i+1|F_{j-1}} = \omega_{i|F_{j-1}}$ (i.e. set current sample as next sample if new 'pre-candidate' was rejected before)
- Else, if $\tilde{\omega}_{i+1} \in F_{j-1}$
  - then $\omega_{i+1|F_{j-1}} = \tilde{\omega}_{i+1}$ (i.e. set sample 'candidate' as a new sample if 'candidate' lies in the intermediate failure domain)
- Otherwise $\omega_{i+1|F_{j-1}} = \omega_{i|F_{j-1}}$ (i.e. set current sample as new sample if 'candidate' is not conditional on the intermediate failure domain)

It can be shown that the acceptance ratio (6.18) goes to zero as the dimension of the parameter space Ω becomes large. This limits the applicability of MH-MCMCS to problems with low dimensions. If (groups of) parameters are however independent from each other (e.g. by NATAF transformation), then these can also be sampled independently from each other. This can keep the effective dimension in which MH-MCMCS is applied low. For example, if a problem consists of many independent parameters, then steps 1) a-c in algorithm A-1 can be applied to every dimension in Ω individually. No use has to be made of a ten-dimensional sampling and proposal distribution. Instead, every dimension can be sampled individually by a one-dimensional sampling and proposal distribution. This variant of MH-MCMCS is called Modified Metropolis-Hastings Markov Chain Monte Carlo Simulation (MMH-MCMCS).

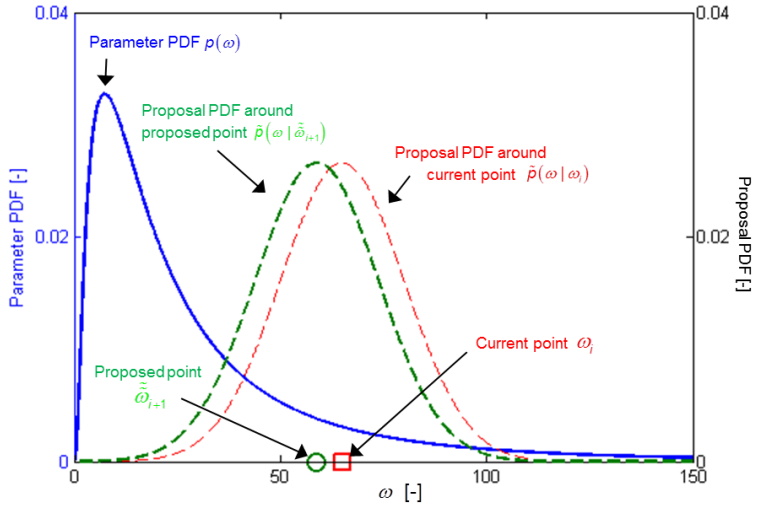


Figure A.9: One-dimensional illustration of the variables in (6.18), for a lognormal parameter distribution and normal proposal distribution.

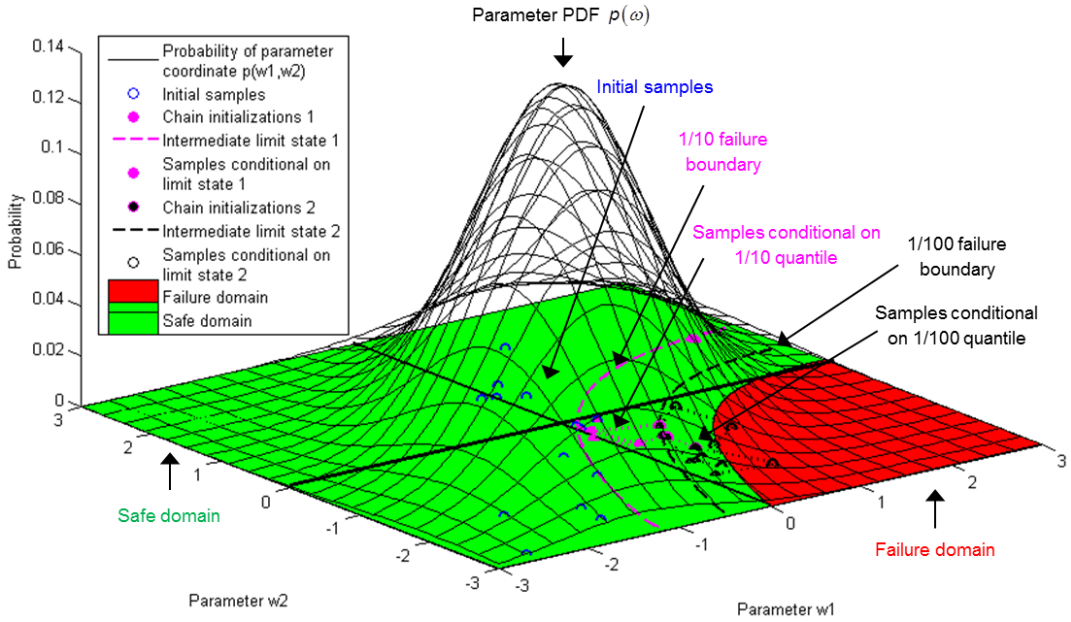


Figure A.10: Example application of Subset Simulation (SS). A top-down view and additional explanation are provided with Figure A.11:

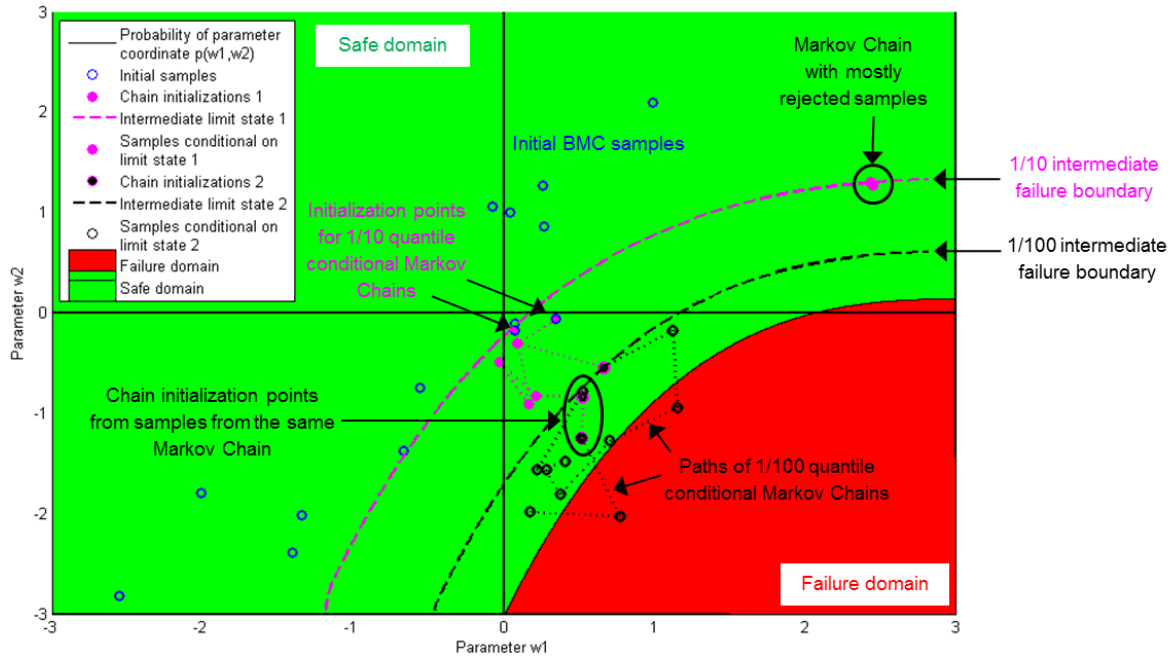


Figure A.11: Example application of Subset Simulation (SS) where a 1/750 probability of failure is estimated. The figure is a top-down view of the example in Figure A.10. The following list is a step-by-step explanation of the illustrated Subset Simulation:

- The first step of SS consists of drawing a small sample from the parameter PDF. These samples are shown with thick blue open circles. The boundary between the 10% samples closest to the failure domain and the rest of the initial samples is shown by a thick dashed purple line. The initial samples beyond this boundary are shown as a thick blue circle filled with purple. These points are the initiation points of the Markov chains in the second step.

- The Markov chains in the second step generate new samples that are conditional on being part of the 10% 'worst' population samples (in terms of failure). These new samples are shown by filled purple circles and the path of the Markov chains is shown by thin dotted black lines. The Markov chain shown in the upper right corner, which seems to consist of only two points, is an example where all new samples are rejected and where the next point in the Markov chain always equals the current point. Once all samples of the second stage are generated, the 10% 'worst' samples, out of the second stage samples, are selected. These are displayed by pink circles filled with black and are the starting points of the Markov chains in the third stage that generate samples conditional on being part of the 1% 'worst' population samples. The boundary indicating the 10% x 10% = 1% 'worst' sample boundary is displayed by a thick dashed black line.

- The samples generated in the third stage are shown as thick open black circles. It can be seen that two of the samples generated in the third stage are within the failure domain.

- The total probability of failure would now be estimated as: 1/10 x 1/10 x 2/15 = 1/750. The failure probability that Subset Simulation estimates is generally not perfect, i.e. due to the use of a limited number of samples per subset. It is possible to compute a confidence interval on the estimated probability of failure.

Before reaching the critical capacity C, intermediate quantiles C_j are set such that they correspond to the γ^j quantile of the distribution of the critical parameter Y, i.e. $P\left(Y_j > C_j \,|\, Y_j \in F_{j-1}\right) = \gamma$. However, these quantiles are the quantiles given the available subset sample Y_j and only approximate the true quantiles of Y. This means that the conditional probability of failure $P\left(Y_j > C_j \,|\, Y_j \in F_{j-1}\right)$ is truly only an approximation of γ.

The coefficient of variation δ of an estimated probability of failure $P_{fail} = P\left(Y > C\right)$ can, according to Au [112], be estimated by the squared sum of the coefficients of variation δ_j of the estimated intermediate conditional failure probabilities $P_{fail,j} = P\left(Y_j > C_j \,|\, Y_j \in F_{j-1}\right)$:

$$\left(\frac{\sigma_{P_{fail}}}{P_{fail}}\right)^2 = \quad \delta^2 = \sum_{j=1}^{m} \delta_j^2 \tag{6.19}$$

where $\sigma_{P_{fail}}$ denotes the standard deviation of the estimated P_{fail}. Equation (6.19) is valid under the assumption that there is no dependence between the accuracy of quantile estimates C_j. This assumption is not generally valid since samples from the $j+1^{th}$ subset are generated by Markov Chains that start from samples of the j^{th} subset that lie in the failure domain F_j. This causes some dependence between the samples of different subsets. However, Au [112] demonstrates that this may be neglected and that (6.19) is nevertheless an accurate approximation.

The coefficient of variation of the estimated intermediate conditional probability of failure $P_{fail,j} = P\left(Y_j > C_j \,|\, Y_j \in F_{j-1}\right)$ can be estimated following Au [112]:

$$\delta_j = \sqrt{\frac{1 - P_{fail,j}}{P_{fail,j} \cdot N_{samp}} \cdot \left(1 + \kappa_j\right)} \tag{6.20}$$

where N_{samp} denotes the number of samples in the in the subset sample Y_j. Note that this estimator is similar to the estimator (6.10) of the coefficient of variation of the BMC estimator with added 'efficiency factor' κ.

It is inherent to Markov Chains that samples within each chain are correlated. This reduces the number of effective statistically independent samples in the chain. This is accounted for by means of the 'efficiency factor' κ. In the case of Subset Simulation, Markov Chains that generate the j^{th} subset Y_j may originate from samples that have been generated by same Markov Chain in the $j-1^{th}$ subset simulation. This means that samples from Markov Chains in the j^{th} subset simulation may not be independent. Nevertheless, it is assumed that this dependence may be neglected and that samples from different chains in the j^{th} subset are independent through the binary indicator function:

$$\mathrm{I}_j\left(\omega\right) = \begin{cases} 1 \text{ if } \omega \in F_{j-1} \\ 0 \text{ otherwise} \end{cases} \tag{6.21}$$

Also under the assumption that the number of samples l_{chain} in each of the n_{chain} Markov Chains (i.e. one for each sample in the previous subset sample Y_{j-1} that is conditional on the previous intermediate failure event F_{j-1}) is the same, i.e. $N_{samp} = l_{chain} \cdot n_{chain}$, according to Au [112] the efficiency factor κ can be computed by:

$$\kappa_j = 2 \cdot \sum_{k}^{l_{chain}-1} \left(1 - \frac{k \cdot n_{chain}}{N_{samp}}\right) \cdot \frac{R_j\left(k\right)}{R_j\left(0\right)} \tag{6.22}$$

Where $R_j\left(0\right)$ can be estimated by:

$$R_J\left(0\right) = P_{fail,j}\left(1 - P_{fail,j}\right) \tag{6.23}$$

and $R_j\left(k\right)$ can be computed as:

$$R_j\left(k\right) = \left(\frac{1}{N_{samp} - k \cdot n_{chain}} \cdot \sum_{q=1}^{n_{chain}} \sum_{r=1}^{l_{chain}-1} \mathrm{I}_j\left(\omega_{q,r}\right) \cdot \mathrm{I}_j\left(\omega_{q,r+k}\right)\right) - P_{fail,j} \tag{6.24}$$

Au [112] found that setting the intermediate failure events F_j such that 10% of the samples in the j^{th} subset sample lie in the failure domain F_j is a reasonable design choice (i.e. $\gamma = 0.1$). If γ is set too low, then it requires many samples to accurately estimate the intermediate quantiles C_j and Subset Simulation loses its efficiency gain with respect to BMC simulation. If γ is set too large (i.e. approaching one from below) then it will require too many intermediate quantiles C_j to reach the quantile C, for which reliability needs to be known, and Subset Sampling becomes inefficient as well. As long as the subset sample size N_{samp} and the intermediate probability of failure γ are set at reasonable values, then it suffices to compute confidence intervals by means of (6.20) in order to ensure the validity of simulated reliabilities.

The choice of the proposal distribution $\tilde{p}(\omega)$ influences the efficiency of the Subset Simulation estimator through the 'efficiency factor' κ. Ideally, the proposal distribution equals the local conditional PDF, i.e. $\tilde{p}_j(\omega) = p_j(\omega \mid \omega \in F_{j-1})$. If the proposal distribution is chosen improperly, then many new sample candidates in the Markov Chains will be rejected. If this happens then the samples in the Markov Chains will be highly correlated and the 'efficiency factor' κ will be high as well, causing the Subset Simulation estimator to be inefficient.

In summary, Subset Simulation by Au & Beck [112, 54] provides a reliability estimation method that features distinct advantages:

- Suitable for simulating very low probabilities of failure (i.e. $<10^{-4}$)
- Capable of handling problems of high dimension (provided that the full parameter set can be subdivided into small sets of correlated parameters that are uncorrelated with the other parameter subsets)
- The possibility to estimate a confidence interval ensures the validity of a simulated probability of failure
- There are no crucial 'tuning' parameters that need to be set (i.e. by expert judgement) to ensure the validity of a simulated probability of failure

A.7 Other methods

Numerical reliability modelling is an active field of research. Many other alternatives including also more recent methods than already discussed in Appendix A are thus available to estimate probabilities of failure at minimal computational costs. Many of these combine Subset Simulation with surrogate modelling, or apply concepts similar to Subset Simulation and/or Importance Sampling, often combined with Kriging surrogate modelling. Recent examples of such work include [131, 52, 132, 53, 133].

For most of these recent methods the following considerations apply:

- Surrogate Modelling is difficult to apply for problems in high dimensions (as discussed in Appendix A.5)
- Often, no suitable method to establish confidence or tolerance intervals is established.
- The rate at which new reliability modelling methods are developed currently seems extraordinarily high. It may, therefore, be inefficient to invest in complex methodologies until their development and validation has matured and a clearly preferred methodology can be identified

Appendix B. Details of methods to estimate tolerance intervals

Several frequentist (i.e. non-Bayesian) methodologies to estimate a tolerance interval of a lognormal distribution are summarized in this chapter. The methods are benchmarked in section 3.3. The last section of this appendix details the likelihood function for a lognormal distributed quantity, which has special importance to the application of Bayesian statistics in sections 3.4 and 3.5.

B.1 Approximate analytical (Wald & Wolfowitz)

Based on Wald & Wolfowitz [62, 63], the lower single-sided tolerance interval of a lognormal distributed quantity *S* can be approximated analytically by the following relation:

$$S_{\gamma,\chi} = \exp\left\{\hat{\mu}_s - \sqrt{\frac{n-1}{inv\chi^2\left(P=1-\chi \mid \nu=n-1\right)}}\cdot r_\gamma \cdot \hat{\sigma}_s\right\} \quad \text{with} \quad r_\gamma = \frac{1}{\sqrt{n}} - \mathrm{N}^{-1}\left(P=\gamma \mid \mu=0,\sigma=1\right) \tag{7.1}$$

Where:

- $\hat{\mu}$ and $\hat{\sigma}$ are the sample estimates of the mean and standard deviation of the associated normal distribution respectively
- *n* designates the sample size
- γ denotes the required quantile (e.g. 10^{-3} when the one-in-a-thousand lower quantile is desired)
- χ denotes the required lower quantile for the lower single-sided confidence interval (e.g. 0.95 for a 95% lower single sided confidence level)
- $inv\chi^2$ designates the inverse cumulative distribution function of the chi-square distribution with ν degrees of freedom
- N^{-1} stands for the inverse cumulative distribution function of the normal distribution:

$$\mathrm{N}^{-1}\left(x \mid \mu,\sigma\right) = \frac{1}{\sigma\cdot\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left\{\frac{-\left(y-\mu\right)^2}{2\sigma^2}\right\}dy \tag{7.2}$$

B.2 Approximate analytical (ESDU 91041)

Alternatively, ESDU 91041 [64] also provides an analytical approximation to the tolerance interval of a lognormal distributed quantity *S* for $n \geq 4$:

$$S_{\gamma,\chi} \cong \exp\left\{\frac{\hat{\mu} + \mathrm{N}^{-1}\left(\gamma\right) - \frac{\mathrm{N}^{-1}\left(\chi\right)}{\sqrt{n}}\sqrt{1 - \frac{\mathrm{N}^{-1}\left(\chi\right)^2}{2\left(n-1\right)} + n\frac{\mathrm{N}^{-1}\left(\gamma\right)^2}{2\left(n-1\right)}}}{1 - \frac{\mathrm{N}^{-1}\left(\chi\right)^2}{2\left(n-1\right)}}\cdot\hat{\sigma}\right\} \tag{7.3}$$

Where $\mathrm{N}^{-1}\left(z\right) = \mathrm{N}^{-1}\left(P=z \mid \mu=0,\sigma=1\right)$ abbreviates the inverse cumulative distribution function (7.2) of the standard normal distribution.

B.3 Approximate analytical (AGARD-AG-292)

In chapter 4.1 of the NATO AGARD-AG-292 Helicopter Fatigue Design Guide [16] the following tolerance estimator for lognormal distributed fatigue strength is proposed:

$$S_{\gamma,\chi} = \hat{\mu}\cdot\exp\left\{\hat{\sigma}\cdot\mathrm{N}^{-1}\left(\gamma \mid \mu=0,\sigma=1\right)\right\}\cdot\exp\left\{\hat{\sigma}\cdot\frac{\mathrm{N}^{-1}\left(1-\chi \mid \mu=0,\sigma=1\right)}{\sqrt{n}}\right\} \tag{7.4}$$

Note that this tolerance interval estimator implicitly assumes that the sample estimate of the population standard deviation is a perfect, or at least conservative, estimate, i.e. $\hat{\sigma} \geq \sigma$.

B.4 Observed likelihood

Confidence intervals for the quantile of a distribution may also be estimated by means of the observed Fisher's information and the assumption of normally distributed likelihood. Fisher's information is the Hessian of the likelihood function evaluated at the Maximum Likelihood Estimate (MLE) of the desired distribution quantile. Efron & Hinkley [65] provide more details and references on this method. In this paper, the default MATLAB implementation is used [49].

B.5 Likelihood profile

A single sided lower tolerance interval may also be estimated according to the likelihood profile of the inverse cumulative distribution and under the assumption that the likelihood profile follows a chi-square distribution (see also Meeker & Escobar [66]):

$$\underset{\mu,\sigma}{\operatorname{argmin}}\ F_{LN}^{-1}(\gamma \mid \mu,\sigma) \quad \text{such that} \quad \lambda\left(\bar{S} \mid \mu,\sigma\right) - \left[\lambda\left(\bar{S} \mid \hat{\mu},\hat{\sigma}\right) - \frac{1}{2} inv\chi^2\left(2 \cdot \chi - 1 \mid \nu = 1\right)\right] = 0 \tag{7.5}$$

Where:

- F_{LN}^{-1} denotes the lognormal inverse cumulative distribution, i.e. $F^{-1}(\gamma \mid z)$ specifies the γ -quantile of the distribution set by parameter *z*
- $\lambda\left(\bar{S} \mid \mu,\sigma\right)$ is the negative log-likelihood of the sample $\bar{S}$ given the distribution parameters μ and σ of a lognormal distribution. The actual likelihood function *L* is defined later on by equation (7.10) in section B.8.
- $\hat{\mu}$ and $\hat{\sigma}$ designate the Maximum Likelihood Estimates (MLEs) of the parameters of the lognormal distribution, given sample $\bar{S}$.

Advantageously, this method can estimate confidence intervals while accounting for the presence of right-censored data, i.e. run-outs. This is not readily the case for all other methods in Appendix B.

B.6 Parametric bootstrapping

If the uncertainty distributions of the parameters of the estimated population parameters are known, then the influence of estimation uncertainty on the desired statistic can be simulated by Monte-Carlo simulation. For the (associated) normal distribution, the uncertainty distribution of the population distribution parameters μ and σ, given a sample at hand, can be derived analytically (see Meeker & Escobar [66] and Moore, *et.al.* [134]):

$$p(\mu \mid \hat{\mu},\hat{\sigma},n) = t\left(\mu = \hat{\mu},\sigma = \frac{\hat{\sigma}}{\sqrt{n}},\nu = n-1\right) \tag{7.6}$$

$$p(\sigma \mid \hat{\sigma},n) \propto \hat{\sigma} \cdot \sqrt{\frac{n-1}{\chi^2(\nu = n-1)}} \tag{7.7}$$

Where $t(\ldots)$ denotes the student t-distribution.

Note that the associated normal distribution is symmetric and that the distributions (7.6) and (7.7) are thus independent, as follows from Shanmugam [135].

The uncertainty distribution of the quantile S_γ can be simulated by computing S_γ for each element in a (large) sample from the joint distribution parameter uncertainty distribution that is formed by equations (7.6) and (7.7). Finally, the required tolerance interval $S_{\gamma,\chi}$ can be approximated by a quantile estimate of the simulated lognormal quantile uncertainty distribution:

$$S_{\gamma,\chi} = F^{-1}\left\{\chi \mid \bar{S}_\gamma\right\} \quad \text{with} \quad \bar{S}_\gamma = \left[F_{LN}^{-1}\left(\gamma \mid \bar{\mu}_1, \bar{\sigma}_1\right) \ldots F_{LN}^{-1}\left(\gamma \mid \bar{\mu}_k, \bar{\sigma}_k\right)\right] \quad \text{as} \quad k \to \infty \tag{7.8}$$

Where:

- F^{-1} denotes the inverse cumulative distribution of the distribution of the parameter S_γ (e.g. an inverse empirical cumulative distribution function)
- $\bar{S}_\gamma$ is a sample of size *k* of a γ-quantile of a lognormal distribution, given sampled parameters $\bar{\mu}_k$ and $\bar{\sigma}_k$; where $\bar{\mu}_k$ and $\bar{\sigma}_k$ are sampled from (7.6) and (7.7) respectively

B.6.1 Application example

Figure B.1 provides a process overview for tolerance interval estimation by parametric bootstrapping. Next, Figure B.2 shows an exemplary case to illustrate the process. The true PDF is estimated here by a sample of size six from the population; the individual sample instances are not shown. In this case, the sample does not provide a good representation of the population and the estimate of the most likely 10^{-3} quantile is too high. If some critical design parameter would be set according to this estimated quantile, then its true reliability would be much lower than 0.999. To mitigate this possibility, it can be considered that given the six available samples, a range of alternative PDFs, other than the Maximum Likelihood Estimate (MLE), are actually likely too.

The probability of such alternatives can be mathematically specified, e.g. as in Equations (7.6) and (7.7). Their corresponding alternative estimates of the 10^{-3} quantile, are distributed and taking the 95^{th} percentile gives the 95% single sided lower confidence interval for the 10^{-3} population quantile. In the example in Figure B.2 however, even the quantile at 95% confidence does not satisfy a 0.999 reliability requirement. This can occur, as a 95% confidence level only indicates that, on average, 95 out of 100 (hypothetical) samples yield a correct or conservative estimate of the true population quantile. In this particular example, the sample at hand thus 'unfortunately' misrepresents the true population quite severely and is one of the remaining (hypothetical) five samples. In reality, this would be completely unknown though.

Figure B.3 shows an example similar to the one in Figure B.2, but now with a sample of size twenty at hand. Clearly, the effect of imprecision due to the limited number of available samples is reduced significantly and the confidence intervals are narrower.

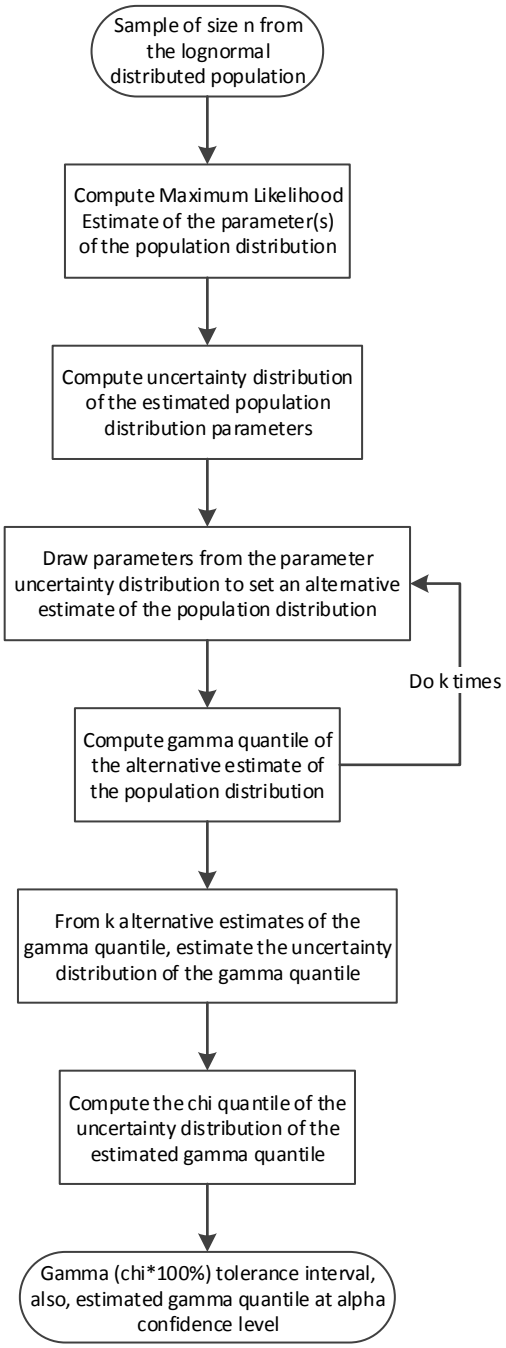


Figure B.1: Parametric bootstrapping

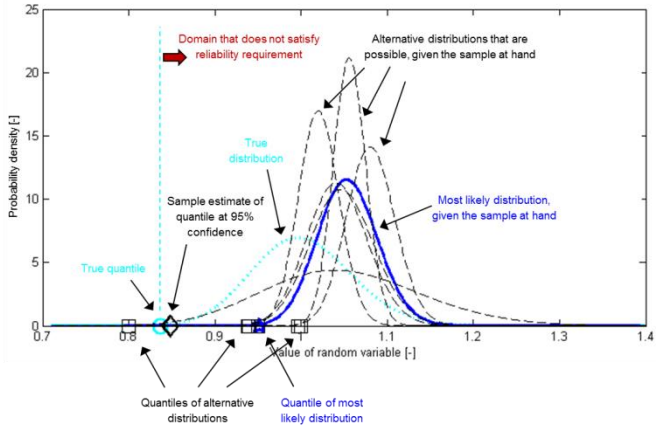


Figure B.2: Tolerance interval estimation by bootstrapping. The illustrated case uses bootstrapping to estimate a 10-3 quantile of a lognormal distribution by means of a sample with size six and with 95% confidence.

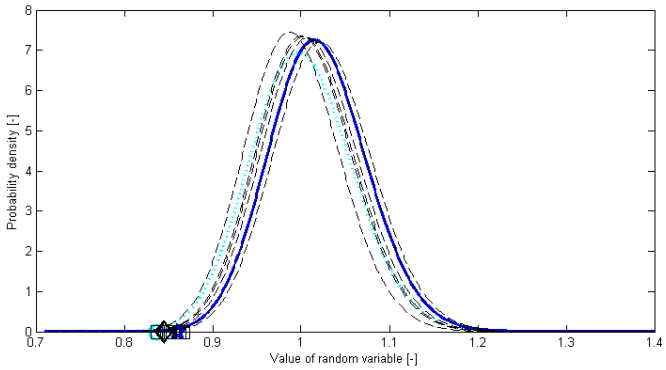


Figure B.3: Uncertainty in distribution estimate under 'medium' sample size conditions**.** Another illustrative case for using bootstrapping to estimate a 10^{-3} quantile but now with a sample of size twenty at hand (Legend as in Figure B.2)

B.7 Non-parametric bootstrapping

Bootstrap confidence intervals can also be obtained without prior knowledge of the joint uncertainty distribution of the population distribution parameters by repeated resampling of the sample at hand, $\overline{S}$ (see DiCiccio & Efron [67] and Hesterberg *et.al*. [55]):

$$S_{\gamma,\chi} = F^{-1}\left\{\chi \mid \overline{S}_\gamma\right\} \quad \text{with} \quad \overline{S}_\gamma = \left[F_{LN}^{-1}\left\{\gamma \mid \bar{\mu}\left(\overline{S}_1\right),\bar{\sigma}\left(\overline{S}_1\right)\right\} \ldots F_{LN}^{-1}\left\{\gamma \mid \bar{\mu}\left(\overline{S}_k\right),\hat{\bar{\sigma}}\left(\overline{S}_k\right)\right\}\right] \quad \text{as} \quad k \to \infty \tag{7.9}$$

Where:

- $\overline{S}_k$ is the kth bootstrap sample of the sample $\overline{S}$. A bootstrap sample $\overline{S}_k$ is a random combination, with replacement, of the original sample $\overline{S}$ with the same size as the original sample
- $\hat{\bar{\mu}}$ and $\hat{\bar{\sigma}}$ are the MLE of the lognormal distribution parameters, given bootstrap sample $\overline{S}_k$

B.8 Lognormal distribution fitting by Maximum Likelihood Estimation

The parameters of the associated lognormal distribution can be estimated by Maximum Likelihood Estimation. The distribution parameters $\{\hat{\mu},\hat{\sigma}\}$ are then set such that they maximize the likelihood function of the normal distribution *L* (Meeker & Escobar [66]):

$$\hat{\sigma},\hat{\mu} = \underset{\tilde{\mu},\tilde{\sigma}\in\square}{\operatorname{argmax}}\ L(\tilde{\mu},\tilde{\sigma}) \quad \text{with} \quad L(\mu,\sigma) = \prod_{i=1}^{n}\left\{\frac{1}{\sigma\cdot S_i}\cdot \mathrm{N}\left(\frac{\log_e(S_i)-\mu}{\sigma}\right),1\right\}^{\delta_i}\cdot\left\{1-\Phi\left[\frac{\log_e(S_i)-\mu}{\sigma}\right]\right\}^{1-\delta_i} \tag{7.10}$$

$$\text{with } \delta_i \begin{cases} 1 & \text{if } S_i \text{ is an exact observation} \\ 0 & \text{if } S_i \text{ is a right-censored observation} \end{cases}$$

where $\Phi(\mu)$ denotes the Cumulative Distribution Function (CDF) of the semi-standard normal distribution:

$$\Phi(\mu) = \int_{-\infty}^{\mu} \mathrm{N}(\mu,\sigma=1)dx \tag{7.11}$$

In the context of fatigue testing, a right-censored observation corresponds to a run-out, i.e. the fatigue test was halted before failure of the part was observed.

Appendix C. Application & verification of Bayesian statistical analysis

This appendix presents a range of synthetic verification exercises to verify that:

- a tolerance interval estimated while using a non-informative prior is the same as a tolerance interval estimated by a classic estimator (Appendix C.1)
- an analytical uncertainty distribution about σ in the form of equation (7.7) in Appendix B.6 can be used as a prior and that this uncertainty distribution is fit to transfer information from one test to another (Appendix C.2)
- the influence of a prior diminishes as more actual test results become available and that the influence of a biased prior expectation is small as long as the prior is relatively unspecific (Appendix C.2)
- an average of uncertainty distributions in the form of equation (3.10) in chapter 3 can be used to formulate a relatively unspecific prior based on multiple previous and comparable test results (Appendix C.3)
- the use of a prior in the form of equation (3.10) can be used to significantly and reliably increase substantiated fatigue strength under small sample size conditions (Appendix C.3)

C.1 Posterior σ distribution with non-informative prior

To perform a basic verification test for the practical implementation of Bayes' Theorem, a posterior distribution can be computed with a non-informative prior, i.e. the prior α conveys no information about the population standard deviation σ. This posterior distribution should then be equal to the analytical uncertainty distribution of σ according to equation (7.7) in Appendix B.

To set a non-informative prior on the population standard deviation σ, Box & Tiao [124] show that this can be done by the inverse function of σ:

$$p(\sigma \mid \alpha \equiv 0) \propto \frac{1}{\sigma} \tag{8.1}$$

A successful comparison between a computed Bayesian posterior and a corresponding analytical uncertainty distribution is finally demonstrated in Figure C.1.[31]

[31] Comparison with an empirical distribution estimate of $p(\hat{\sigma} \mid \sigma, n)$ by BMC simulation was also carried out successfully but is not presented here.

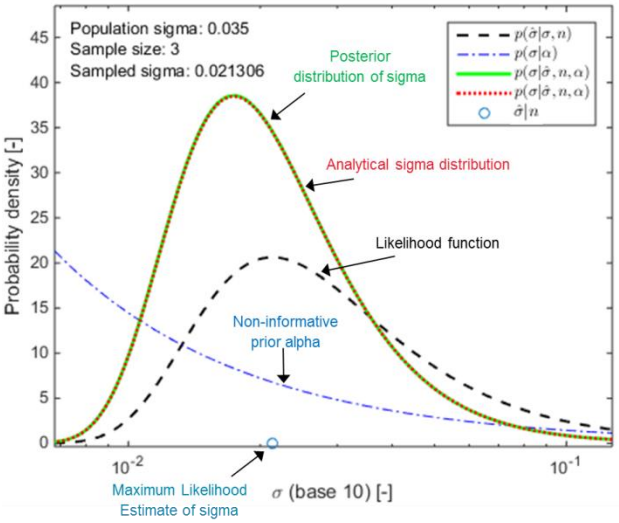


Figure C.1: Verification of the practical implementation of Bayes' Theorem. Verification by comparison between the analytical solution for $p(\sigma|\hat{\sigma},n)$ and the Bayesian posterior with a non-informative prior. (Displayed sigma values are for a base-10 lognormal distribution)

C.2 Posterior σ distribution with informative prior

Setting an appropriate informative prior generally depends on expert judgment. Any sensible function conveying appropriate information concerning the prior expectation on the posterior may be used. To support understanding of the influence of a prior on the posterior distribution, a range of verification exercises is presented here.

A basic test to verify the use of the analytical uncertainty distribution (7.7) in Appendix B to convey prior knowledge concerning a lognormal distribution's standard deviation is presented first. The test considers a hypothetical situation in which the same test is carried out twice and where an identical test result is obtained for both tests. If the uncertainty distribution on σ that follows from the first test is used as a prior during the second test, then the resulting posterior distribution on σ should be the same as when the two test samples are simply concatenated and an uncertainty distribution on σ is computed directly. This test process is presented in a more generic form in Figure C.2.

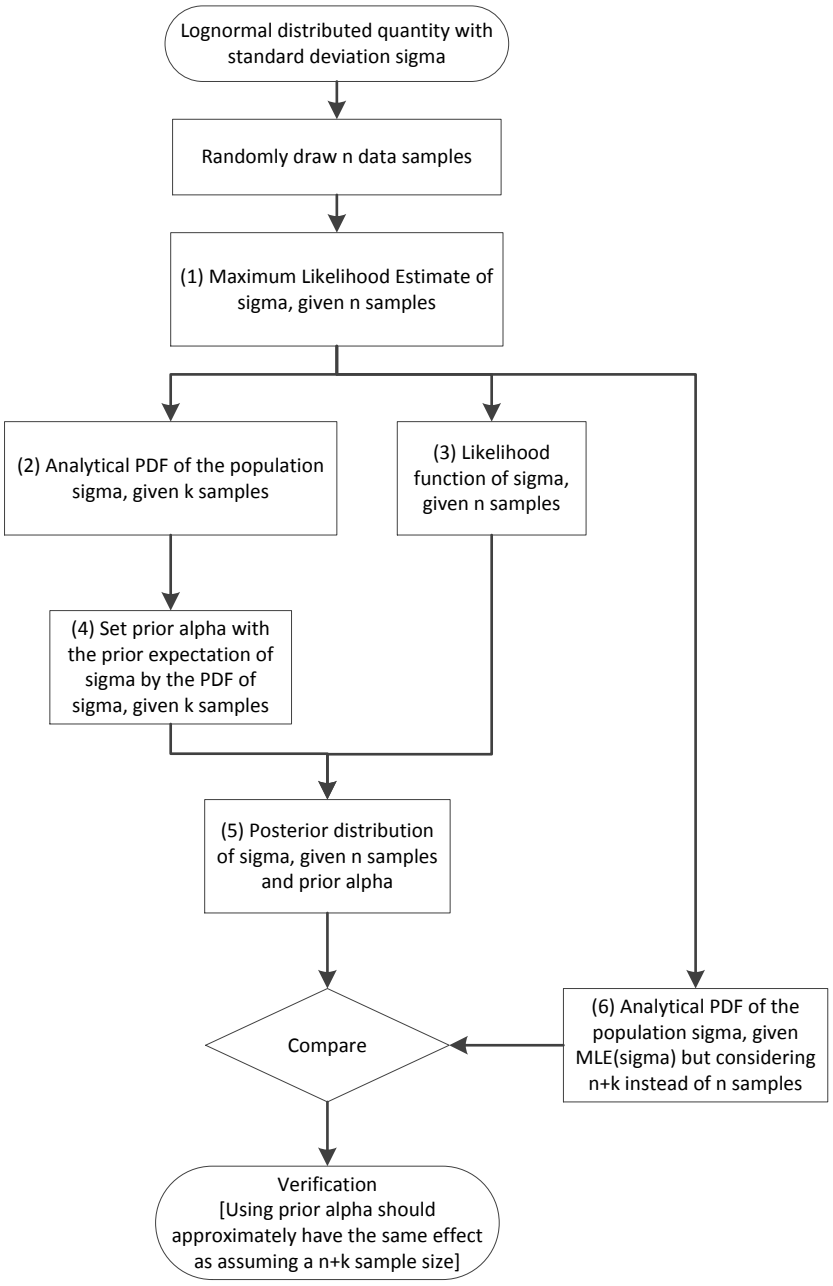


Figure C.2: Verification test for informative priors.

The test results presented in Figure C.3 to Figure C.7 all demonstrate that an uncertainty distribution according to equation (7.7) in Appendix B can indeed be used to transfer prior information. In all of these cases, a prior based on *k* samples adds almost the same amount of information as enlarging the sample size by *k* samples. The information transfer is efficient and only a slight 'information loss' is observed. Additionally, Figure C.4 illustrates that when the prior is more specific than the actual test result at hand, the prior successfully diminishes the otherwise unrealistic heavy tails. The other way around, Figure C.5 illustrates the situation where the prior is much less informative than the actual sample at hand. In that case, the prior indeed has little effect, other than slightly reinforcing the 'power' of the test at hand.

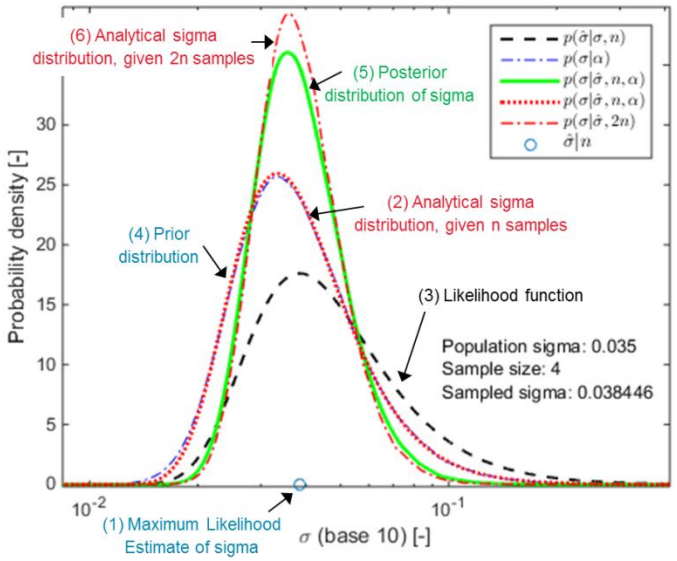


Figure C.3 Verification test result[32] for n = 4 and k = 4. (Element numbers correspond to verification test process in Figure C.2)

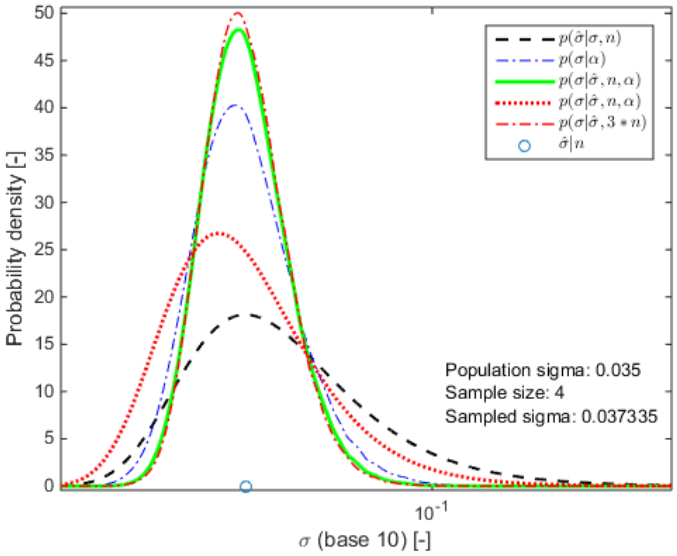


Figure C.4: Verification test result with n = 4 and k = 8.

[32] Prior distribution (4) and the analytical solution for the uncertainty distribution of sigma (2) are not exactly equal as is expected from theory. This is because the prior distribution (4) actually stems from a Gaussian kernel distribution fit through a large BMC sample whereas distribution (2) is semi-analytic. The prior actually used to compute the posterior distribution (5) is not shown and stems from semi-analytical analysis according to equation (7.7) in Appendix B. This approach was chosen to increase the scope of the verification test.

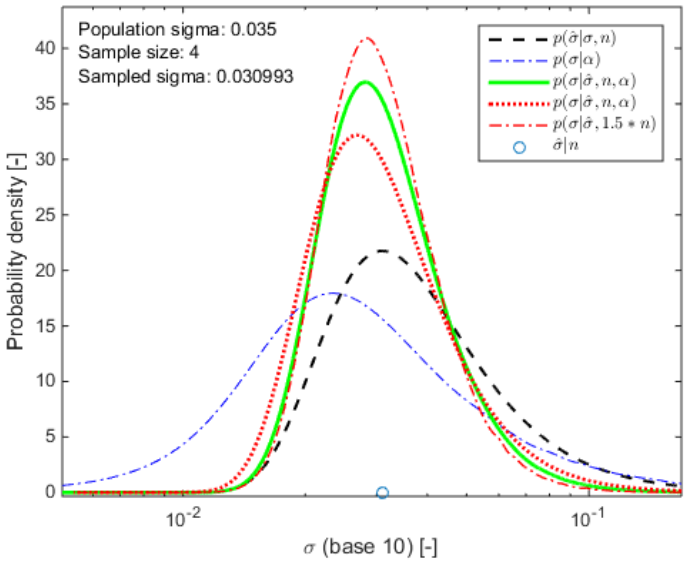


Figure C.5: Verification test result with n = 4 and k = 2.

The previous test cases all simulate that the prior conveys correct and unbiased information. In reality, this assumption may not be valid. Importantly, as the 'power' of the actual test result increases with sample size, the more it 'overrules' the prior. The amount of influence that the prior has also depends on its specificity, i.e. its 'width'. In general, it is recommended to limit the strength of the prior expectation. As an illustration, Figure C.6 demonstrates that a 'wrong' though relatively 'weak' prior has limited influence and is mostly 'overruled' by actual test results. Conversely, Figure C.7 illustrates a case where the prior is rather specific and quite off, resulting in a relatively large and erroneous expectation bias.

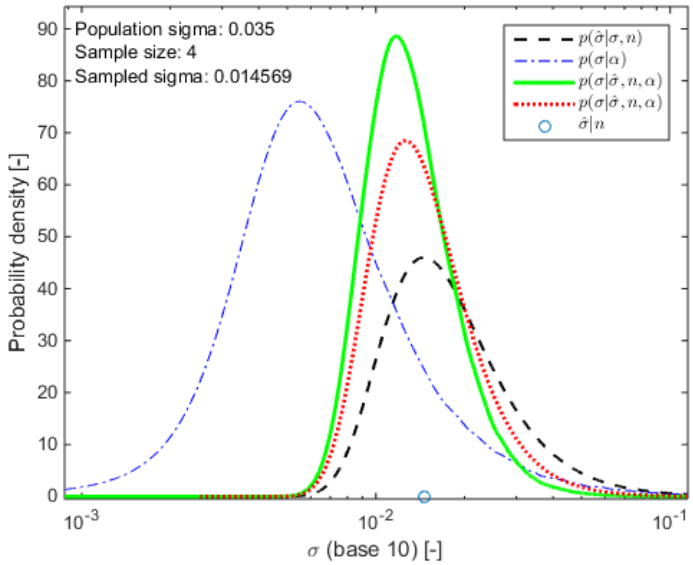


Figure C.6: Verification test result with n = 4 and k = 2. The mean μ of the prior distribution α is set to $\frac{1}{2} \cdot \sigma$.

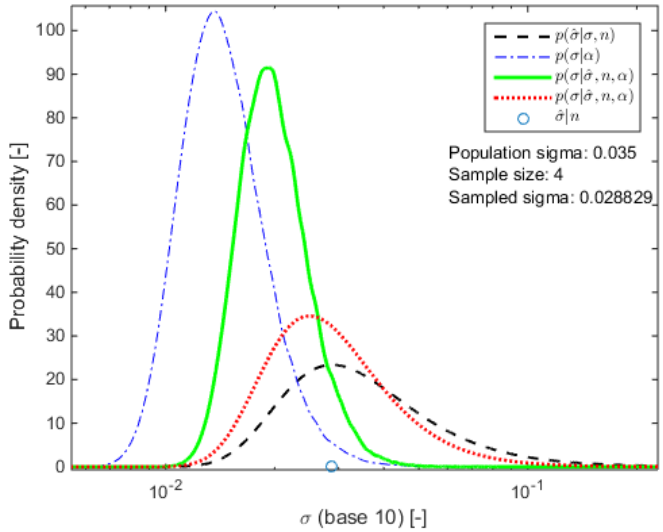


Figure C.7: Verification test result with n = 4 and k = 8. The mean μ of the prior distribution α is set to $\frac{1}{2}\cdot\sigma$.

C.3 Posterior σ distribution with uncertainty distribution averaged prior

The use of a prior in the form of equation (3.10) in chapter 3 is here studied for its appropriateness. A comprehensive overview of the test procedure that is followed to test this type of prior is presented in Figure C.8. The basic test condition comprises ten random samples, each with a random size ranging from two to twelve and drawn from the same lognormal distribution:

$$p(S) = 10^{\sigma_{10}\cdot N(0,1)+\mu_{10}} \quad \text{with} \quad \{\mu_{10},\sigma_{10}\} = \{0,0.035\} \tag{8.2}$$

This test condition reflects common situations in fatigue life prediction in industry. In practise though, each set of test data will come from a (slightly) different population. However, this idealized set-up allows to accurately study the relative efficiency of information transfer by a prior set by equation (3.10).

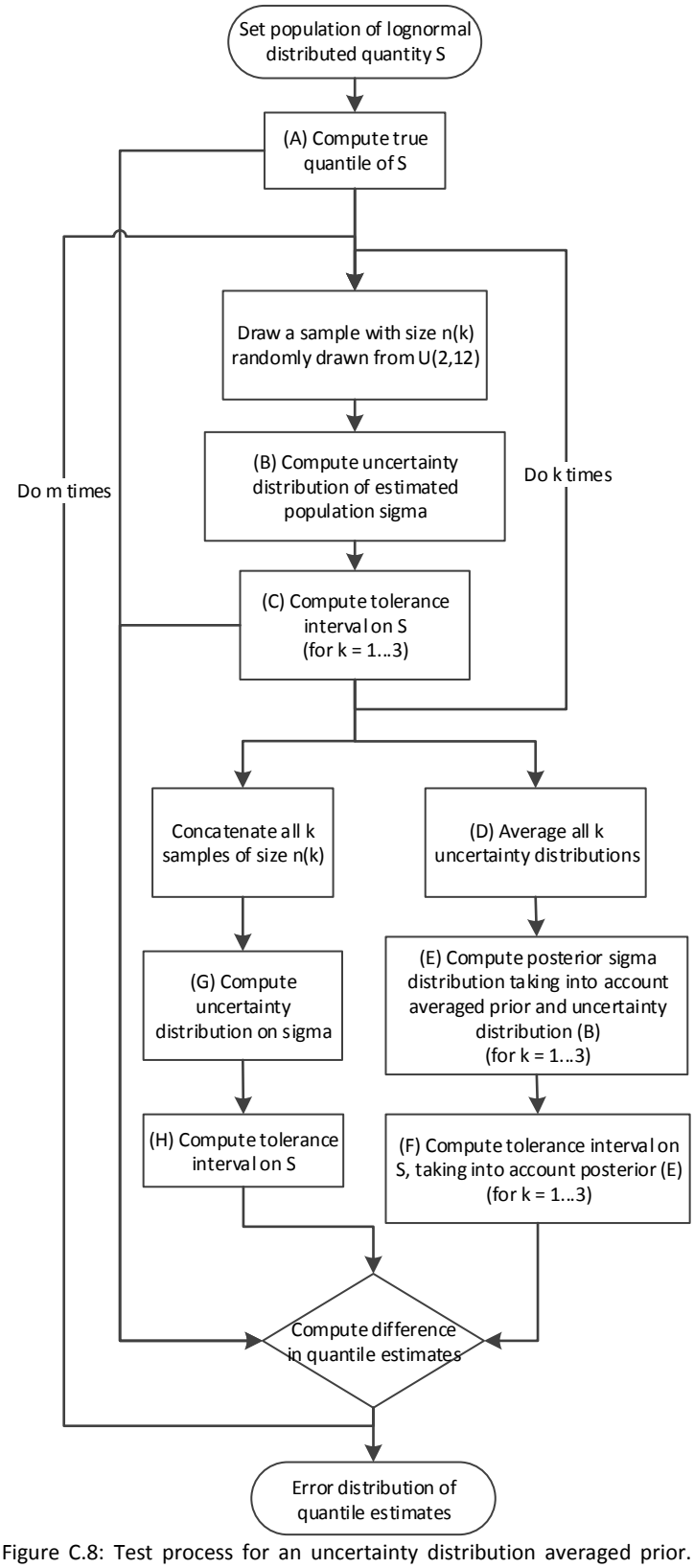


Figure C.8: Test process for an uncertainty distribution averaged prior. U(2,12) denotes a random integer drawn from the domain [2 12].

The thin dashed distributions in Figure C.9 illustrate uncertainty distributions for the population parameter σ, with one distribution for each estimation attempt. The thick blue distribution shows the average of these ten uncertainty distributions and this distribution is thus used as a prior assumption on the population standard deviation σ. Finally, for three randomly picked data sets, the posterior distribution of σ is computed while also taking into account this prior. The three data sets for which these thick red posteriors are computed are marked in red. The thick green line shows the ideal uncertainty distribution that results from concatenating all ten sample sets. This distribution represents the full amount of information that can theoretically be obtained by making use of all prior information.

Clearly, the Bayesian posterior distributions making use of the averaged prior (3.10) are more accurate and precise than the uncertainty distributions from individual estimation attempts. However, the comparison between the 'ideal' (in green) and actually demonstrated uncertainty distributions (continuous red) indicates that the prior is not very efficient. In effect, the prior only corrects estimates that it expects to be highly unlikely but does not have much influence in the domain it considers relatively likely. As discussed before, this can behaviour can be considered as an advantage. A prior that is only modestly informative and 'diffuses' the information from different sample sets, greatly reduces the probability that the prior can unintentionally have a large non-conservative influence.

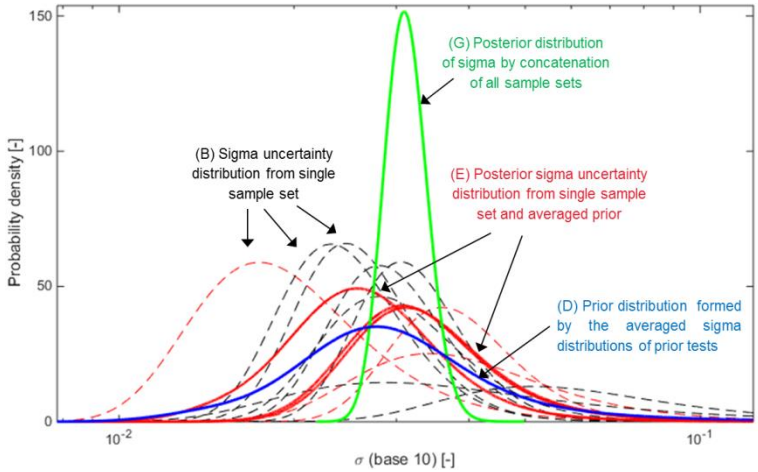


Figure C.9: Example of a single basic test result (i.e. for one m) according to the test procedure in Figure C.8, to which the element references (X) also correspond.

The final and complete test result is presented in Figure C.10. This result demonstrates that the use of a prior according to equation (3.10) indeed successfully improves the precision of the tolerance interval estimates. Low-tail quantile estimates that are very low relative to the prior are corrected upwards. This results in significant upward corrections in strength for cases in which an estimate only based on the individual sample set itself is unexpectedly low. Upper-tail quantile estimates that are very high relative to the prior are however corrected downwards. These (unspecified) strength reductions at the upper tail are minor but actually cause a very significant increase in the effective confidence level. The level of this increase can however not be predicted.

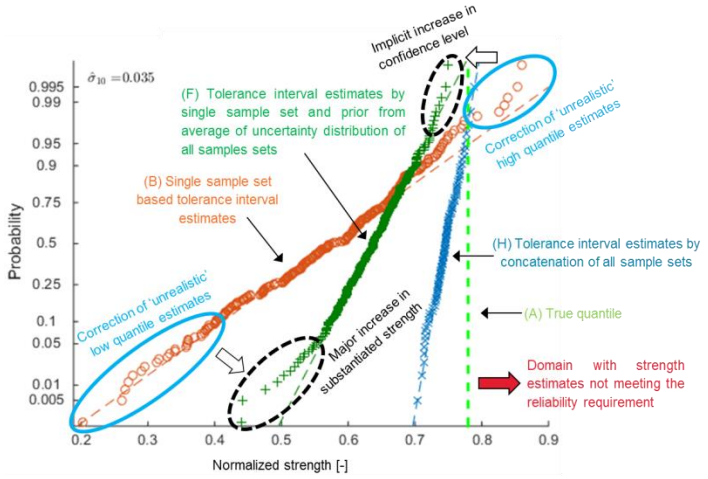


Figure C.10: Comparison between tolerance interval estimates with and without an (averaged) prior. 10^{-3} quantiles are estimated with 95% confidence. Element references (…) are corresponding to element identifiers in Figure C.8. (The probability plots are for a normal distribution)

Appendix D. Alternative simulation-based prior

This appendix details a simulation-based method using likelihood optimization to formulate a prior, instead of simple averaging according to equation (3.10) in chapter 3 and as studied in Appendix C. This simulation-based method to formulate a prior is used to cross-check the results between a prior from simple averaging and likelihood optimization two methods. Similarity between priors resulting from these two methods provides increased confidence in the validity of the generated priors.

As an alternative to the straightforward prior in the form of equation (3.10), it can be postulated that, in line with the earlier made engineering assumption, that realizations of $\hat{\sigma}_i$ are actually samples from a single distribution. The parameter vector $\bar{z}$ that defines this distribution can be proposed to be the solution to the following likelihood optimization problem:

$$p(\sigma \mid \alpha) \xleftarrow{\alpha \leftarrow \bar{z}} \underset{\bar{z}}{\operatorname{argmin}} \left(\sum_{i=1}^{N_{samp}} \sum_{j=1}^{q} -\log\left(p\left(\hat{\tilde{\sigma}}_{i,j} \mid \hat{\tilde{\sigma}}_{i,j} \right) \right) + \Psi(\bar{z}) \right) \quad \text{as} \quad q \to \infty \tag{9.1}$$

Where $\hat{\tilde{\sigma}}_{i,j}$ is a sample from the σ-uncertainty distribution associated with the i^{th} test program in which $\hat{\sigma}_i$ was observed based on $N_{samp,i}$ samples:

$$\hat{\tilde{\sigma}}_{i,j} \sim p\left(\sigma \mid \hat{\sigma}_i, N_{samp,i}\right) \tag{9.2}$$

And where $\hat{\tilde{\sigma}}_{i,j}$ is the sample estimate of the standard deviation of the j^{th} sample, of size $N_{samp,i}$, from a lognormal distribution that has its standard deviation set by a sample from the proposed 'super-distribution' $p(\sigma \mid \bar{z})$:

$$\hat{\tilde{\sigma}}_{i,j} = \sqrt{\frac{1}{n-1} \cdot \sum_{k=1}^{N_{samp,i}} \left[\hat{\mu}\left(S_{i,j}\right) - \log_e\left(S_{i,j,k}\right) \right]} \quad \text{with} \quad S \sim \exp\left(N\left(0, \tilde{\sigma}_{i,j}\right)\right) \quad \text{and with} \quad \tilde{\sigma} \sim p(\sigma \mid \bar{z}) \tag{9.3}$$

And where $\Psi(\bar{z})$ finally is a regularization function that penalizes the complexity and magnitude of the parameter vector $\bar{z}$.

The prime motivation behind this likelihood-optimization based method to formulate a prior is its ability to account for uncertainty distributions of σ that are very 'wide' just due to very small sample size conditions. The presence of wide uncertainty distributions does however not necessarily imply that the hypothetical distribution of σ is also very wide, as observed 'width' may just be the result of the uncertainty that comes from drawing small samples from an on itself rather specific distribution.

The likelihood-optimization based method is not advocated as a method of choice to form a prior due to its relative complexity. Nevertheless, from Figure D.1 it follows that its resulting prior is quite comparable to the prior derived by simple averaging by equation (3.10). This can be verified in more detail by comparing the quantiles of the priors shown in Figure D.1 and Figure 3.10. The similarity between the results from the two independent methods that generated the priors is considered to verify their applicability and correctness of the implementation.

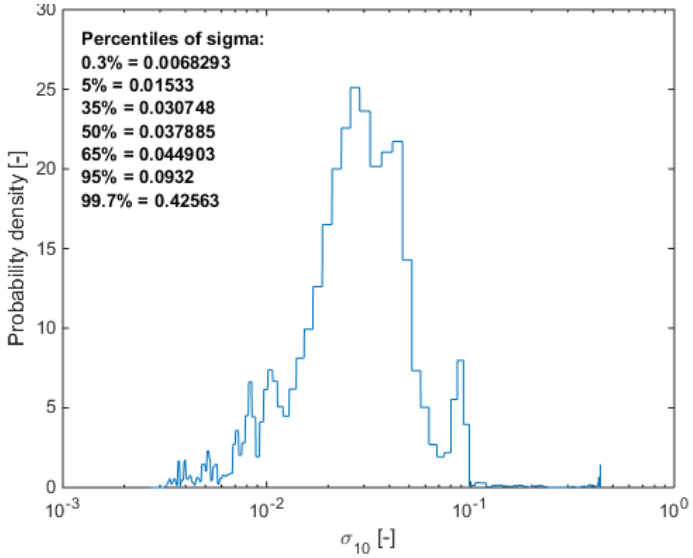


Figure D.1: Generic prior expectation on the standard deviation of fatigue strength according to the data in Figure 3.8 and an alternative simulation-based method.

Appendix E. Consistency verification of data sample with fatigue strength variance

The data in the scatterplot in Figure 3.8 demonstrates a clear and significant positive correlation between sample size and MLE estimate of σ, especially for small sample sizes. It is assumed that the decision on how many fatigue tests are done is independent of the (prior) expected standard deviation in fatigue strength. The prior uncertainty distribution must then be independent of sample size, i.e. the expected population standard deviation should be independent of sample size. There may then be two explanations for the positive correlation:

- Normal statistical behaviour. The estimator of the standard deviation, see also equation (3.3), is only an asymptotically unbiased estimator, i.e. as $n \to \infty$. For small sample sizes, the estimator may be significantly biased. The bias is proportional to $\sigma/(4n)$. This may easily be verified, e.g. by simulation.
- Over-fitting of an S-N curve. The estimates in Figure 3.8 are based on the residual distance of fatigue test points to an S-N curve that is defined to be most likely. This S-N curve is set by a four-parameter Weibull curve. Especially when few test points are available, it may be easy to fit the S-N curve unrealistically well through the available test points. An estimate of the standard deviation of fatigue strength may then be unrealistically low. This may be avoided though by using custom and expert judgement based methods (such as enforcing the shape of a material S-N curve).

A custom test procedure is defined to test if, given the prior in Figure 3.10 and given the data distribution in Figure 3.8, the correlation between sample size and estimated standard deviation is according to normal statistical behaviour. The specific design of the test is summarized in Figure E.1. In essence, a stochastic simulation is carried out to compare the statistically expected dependence between sample size and estimated standard deviation with the dependency that is observed in the database actually at hand.

The test's result in Figure E.2 indicates that for sample sizes two and three the median is significantly lower than may statistically be expected. It is therefore concluded that these parts of the dataset are significantly and non-conservatively biased and should, therefore, be removed from the dataset that is used to form a prior expectation. Additionally, it can be seen that the medians of the estimated standard deviations are consistently on the low side. As the prior is mostly generated by means of very small sample-sized data, the prior is therefore also biased towards underestimation of the standard deviation. This is especially notable in the last sampling size class with sizes ranging from 10-30. This bias can be a consequence of two modelling assumptions: first, by assuming a perfect estimate of the mean S-N curve and second by using estimator (3.3) to estimate the standard deviation of fatigue strength.

As the standard deviations for fatigue strength in the here used database were estimated by the residuals around a four-parameter Weibull-curve, there was a relatively high risk of overfitting, especially in comparison to using the well-known Basquin relation, which models the number of cycles to failure N_{cycle} as a function of the loading amplitude Λ with only two curve-fitting parameters *A* and *B* [1]:

$$N_{cycle} = A \cdot \Lambda^{-B} \tag{10.1}$$

The use of two curve-fitting parameters instead of four can significantly increase the risk of S-N curve overfitting and consequent underestimation of scatter.

Scatter around the S-N curve actually, takes place in two dimensions, i.e. in the load domain and in the cycle domain. Therefore, instead of equation (3.3) the following estimator is more appropriate and can further reduce the non-conservative and consistent bias in estimated standard deviations:

$$\hat{\sigma} = \sqrt{\frac{1}{n-2} \cdot \sum_{i=1}^{n} \left[\hat{\mu} - \log_e(S_i) \right]} \qquad (10.2)$$

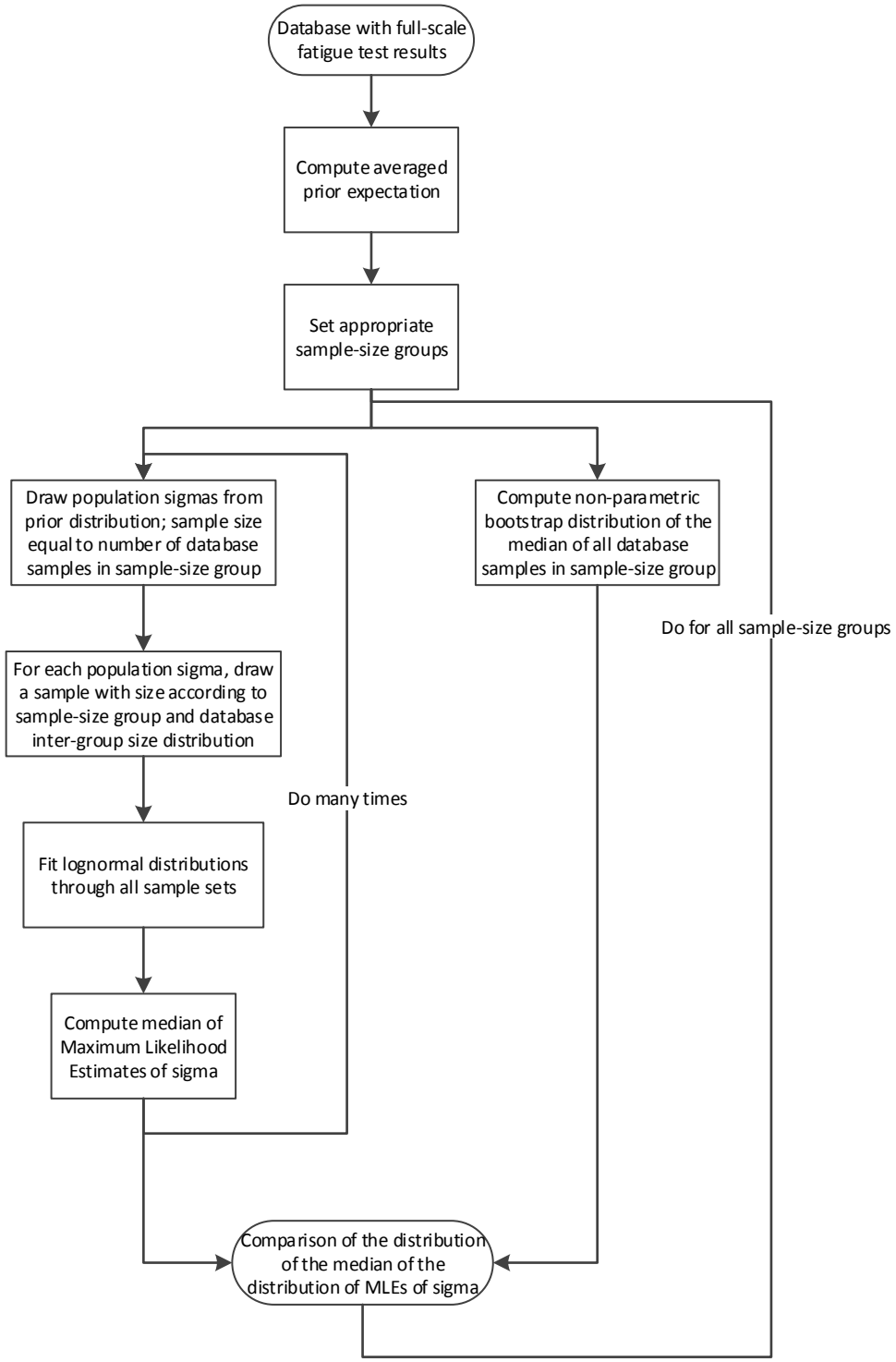


Figure E.1: Test procedure to verify if the average of MLEs of a standard deviation depends on the sample size according to normal statistical behaviour.

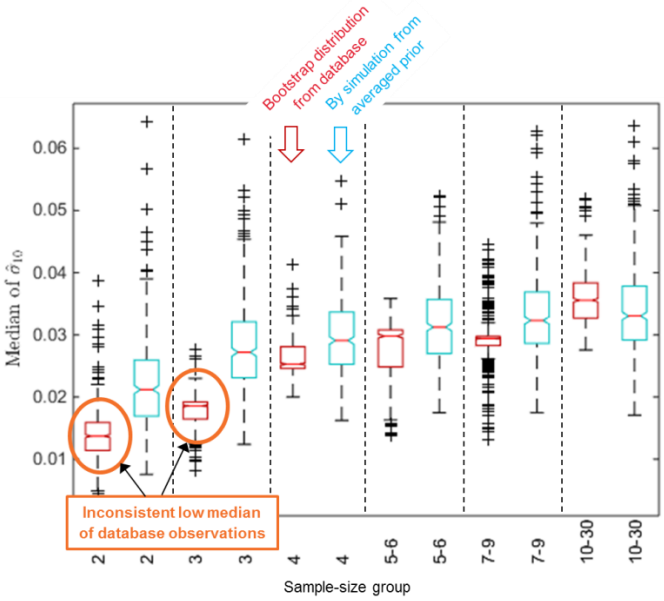


Figure E.2: Boxplot[33] comparison between the median of $\hat{\sigma}_{10}$ in the component fatigue test database and according to the prior in Figure 3.10

[33] Box limits correspond to the 25th and 75th percentiles. The whiskers have an approximate coverage of 2.7σ. (Full specification according to MATLAB R2014b *boxplot.m* default settings [42])

Appendix F. Manoeuvre extreme load and damage distributions

Real test flight data exhibits a significant variance in flight regime loads. Even when repeatedly flying a manoeuvre with about the same weight, centre-of-gravity and altitude, there is significant scatter in the resulting component loads. Furthermore, it is illustrated that the Generalized Extreme Value distribution can be used to model manoeuvre damage and manoeuvre extreme loads with reasonable accuracy. A short discussion on the necessity of accurate modelling of manoeuvre damage distributions and extreme load distributions is finally presented as well.

F.1 Extreme load distributions from test flight data

Only a few flight regimes have been repeatedly flown so many times during dedicated manoeuvre load test campaigns (load classification flights (LCF)) that at least some indication of the distribution of manoeuvre loads within a flight regime can be obtained. Distributions of the minimum and maximum loads while flying a regime are shown for a few gentile regimes and a set of critical dynamic components in sections F.1.1 to F.1.5. The fatigue damage corresponding to the regime loads and the appropriate working curves was always computed to be zero; regime damage distributions are therefore not shown.

A flight regime is the lowest sub-classification that is made in the Design Mission Profile and load classification flights, and is defined by a manoeuvre, e.g. a left turn at 1.2g, and a configuration specifying the weight, centre-of-gravity, and altitude class, e.g. median weight in a certain range, density altitude between 1500m and 3000m and a forward centre of gravity. Additional subclasses are added when the operation of external equipment such as a load hook or rescue hoist is additionally considered. Usually, flight tests are all performed at the same boundaries of the prescribed weight, centre-of-gravity (c.g.) and altitude classes. So even though the configuration classes may be relatively coarse, variance in load data from tests flights of the same regime should not come from large differences in weight, c.g. or altitude.

The example load data presented in sections F.1.1 to F.1.5 support the following two conclusions:[34] [35]

- The variance in the minimum and maximum load that occurs while flying a regime is very significant. It may, therefore, be expected that this variance has a significant influence on predicted fatigue lives and that the variance in regime loads should be taken into account in reliability substantiations of predicted fatigue lives.
- The distribution of the minimum and maximum load within a regime may with reasonable accuracy be described by a Generalized Extreme Value distribution (2.13).

Analysis of the appropriateness of the fitted distributions was done by visual inspection. Other methodologies to assess and compare the quality of a distribution fit do exist, however, such as the Bayesian or Akaike Information Criterion or dedicated goodness-of-fit tests, e.g. Wang [136] or Stedinger & Lu [137].

Analysis of test flight data presented here shows reasonable agreement with the assumption that manoeuvre extreme loads follow an unbounded generalized extreme value distribution. It can be argued though, that the extreme load of a manoeuvre should, in reality, be bounded, e.g. by manoeuvre physics. However, as argued in section F.3, the necessity of accurate tail modelling of extreme load distributions is not clear and likely not important. In any case, considering manoeuvre extreme load as an unbounded variable should be conservative

[34] Note that the load signal data in sections F.1.1 to F.1.5 was only subject to limited or no outlier filtering. The data distributions clearly show the presence of few outliers. However, it is not expected that these outliers affect any of the above conclusions.

[35] Note that the minimum flight regime loads are multiplied by minus one before fitting a Generalized Extreme Value distribution. This transformation is used to represent minima as maxima. The results in Section F.2 indicate that this transformation may better be reversed as the current implementation of the Generalized Extreme Value Distribution (2.13) actually appears better suited for the modelling of minima.

as it increases the probability of more extreme load situations. There is another expected advantage of modelling manoeuvre extreme load as an unbounded random variable. In most practical cases there is not enough test flight data available such that the few manoeuvre load tests can reasonably be expected to statistically cover the full range of possible loads. A model based on observed top-of-scatter loads, e.g. by Thompson & Adams [18], is more unlikely to include the actually possible most extreme loads.

F.1.1 FBTHETA - Main rotor collective booster load in axial direction

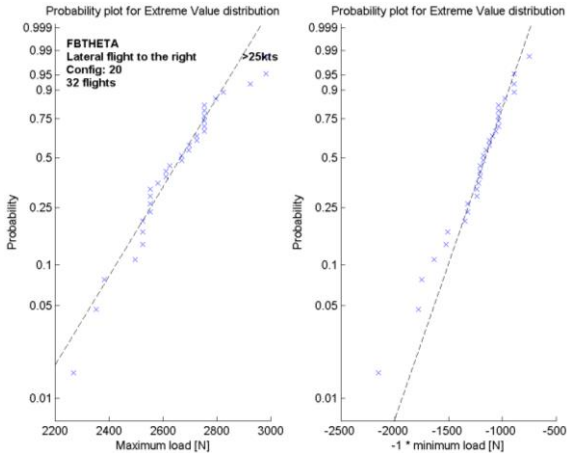


Figure F.1

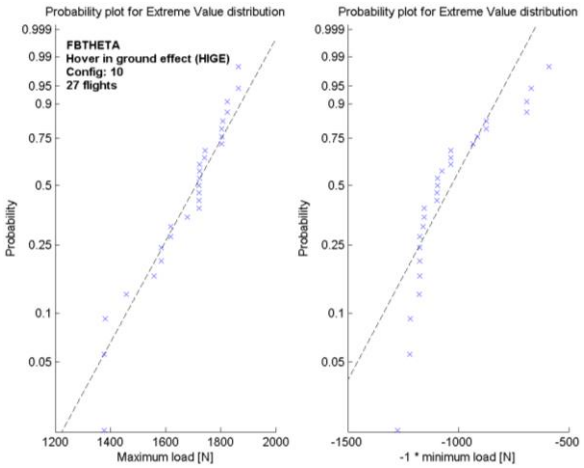


Figure F.2

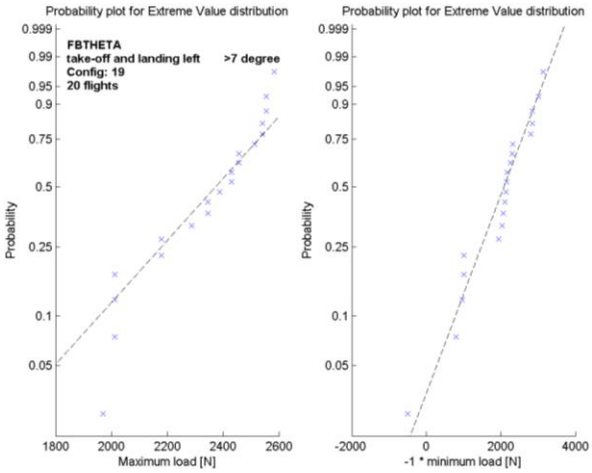


Figure F.3

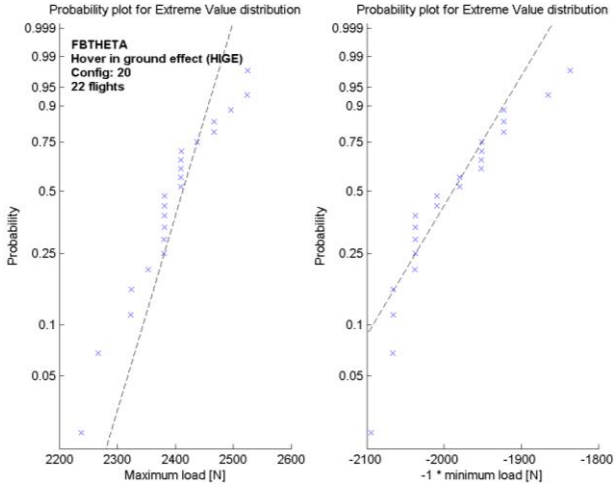


Figure F.4

F.1.2 FKAR – Composite load signal for cardan ring

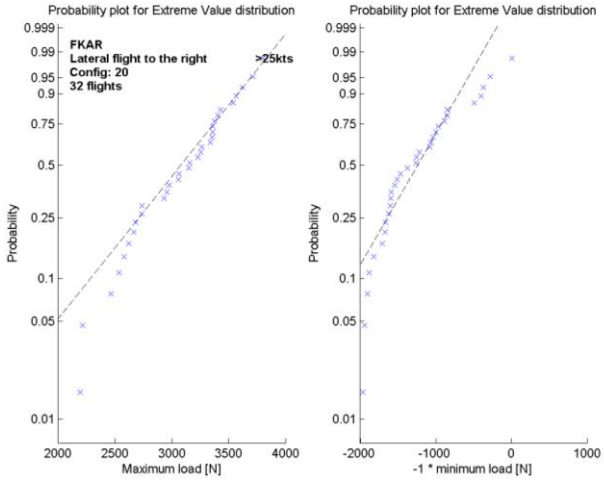


Figure F.5

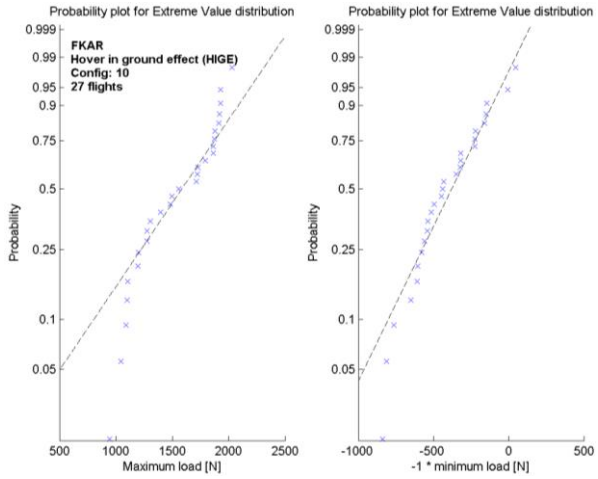
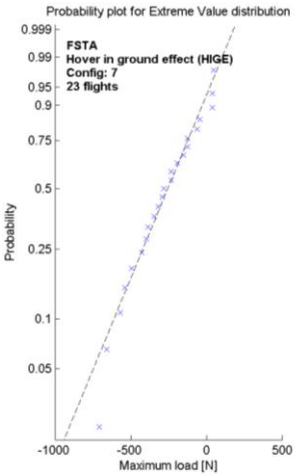


Figure F.6

F.1.3 FSTA – Composite load signal for forked lever



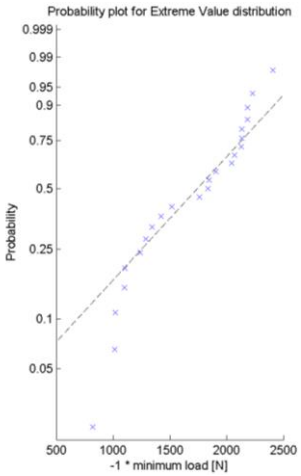
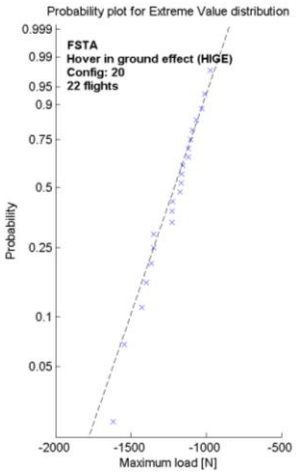


Figure F.7



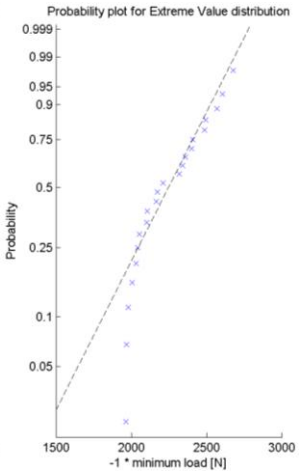
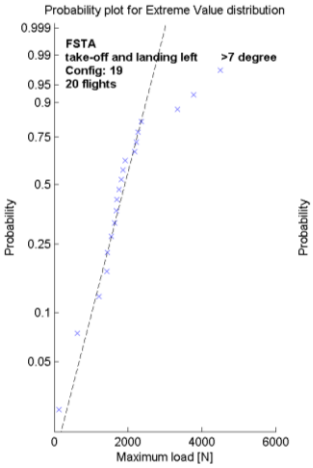


Figure F.8

Figure F.9

Figure F.10



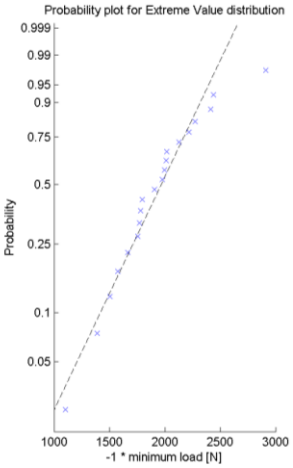
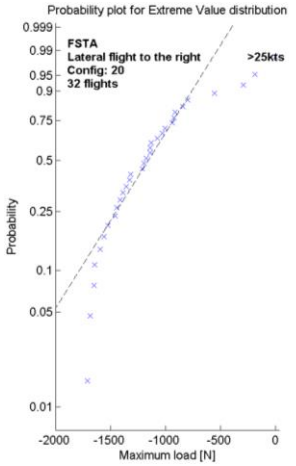


Figure F.11



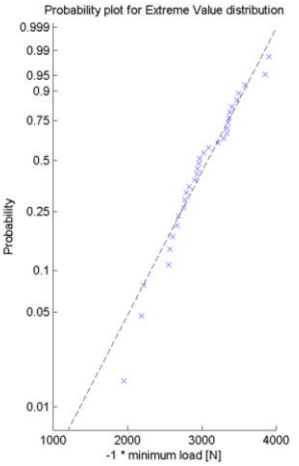
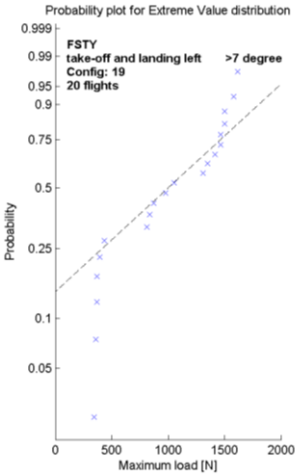


Figure F.12

F.1.4 FSTY – Load on main gearbox side strut



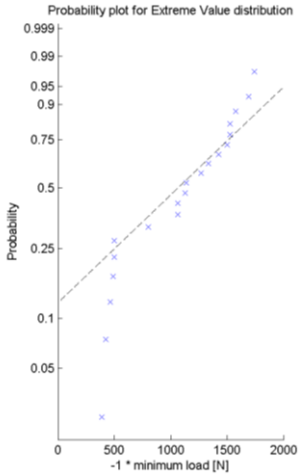
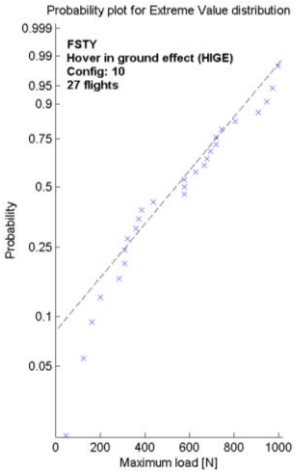


Figure F.13



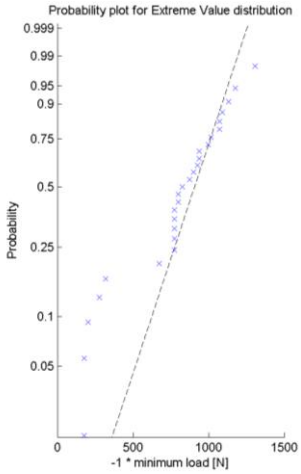


Figure F.14

F.1.5 MQF – Fenestron torque

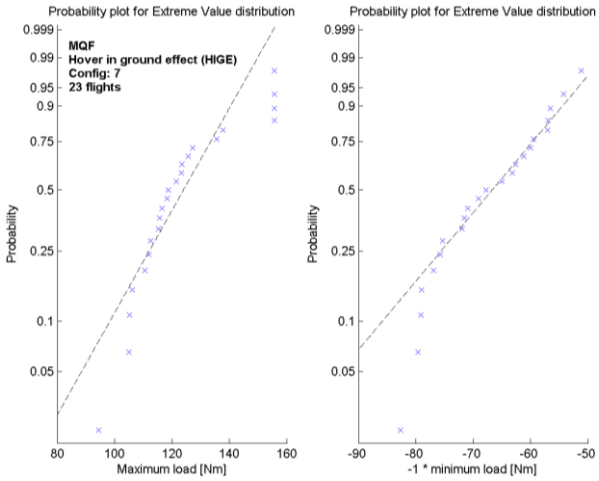


Figure F.15

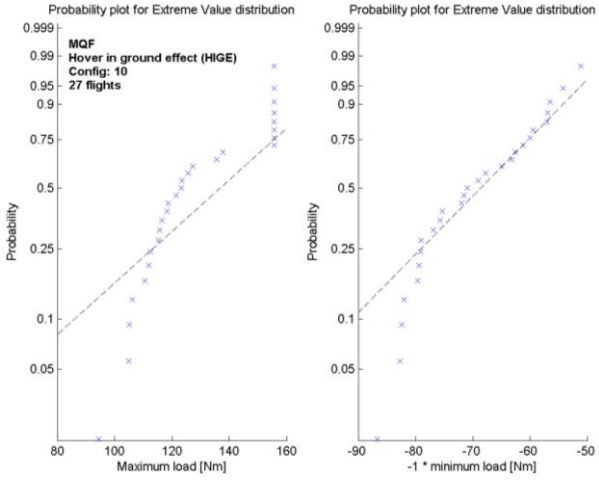


Figure F.16

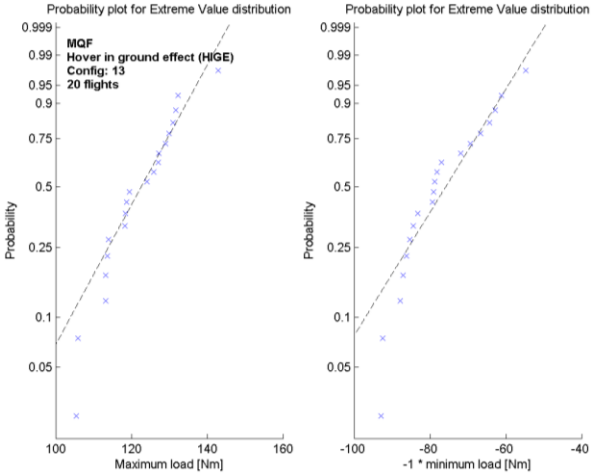


Figure F.17

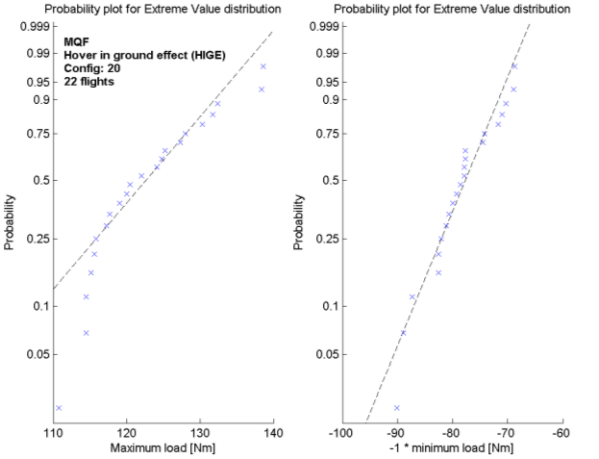


Figure F.18

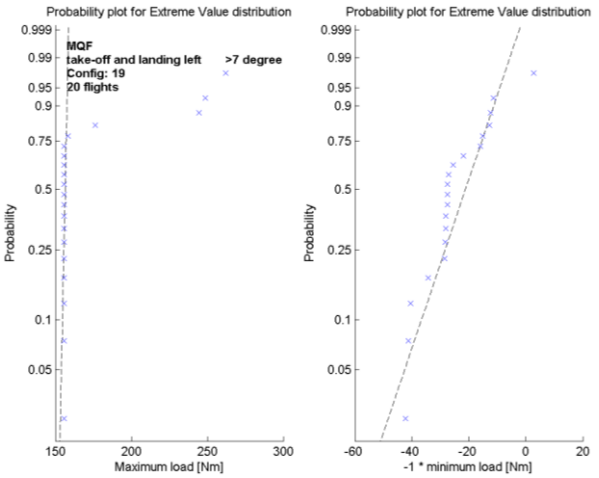


Figure F.19

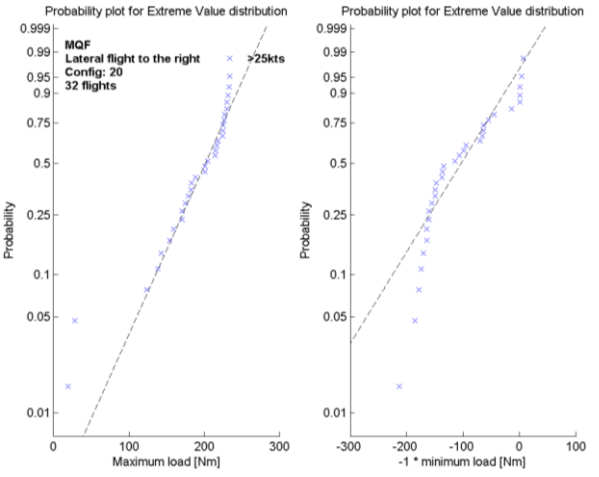


Figure F.20

F.2 Synthetic manoeuvre extreme load and damage distributions

The flight data in Section F.1 demonstrates a relatively small amount of extreme manoeuvre load tests per flight regime. To obtain more samples from flight tests is not realistic due to the associated high costs. Section 2.6.2 introduces synthetic flight regimes. The random loads in these flight regimes are simulated and virtual manoeuvre load tests can, therefore, be done as many times as desired.

A large amount of virtual manoeuvre load tests for fifteen different (randomly defined) synthetic manoeuvre types are simulated. The distributions of the maximum load in these synthetic manoeuvres are shown in Figure F.21. A Generalized Extreme Value (GEV) distribution is fitted to the observed data. The GEV distribution describes the observed data reasonably well but can certainly not be seen as a perfect model for this type of data.

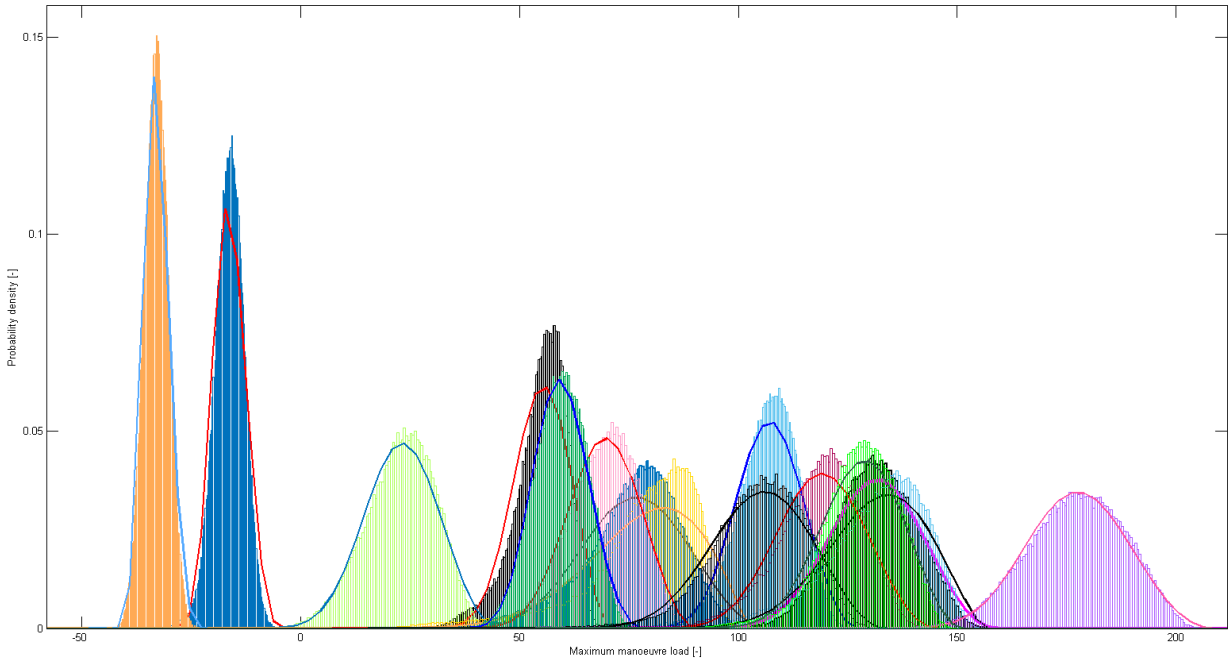


Figure F.21: Maximum manoeuvre load distributions for fifteen synthetic manoeuvre types. Generalized Extreme Value distributions are fitted.

Figure F.22 shows distributions for the manoeuvre minimum load. Note that the minimum load is multiplied by minus one to transform the minima to maxima. This transformation is used in the implementation of the stochastic load model as introduced in section 2.5.2.2. Again, it shows that the GEV distribution manages to describe the extreme load distribution reasonably well but not with full accuracy.[36]

[36] Figure F.22 shows the minimum manoeuvre load distributions without sign change transformation. It shows that the fitted GEV distributions capture the load data with much higher accuracy. It is thus concluded that the sign change transformation procedure is not implemented appropriately. It is thus recommended to change the sign change procedure to transform maxima into minima instead, i.e. to multiply the manoeuvre maximum load by minus one and not the manoeuvre minimum load. This suspected implementation error applies to thw work in chapter 2 but is not expected to have any significant effect on any of the validation results in Section 2.6. The outcome of the verification test of the simulation-based model indicates that inappropriately fitted GEV distributions do not result in a significant bias in the modelled reliabilities. The validity of the sampled fatigue life reference distributions is obviously not affected by this likely implementation error.

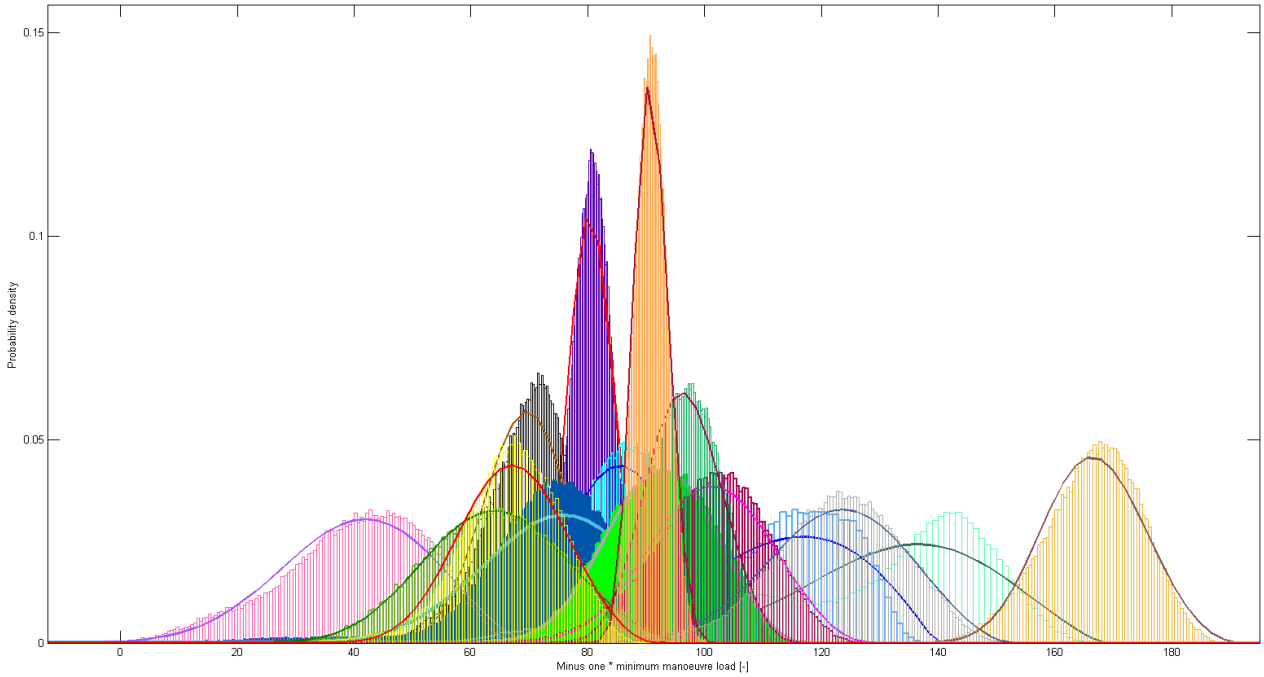


Figure F.22: Minimum manoeuvre load distributions for fifteen synthetic manoeuvre types. Generalized Extreme Value distributions are fitted. Note that the minima as transformed to maxima by a sign change.

Figure F.24 shows distribution fits through manoeuvre damage data for fifteen synthetic manoeuvre types. It can readily be seen that the fitted GEV distributions do not always accurately model the data. However, modelling in the domain of the highest and most influential manoeuvre damage is done with reasonable accuracy. It also shows that in most cases where the GEV distribution fit fails to provide an adequate model, the resulting model overestimates the probability mass in the upper tail, i.e. the probability of large manoeuvre damage is too high.

The endurance limit σ_{a_∞}, see also Section 2.2.1.1, makes that there is a hard lower limit of the manoeuvre damage. This can cause a strong discontinuity in the observed manoeuvre damage distribution. It is expected that some of the problems in fitting manoeuvre load distributions, given a particular strength value, do arise from the presence of such a discontinuity.

Again, the validation results in Section 2.6 indicate that these problems do not have a significant impact on the validity of the simulation-based reliability model. However, it cannot be excluded that such problems may become more significant if the general definition of the (synthetic) fatigue life prediction problems is changed, for example when the influence of manoeuvre damage on fatigue life becomes greater. (Up to approximately 5% of the fatigue life is determined by manoeuvre damage in the currently tested family of synthetic fatigue life prediction problems.)

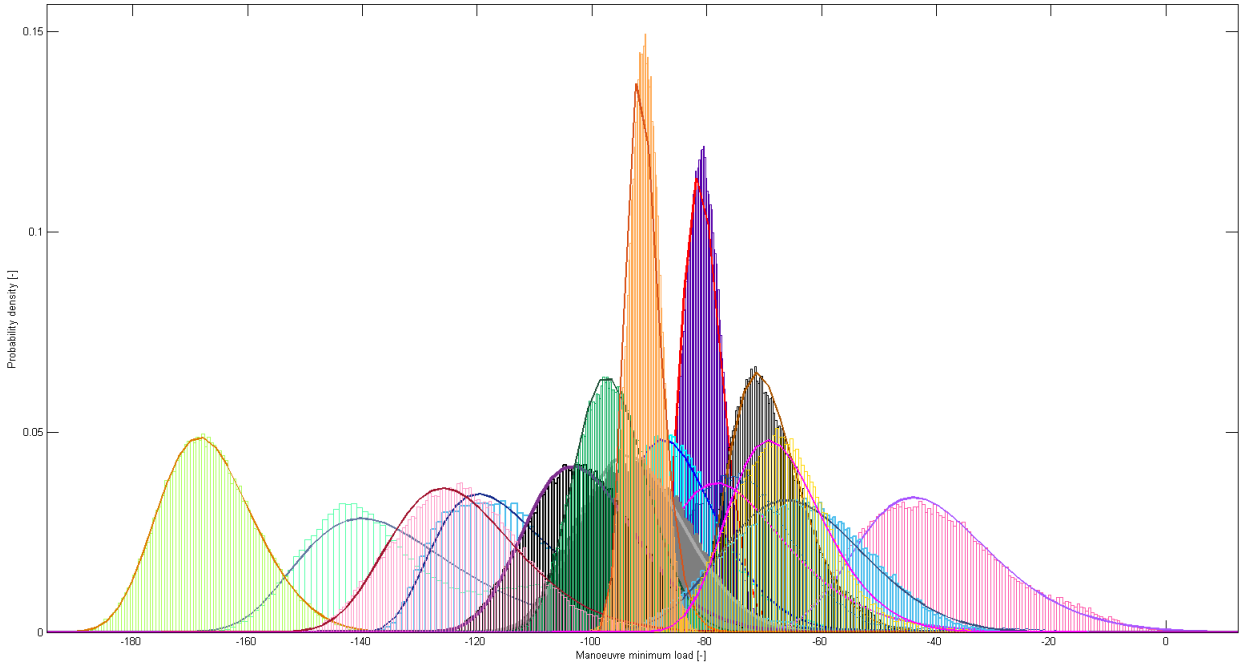


Figure F.23: Minimum manoeuvre load distributions for fifteen synthetic manoeuvre types. Generalized Extreme Value distributions are fitted.

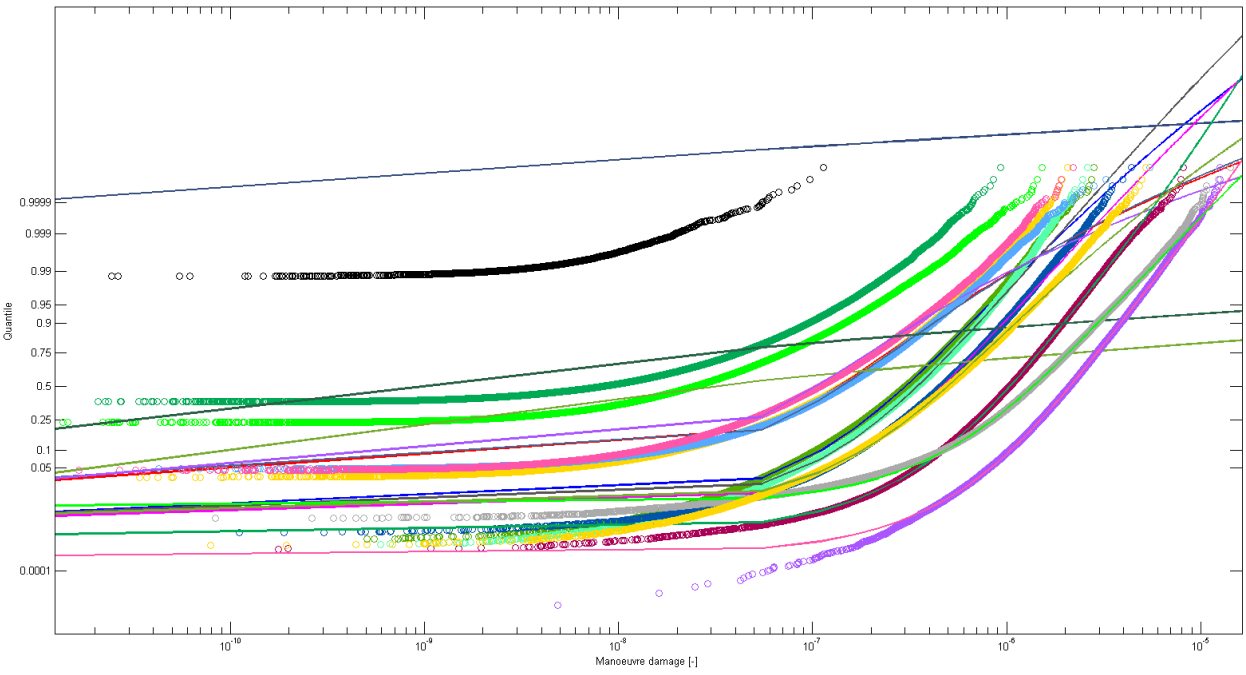


Figure F.24: Lognormal probability plot of the manoeuvre damage distributions, given the 10^{-6} strength quantile, for fifteen synthetic manoeuvre types. Generalized Extreme Value distributions are fitted through the samples. When the data (open circles) or distribution fits (continuous lines) follow a straight line, this indicates a lognormal distribution. A distribution fit represents the data with high accuracy when the fit goes through the sample points.

F.3 Discussion of requirements on accurate distribution modelling

The earlier research from Thompson & Adams [18] or Zhao & Adams [37, 38] (see also Section 2.4) indicates that component strength is generally the dominant parameter in the prediction of fatigue lives. If this is the case then the distributions of manoeuvre (extreme) loads may not need to be represented very accurately. It is reasonable to assume that approximate modelling of the random behaviours is sufficient.

The load spectrum according to the Design Mission Profile is determined by a high dimensional multivariate distribution, e.g. as in section 2.5.2.2. When it is further assumed that the (extreme) manoeuvre loads of each manoeuvre are of comparable influence, then accurate modelling of the tails of the marginal distributions of such a high dimensional distribution may not be of high importance. The relative probability mass located at the extremities of the multivariate load distribution becomes increasingly smaller as the dimensionality increases. The relative importance of the tail of any marginal distribution may then diminish.

However, Ground-Air-Ground (GAG) loads typically determine most of the fatigue life. The most influential load cycles in the GAG spectrum may be determined by the loads in only a few manoeuvres. Therefore, the assumption that all extreme manoeuvre loads have equal importance may not be appropriate. Also, due to the non-linearity of typical S-N curve, a small change in loads could have a large effect on fatigue lives. Thus, a single 'outlier' load may have a significant influence on fatigue life. This would encourage at least some attention to the modelling of the tails of load distributions.

The validation results in Section 2.6 finally demonstrate that it is likely that the variance in regime loads may be neglected when modelling reliability of predicted fatigue lives. Therefore, limited analysis of regime load distribution types may be sufficient and elaborate tail modelling may be neglected.

Appendix G. Implementation issues of the simulation-based model

The simulation-based numerical reliability model can be regarded as relatively complex. The high-level description in section 2.5 does not cover all aspects that are anticipated to be necessary for practical implementation and replication of the presented experiments. This appendix summarizes some practical issues in the practical implementation of the simulation-based model. These practical issues were addressed by extensions to the idealized model definition in section 2.5. Inadvertent over-specialization of the implemented practical model is prevented as the tested fatigue life cases vary significantly as an effect of random parameter initializations in equation (2.25). (i.e. the configuration of the extensions should not only be applicable to the specific test cases under consideration)

G.1 Proposal distributions

The simulation-based method requires additional design decisions to the configuration settings in section 2.6. Subset-Simulation requires setting Markov-Chain proposal distributions, see also Appendix A.6. These proposal distributions only need to be set at reasonable values when confidence levels are computed. Their setting then only influences the efficiency of the probability estimator and should not introduce a bias. Nevertheless, their setting is given here for completeness.

The proposal distribution to sample strength in Subset Simulation, see also Appendix A.6, is set as follows:

$$\tilde{p}\left(\tilde{\tilde{\omega}}_{i+1} \mid \omega_{i|F_{j-1}}\right) = \mathrm{N}\left(\mu = \omega_{i|F_{j-1}}, \sigma = 0.2 \cdot ds_k\right) \tag{12.1}$$

where $\mathrm{N}(\mu,\sigma)$ denotes a normal distribution and ds_k the width of the k^{th} strength interval under consideration (see also other nomenclature in Appendix A.6).

Extreme loads and manoeuvre damage larger than zero are sampled from multivariate *t*-distributions due to the use of *t*-copulas. The Subset Simulation proposal distributions are set by:

$$\tilde{p}\left(\tilde{\tilde{\omega}}_{i+1} \mid \omega_{i|F_{j-1}}\right) = \mathrm{N}\left(\mu = \omega_{i|F_{j-1}}, \sigma = 0.5 \cdot \mathrm{I}\right) \tag{12.2}$$

where I is the identity matrix.

Finally, the proposal distribution for the probability of non-zero manoeuvre damage is set by:

$$\tilde{p}\left(\tilde{\tilde{\omega}}_{i+1} \mid \omega_{i|F_{j-1}}\right) = \mathrm{B}\left(P = 0.3\right) \tag{12.3}$$

were $\mathrm{B}(P)$ denotes the binomial distribution.

The proposal distributions are set according to engineering judgement and may not be optimal.

G.2 Random noise addition

Non-parametric bootstrapping of small samples can result in a large number of duplicated samples in a sample set. The presence of such duplicated samples may cause numerical problems when fitting a distribution through such a sample set. When such problems are detected by an indicator function *T*, random Gaussian noise is iteratively added to the duplicated samples until the distribution fit is successful. It is assumed that arising inaccuracies may be neglected.

Two tests comprise the indicator function T for unsuccessful distribution fitting:

- T1: Passing of the "ConvergencedToBoundary" parameter test as defined in the commercial MATLAB gevfit.m GEV fitting function (this test indicates non-convergence of the optimization function that searches for the GEV parameters that correspond to a (local) maximum of the likelihood function)
- T2: Exceedance of the (default) upper limit on the number of function evaluation or search iterations in the fminsearch.m commercial MATLAB optimization function. (this is the optimization function that searches for the GEV parameters that correspond to a (local) maximum of the likelihood function)

$$\text{unsuccessul distribution} := \text{T1 OR T2} \tag{12.4}$$

At the j^{th} distribution fitting attempt and upon detection of an unsuccessful distribution fit, noise ϵ_i is added to the i^{th} duplicated variable from a uniform distribution:

$$\varepsilon_i = U\left(-0.9 \cdot \mathrm{M}, 0.9 \cdot \mathrm{M}\right) \text{ with } \mathrm{M} = 10^{-7+(j-1)} \text{ and } j \leq 7 \tag{12.5}$$

If the iteration counter j exceeds its permissible limit, i.e. 7, then the distribution fit is always accepted, regardless of the outcome of the tests indicating an unsuccessful distribution fit.

G.3 Addition of artificial samples

A distribution fit through a small number of samples may not yield a reasonable distribution. For example, the probability of extremely high manoeuvre damage may be grossly overestimated by an unrealistic distribution fit through few samples. Custom measures are implemented to detect the presence of such unrealistic distributions. Upon detection, random noise or artificial sampling points are added until a more realistic distribution is observed. The addition of artificial sampling points is done when all samples are transformed to a standard normal distribution. Artificial points are added from a normal distribution with a standard deviation of 1.2. These new points are sampled from a distribution with the actually observed mean but with a higher conservative variance.

$$data_{j,k} = \left\{ data_{j,\max(1,k-1)}, \mathrm{N}\left(\mu = average\left(data\right), \sigma = 1.2\right) \right\} \text{ with } k \leq 5 \text{ and } \mathrm{data}_{j,k=1} = data \tag{12.6}$$

The artificial support points are added iteratively until the distribution fitting test (12.4) is passed, or until the maximum number of additional support points is reached, i.e. 5.

If the iteration counter k exceeds its permissible limit, i.e. 5, then the distribution fit is always accepted, regardless of the outcome of the tests indicating an unsuccessful distribution fit.

The addition of artificial support points is combined with the addition of random noise to duplicated samples originating from bootstrapping, i.e. see Figure G.1.

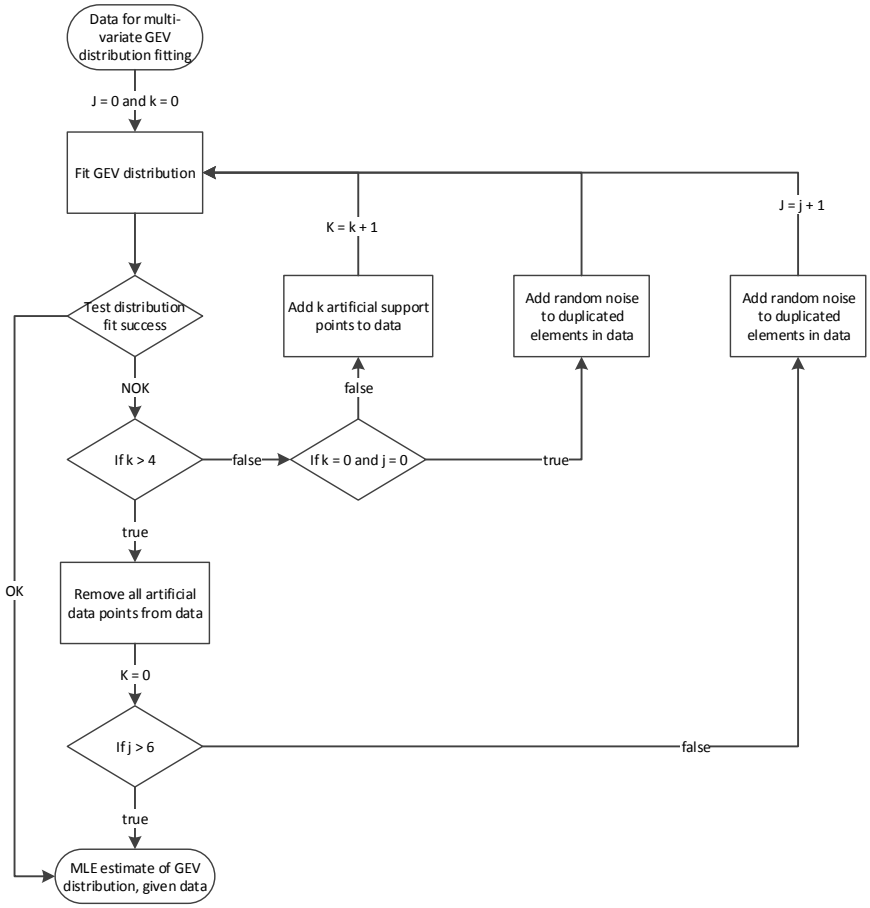


Figure G.1: Distribution fitting procedure under small sample size conditions

G.4 Filtering of MMH-MCMS samples

Another feature that is implemented to limit the effects of improper distribution estimates is filtering of samples drawn from the stochastic load spectrum model. Custom filters are implemented to detect and reject unrealistic samples that may have a major (unrealistic) impact on modelled reliabilities. See also algorithm A-1 on how an individual sample is generated.

Drawn minimum manoeuvre loads are filtered as follows:

$$\begin{aligned} &if\ sample_i < 2\cdot\mu(L_5)\ then\ sample_i = 2\cdot\mu(L_5)\\ &if\ sample_{i,j} > 8\cdot\left|sample_{i,j-1}\right|\ then\ sample_{i,j} = 2\cdot sample_{i,j-1}\\ &if\ sample_{i,j} < 8\cdot\left|sample_{i,j-1}\right|\ then\ sample_{i,j} = -2\cdot sample_{i,j-1} \end{aligned} \qquad (12.7)$$

where $sample_{i,j}$ is the i^{th} sample of the j^{th} subset and the function $\mu(L_5)$ computes the average of the lowest five samples the subset. Note that in all test cases in section 2.6, at least 150 samples are drawn per subset.[37]

Maximum loads are filtered equivalently:

$$\begin{aligned} &if\ sample_i > 2\cdot\mu(U_5)\ then\ sample_i = 2\cdot\mu(U_5)\\ &if\ sample_{i,j} > 8\cdot\left|sample_{i,j-1}\right|\ then\ sample_{i,j} = 2\cdot sample_{i,j-1}\\ &if\ sample_{i,j} < 8\cdot\left|sample_{i,j-1}\right|\ then\ sample_{i,j} = -2\cdot sample_{i,j-1} \end{aligned} \qquad (12.8)$$

[37] Although these may be divided into two groups according to the existence of manoeuvre damage.

where the function $\mu(U_5)$ computes the average of the highest five samples of the subset.

Non-zero manoeuvre damage is filtered as follows:

$$\begin{aligned} &if\ sample_i > 500 \cdot \mu(U_5)\ then\ sample_i = \mu(U_5) \\ &if\ sample_i \geq 1\ then\ sample_i = \mu(U_5) \\ &if\ sample_i \geq 1\ then\ sample_i = 0.1 \\ &if\ sample_i < 0\ then\ sample_i = 0 \end{aligned} \tag{12.9}$$

Further development of these sampling filters is recommended. The practical significance of the filters can be minimized by ensuring proper distribution fits, see also G.2 and G.3. When it can be ensured that Subset Simulation is executed with proper distributions, then filtering of unrealistic subset samples should not be necessary.

G.5 Local adjustment of strength

The stochastic load spectrum model in section 2.5.2.2 can require splitting of the available manoeuvre load tests if separate multivariate distribution fits must be made for manoeuvres that do and do not cause manoeuvre damage. The number of samples in each of the two groups thus depends on the fatigue strength by which manoeuvre damage is computed. When it is observed that one of the two groups contains too few samples to allow a proper distribution fit, the strength value to compute manoeuvre damage is reduced by one percent until the sample division is such that distribution fitting can be expected to be successful, i.e. at least six samples are available for distribution fitting. This method is conservative as it causes overestimation of manoeuvre damage. See also Figure G.2.

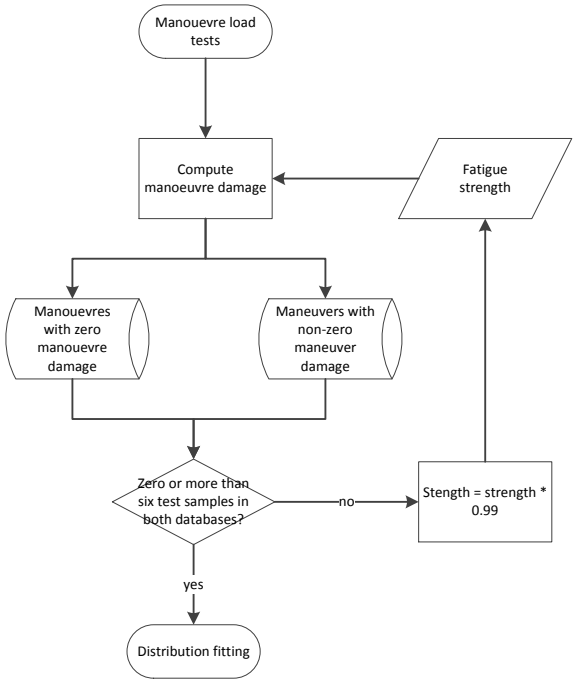


Figure G.2: Multivariate distribution fitting. The procedure in executed independently per manoeuvre.

G.6 Relevant strength domain

The probability of failure of a given Service Life Limit at a given strength interval s_i, $P_{fail}(\mathrm{SLL}, s_i)$, only needs to be evaluated for a small number of strength discretization points s_i. The graphical parameter analysis of

equation (2.21) in Figure 2.19 clearly demonstrates that $P_{fail}(\mathrm{SLL}, s_i) \cdot P(s_i)$ is not significant in large portions of the strength domain. Computational costs can thus be reduced by only computing $P_{fail}(\mathrm{SLL}, s_i)$, which has relative high computational costs when it has a significant influence on the total reliability integral (2.21). A crude but simple implementation of this feature can be realized by computing $P_{fail}(\mathrm{SLL}, s_i)$ for increasing strength intervals (i.e. starting at the lowest strength value) until the value of $P_{fail}(\mathrm{SLL}, s_i)$ falls below a certain threshold. $P_{fail}(\mathrm{SLL}, s_i)$ for the remaining strength intervals is set to the last computed value of $P_{fail}(\mathrm{SLL}, s_i)$. This is generally conservative as it is reasonable to assume that $P_{fail}(\mathrm{SLL}, s_i)$ decreases with increasing fatigue strength. However, under small sample size conditions, when using a small number of samples per subset, and when modelling low-reliability requirements, then $P_{fail}(\mathrm{SLL}, s_i)$ may feature irregular behaviour (i.e. $P_{fail}(\mathrm{SLL}, s_i)$ increases with increasing strength) and care must be taken in the selection of relevant strength domains.

$P_{fail}(\mathrm{SLL}, s_i)$ is evaluated for increasing strength domains s_i until the following condition is met (or until $P_{fail}(\mathrm{SLL}, s_i)$ has been evaluated for all i):

$$\left\{R < 0.05\right\} AND \left\{P_{fail}(SLL) \le 10^{-3} \text{ OR } P_{fail}(SLL, s_i) \le 10^{-8}\right\}$$
$$with: \; R = \frac{\sum_{j=i+1}^{I} P_{fail}(SLL, s_i) \cdot P(s_j)}{\sum_{j=1}^{i} P_{fail}(SLL, s_j) \cdot P(s_j)} \tag{12.10}$$

where I denotes the number of discretized or truncated strength distribution domains.

When $P_{fail}(\mathrm{SLL}, s_i)$ is evaluated up to and including the k^{th} strength interval s_k and condition (12.10) is met and $k < I$, then $P_{fail}(\mathrm{SLL})$ is estimated by:

$$P_{fail}(SLL) = \sum_{j=1}^{k} P_{fail}(SLL, s_j) \cdot P(s_j) + \sum_{l=k+1}^{I} P_{fail}(SLL, s_k) \cdot P(s_j) \tag{12.11}$$

G.7 Inverse Subset Simulation

The importance of very accurately estimating $P_{fail}(\mathrm{SLL}, s_i)$ when this value is close to one (i.e. > 0.1) is usually limited. Nevertheless, the regular Subset Simulation method was extended to also enable accurate modelling of high probabilities of failure (e.g. to be able to distinguish between P_{fail} = 1-10^{-2} and 1-10^{-3}). Inverse Subset Simulation can be realized by reversing the subset directions, effectively changing Subset Simulation into Superset Simulation:

$$F_1 \subset F_2 \subset \ldots \subset F_m = F \tag{12.12}$$

Inverse Subset Simulation or Superset Simulation is not further detailed, mainly as it is not critical and of small significance to the practical implementation of the simulation based reliability model. The procedure may however straightforwardly be derived from 'regular' Subset Simulation, as in appendix A.6. Superset Simulation is initiated if the probability of failure estimated by the initial Basic Monte Carlo sample is higher than 0.9.

$P_{fail}(\text{SLL}, s_i)$ values higher than 1-10^{-3} are non-conservatively truncated to 1-10^{-3}. This truncation is not conservative but improves the accuracy and precision of the estimate of $P_{fail}(\text{SLL}, s_i)$. Figure 2.19 demonstrates that this truncation has negligible significance on the estimate of $P_{fail}(\text{SLL})$.

Development and implementation of Inverse Subset Simulation was mainly driven by the use of gradient-based algorithms to estimate the relevant strength integration domain for $P_{fail}(\text{SLL}, s_i)$. The use of such gradient-based was however abandoned in favour of simpler methods.

G.8 Truncated probabilities

The strength domain is only discretized and evaluated between the upper and lower 10^{-12} quantiles. See also Figure G.3.

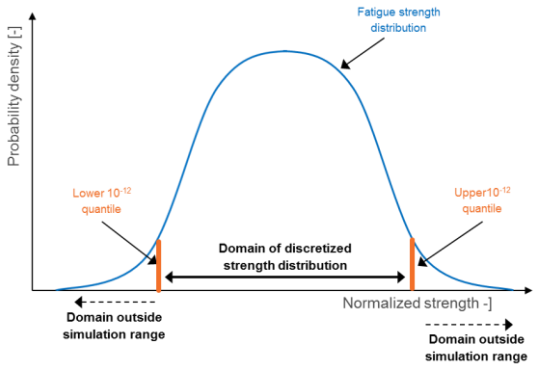


Figure G.3: Schematic representation of the truncation of a strength distribution

The probability mass of the strength interval with the highest strength (i.e. outermost left in Figure 2.19) is then increased by 10^{-12}. This is conservative as it assigns the probability mass of higher strength values to lower strength values. At the upper quantile this practise is not conservative but also not significant, as demonstrated in Figure 2.19.

Subset Simulation is executed using a subset probability of 0.1, i.e. $\gamma = 0.1$ in (6.17). Subset Simulation is considered converged as soon as the intermediate failure boundary lies beyond the SLL under consideration. However, the number of subsets is limited to fifteen, i.e. probability estimates of $P_{fail}(\text{SLL}, s_i)$ lower than 10^{-15} are truncated to 10^{-15}. This truncation is conservative and of negligible significance to $P_{fail}(\text{SLL})$, as also demonstrated in Figure 2.19.

G.9 Aborted reliability estimates

The currently implemented simulation-based numerical reliability estimator may abort its estimate before $P_{fail}(\text{SLL})$ can be estimated. This is only observed for reliability estimations of bootstrapped samples. The cause is suspected in a rare combination of highly distorted data samples due to non-parametric bootstrapping of a small sample and a logical implementation error in the tested implementation of the simulation-based reliability model.

To prevent that a small number of unevaluated bootstrap evaluations can cause complete validation tests to abort, it is assumed that these aborted bootstrap evaluations can be removed from the full set of bootstrap estimates without influencing the confidence level estimator. It is thus assumed that aborting of a reliability estimate is random and does not only occur in cases where, for example, reliability would be overestimated, as this would then cause an overoptimistic estimate of the realized confidence level.

G.10 Conclusion

Several extensions to the simulation-based model as presented in section 2.5 were necessary to allow practical testing of the model. All the extensions are however fully reproducible algorithms and procedures. This means that all the validation test results in section 2.6 thus refer to the same 'as-is' practical implementation of the ideal simulation-based model. None of the validation results indicated an inappropriate non-conservative influence of any of the extensions or custom configurations.

Appendix H. Machine Learning

The practical implementation of Virtual Fatigue Life Monitoring in chapter 4 makes heavy use of machine learning or non-linear statistical data modelling to predict in-flight loads or values for timeframe fatigue. The present appendix, therefore, provides a brief summary of some of the concepts of linear regression, principle component analysis, artificial neural networks and relevance vector machines.

H.1 Regression

Some fundamentals on the estimation of functions are introduced. In particular, how function parameters can be estimated by means of observed data points and how noise models and function uncertainty can be described.

H.1.1 Function fitting

A function *f()* of *x* with parameters A and Gaussian noise *N()* can be defined as follows:

$$f(x) = f(A,x) + N(\mu = 0, \sigma) \tag{13.1}$$

Its function parameters can be estimated from a sample of observed function values $\bar{f}(\bar{x})$ by least-squares error minimization, i.e. minimizing the average absolute distance between the estimated function and *k* available function samples:

$$\underset{A}{\arg\max} \sum_{n=1}^{k} -\left[f(A,x_n) - \bar{f}(x_n) \right]^2 \tag{13.2}$$

It can be shown that the result of least-squares error minimization is equivalent to the result of Maximum Likelihood Estimation (MLE):

$$\underset{A,\bar{\sigma}}{\arg\min} \sum_{n=1}^{k} \left[-\log_{10} p(z_n, A, \bar{\sigma}) \right] \quad \text{with} \quad p(z_n, A, \bar{\sigma}) = N\left(\bar{f}(x_n), \mu = f(\bar{x}, A), \sigma = \bar{\sigma}\right) \tag{13.3}$$

H.1.2 Noise models

Commonly, regression models assume that function noise is constant in the function's domain. This situation is referred to as homoscedastic noise. In practice, however, it can be that the degree of noise is variable. For example, noise can be increasing with increasing function value, e,g. temperature, or vary in time. Variable noise is referred to as heteroscedasticity and often needs to be estimated by a separate noise model. The two different situations for function noise are schematically illustrated in Figure H.1.

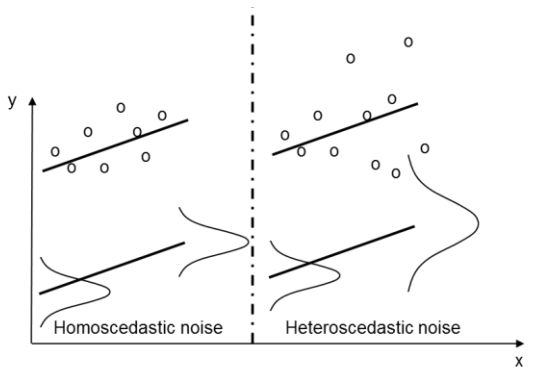


Figure H.1: Constant and non-constant function noise

H.1.3 Prediction and confidence intervals

If a regression model describes a noisy function, then it usually only makes a maximum likelihood estimation of the function value. If, additionally, the function's noise is estimated, then it is also possible to predict the probability of how far a future sample can lay from the most likely estimated function value. For example, given the following linear function with homoscedastic and Gaussian noise:

$$y(x) = a \cdot x + b + N(\mu = 0, \sigma) \tag{13.4}$$

then, while assuming that that function parameters *a,b* and σ have been estimated perfectly, the probability of observing a future value $\bar{y}(x)$ away from the most-likely function value $\hat{y}(x)$ can be described by the following probability density function (PDF):

$$p(\hat{y}(x) - \bar{y}(x)) = N(0, \hat{\sigma}) \quad \text{if} \quad \hat{a} = a, \hat{b} = b, \hat{\sigma} = \sigma, \hat{\mu} = 0 \tag{13.5}$$

A prediction interval for a future observation of the function *y(x)* can be computed using this PDF. For example, a prediction interval can specify the proportion of samples that, on average, will remain within a distance *PI* from the most likely function value:

$$P(|\hat{y} - \bar{y}| \le PI) = \int_{-PI}^{PI} \left[p(\hat{y} - \bar{y}) \right] dx \tag{13.6}$$

If the function parameters are estimated while the performance function could only be evaluated using a small number of samples, then there is often significant uncertainty concerning the true value of the function parameters. It could be, for example, that due to the randomness of the available data sample, an unusually steep and high-noise function is observed and deemed most likely, as in Figure H.2. Estimation uncertainty of function parameters can be described by confidence intervals and their associated PDF can be estimated by various techniques, for example by their normalized likelihood distribution [30]. Tolerance intervals, in turn, consider the combined effect of inherent function noise and imprecision in the estimation of the function parameters, due to a small sample size, on the probability of where a future function value could be observed.

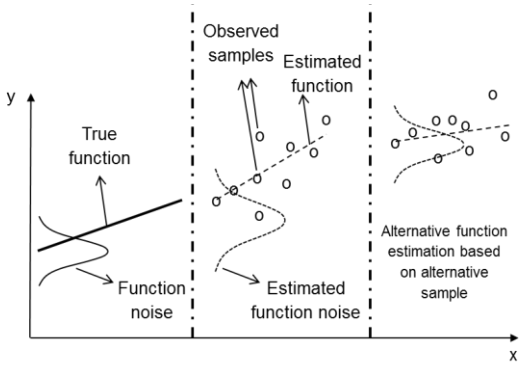


Figure H.2: Alternative function estimation by (hypothetical) resampling.

Prediction and confidence intervals generally assume that the models that have been employed in their estimation are correct. This means that it is assumed that the regression model captures the function appropriately and that the noise and uncertainty models are applicable as well. In practise, this can for example mean that it is assumed that the function is not quadratic when a linear regression model has been used and that it is assumed that function noise is indeed homoscedastic and Gaussian. If it is necessary to account for uncertainties about such modelling assumptions too, then the use of more advanced probabilistic techniques is necessary.

H.2 Data normalization

It can be useful to normalize all data before processing by machine learning algorithms. For example, this may prevent differences in scale or units between parameters in the same dataset to inappropriately influence performance functions.

A simple method to normalize a data vector X is to use a standard-normal translation:

$$\overline{X} = \frac{X - \mu(X)}{\sigma(M)} \tag{13.7}$$

where:

- $\overline{X}$ is the normalized data vector
- $\mu(X)$ and $\sigma(X)$ are the mean and standard deviation of the data vector *X* respectively

This normalization method assumes that the data in $\overline{X}$ follow a Gaussian distribution. In many cases, this assumption does not hold and standard normal translation does not result in perfect normalization. Nevertheless, standard normal translation generally provides robust and sufficient normalization to mitigate primary scaling issues. Standard normal translation according to (13.7) has been used to process all the data presented in chapter 4.[38]

H.3 Principal Component Analysis

Principle Component Analysis (PCA) is a methodology to find dependencies between different data sources within a dataset. Re-expressing the dataset with correlated parameters in it by using new parameters that are a linear combination of the original parameters can be used to effectively reduce the dimensionality of the data, without losing much accuracy. A reduction of the dimensionality can have a considerable positive influence on the convergence properties and computation costs of generating machine learning models. In addition, a reduction in the dimensionality of the feature space effectively reduces the sparseness of the dataset. Also, a denser distribution of samples throughout the sample space makes it easier to fit a robust regression model through the data and to estimate prediction error distributions, see for example Vapnik *et.al.* [138].

For example, the angle of attack of a Fenestron (ducted tail rotor) blade (DTHETA) can be expected to negatively correlate with indicated airspeed (IAS, helicopter forward speed). The aerodynamic load on the tail fin increases with forward speed, therefore, at higher speeds, the Fenestron needs to provide less anti-torque to compensate for the torque moment the main rotor provides. Expressing DTHETA and IAS as a linear combination of one another can thus reduce the number of independent parameters necessary to describe a flight state of a helicopter without losing much information on the overall state of the helicopter.

From a more abstract perspective, it can also be considered that the number of flight states of a helicopter is limited. In practice, helicopter operations can be entirely expressed by a limited number of 'eigen-motions'. The machine learning models that are generated in chapter 4 make use of 15 flight parameters, each sampled with 10Hz, to base their predictions on. With a timeframe duration of 1s, the feature space of these predictive

[38] Other normalization techniques, such as by Gaussian Kernel Density fitting and subsequent normalization by the inverse cumulative density function have been experimented with but did not provide clear benefits and came with significantly increased computational costs when processing large quantities of timeframe data. Moreover, this method was observed to provide less robust normalization results when faced with outliers or low-quality data.

machine learning models is thus 150 dimensional. Using PCA there are 150 normalized 'eigen-motions' available to describe the movement of a helicopter, as defined by equations (13.8) and (13.9).

The first principle component u_1 of a dataset X consisting of k data vectors x with length j follows from the following optimization problem:

$$u_i = \underset{\|u=1\|}{\arg\max}\left\{\sum_{i=1}^{k}\left(x_k.u\right)^2\right\} \tag{13.8}$$

The other j-1 principle components are defined by:

$$u_j = \underset{\|u=1\|}{\arg\max}\left\{\sum_{i=1}^{k}\left[\left(\hat{X}_i \cdot u_j\right)^2\right]\right\} \quad \text{with} \quad \hat{X} = X - \sum_{s=1}^{j-1} X u_s u_s^T \tag{13.9}$$

As illustrated in Figure H.3, the projection of the original dataset on its principle components reduces the variance of the data. Each principle component can be considered to 'explain' a percentage of the total variance in the original dataset. Specifying a percentage of retained variance can be used to effectively select the number of principle components upon which the original data is projected and how much data compression is applied.

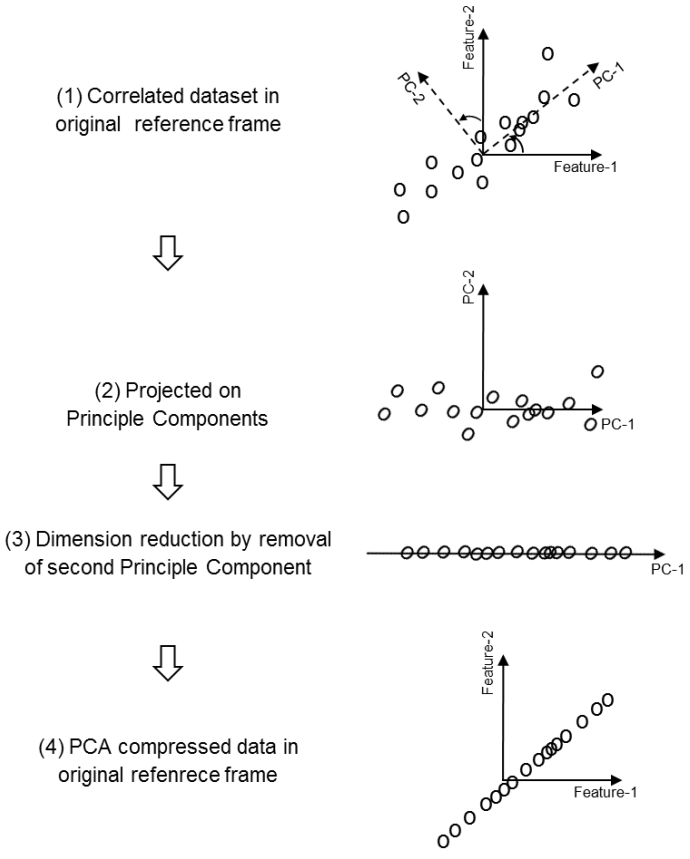


Figure H.3: Principle component analysis of correlated dataset and reduction of the dimensionality of the feature space.

The fraction of explained variance V_q by using the q^{th} principle component can be computed by:

$$V_q = \frac{\lambda_q}{\sum_{i=1}^{j}\lambda_j} \quad \text{with} \quad \lambda_q = \text{var}\left(\sum_{i=1}^{j} u_{q,j} X_i\right) \tag{13.10}$$

More in-depth and introductory treatments of Principle Component Analysis can be found in literature, for example by Hastie *et.al.* [106], Shlens [139] or Ng [140].

H.4 Non-linear statistical data modelling

Non-linear statistical data models or machine learning models mainly distinguish themselves in terms of the complexity of a response function they can approximate by regression. For example, a linear regression model can only describe a target function as a linear combination of its *n* inputs x_n :

$$y(x_1 \ldots x_n) \approx \sum_{i=1}^{n} (a_n \cdot x_n) \tag{13.11}$$

A non-linear regression model instead can take more complex forms and may even aim to capture functions $f(\ )$ of any complexity:

$$y(x_1 \ldots x_n) \approx f(x_1 \ldots x_n) \tag{13.12}$$

In general, the more complex the function that one attempts to model, the more useful such advanced regression models can be. However, as the complexity of the regression model increases, this often also requires more data for applicable and precise model generation.

There exist a large variety of models and techniques for non-linear statistical data models. A selection of models has been reviewed by Dekker [47] in light of previous work on Flight Regime Recognition. The suitability of Artificial Neural Networks and Relevance Vector Machines for Direct Load and Damage Modelling (DLDM) was demonstrated in Dekker [47]. These machine learning models have been applied successfully in chapter 4 for Probabilistic Load and Damage Modelling (PLDM) as well. Nevertheless, it should be noted that many other models and modelling variants exist and that these may be applied successfully for DLDM and PLDM also.

H.5 Artificial Neural Networks

An Artificial Neural Network (ANN) is a non-linear statistical data model with the ability to represent any arbitrary function, given some minor restrictions on the function and network design. ANNs can be complex to design and generate and a large variety of models and training methodologies have been developed. There currently exists no algorithm that can guarantee the optimal design and training of an ANN. The ANN regression models used in chapter 4 all make use of the same relatively simple and robust methodology.

H.5.1 Classic feedforward neural network with supervised backpropagation learning

An ANN consists of a large number of computational nodes referred to as neurons. A neuron computes a weighted sum of input vector *x* with length *I*, adds a bias *b* and generally performs a non-linear transformation of the weighted sum according to a transfer function *f()* before passing on its output *h*:

$$h = f\left(\sum_{i=1}^{I} \{W_i x_i\} + b \right) \tag{13.13}$$

A classic choice for the 'activation function' *f()* is the sigmoid function:

$$f(z) = \frac{1}{1 + \exp(-z)} \tag{13.14}$$

Neurons can be placed in a layered and interconnected array to form a computational network as exemplified in Figure H.4.

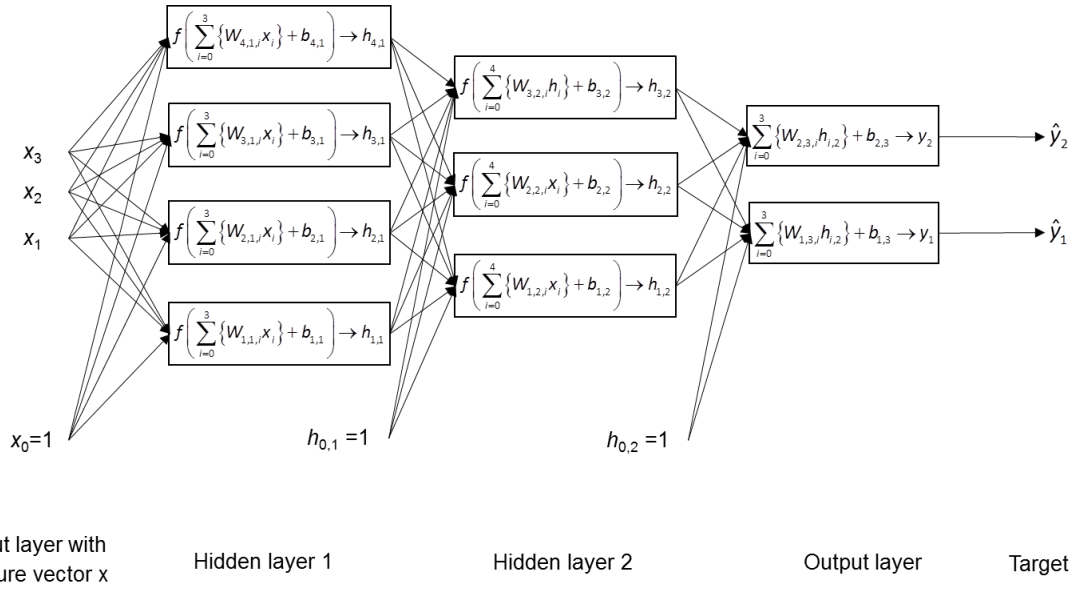


Figure H.4: Example of a two-layer artificial neural network with linear output layer.

Given a dataset with a large number of *m* examples in the form of feature or parameter vectors *x* and corresponding target vectors *y,* the weights *W* and biases *b* of a network with *a* layers and *d* neurons per layer can be optimized to maximize a performance function *J*:

$$\underset{W,b}{\arg\max} J(W,b) = -\left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|y_m - \hat{y}_m\right\|^2\right)\right] - \frac{\lambda}{2}\sum_{r=1}^{d}\sum_{q=1}^{a}\sum_{s=1}^{l} W_{r,q,s}^2 \tag{13.15}$$

The hyperparameter λ is a weight decay or regularization parameter, penalizing the magnitude of the design parameters *W*, thus encouraging sparse models and preventing overfitting. In this case the bias terms *b* are not included in the regularization function.

Another measure to counter overspecialization of the design parameters *W* and *b* is to divide the available examples into three separate subsets: a training set; a validation set and a testing set. The training set is used to iteratively evaluate the performance function *J(W,b)* and to estimate the gradient of the performance function with respect to updates of the design parameters. The performance function is however also computed for the validation set after each iteration. Once the performance function of the validation is observed to have reached a local optimum, further training is cancelled in order to prevent over-specialization. The testing set is finally used to obtain an independent measure of the quality of the regression model and is not used in the generation of the model itself.

Iterative adjustment of the design parameters *W* and *b* can be done by a backpropagation gradient-descent methodology to find a local optimum of *J*. For the output layer the required gradients can be computed as:

$$\begin{aligned}\frac{\partial}{\partial W_{r,l,s}} J(W,b) &= \left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{\partial}{\partial W_{r,l,s}} J(W,b,x_i,y_i)\right)\right] + \lambda W_{r,l,s} \\ \frac{\partial}{\partial b_{r,l}} J(W,b) &= \left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{\partial}{\partial b_{r,l}} J(W,b,x_i,y_i)\right)\right]\end{aligned} \tag{13.16}$$

To estimate the effect that the parameters of the hidden layers have on the cost function, the error function is traced back through the computational nodes and averaged until the node under consideration is reached. Corresponding implementation details can be found in Ng *et.at.* [107].

Once the regression model has been trained, the design parameters *W* and *b* are frozen and the network can be used as a normal regression model, and can be operated with relatively low computational costs.

H.5.2 Deep Learning

It can be shown that training becomes more difficult and unstable with an increasing number of hidden layers. Generally, the effective use of more than two hidden layers requires specialized training methodologies. Deep Learning is a technique in which successive hidden-layers are pre-trained or pre-conditioned to represent and compress the feature data itself before actual training to become a regression model commences. This is a process also referred to as Sparse Auto Encoding. Deep Learning has been experimented with during the work to generate regression models for Direct Load & Damage Modelling in chapter 4. However, it was found that this technique did not lead to substantial improvements but at the cost of considerably increased complexity and computational costs. It is presumed that the size of the available dataset for training from test flights is not sufficiently large to effectively support the generation of a Deep Neural Network with a large number of design parameters.

H.5.3 Further references

For a more comprehensive and in-depth overview of Artificial Neural Networks, refer to Hastie *et.al.* [106] and Ng *et.al.* [107] Or refer to Dekker [47] for the principle design considerations that are used in the selection and implementation of ANNs in chapter 4.

H.6 Relevance Vector Machines

The Relevance Vector Machine developed by Tipping [108] is similar to the more common and popular Support Vector Machine (SVM) [106]. However, an RVM holds distinctive advantages over an SVM [109]:

- An RVM does not utilize hyper-parameters whose value must be set heuristically or by cross-validation
- An RVM generally results in less complex regression models, utilizing significantly less data and kernel functions to build its model
- For classification, an RVM directly provides a probabilistic estimate of class membership
- The kernel function that can be used does not need to satisfy Mercer's condition and can, therefore, be chosen more easily

Relevance Vector Machines (RVM) may be used for both regression and classification. However, since the practical use of RVMs for DLDM in chapter 4 is restricted to classification, only the classification variant is summarized here.

H.6.1 Bayesian predictive modelling for classification

Given a dataset with *N* parameter vectors $\bar{x}$ and corresponding binary class membership targets $\bar{t} \in \{0,1\}$, a function can be defined a mapping $\bar{x}$ to a scalar value *y* using the sum of *M* linearly weighted basis-functions $\phi(\bar{x})$ with weights w_m:

$$y(\bar{x},\bar{w}) = \sum_{m=1}^{M} w_m \phi(\bar{x}) \tag{13.17}$$

For binary classification, this mapping is scaled to $y \in [0,1]$ using a sigmoid function:

$$\sigma(y[\bar{x},\bar{w}]) = \frac{1}{1+e^{-y[\bar{x},\bar{w}]}} \tag{13.18}$$

In the Bayesian framework of the relevance vector machine, each of the *M* weights is subject to a zero-mean Gaussian prior with variance proportional to hyper-parameters α_m . The likelihood function of the weights can then be shown to take the following form:

$$p(\bar{w}\,|\,\bar{\alpha}) = \prod_{m=1}^{M}\left(\frac{\alpha_m}{2\pi}\right)^{1/2}\exp\left(-\frac{\alpha_m}{2}w_m^2\right) \tag{13.19}$$

Following Bayes' theorem, the following relation can now be established for the probability of the function weights and hyper-parameters, given the available data:

$$p(\bar{w},\bar{\alpha}\,|\,\bar{t}) = \frac{p(\bar{t}\,|\,\bar{w})p(\bar{w}\,|\,\bar{\alpha})p(\bar{\alpha})}{p(\bar{t})} \tag{13.20}$$

Since the denominator of this relation can be regarded as an integration constant, it follows that:

$$p(\bar{w},\bar{\alpha}\,|\,\bar{t}) \propto p(\bar{t},\bar{w})p(\bar{w}\,|\,\bar{\alpha})p(\bar{\alpha}) \tag{13.21}$$

The weights and hyper parameters can be found by maximizing their log-likelihood given the available examples:

$$\underset{\bar{w},\bar{\alpha}}{\arg\max}\, p(\bar{w},\bar{\alpha}\,|\,\bar{t}) = p(\bar{t},\bar{w})p(\bar{w}\,|\,\bar{\alpha})p(\bar{\alpha}) \tag{13.22}$$

Using a Bernoulli distribution to describe class membership probability given a parameter vector $\bar{x}$, the likelihood of the targets given a set of weights can be computed as:

$$P(\bar{t}\,|\,\bar{w}) = \prod_{n=1}^{N}\sigma\{y(\bar{x}_n,w)\}^{t_n}\left[1-\sigma\{y(\bar{x}_n,w)\}^{1-t_n}\right] \tag{13.23}$$

Using a special case of the gamma distribution such that the priors on $\bar{\alpha}$ themselves are uninformative and flat in $\log(\alpha_n)$, simplifying the prior distributions to delta functions at their mode (i.e. $p(\bar{\alpha}\,|\,\bar{t}) \approx \delta(\bar{\alpha}_{MP})$), and for given values of the hyper parameters $\bar{\alpha}$, the optimization problem (13.22) can be developed as:

$$\begin{gathered}\underset{\bar{w}}{\arg\max}\log\{P(\bar{t}\,|\,\bar{w})p(\bar{w}\,|\,\bar{\alpha})\} = \sum_{n=1}^{N}\left[t_n\log y_n + (1-y_n)\right] - \frac{1}{2}\bar{w}^T\bar{\bar{A}}\bar{w} \\ with: \\ y_n = \sigma\{y(\bar{x}_n;\bar{w})\} \\ \bar{\bar{A}} = diag(\alpha_1,\ldots,\alpha_M)\end{gathered} \tag{13.24}$$

Using Laplace's method it can be shown that an approximate analytic solution to this optimization problem is given by:

$$\begin{gathered}\bar{w}_{MP} = \bar{\bar{\Sigma}}\bar{\bar{\Phi}}^T\bar{\bar{B}}\bar{t} \\ \text{with} \\ \bar{\bar{\Sigma}} = \left(\bar{\bar{\Phi}}^T\bar{\bar{B}}\bar{\bar{\Phi}} + \bar{\bar{A}}\right)^{-1} \\ \Phi_{n,m} = \phi_m(\bar{x}_n) \\ \bar{\bar{B}} = diag(\beta_1,\beta_2,\ldots,\beta_N) \\ \beta_n = \sigma\{y(\bar{x}_n,\bar{w}_{MP})\}\left[1-\sigma\{y(\bar{x}_n,\bar{w}_{MP})\}\right]\end{gathered} \tag{13.25}$$

The hyper parameters $\bar{\alpha}$ can be adjusted iteratively:

$$\alpha_i^{new} = \frac{\gamma_i}{w_{MP_i}^2} \quad \text{with} \quad \gamma_i \equiv 1 - \alpha_i \bar{\bar{\Sigma}}_{ii} \quad \text{and} \quad \bar{\bar{\Sigma}} = \left(-\bar{\bar{B}}\right)^{-1} \tag{13.26}$$

Tipping *et.al.* demonstrate that generally many of the hyper-parameters α_m converge towards infinity, effectively forcing the corresponding weight w_m to zero and removing the influence of the basis function from the function mapping function *y.* This property makes that this method of sparse Bayesian learning leads to very sparse prediction models.

H.6.2 Implemented Method for present work

The practical implementation of the Relevance Vector Machine used for Direct Load & Damage Modelling in chapter 4 is realized by a slightly adapted version of the *SparseBayes V2.0* software package developed by Tipping [111]. Note that this implementation contains improvements to the iterative optimization algorithm used to estimate the hyper parameters α_i and σ^2 that were developed by Tipping and Faul [110, 141]. These improvements are not treated in section H.6.1.

H.6.3 RVM for DLDM regression

The use of RVM models for regression of maximum and minimum loads was experimented with but did not result in encouraging results. The classic regression process for RVMs makes use of one basis function per example data point. However, it is difficult to provide such basis-functions with a reasonable coverage, given the high dimensionality of the feature space. Also, large coverage of the basis-functions initiates increased and undesired complexity of the practical implementation of the RVM algorithm due to memory limitations. Alternatively, the use of a relatively small number of heuristically defined basis-functions has been experimented with as well, however with unsatisfactory results.

Appendix I. Specific PLDM implementation

The conceptual introduction to PLDM in chapter 4 omits many details specific to the implementation by which PLDM has been tested. These design choices and implementation details are clarified in this appendix.

I.1 Specification of machine learning models

If the example database with data from test flights sufficiently covers the permissible flight envelope, then non-linear data models can use this database to 'learn' prediction models for timeframe maximum load, minimum load and timeframe damage. In the example implementation of chapter 4, Artificial Neural Networks (ANN) are used to predict minimum and maximum load during a timeframe. Timeframe damage is predicted in two stages as shown in Figure 4.4. First, a binary decision is made on whether timeframe damage is more than zero. This decision is made by a Relevance Vector Machine (RVM). If timeframe damage is predicted to be above zero, then its value is further detailed by an Artificial Neural Network. All the used ANNs are generated by standard back-propagation. Their configuration comprises of two hidden layers, each with sixteen neurons and sigmoid activation functions. For frame damage regression though, only eight neurones per layer are used.

I.2 Timeframe and feature specification

Timeframe duration is set to one second in the example implementation for chapter 4 and timeframes are cut out of the example database with an 80% timewise overlap, as shown in Figure I.1. The basic feature vector is computed by concatenating the recorded flight parameters (e.g. resulting in a 150-dimensional feature vector in the case of signal sampling to 10Hz and summation of engine torque #1 and #2). The dimensionality of the resulting vector is subsequently reduced by using standard Principle Component Analysis where 99.9% of information is kept.

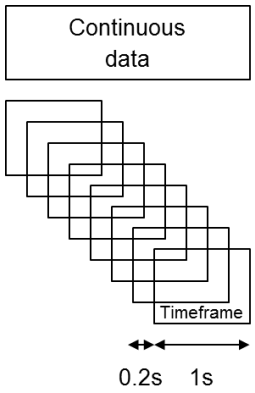


Figure I.1: Dividing Load Classification Flight data into timeframes for model generation

I.3 Database division

The example database is divided into a stage-1 dataset for point-prediction model generation, a stage-2 dataset to generate probabilistic prediction error models and finally a third and semi-independent dataset for testing. This division is done by standard interleaving with ratios 0.475, 0.425 0.1 respectively. During stage-1 model generation, e.g. for ANN maximum likelihood prediction of timeframe extreme load, the stage-1 dataset is again divided into a training and validation dataset by random shuffling with ratios 0.8 and 0.2 respectively. This last division is however not done for RVM training, as this method does not need a separate dataset for validation to prevent overspecialization.

I.4 Details for timeframe fatigue damage prediction

The RVM makes use of a non-linear kernel function. Each row in the kernel matrix is composed out of:

- a unit-value, i.e. one, the basic feature vector of the timeframe the row corresponds to

- the ANN-predicted maximum load and minimum load of the timeframe
- the difference between the ANN-predicted maximum and minimum load

If too few timeframes with positive timeframe damage are available to train the RVM and ANN for its prediction, then timeframe damage is recomputed with an artificially reduced value of fatigue strength, until enough samples are available for model generation. In the example application, a threshold of at least 0.5% of the total number of available frames is taken, provided that at least one frame features non-zero fatigue damage, given the starting value of fatigue strength. This practise is conservative since it can only lead to overestimation of timeframe damage.

I.5 Discretization of fatigue strength distribution

Timeframe damage depends on fatigue strength and PLDM considers both timeframe damage and fatigue strength as a random variable. Ideally, the prediction of timeframe damage can be adjusted according to each new sampled value of fatigue strength; however this would induce very high computational costs. Instead, a practical implementation was chosen in which the prediction of timeframe damage is only updated in a discrete fashion, i.e. if fatigue strength crosses pre-defined boundary values. The distribution of fatigue strength is divided into a finite number of intervals as illustrated in Figure I.2. Within each interval, high-frequency fatigue damage is always predicted conservatively while assuming fatigue strength equals the lower boundary of the interval.

This practice is conservative as it leads to consistent overestimation of timeframe damage. The boundaries upon which fatigue damage is re-estimated are determined dynamically in order to minimize the number of discrete intervals. The number of prediction models necessary is thus limited, but while also reducing the conservative prediction bias due to discretization. The boundaries are set according to the change of total timeframe damage in the example LCF dataset between the upper and lower boundaries of an interval. Since fatigue strength is continuously distributed and cannot be predicted for an infinitely low value, discretization is truncated at a reasonably lower quantile, e.g. the lowest value of fatigue strength according to which timeframe damage is predicted corresponds to the 10^{-12} quantile of fatigue strength.

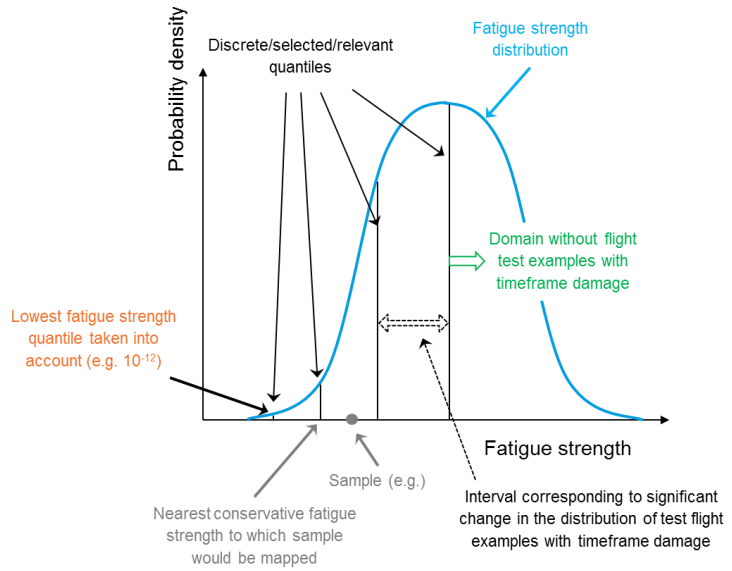


Figure I.2: Discrete and relevant quantiles of fatigue strength for which timeframe damage is modelled.

As was the case for DLDM, timeframe damage is actually predicted according to the two steps summarized in Figure J.3: first, a binary regression model predicts if timeframe damage is more than zero, and if so, then the amount of timeframe damage is further detailed by a continuous regression model. As a difference with DLDM

though, PLDM repeats the prediction of timeframe damage for all considered values for fatigue strength, instead of just making one prediction using a conservative working S-N curve as DLDM does.

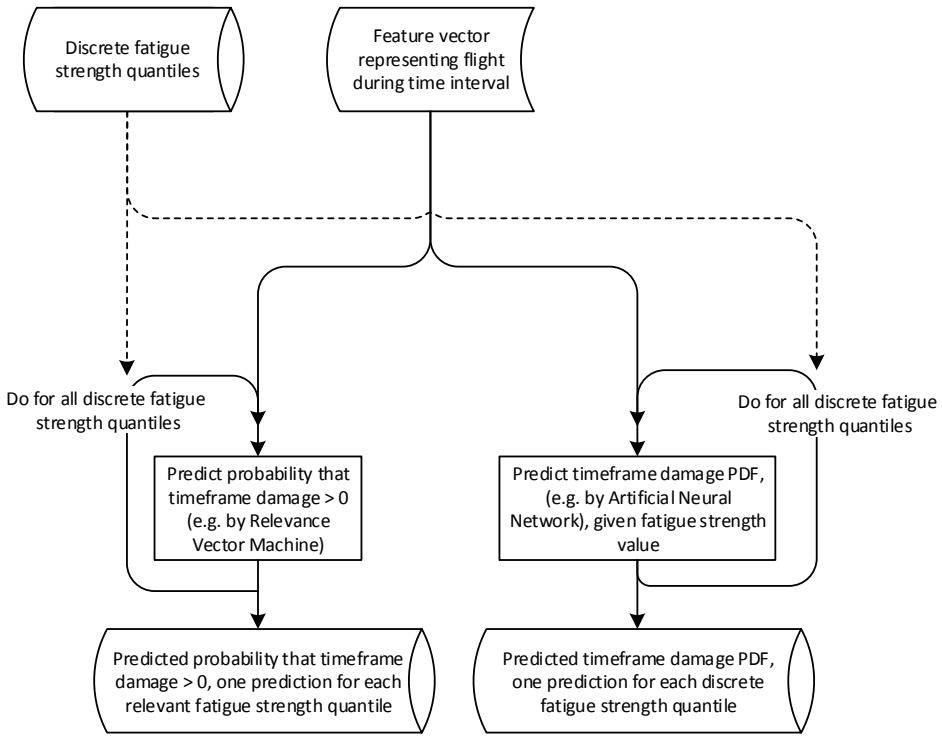


Figure I.3: PLDM timeframe damage modelling, for a single timeframe.

I.5.1 Implementation details for the discretization of fatigue strength distributions

Timeframe damage is only predicted for several discrete intervals of the fatigue strength distribution. These intervals are chosen such that they constitute a significant difference in frame fatigue damage. In the example implementation of chapter 4, these significant intervals are chosen such that the sum of the timeframe damage of all available example frames at least halves when moving to the next discretization point, starting from a fatigue strength value corresponding to the 10^{-12} quantile of the distribution of fatigue strength. Target discretization points are the following quantiles: 1e-12, 1e-11, 1e-10, 1e-9, 1e-8, 1e-7, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 0.25, 0.5, 0.75, 0.9, and 0.99. A discretization point is 'skipped-over' if it does not correspond to a significant change in the sum of frame damage.

I.6 Probabilistic model for prediction error

PLDM makes use of a probabilistic prediction model that makes a distribution estimate of a minimum or maximum load that occurred during a timeframe, instead of a point estimate as done by DLDM. In the example implementation of chapter 4, this probabilistic prediction is realized by relatively straightforward means whose implementation steps are summarized below:

- First, an ANN regression model is trained based on the stage-1 example dataset.
- Then, the stage-2 example dataset is used to fit a heteroscedastic prediction-error distribution model.

The implemented process to fit the heteroscedastic prediction-error distribution model consists of several steps:

- first, the stage-2 dataset is fed to the point prediction model generated by the stage-1 dataset.
- Then, these predictions are sorted according to magnitude.
- Finally, an error distribution is fitted through sequential batches of 1000 samples. Subsequent batches have 50% overlap and the sequential fitting process is carried out twice:

- once from the centre of the sorted prediction values to the lower end
- and once again from the centre to the upper end.

The doubling of the fitting process ensures that the final batch at the outer end of the range can be artificially enlarged to encompass a remainder of samples smaller than a full batch size.

If the distribution fitting process fails to converge, then the particular batch is skipped and the associated error distribution is replaced by the neighbouring error distribution, from the closest 'outside-in' direction.

Each thus fitted error distribution is associated with the mean of the point prediction values in the associated batch. Future predictions are assigned to a batch, i.e. fitted error distribution, according to the minimum geometric distance to these mean values. Four different distribution types are fitted:

- a Gaussian distribution,
- a Generalized Extreme Value distribution
- a Gaussian distribution with dedicated Pareto tail models
- a Gaussian kernel density function.

The actually employed error distribution is selected by the Akaike and Bayesian Information Criteria according to the following selection logic:[39]

$$
\begin{aligned}
&if\ BIC(Gauss) \leq BIC(GEV)\ AND\ BIC(Gauss) \leq BIC(GaussWithParetoTails)\\
&\quad then\ ErrorDistribution = Gauss\\
&else\ if\ BIC(GEV) \leq BIC(GaussWithParetoTails)\ AND\ AIC(GEV) \leq AIC(GaussWithParetoTails)\\
&\quad then\ ErrorDistribution = GEV\\
&else\ ErrorDistribution = GaussWithParetoTails
\end{aligned}
$$

Where:

- $BIC(\dots)$ and $AIC(\dots)$ represents the parameter-corrected likelihood according to the Bayesian and Akaike information criteria respectively
- $Gauss$ represents the likelihood of a fitted Gaussian distribution
- GEV represents the likelihood of a fitted Generalised Extreme Value Distribution
- $GaussWithParetoTails$ represents the likelihood of a fitted Gaussian distribution with Pareto tail models
- $ErrorDistribution$ represents the selected distribution model for the expected prediction error

The decision logic has thus been set to prioritize distributions with fewer distribution parameters and only choose distributions with more parameters if statistical evidence encourages doing so.

The binary frame damage classification model is realized by a binary Relevance Vector Machine classifier, which automatically entails a fully probabilistic prediction. It is assumed that the RVM prediction is unbiased and perfect. Its prediction is thus not adjusted according to an empirical error distribution obtained from the stage-2 dataset.

I.7 Details of Subset Simulation execution

Several implementation details for the Subset Simulation process used in chapter 4 are summarized below:

[39] The use of a Gaussian kernel density function has been tested in the selection logic as well. However, this option was never selected in practise and has been removed from the selection logic in order to reduce computational costs.

- Sample size: In the example implementation of chapter 4, Subset Simulation is carried out using 80 samples per subset and a target probability of failure per subset of 0.1. Bootstrapping is carried out using 100 bootstrap samples unless otherwise mentioned.
- Chain initiation: All samples falling within a set intermediate failure domain are used as chain initiation coordinates during the next subset simulation step. The chain length during subset simulation is thus variable.
- Proposal distribution for binary timeframe damage prediction: The proposal distribution employed during Modified Metropolis-Hastings Markov Chain Monte Carlo Simulation (MMH-MCMS) for the binary frame damage decision - i.e. frame damage equal to zero, or not - is set dynamically to equal the currently sampled binary prediction.
- Proposal distribution for non-zero timeframe damage: For the error distribution of predicted larger-than-zero timeframe fatigue damage, the proposal distribution is a Gaussian distribution with mean equal to the current coordinate in the chain and a standard deviation of two, while sampling occurs in a standard normalized space.
- Proposal distribution if binary timeframe damage prediction changes: If the current coordinate in the chain corresponds to zero-damage, and the next coordinate is sampled to have non-zero damage, then the proposed value of non-zero timeframe damage is directly drawn from the associated probabilistic prediction.
- Proposal distribution for fatigue strength: When sampling from error distributions for fatigue strength, which is also performed in a standard normalized space, the proposal distribution is set to be a Gaussian distribution with its mean equal to the current strength coordinate, and a standard deviation equal to 0.3 or the standard deviation of the fatigue strength sample from the previous step, whichever is highest.
- Proposal distribution for timeframe extreme load: The proposal distribution to sample timeframe minimum and maximum load samples during MMH-MCMS for the next intermediate failure domain during Subset Simulation is set dynamically and consists of a GEV distribution fit through the current subset sample.

I.8 Surrogate or proxy damage for Subset Simulation

When simulating a distribution of accumulated fatigue damage, then this distribution is not always smooth and continuous since load cycles can all fall below a sampled fatigue limit. A discontinuity in the distribution can thus appear at the point where accumulated damage switches from zero to a positive quantity.

Subset simulation needs to be able to differentiate and compare the severity of different combinations of sampled fatigue strength, timeframe damage, and extreme loads, even if they all do not cause any fatigue damage. If samples cannot be ranked by their severity, then any practical implementation of Subset Simulation cannot sample progressively more severe subsets and for example cannot implement the Metropolis-Hastings Markov Chain Monte Carlo Simulation algorithm introduced in Appendix A.6.

Therefore, a continuous proxy indicator for fatigue damage has been created. This artificial proxy for fatigue damage severity is computed using the difference between the fatigue limit and a quantile of the distribution of sampled timeframe maximum loads.

I.8.1 Application examples

As demonstrated in Figure I.4, the employed proxy quantity of accumulated fatigue damage correlates well with real damage and increases monotonically with decreasing probabilities of accumulated fatigue damage, as expected. In the practical implementation of subset sampling for PLDM, proxy values for fatigue damage are only used to rank samples resulting in zero accumulated damage. Where possible, real damage values are always used. Figure I.6 provides an example where actual fatigue damage was zero for many subset samples

but where subset simulation could be carried out successfully nevertheless. Distributions of corresponding proxy or surrogate damage values in Figure I.5 illustrate simultaneously sampled values of proxy damage.

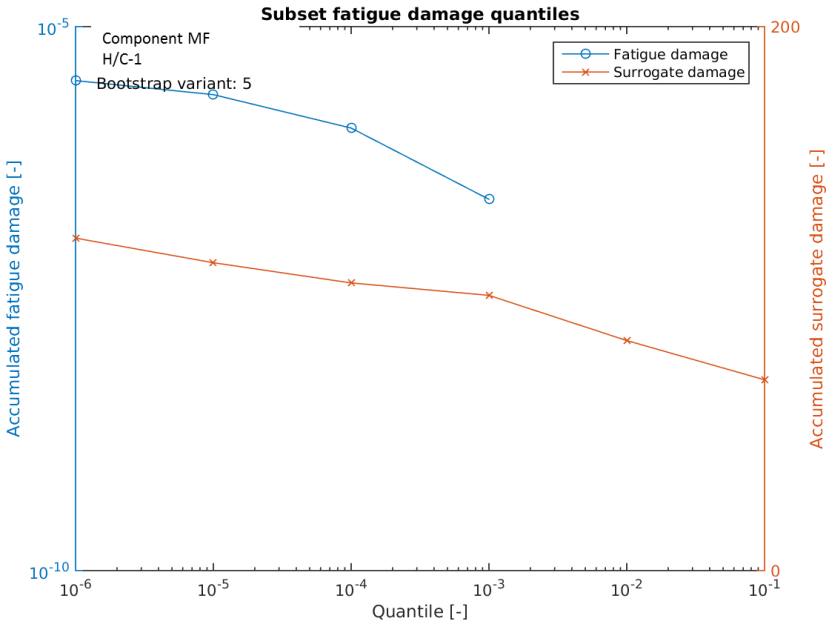


Figure I.4: Fatigue damage corresponding to subsequent (intermediate) limit state conditions during Subset Simulation

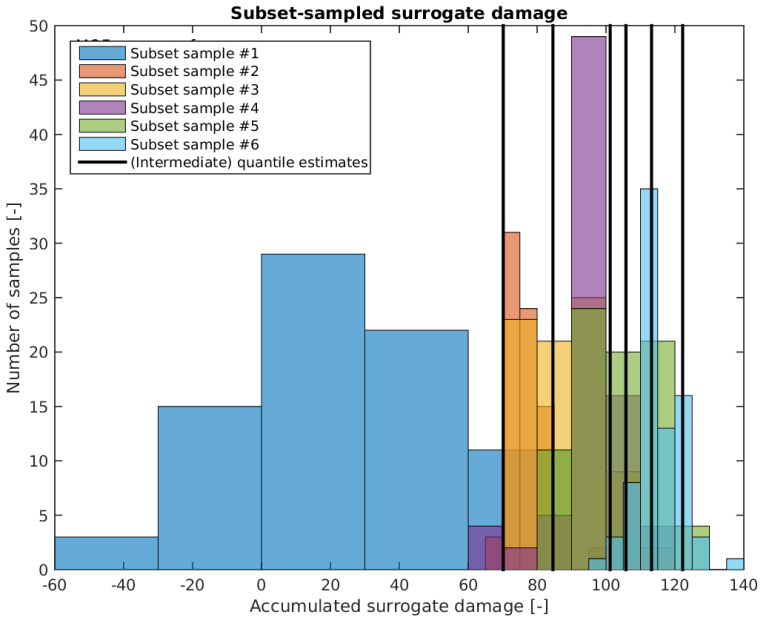


Figure I.5: Example of increasing surrogate damage with more unlikely events and severe subsets during Subset Simulation.

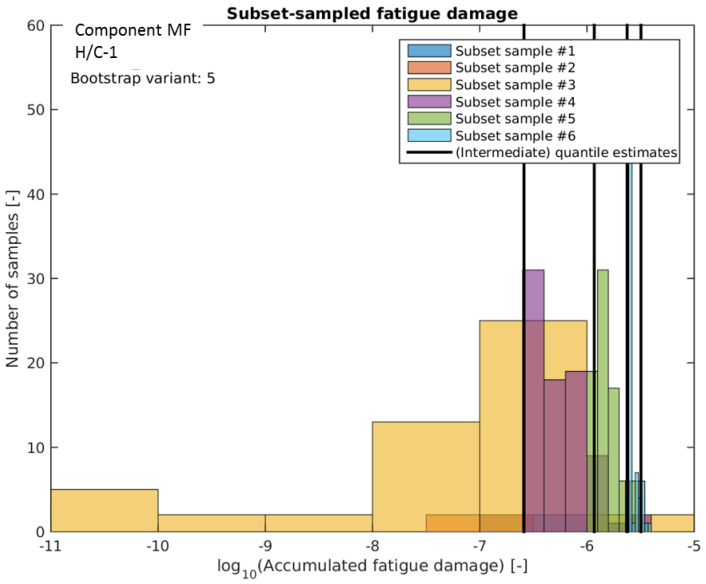


Figure I.6: Sampled accumulated fatigue damage increases as subsets become more severe and correspond to more unlikely events during Subset Simulation.

I.8.2 Implementation details

If a fatigue damage model is employed that features a fatigue limit, as is the case in the example application shown in section I.8.1, then it may be that many sampled instances correspond to zero fatigue damage. To perform Subset Simulation it is however required to sort all instances in a sample in order to determine an intermediate failure condition and to select a percentage of 'worst' or 'most-severe' samples to serve as initiation points for the chains in the next subset.

To circumvent this problem, use is made of a synthetic damage quantity to rank the severity of sample instances. In the example implementation of chapter 4, a surrogate quantity for fatigue damage D_{sur} is computed according to:

$$D_{surr} = -\left(\sigma_{\infty} - F^{-1}\left(I_1 \ldots I_m, 1 - \frac{\min[m,10]}{m}\right)\right) \tag{14.1}$$

Where σ_{∞} denotes the fatigue or endurance limit, I frame damage maximum load, m the number of timeframes and $F^{-1}\left(\bar{x},P\right)$ the empirical inverse cumulative density function specifying the P quantile of the sample $\bar{x}$.

I.9 Prediction filters and sampling filters during Subset Simulation

Samples generated during Subset Simulation are subject to several filters:

- Extreme load pre-filtering: Stage-1 ANN point-predictions are pre-filtered before starting the Subset Simulation process:
    - Stage-1 predicted minimum and maximum loads that are inconsistent, i.e. minimum load higher than maximum load, are interchanged.
    - Initial Stage-1 point predictions that fall outside statically defined feasible limits are truncated.

        - For extreme load predictions, these limits correspond to the extreme values observed in the complete example database from flight testing, conservatively enlarged with a margin of at least 10%.
- Extreme load sampling filters: sampled extreme loads that fall outside a dynamically determined feasibility limit are resampled until they fall inside a dynamic feasibility limit.
    - The basic feasibility limit consists of the earlier mentioned pre-set generic limit for the minimum load.
    - However, the maximum load limit is dynamically adjusted to correspond to the ultimate load of the currently sampled value of fatigue strength.
    - If a sampled minimum is not consistent with the maximum load sampled in the same timeframe, then both values are redrawn too.
    - Additionally, there is also a random chance that the load is resampled if the sampled load exceeds the pre-set generic maximum load limit. The probability of random resampling is proportional to:

$$\mathrm{B}\left(1,2\cdot\max\left[\Phi\left(\frac{l}{L_{\max}},\mu=1,\sigma=0.5\right)-0.5,0\right]\right) \tag{14.2}$$

        - Where $\mathrm{B}(1,P)$ denotes a binomial distribution with a single trial and probability of success equal to *P*. The currently sampled timeframe maximum load is *l* and L_{max} represents the pre-set generic maximum load limit, e.g. maximum load observed during test flights plus a 10% margin. $\Phi(z,\mu,\sigma)$ denotes the cumulative density function of a normal distribution with mean μ and standard deviation σ. This probabilistic resampling procedure is necessary since if a component, for example, accumulates ten thousand flight hours, then this corresponds to $10^4\cdot 3600 = 3.6\cdot 10^7$ timeframes, using timeframes of one-second duration. It can, therefore, be expected that some randomly sampled prediction errors will come from the extreme tail of the error distribution, suggesting unrealistically large loads.
    - If continued resampling of extreme loads fails to satisfy the feasibility limits, then the number of resampling attempts exceeds 250, then the maximum load is set to 95% of the ultimate load and the sampled minimum load too, if necessary.
- Timeframe damage filtering: For the prediction of timeframe damage, dynamic feasibility limits are pre-computed. These feasibility limits are a function of sampled maximum and minimum timeframe load, as well as sampled fatigue strength. In detail, the feasibility limits are computed as follows:
    - For discrete quantiles of the fatigue strength distribution at hand, e.g. [1e-9 1e-6 1e-5 1e-4 1e-3 1e-2 1e-1 0.75 0.5 0.9], and for a dynamically generated load discretization grid, a particle swarm optimization algorithm determines the worst-case load cycle and corresponding frame damage corresponding to a load cycle falling within a selected minimum and maximum load interval.
        - These intervals are bounded below by the overall minimum load and above by a selected discretization point.
        - The maximum load discretization points consist of 11 equally spaced points between the overall minimum load and the ultimate load corresponding to the fatigue strength of the current (truncated) fatigue strength value.
        - During Subset Simulation, sampled extreme loads and fatigue damage are conservatively truncated to the nearest discretization point and the corresponding feasibility limit is selected.

- The computed worst-case value for fatigue damage, given a load interval and truncated fatigue damage value, is multiplied by 5, and then sets an upper feasibility limit for frame fatigue damage.
- If a stage-1 prediction of timeframe fatigue damage falls outside these feasibility limits, then the prediction is truncated towards the appropriate limit, which is thus a function of timeframe maximum load and fatigue strength.
- For frame fatigue damage, if, despite continued resampling of the next MMH-MCMC coordinate, the next coordinate continues to lie outside the feasible domain, the coordinate value is truncated to fall within the feasible domain. This is done after one-thousand unsuccessful sampling repetitions.
- Coordinates for which it is known that given the sampled value of fatigue strength, and given the sampled values for minimum and maximum load, the frame fatigue damage must be zero because of a particle-swarm optimization algorithm pre-determined that the worst-case load cycle frame damage is zero are set to zero.

Appendix J. Other methods for Virtual Fatigue Life Monitoring

In the framework of the research presented in the present dissertation, two alternative methods to Direct- and Probabilistic Load & Damage Modelling have been investigated or newly developed as well. Both models are especially aimed at real-time application and to feature low computational costs. The two methods that are introduced and tested in the following sections are Design Spectrum Discretization (DSD) and Top-of-Scatter Load Block Modelling (ToSLBM). Application of both DSD and ToSLBM on real data, i.e. using the same datasets as in chapter 4, did not demonstrate sufficient performance to prioritize their further development in favor of probabilistic load and damage modelling. The results suggest that these models are not suitable for the situations of complex helicopter fatigue loading that they are tested on.

J.1 Design Spectrum Discretization

Previous work by Trigui [142] and Wurzel targets to simplify the fatigue life prediction process by reducing the number of flight regimes that must be taken into account. The work then continues to focus on the recognition of only these few but highly relevant flight regimes by simple means, i.e. by using compact, shallow, and binary decision trees. Based on this approach, present work postulates that only a few flight regimes in the entire Design Mission Profile (DMP) are relevant to substantiate fatigue life. In addition, present work also assumes that these regimes can be-recognized by simple design logic, such as binary decision trees [106].

Figure J.1 shows the minimum and maximum loads that occur during all the flight regimes that the DMP foresees for the casing of a hydraulic control actuator for the main rotor. From the figure, it can be seen that indeed only a limited number of flight regimes cause very high loads. The DMP for this component consist of a combination of three load sequences: a normal load sequence, a medium load sequence, and a severe load sequence. The figure marks the maximum and minimum load that occurs during these load sequences with dotted horizontal lines. The flight regimes that DSD can monitor using the decision logic developed by Trigui are color-coded in red and with filled dots at the extremes of their load ranges. Since DSD monitors the occurrence of these regimes, they can be removed from the baseline fatigue life prediction that still makes use of a DMP. The figure shows that by applying DSD, the most extreme loads can indeed be removed from a DMP-based load spectrum, and are only taken into account if their occurrence is actually detected. By avoiding the occurrence of these high-load flight regimes, the life of the component can thus be extended significantly.

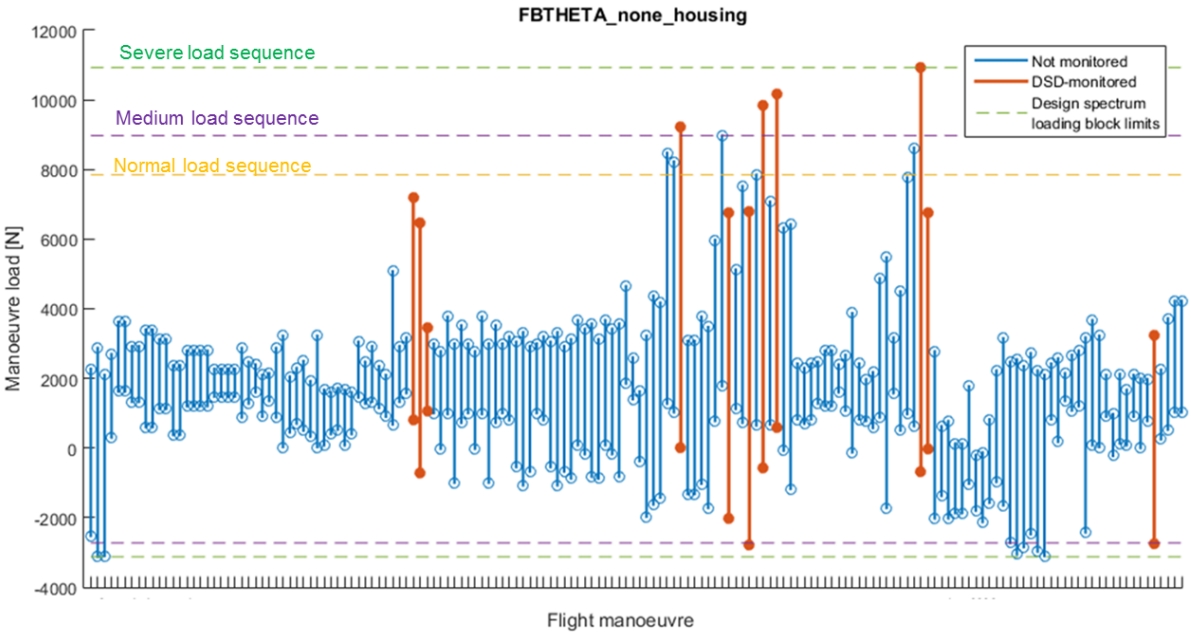


Figure J.1: Overview of maximum and minimum loads for all flights regimes observed during load classification flights.

Just as for Direct Load & Damage Modelling (DLDM) in chapter 4, DSD assumes that all reliability can be substantiated by the reliability of the working S-N curve. Therefore, all damage values are computed using the working S-N curve. Also, as for DLDM as well, it is assumed that average (extreme) loads may be used and that the effects of statistical scatter on regime loads have a negligible effect on overall reliability.

The DMP of the helicopter components listed in section 4.2.3 is a combination of three standard sequences: a normal, medium, and severe sequence of flight regimes. A severe sequence of flight regimes in Figure J.2 shows that for the exemplified component non-DSD monitored flight regimes still considerably contribute to the low-frequency load spectrum. Since DSD modelling is intended for highly resource constrained and real-time application, it is not possible to store a long regime recognition sequence in a computer memory and post-process the result to compute an actual extreme load curve.

Thus, if a flight does not contain any DSD-monitored regimes, then the flight is assumed to have caused damage according to the DMP with all DSD-monitored high-load events removed. If however, one or more DSD-monitored regimes are detected, then the entire flight is assumed to have caused damage equivalent to the DMP sub-sequence containing most severe detected DSD-monitored flight regime.

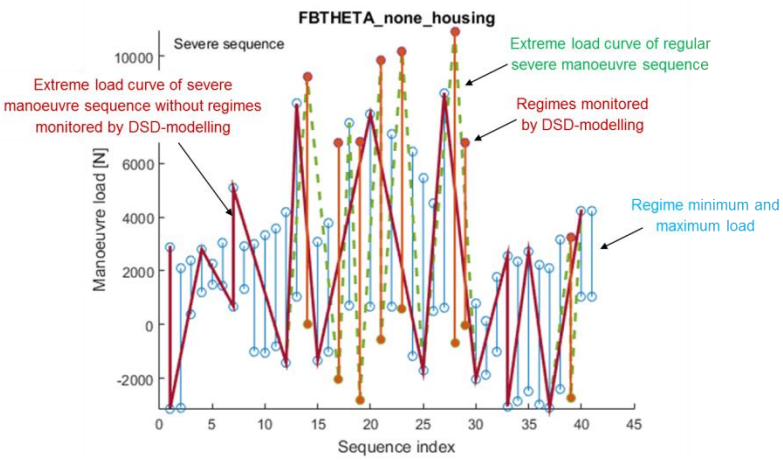


Figure J.2: Low-frequency load spectrum of severe flight regime sequence for classic fatigue life prediction with and without DSD-monitored flight regimes.

Since DSD only monitors a limited amount of flight regime and models the accumulation of fatigue damage due to all other regimes classically and according to a conservative Design Mission Profile. This means that the ideal SLL-extensions that can be substantiated by DSD-modelling are limited. As an example, in Figure J.3 it is demonstrated that the ideal SLL-extension that DSD can justify is capped at 12.9% because flight regimes not covered by DSD still cause significant fatigue damage according to the DMP.[40]

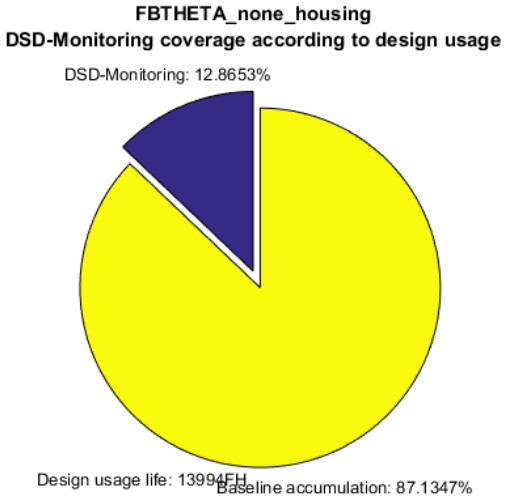


Figure J.3: Pie chart showing the proportion fatigue damage that is caused by DSD-monitored flight regimes, given the DMP and for the casing of a hydraulic actuator for a main rotor control rod.

In general, DSD assumes that the DMP is applicable to short time durations as well and that all SLL reliability can be substantiated by the working S-N curve. Both of these assumptions may not hold though, as follows from the irregular damage accumulation observed in Figure 4.22 and the reliability testing of Direct Load & Damage Modelling in section 4.2.6. Although it could be envisioned that DSD could be applied successfully to components whose SLL is governed by few and easy-to-recognize regimes, the method is not generally applicable. Moreover, additional the development of a reliability substantiation model and testing is required.

[40] Unpublished application tests revealed that practical application of DSD monitoring to all other components listed in section 4.2.3 did not yield encouraging results either, and showed negligible potential for the extension of SLLs.

J.2 Top-of-Scatter Load Block Modelling

The primary reason why DSD must assign a single damage value to an entire flight for low-frequency damage is that the low-frequency extreme load curve can potentially be formed by connecting load extremes with long time-wise separation. Since DSD aims to prevent having to perform in-flight and real-time rainflow counting, the entire flight is thus effectively assigned a DMP-based conservative low-frequency load spectrum.

Top-of-Scatter Load Block Modelling (ToSLBM) instead assigns a damage value to each individual timeframe in real time and models the damage of an entire flight by a running sum of these timeframe damage values, thus leading to an algorithm that can be implemented while using very small computer memory.

As for DSD, ToSLBM's practical value relies on the assumption that only few and rare load events determine a major part of a component's fatigue life. In another similarity to DSD, ToSLBM computes all damage values according to the working S-N curve. However, in ToSLBM's case, it is not necessary to assume that reliability influences of load scatter can be neglected. Assuming a large and representative test set, the damage is always computed according to the worst-case loads.

ToSLBM classifies each timeframe to one of a number of load classes and assigns damage according to the worst-case low-frequency load cycle the timeframe could be part of and the worst-case high-frequency damage that could occur for the class. As illustrated in Figure J.4, the minimum and maximum load of the worst load class encompasses the next most conservative load class, and so on. The number of classes is the result of an optimization algorithm trying to minimize the number of discretized classes as well as unnecessary conservatism. Using this staged approach to classifying the loads in a timeframe, one can be sure that an extreme load curve of one cycle, bounded by the class' minimum and maximum load, conservatively accounts for any extreme load curve the timeframe could be part of.

It must be noted that the most conservative load cycle encompassed by a minimum and maximum load may not use this minimum load. Due to the non-linear nature of the fatigue surface, a reduction in load cycle amplitude but an increase in mean load may result in higher damage than a load cycle through the class's minimum and maximum load. The practical implementation of ToSLBM in present work uses a particle swarm optimization algorithm to ensure that the actual worst-case load cycle is computed.

As illustrated in Figure J.6, loading blocks of increasing severity correspond to increasingly severe damage. And as demonstrated by Figure J.5, the occurrence of high-damage timeframes is indeed rare. However, in the illustrated case, the damage assumed for the least-damaging class is less than two orders of magnitude smaller than the most severe one. Since every frame must be assigned a damage value, the application of ToSLBM can is this case result in significant overestimation of accumulated fatigue damage. For ToSLBM to work in practice, the difference in damage between high and low classes should be as high as possible. Ideally, most timeframes can be assigned a zero-damage value.

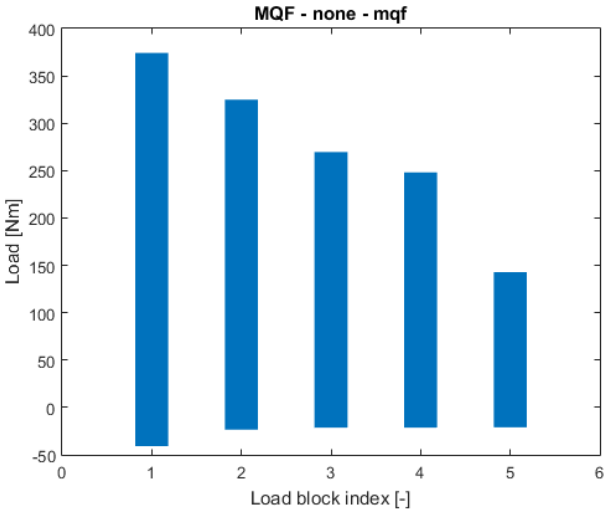


Figure J.4: Top-of-Scatter Load Block Modelling loading blocks for component-6.

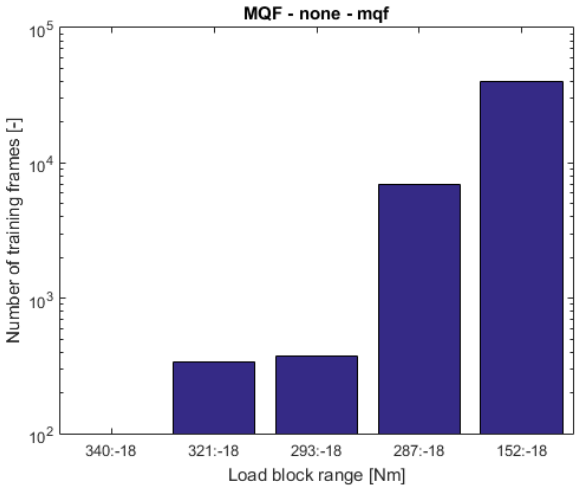


Figure J.5: Relative occurrence of load classes in a training set extracted from Load Classification Flights and used to generate the Top-of-Scatter Load Block Model..

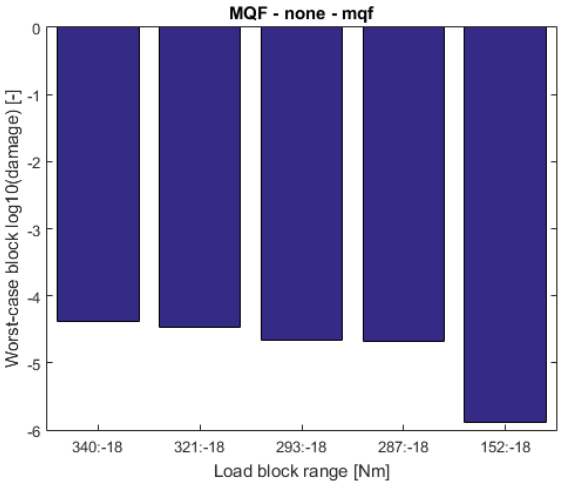


Figure J.6: Worst-case low-frequency damage that can be caused by a load block.

To make sure that the effect of classification errors can be neglected, the load class boundaries are enlarged after the machine learning process to generate the ToSLBM classifier has finished. Widening of the classes is done such that they encompass the worst-case test errors. This method is illustrated in Figure J.7. Colour coding shows how the ToSLBM classifies divides timeframes from a semi-independent test set into 5 load classes. Solid lines show the load boundaries that ToSLBM determined for the defined classes during the training phase. Dashed lines show how these boundaries of these classes are enlarged after testing with an independent test set revealed that load cases of timeframes from the test set fell outside of the original load boundaries defined for the class they were classified in.

For the practical test conducted in the framework of present work, classification is done by automatically generated binary decision trees. The use of binary decision trees is due to their explicit nature and relative ease of testing and certification for high Design Assurance Levels according to RTCA DO-178. However, other classification methods can be used as well and may yield better and more discretized performance. The present implementation features considerable recognition errors, as demonstrated in Figure J.7 and Figure J.8.

As was the case for DSD, unpublished application tests revealed that practical application of ToSLBM to the components listed in section 4.2.3 did not yield encouraging results, with negligible potential for SLL limit extension.

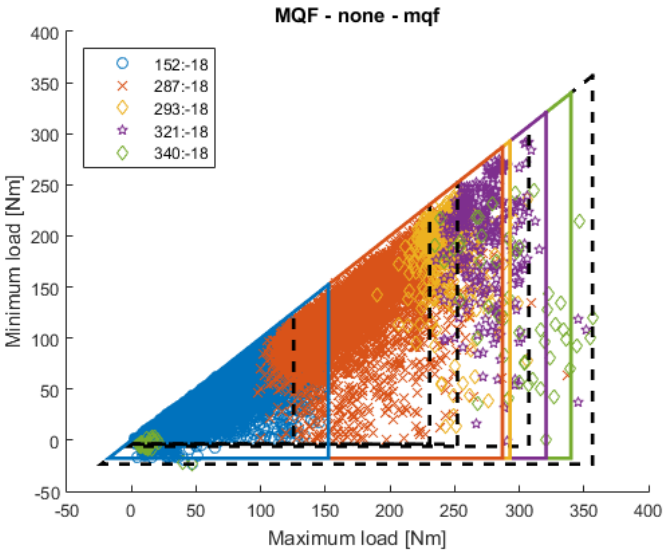


Figure J.7: Thick black dashed lines show load-block limits observed during testing with a semi-independent dataset set aside for testing.

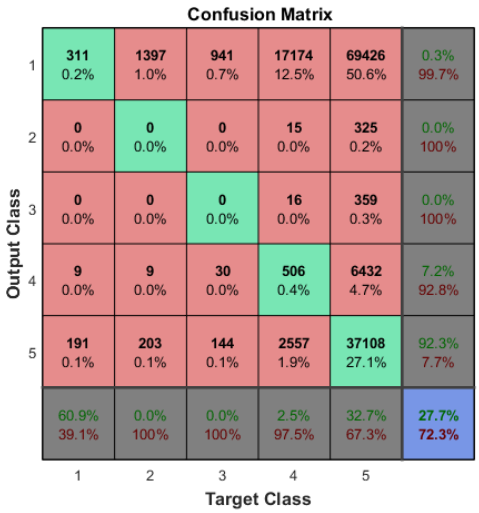


Figure J.8: Confusion of recognized classes according to a semi-independent test set from an LCF dataset.

.

Appendix K. Model generation for additional components

This appendix contains test results for DLDM regression models for components 1-5, and 7, as listed in Table 4-2 and demonstrates that DLDM regression can be achieved accurately for all these components. This indicates the general applicability of the developed PLDM prediction models, which is illustrated in section K.2. The conducted tests are the same as the test results presented and discussed in sections 4.2.5 and 4.3.3 for the main gearbox casing loaded by tail rotor driveshaft torque.

K.1 Direct Load & Damage Models

K.1.1 FBTHETA

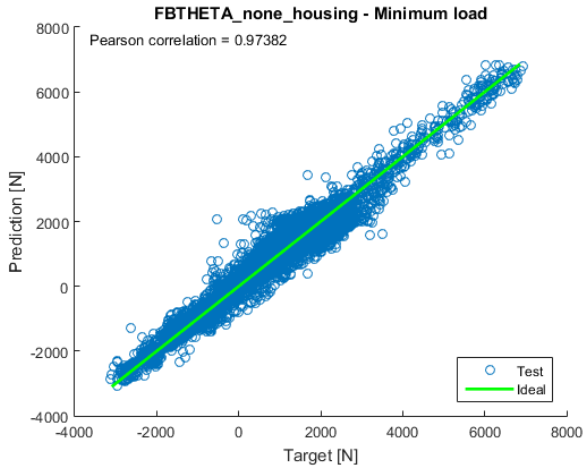


Figure K.1: Regression plot showing the correlation between predicted minimum load and the actually measured minimum load during timeframes in the portion of LCF flight test data set aside for model testing. The predictions are made by a shallow Artificial Neural Network.

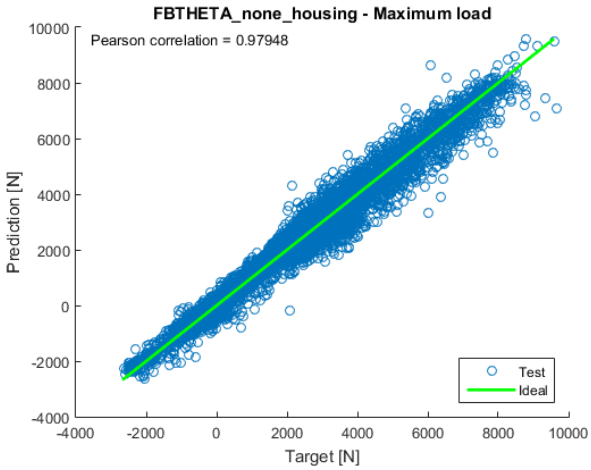


Figure K.2: Regression plot showing the correlation between predicted maximum load and the actually measured maximum load during timeframes in the portion of LCF flight test data set aside for model testing. The predictions are made by a shallow Artificial Neural Network.

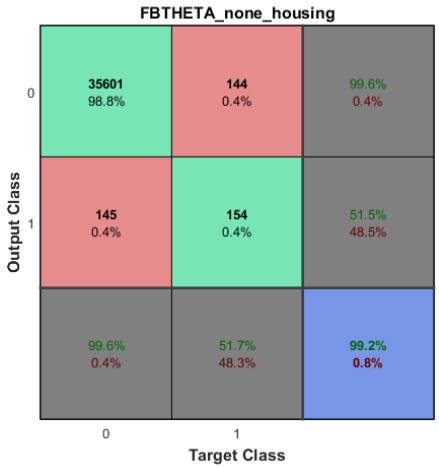


Figure K.3: Confusion matrix showing the accuracy of timeframe damage classification using an RVM classifier and for LCF data set aside for testing.

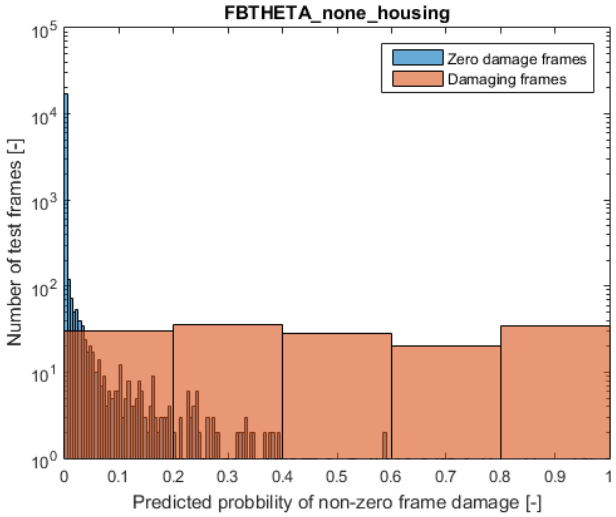


Figure K.4: Chart showing the distribution of predicted timeframe damage probabilities and a comparison with the actual occurrence of timeframe damage. The predictions are made with an RVM classifier for LCF data set aside for model testing.

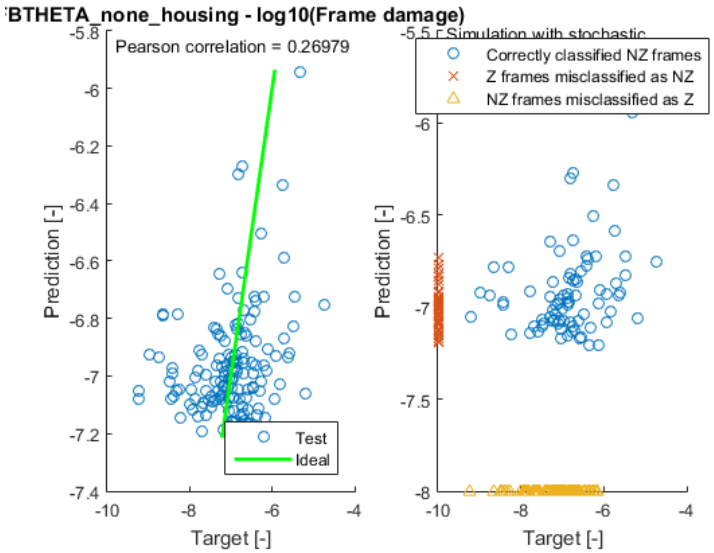


Figure K.5: Regression plot showing the correlation between predicted and actually measured timeframe damage for correctly classified damaging timeframes in the LCF test set. (left)
Regression plot showing the correlation between predicted timeframe damage and actually measured timeframe damage for all timeframes classified as damaging in the LCF test set. (including incorrectly classified timeframes)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes). (right)

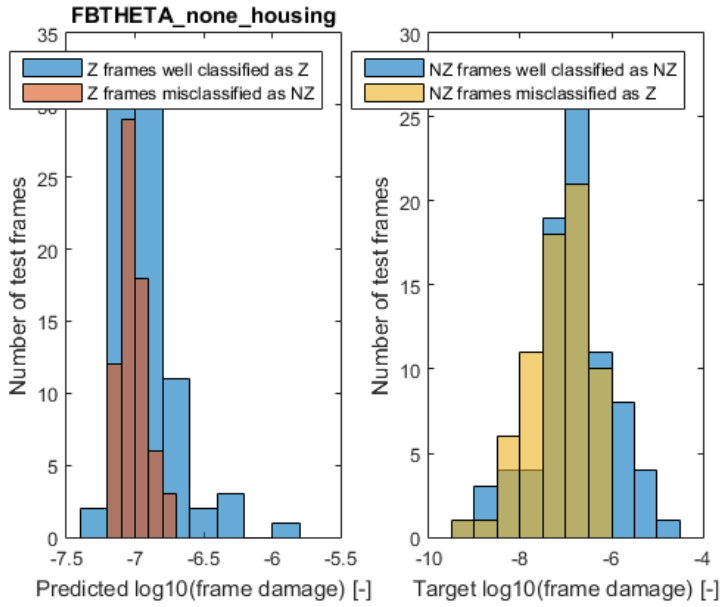


Figure K.6: Regression plot showing the distribution of predicted timeframe damage for frames misclassified as damaging or correctly classified as not damaging in LCF test set (left)
Regression plot showing the distribution of the true timeframe damage of non-damaging timeframes that are misclassified as damaging and damaging timeframes that are misclassified as non-damaging in the LCF test set (right)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes)

K.1.2 FBTHETAP

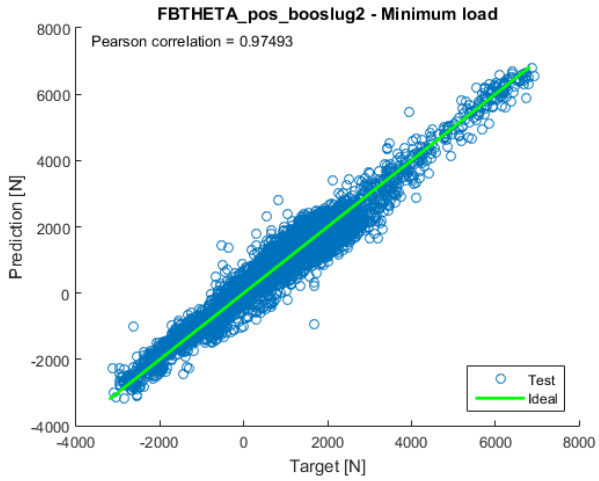


Figure K.7: Regression plot showing the correlation between predicted minimum load and the actually measured minimum load during timeframes in the portion of LCF flight test data set aside for model testing. The predictions are made by a shallow Artificial Neural Network.

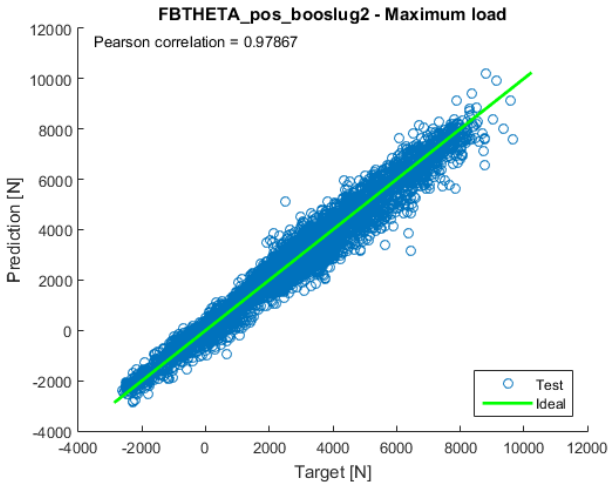


Figure K.8: Regression plot showing the correlation between predicted minimum load and the actually measured minimum load during timeframes in the portion of LCF flight test data set aside for model testing. The predictions are made by a shallow Artificial Neural Network.

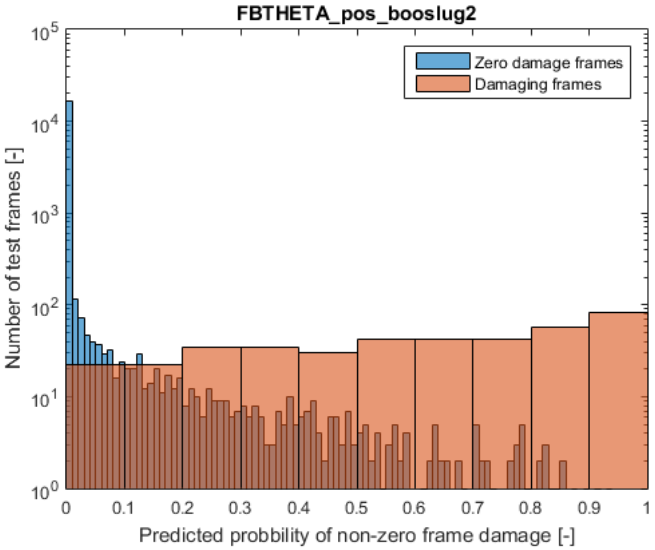


Figure K.9: Chart showing the distribution of predicted timeframe damage probabilities and a comparison with the actual occurrence of timeframe damage. The predictions are made with an RVM classifier for LCF data set aside for model testing.

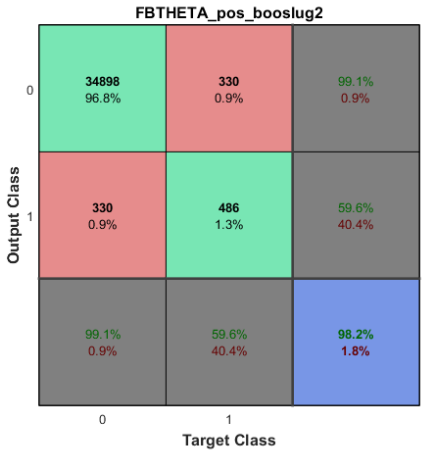


Figure K.10: Confusion matrix showing the accuracy of timeframe damage classification using an RVM classifier and for LCF data set aside for testing.

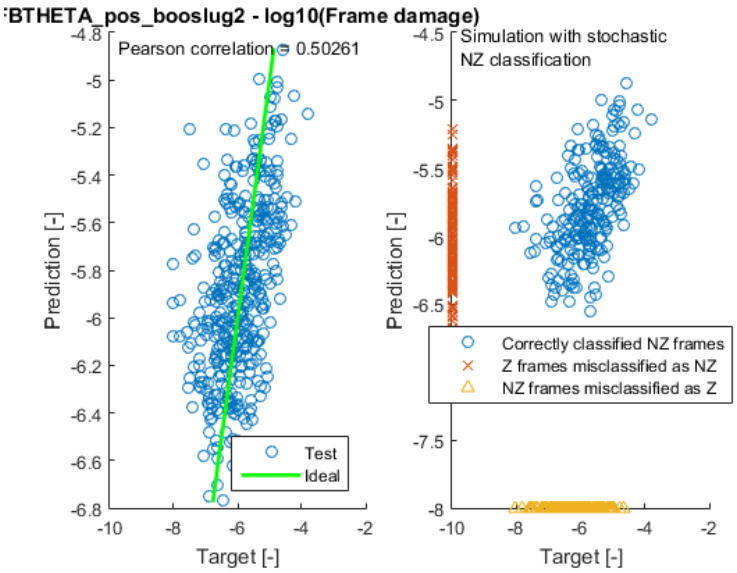


Figure K.11: Regression plot showing the correlation between predicted and actually measured timeframe damage for correctly classified damaging timeframes in the LCF test set. (left)
Regression plot showing the correlation between predicted timeframe damage and actually measured timeframe damage for all timeframes classified as damaging in the LCF test set. (including incorrectly classified timeframes)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes). (right)

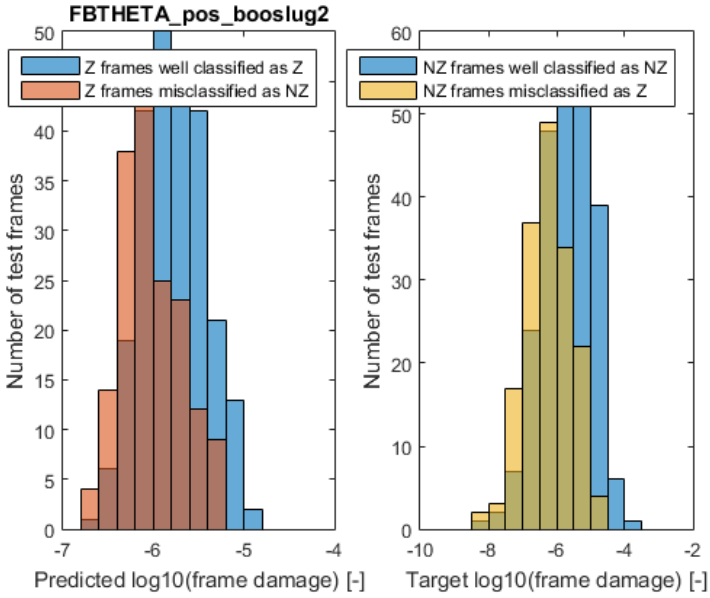


Figure K.12: Regression plot showing the distribution of predicted timeframe damage for frames misclassified as damaging or correctly classified as not damaging in LCF test set (left)
Regression plot showing the distribution of the true timeframe damage of non-damaging timeframes that are misclassified as damaging and damaging timeframes that are misclassified as non-damaging in the LCF test set (right)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes)

K.1.3 FKAR

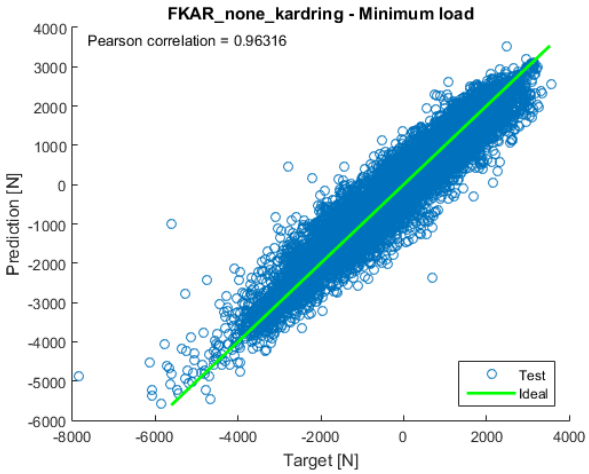


Figure K.13: Regression plot showing the correlation between predicted minimum load and the actually measured minimum load during timeframes in the portion of LCF flight test data set aside for model testing. The predictions are made by a shallow Artificial Neural Network.

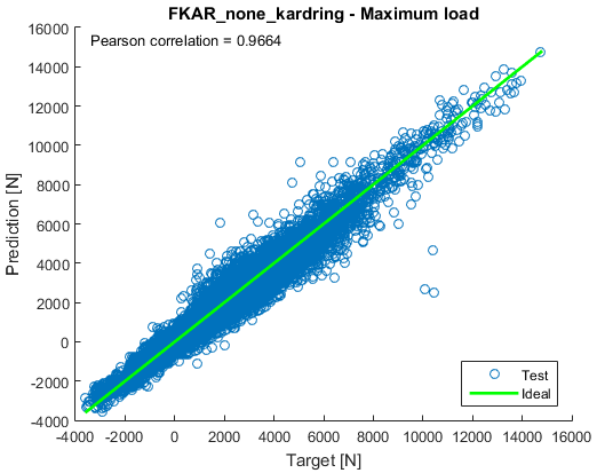


Figure K.14: Regression plot showing the correlation between predicted minimum load and the actually measured minimum load during timeframes in the portion of LCF flight test data set aside for model testing. The predictions are made by a shallow Artificial Neural Network.

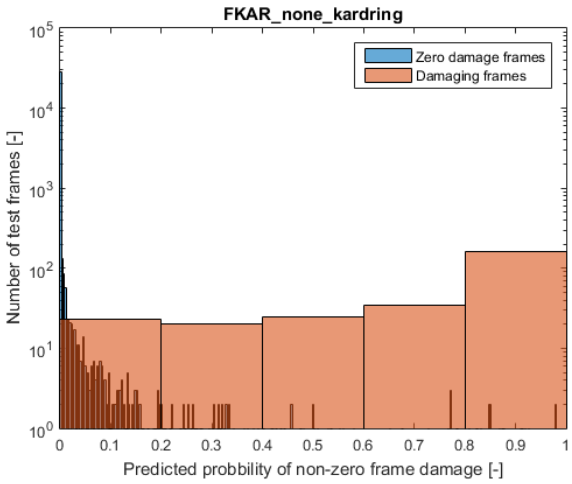


Figure K.15: Chart showing the distribution of predicted timeframe damage probabilities and a comparison with the actual occurrence of timeframe damage. The predictions are made with an RVM classifier for LCF data set aside for model testing.

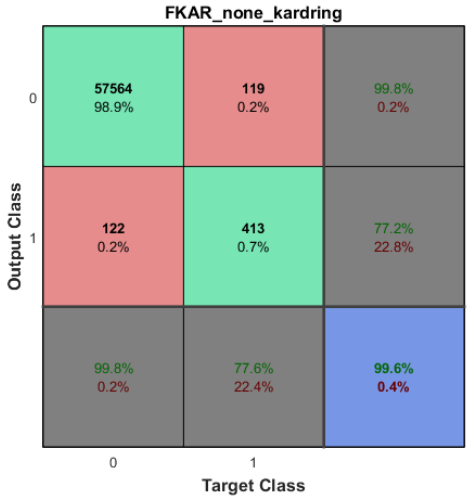


Figure K.16: Confusion matrix showing the accuracy of timeframe damage classification using an RVM classifier and for LCF data set aside for testing.

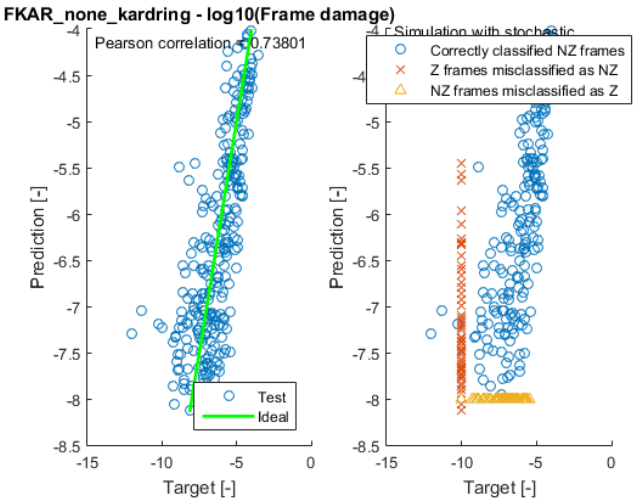


Figure K.17: Regression plot showing the correlation between predicted and actually measured timeframe damage for correctly classified damaging timeframes in the LCF test set. (left)
Regression plot showing the correlation between predicted timeframe damage and actually measured timeframe damage for all timeframes classified as damaging in the LCF test set. (including incorrectly classified timeframes)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes). (right)

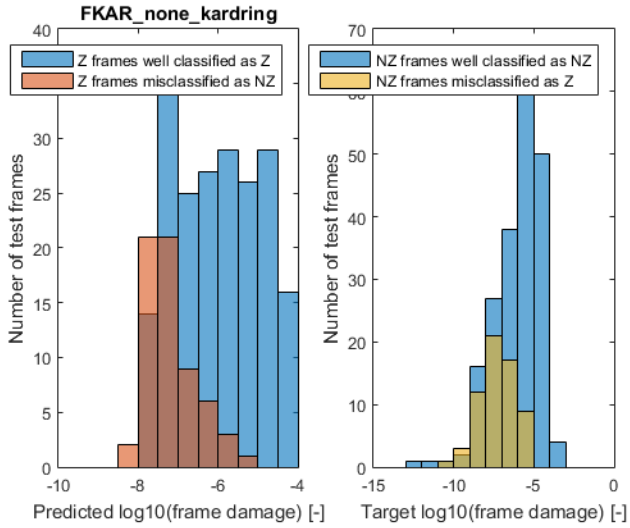


Figure K.18: Regression plot showing the distribution of predicted timeframe damage for frames misclassified as damaging or correctly classified as not damaging in LCF test set (left)
Regression plot showing the distribution of the true timeframe damage of non-damaging timeframes that are misclassified as damaging and damaging timeframes that are misclassified as non-damaging in the LCF test set (right)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes)

K.1.4 FSTA

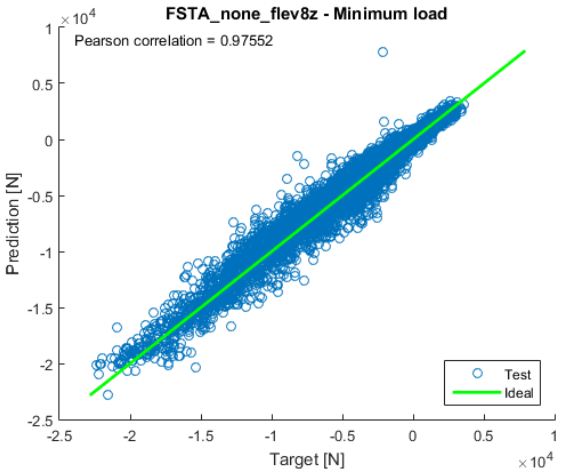


Figure K.19: Regression plot showing the correlation between predicted minimum load and the actually measured minimum load during timeframes in the portion of LCF flight test data set aside for model testing. The predictions are made by a shallow Artificial Neural Network.

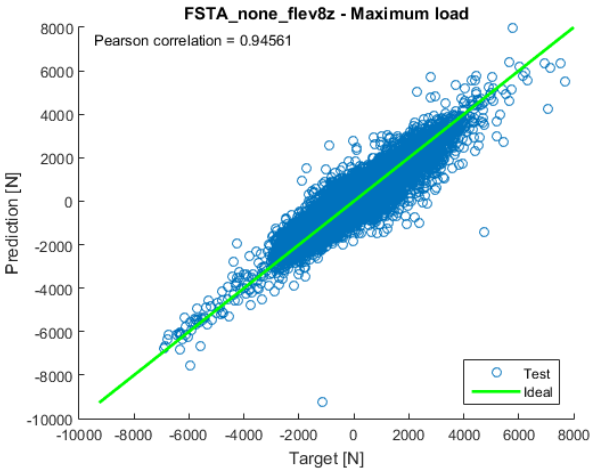


Figure K.20: Regression plot showing the correlation between predicted minimum load and the actually measured minimum load during timeframes in the portion of LCF flight test data set aside for model testing. The predictions are made by a shallow Artificial Neural Network.

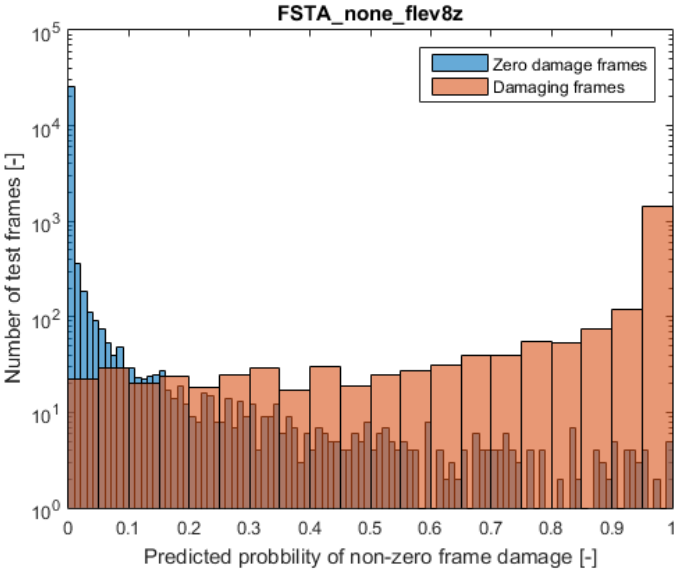


Figure K.21: Chart showing the distribution of predicted timeframe damage probabilities and a comparison with the actual occurrence of timeframe damage. The predictions are made with an RVM classifier for LCF data set aside for model testing.

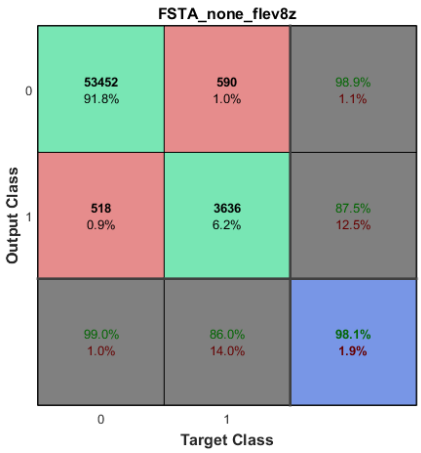


Figure K.22: Confusion matrix showing the accuracy of timeframe damage classification using an RVM classifier and for LCF data set aside for testing.

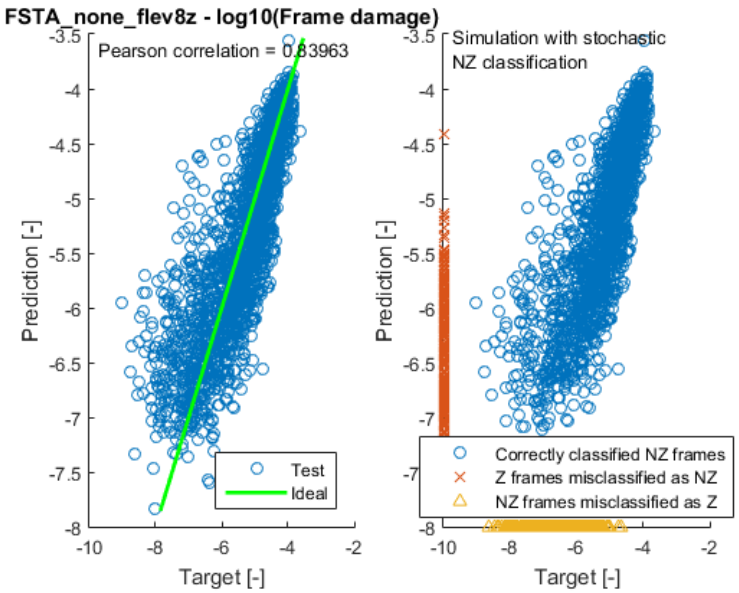


Figure K.23: Regression plot showing the correlation between predicted and actually measured timeframe damage for correctly classified damaging timeframes in the LCF test set. (left)
Regression plot showing the correlation between predicted timeframe damage and actually measured timeframe damage for all timeframes classified as damaging in the LCF test set. (including incorrectly classified timeframes)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes). (right)

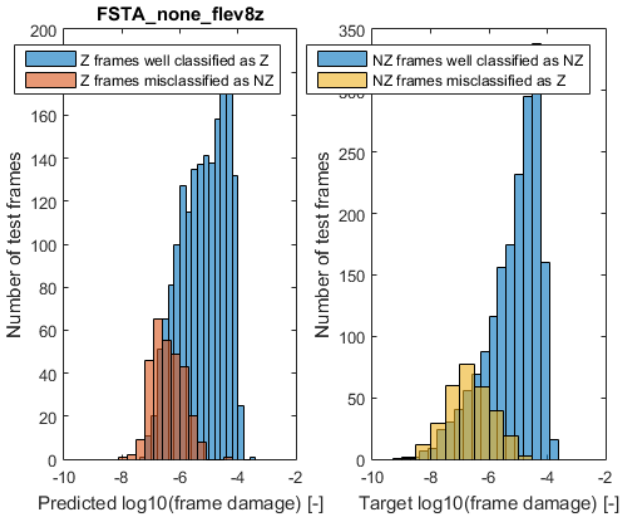


Figure K.24: Regression plot showing the distribution of predicted timeframe damage for frames misclassified as damaging or correctly classified as not damaging in LCF test set (left)
Regression plot showing the distribution of the true timeframe damage of non-damaging timeframes that are misclassified as damaging and damaging timeframes that are misclassified as non-damaging in the LCF test set (right)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes)

K.1.5 FSTY

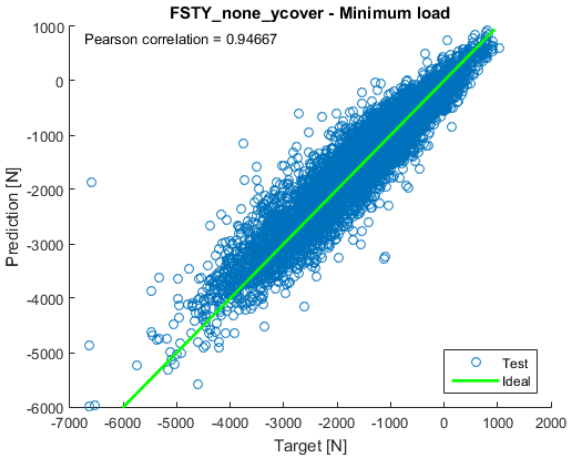


Figure K.25: Regression plot showing the correlation between predicted minimum load and the actually measured minimum load during timeframes in the portion of LCF flight test data set aside for model testing. The predictions are made by a shallow Artificial Neural Network.

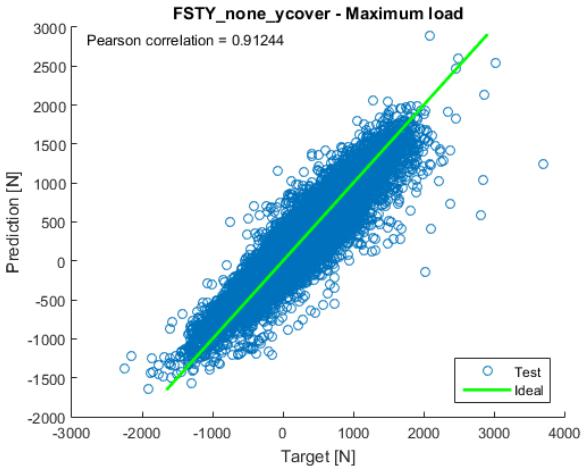


Figure K.26: Regression plot showing the correlation between predicted minimum load and the actually measured minimum load during timeframes in the portion of LCF flight test data set aside for model testing. The predictions are made by a shallow Artificial Neural Network.

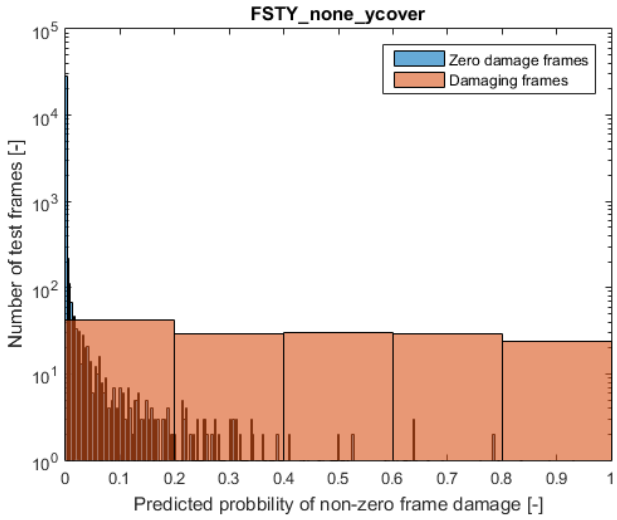


Figure K.27: Chart showing the distribution of predicted timeframe damage probabilities and a comparison with the actual occurrence of timeframe damage. The predictions are made with an RVM classifier for LCF data set aside for model testing.

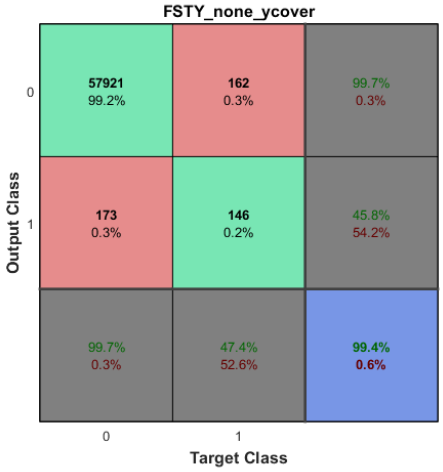


Figure K.28: Confusion matrix showing the accuracy of timeframe damage classification using an RVM classifier and for LCF data set aside for testing.

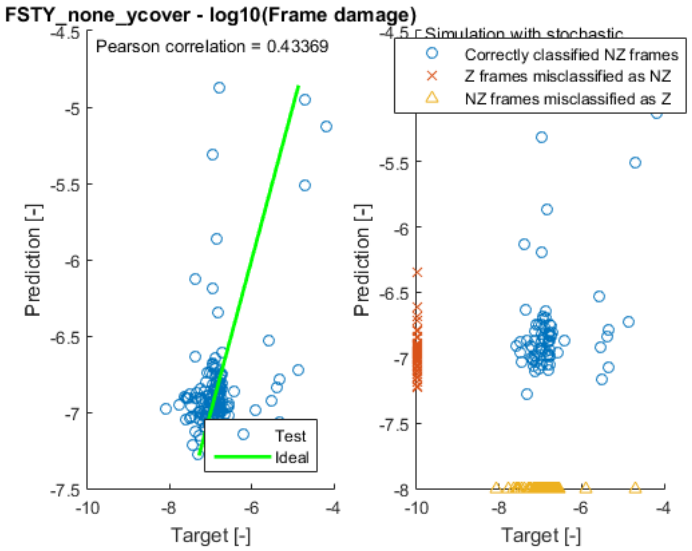


Figure K.29: Regression plot showing the correlation between predicted and actually measured timeframe damage for correctly classified damaging timeframes in the LCF test set. (left)
Regression plot showing the correlation between predicted timeframe damage and actually measured timeframe damage for all timeframes classified as damaging in the LCF test set. (including incorrectly classified timeframes)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes). (right)

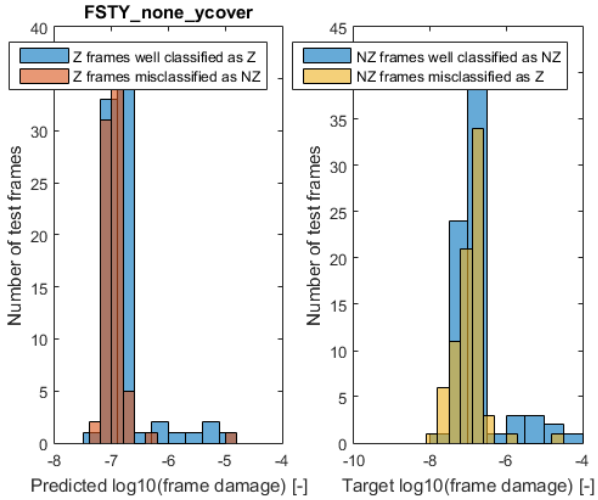


Figure K.30: Regression plot showing the distribution of predicted timeframe damage for frames misclassified as damaging or correctly classified as not damaging in LCF test set (left)
Regression plot showing the distribution of the true timeframe damage of non-damaging timeframes that are misclassified as damaging and damaging timeframes that are misclassified as non-damaging in the LCF test set (right)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes)

K.1.6 MTM

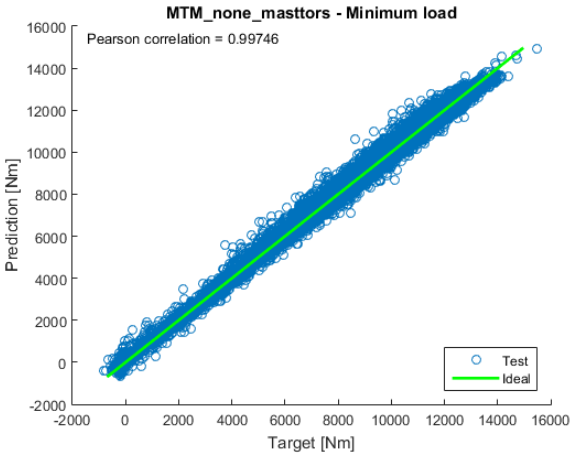


Figure K.31: Regression plot showing the correlation between predicted minimum load and the actually measured minimum load during timeframes in the portion of LCF flight test data set aside for model testing. The predictions are made by a shallow Artificial Neural Network.

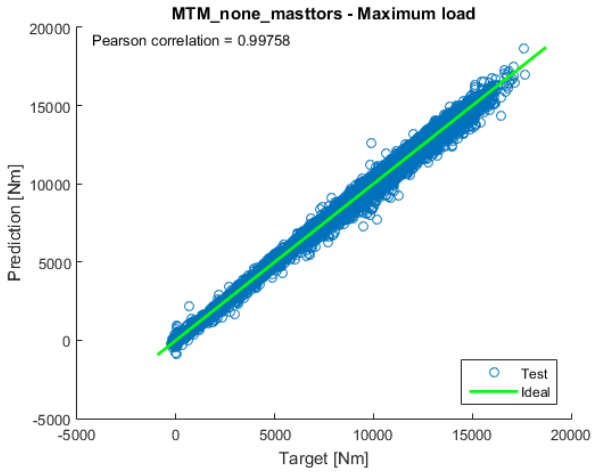


Figure K.32: Regression plot showing the correlation between predicted minimum load and the actually measured minimum load during timeframes in the portion of LCF flight test data set aside for model testing. The predictions are made by a shallow Artificial Neural Network.

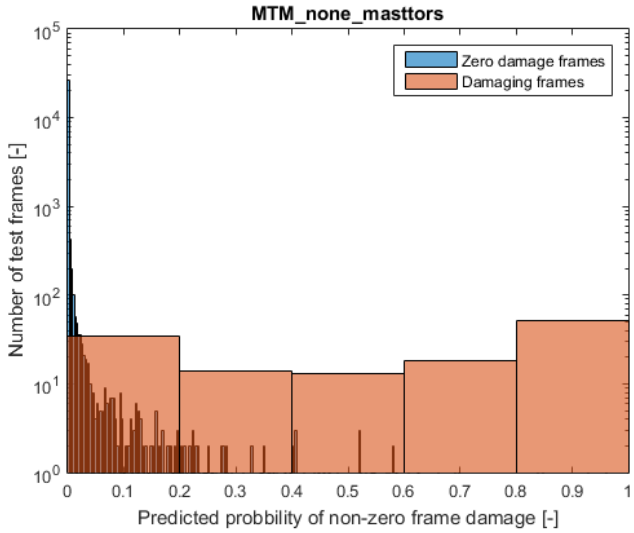


Figure K.33: Chart showing the distribution of predicted timeframe damage probabilities and a comparison with the actual occurrence of timeframe damage. The predictions are made with an RVM classifier for LCF data set aside for model testing.

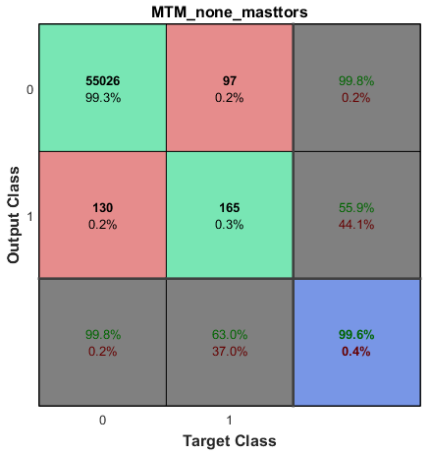


Figure K.34: Confusion matrix showing the accuracy of timeframe damage classification using an RVM classifier and for LCF data set aside for testing.

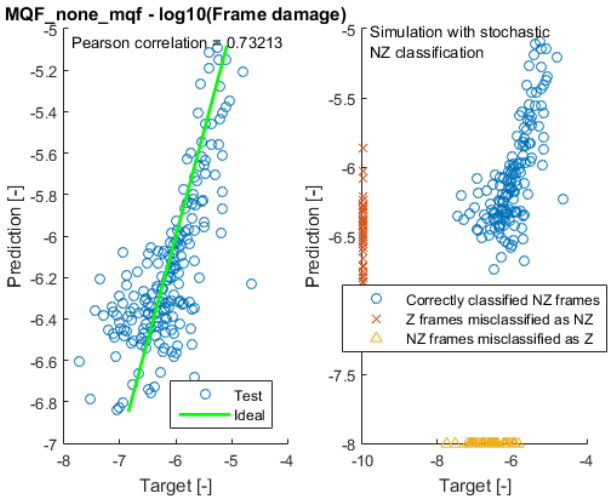


Figure K.35: Regression plot showing the correlation between predicted and actually measured timeframe damage for correctly classified damaging timeframes in the LCF test set. (left)
Regression plot showing the correlation between predicted timeframe damage and actually measured timeframe damage for all timeframes classified as damaging in the LCF test set. (including incorrectly classified timeframes)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes). (right)

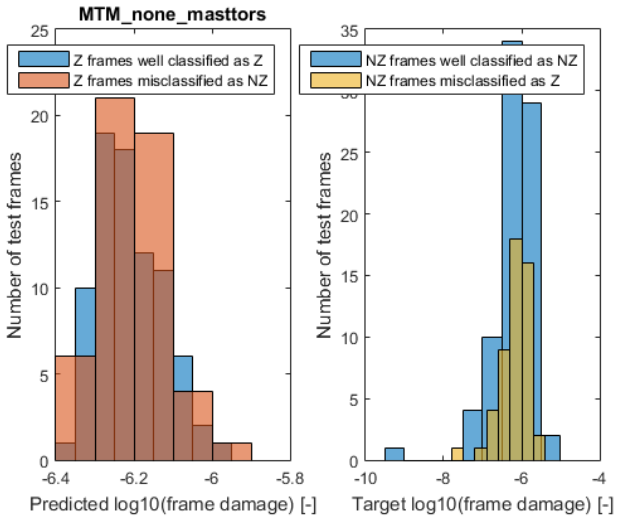


Figure K.36: Regression plot showing the distribution of predicted timeframe damage for frames misclassified as damaging or correctly classified as not damaging in LCF test set (left)
Regression plot showing the distribution of the true timeframe damage of non-damaging timeframes that are misclassified as damaging and damaging timeframes that are misclassified as non-damaging in the LCF test set (right)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes)

K.2 Probabilistic Load & Damage Models

All results given S-N curve with $\gamma=10^{-6}$ reliability, unless mentioned otherwise, and for the same randomly selected bootstrap sample. All tested models are generated with LCF data and all test results are obtained from the portion of LCF data set aside for testing.

K.2.1 FBTHETA

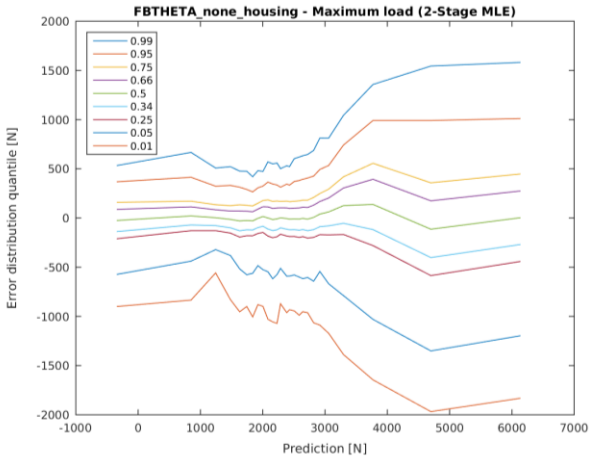


Figure K.1: Graph showing how selected quantiles of a predicted distribution of regression errors for the timeframe maximum load varies with the MLE point prediction.

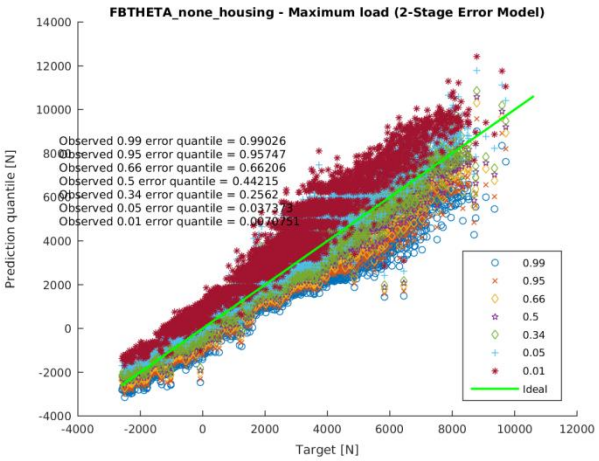


Figure K.2: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe maximum load predictions Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

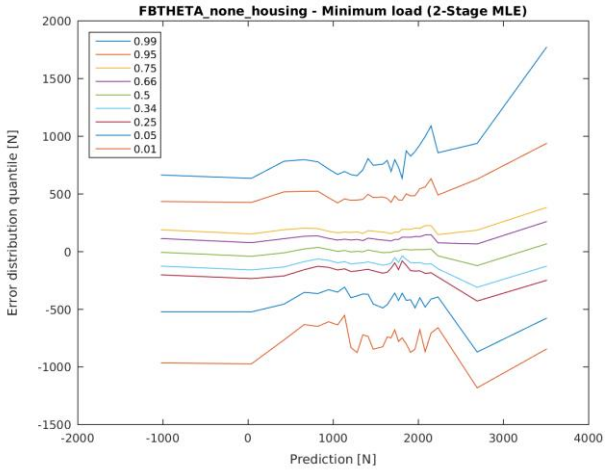


Figure K.3: Graph showing how selected quantiles of a predicted distribution of regression errors for the timeframe minimum load varies with the MLE point prediction.

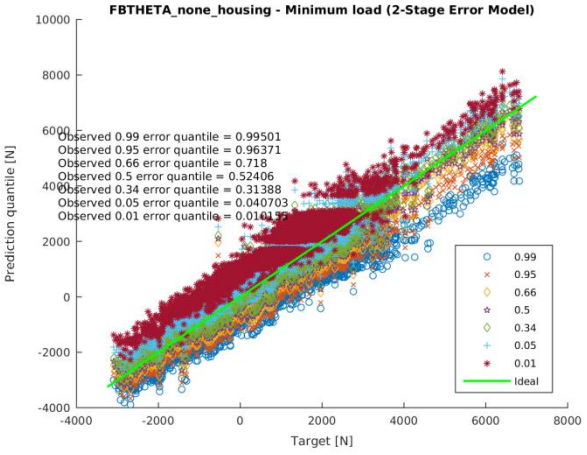


Figure K.4: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe minimum load predictions. Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

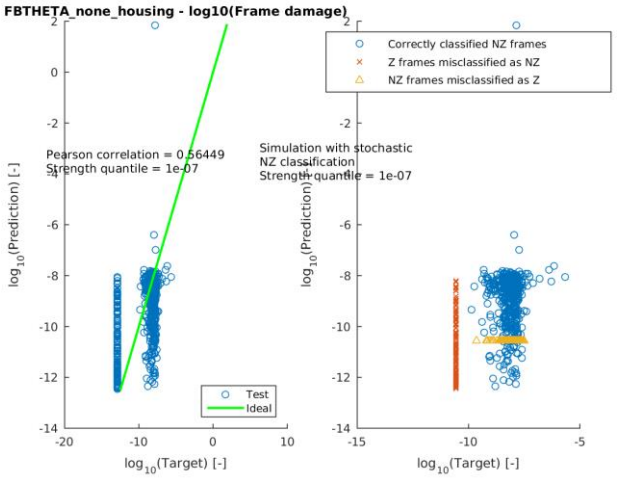


Figure K.5: Regression plot showing the correlation between predicted and actually measured timeframe damage for correctly classified damaging timeframes in the LCF test set. (left)
Regression plot showing the correlation between predicted timeframe damage and actually measured timeframe damage for all timeframes classified as damaging in the LCF test set. (including incorrectly classified timeframes)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes). (right)

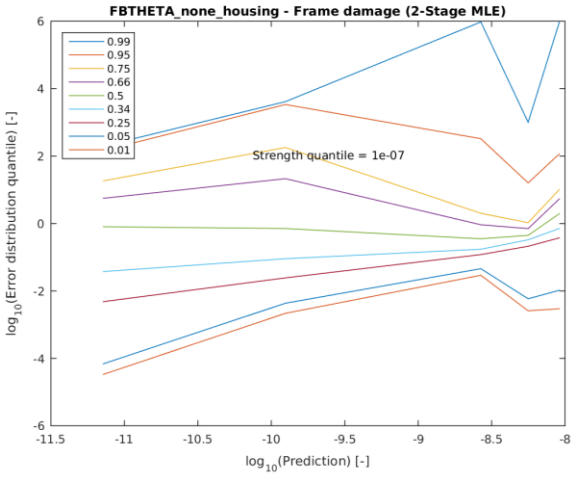


Figure K.6: Graph showing how selected quantiles of a predicted distribution of regression errors for timeframe damage varies with the MLE point prediction.

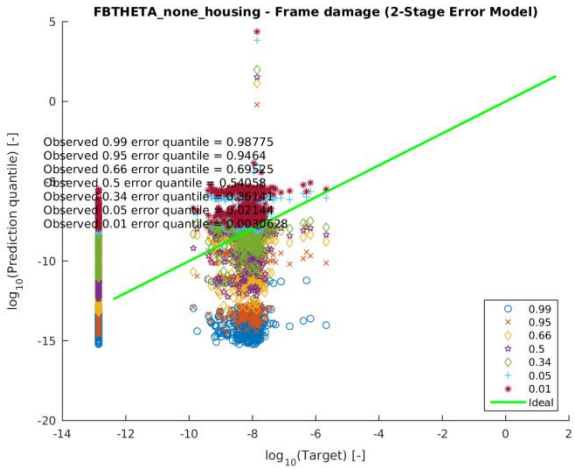


Figure K.7: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe damage predictions. Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

K.2.2 FBTHETAP

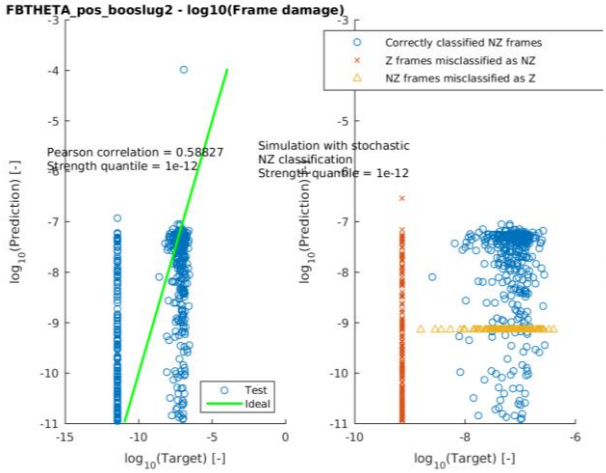


Figure K.8: Regression plot showing the correlation between predicted and actually measured timeframe damage for correctly classified damaging timeframes in the LCF test set. (left)
Regression plot showing the correlation between predicted timeframe damage and actually measured timeframe damage for all timeframes classified as damaging in the LCF test set. (including incorrectly classified timeframes)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes). (right)

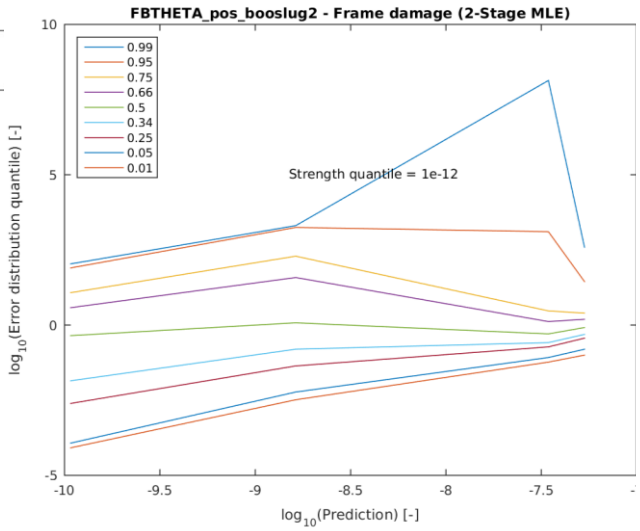


Figure K.9: Graph showing how selected quantiles of a predicted distribution of regression errors for timeframe damage varies with the MLE point prediction.

K.2.3 FKAR

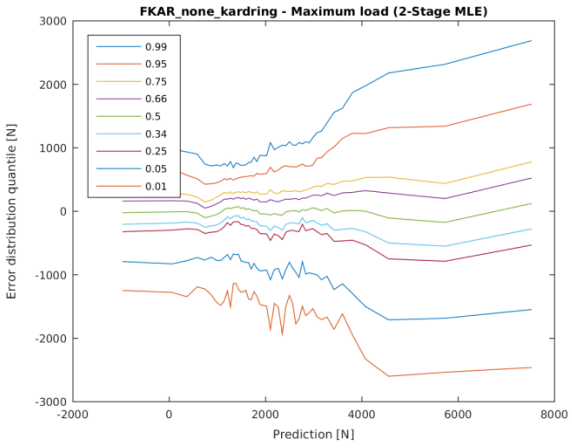


Figure K.10: Graph showing how selected quantiles of a predicted distribution of regression errors for the timeframe maximum load varies with the MLE point prediction.

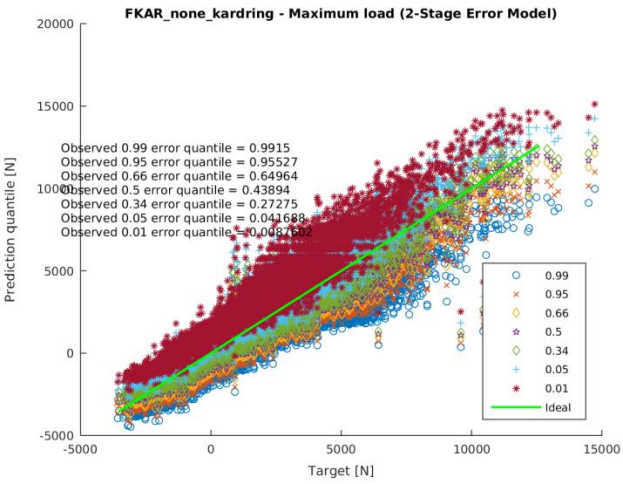


Figure K.11: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe maximum load predictions Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

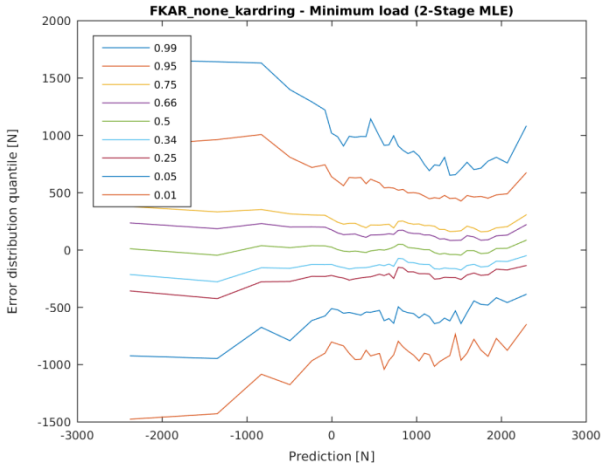


Figure K.12: Graph showing how selected quantiles of a predicted distribution of regression errors for the timeframe minimum load varies with the MLE point prediction.

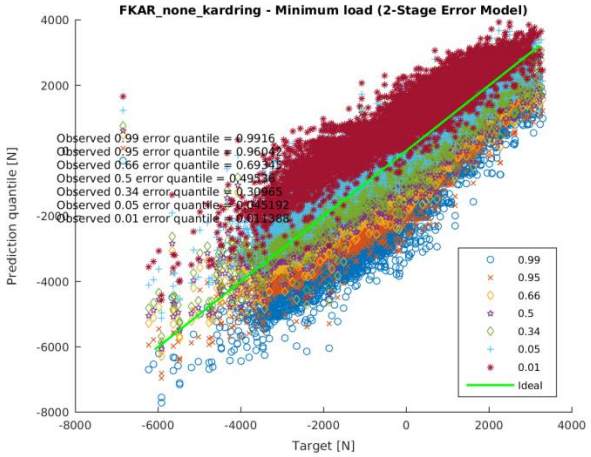


Figure K.13: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe minimum load predictions. Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

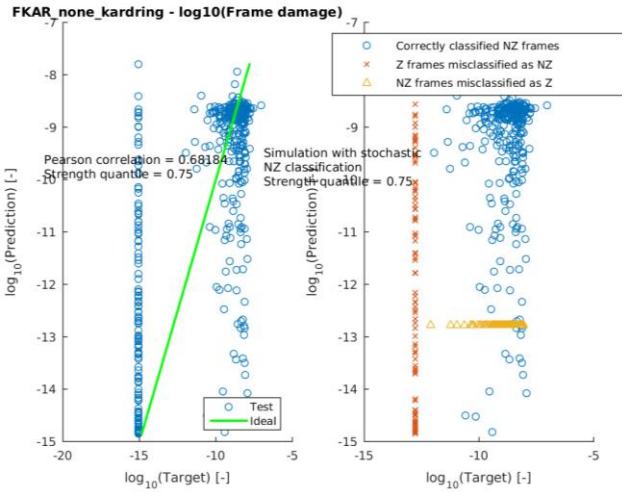


Figure K.14: Regression plot showing the correlation between predicted and actually measured timeframe damage for correctly classified damaging timeframes in the LCF test set. (left)
Regression plot showing the correlation between predicted timeframe damage and actually measured timeframe damage for all timeframes classified as damaging in the LCF test set. (including incorrectly classified timeframes)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes). (right)

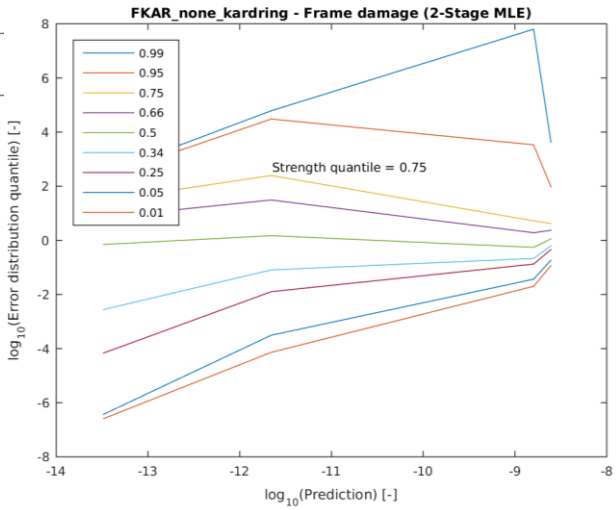


Figure K.15: Graph showing how selected quantiles of a predicted distribution of regression errors for timeframe damage varies with the MLE point prediction.

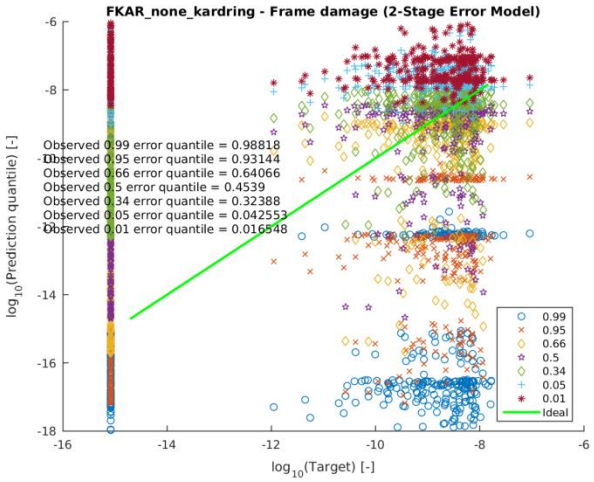


Figure K.16: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe damage predictions. Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

K.2.4 FSTA

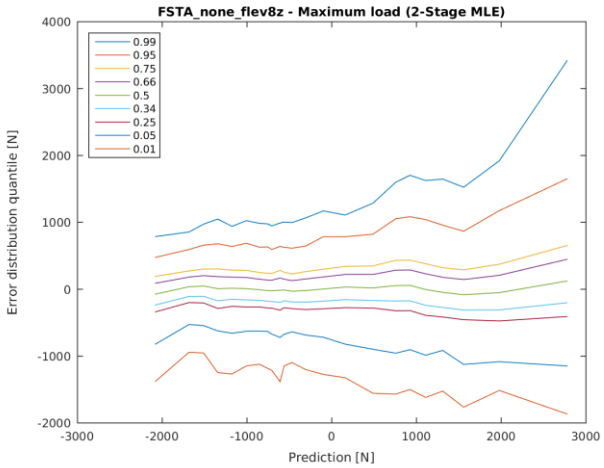


Figure K.17: Graph showing how selected quantiles of a predicted distribution of regression errors for the timeframe maximum load varies with the MLE point prediction.

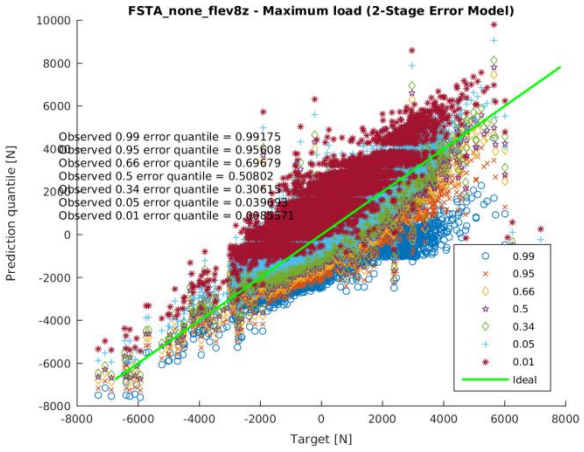


Figure K.18: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe maximum load predictions Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

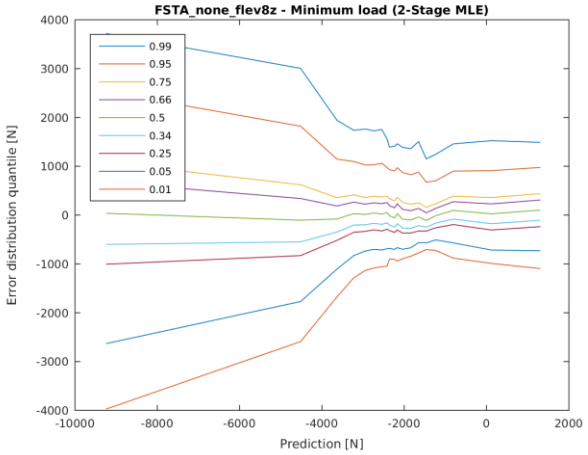


Figure K.19: Graph showing how selected quantiles of a predicted distribution of regression errors for the timeframe minimum load varies with the MLE point prediction.

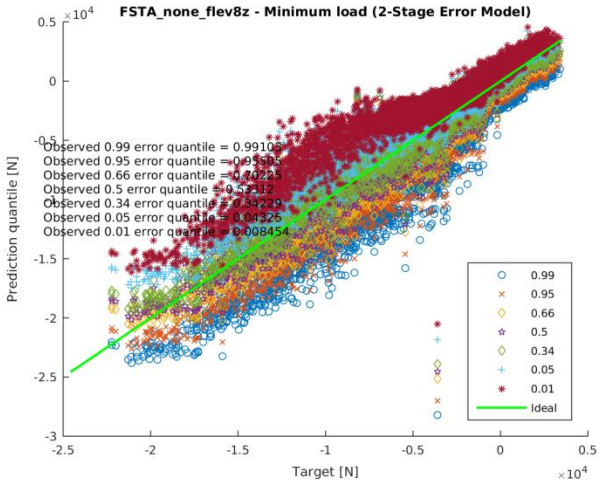


Figure K.20: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe minimum load predictions. Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

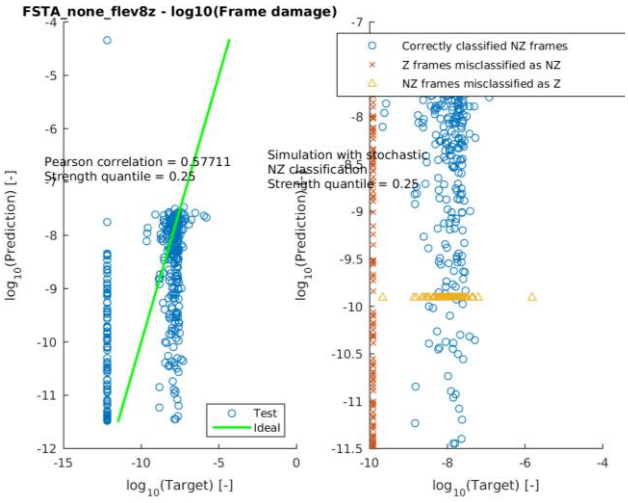


Figure K.21: Regression plot showing the correlation between predicted and actually measured timeframe damage for correctly classified damaging timeframes in the LCF test set. (left)
Regression plot showing the correlation between predicted timeframe damage and actually measured timeframe damage for all timeframes classified as damaging in the LCF test set. (including incorrectly classified timeframes)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes). (right)

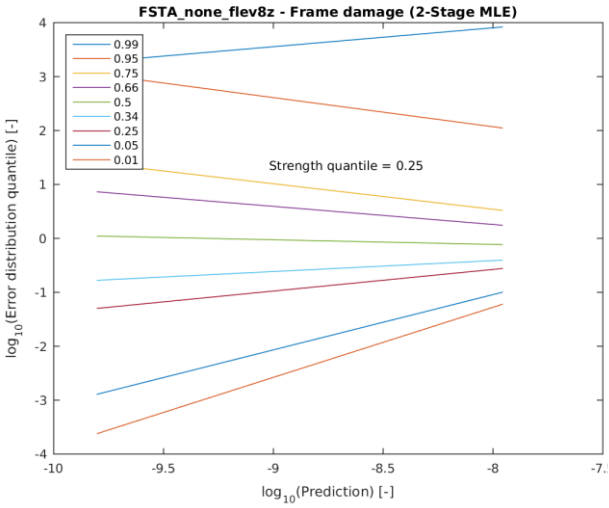


Figure K.22: Graph showing how selected quantiles of a predicted distribution of regression errors for timeframe damage varies with the MLE point prediction.

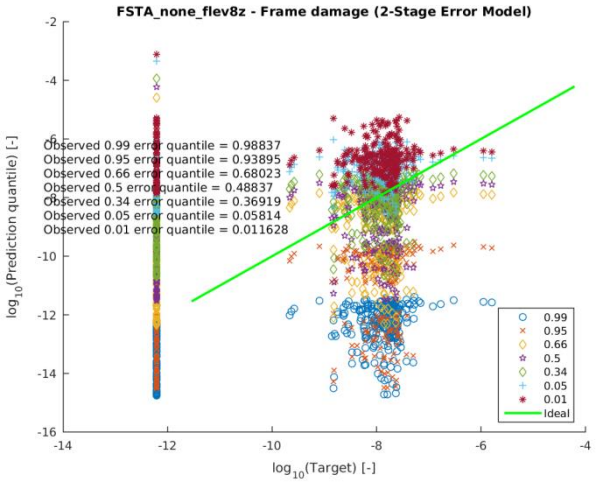


Figure K.23: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe damage predictions. Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

K.2.5 FSTY

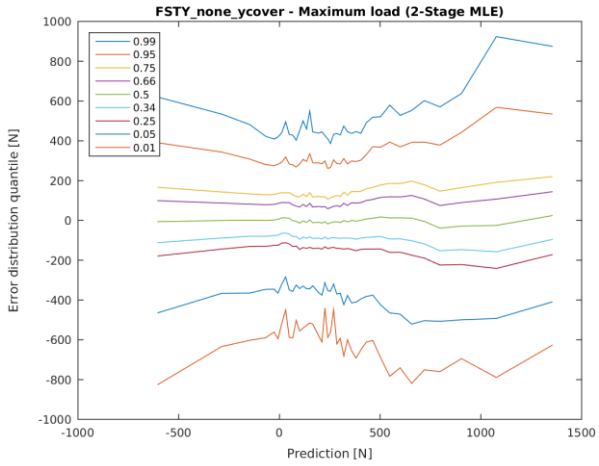


Figure K.24: Graph showing how selected quantiles of a predicted distribution of regression errors for the timeframe maximum load varies with the MLE point prediction.

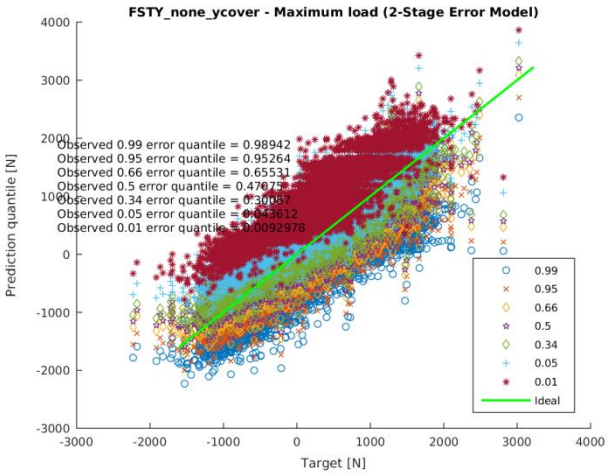


Figure K.25: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe maximum load predictions Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

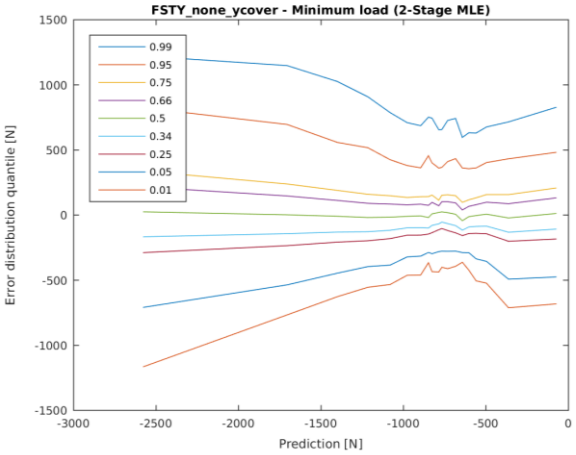


Figure K.26: Graph showing how selected quantiles of a predicted distribution of regression errors for the timeframe minimum load varies with the MLE point prediction.

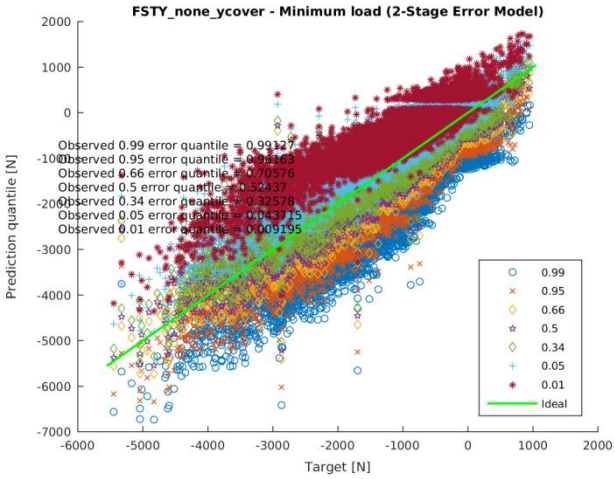


Figure K.27: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe minimum load predictions. Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

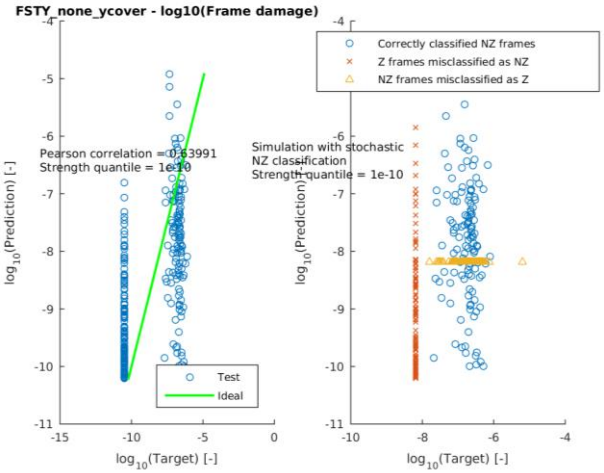


Figure K.28: Regression plot showing the correlation between predicted and actually measured timeframe damage for correctly classified damaging timeframes in the LCF test set. (left)
Regression plot showing the correlation between predicted timeframe damage and actually measured timeframe damage for all timeframes classified as damaging in the LCF test set. (including incorrectly classified timeframes)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes). (right)

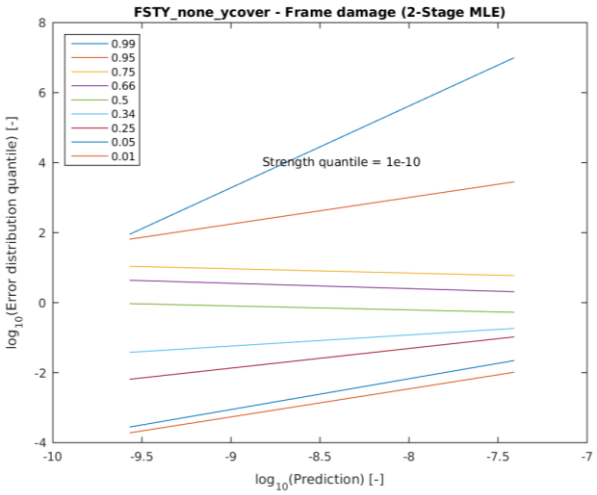


Figure K.29: Graph showing how selected quantiles of a predicted distribution of regression errors for timeframe damage varies with the MLE point prediction.

K.2.6 MQF

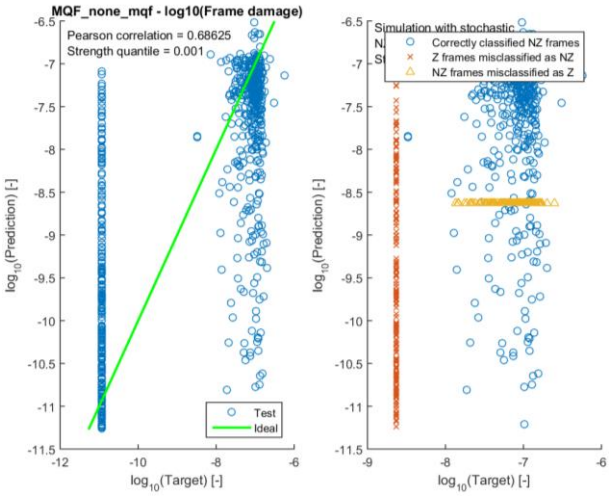


Figure K.30: Regression plot showing the correlation between predicted and actually measured timeframe damage for correctly classified damaging timeframes in the LCF test set. (left)
Regression plot showing the correlation between predicted timeframe damage and actually measured timeframe damage for all timeframes classified as damaging in the LCF test set. (including incorrectly classified timeframes)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes). (right)

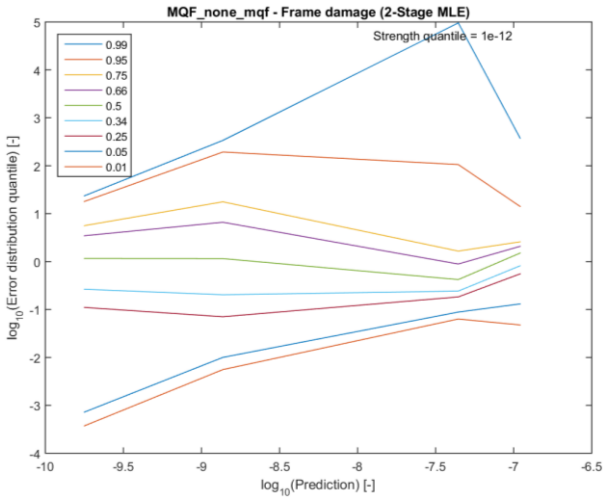


Figure K.31: Graph showing how selected quantiles of a predicted distribution of regression errors for timeframe damage varies with the MLE point prediction.

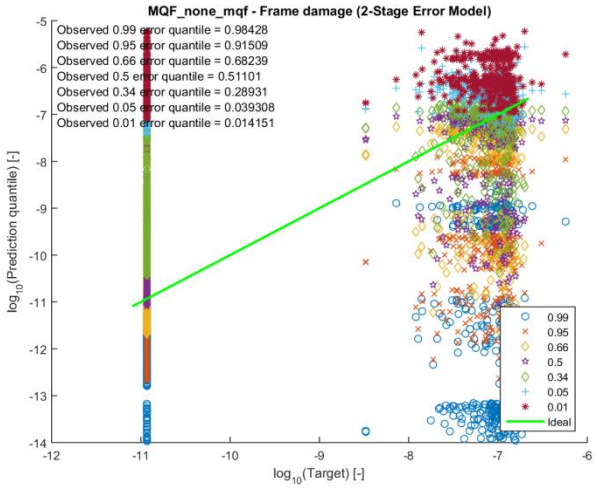


Figure K.32: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe damage predictions. Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

K.2.7 MTM

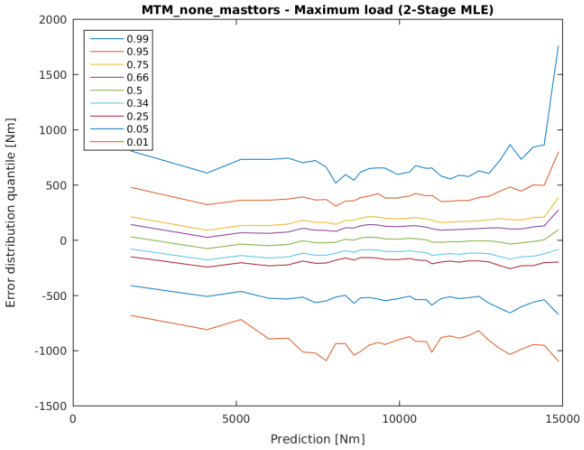


Figure K.33: Graph showing how selected quantiles of a predicted distribution of regression errors for the timeframe maximum load varies with the MLE point prediction.

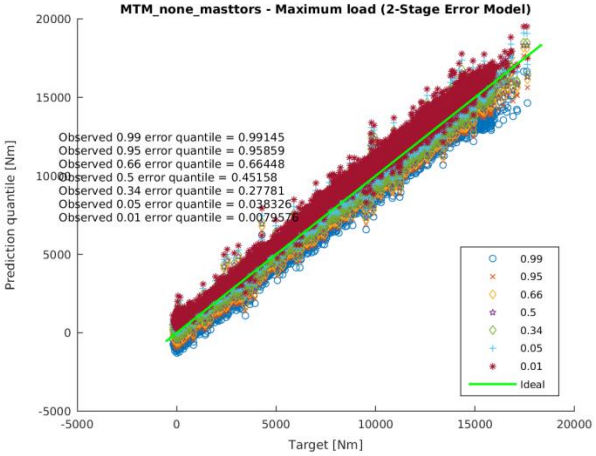


Figure K.34: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe maximum load predictions Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

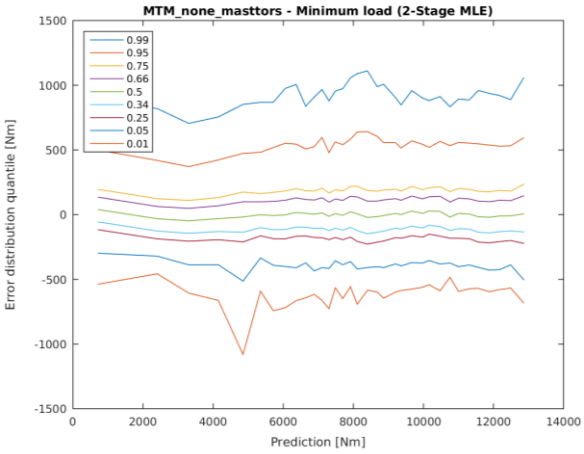


Figure K.35: Graph showing how selected quantiles of a predicted distribution of regression errors for the timeframe minimum load varies with the MLE point prediction.

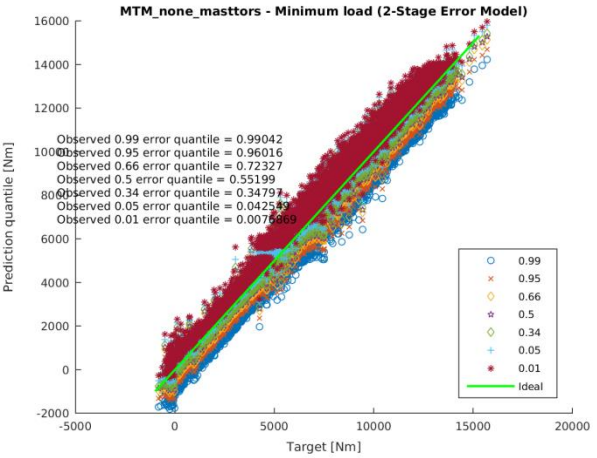


Figure K.36: Regression plot showing how the addition of selected and predicted prediction error quantiles to MLE point estimates affects regression performance on the LCF test set for timeframe minimum load predictions. Tabulated in the upper left corner are the quantiles of MLE point predictions that are more than their true value after the addition of a selected and predicted prediction error quantile.

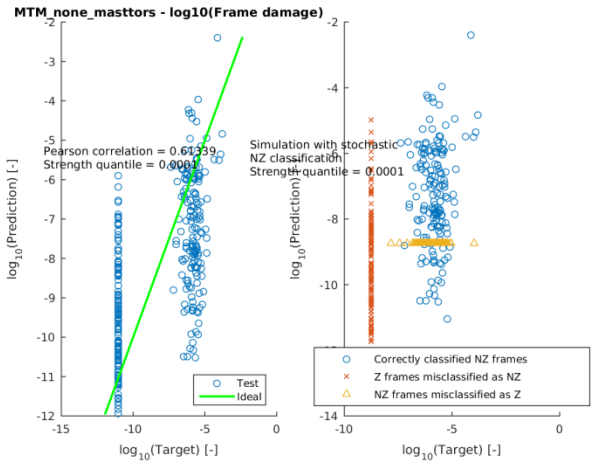


Figure K.37: Regression plot showing the correlation between predicted and actually measured timeframe damage for correctly classified damaging timeframes in the LCF test set. (left)
Regression plot showing the correlation between predicted timeframe damage and actually measured timeframe damage for all timeframes classified as damaging in the LCF test set. (including incorrectly classified timeframes)
(Z denotes zero-damage non-damaging timeframes and NZ denotes non-zero damaging timeframes). (right)

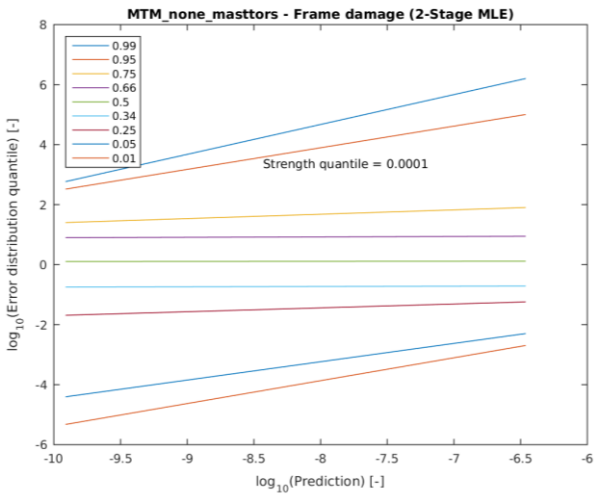


Figure K.38: Graph showing how selected quantiles of a predicted distribution of regression errors for timeframe damage varies with the MLE point prediction.

Appendix L. Database equivalence analysis

The applicability and validity of statistical regression models depend on the correlation of predictors and data features with the prediction target. If this is not the case, there is an increased risk that predictive models respond strongly to noise or unrelated but coincident phenomena and lose their predictive power when applied to future and unknown data. A detailed and systematic overview of the pre-requisites for proper statistical data for machine learning can be found in standard machine learning texts, such as by Hastie *et.al.* [106].

To perform a basic check of the validity and usefulness of the data used to build and validate machine learning models in chapter 4, several methods for exploratory data analysis are applied in sections L.1 to L.4. The used analysis methods include a descriptive analysis of their physical coherence, Principle Component Analysis, linear regression, and a custom feature distribution study respectively. These exploratory data analysis techniques are not sufficient to guarantee the representativeness and relevance of the datasets, but their positive results do considerably raise confidence in the validity of ensuing results from complex models.

The results from these studies all indicate that the datasets that are used in this work are reasonably consistent, relevant, and comparable and that they can thus be used to build and validate predictive statistical models. Nevertheless, significant differences between the properties of the datasets are still observed. Care should thus be taken to understand and mitigate possible adverse effects arising from differences that could influence the representativeness and inter-comparability of data from flight tests and in-service helicopters.

L.1 Physical coherence

The flight data and parameters listed in Table 4-1 are used as predictors and data features to predict load and fatigue damage values for the components listed in Table 4-2. It is assumed that this set of predictors implicitly contains a reasonable amount of information on the predicted loads. To predict MQF, or the torque on the tail rotor driveshaft (Fenestron), it must be predicted how engine torque is distributed between the main rotor and Fenestron. The torque output from the engines is directly measured and thus known. It can be understood that the power taken by a rotor mainly depends on its rotational speed, which is known too, and the angle-of-attack of its blades, which, for the main rotor, is proportional to the position of the cyclic and collective, and for the Fenestron to the pedal position. Variances due to, for example, changes of relative wind speed may be estimated through Indicated Airspeed, vertical speed and body accelerations, which are all recorded as well. Changes in rotor thrust due to air density can also be expected to be predictable by measuring outside air temperature and pressure altitude. Although the sampling rate, accuracy, or precision of the recorded vehicle data may not permit perfect prediction or reconstruction of the reference load signal, sampled at 100Hz, it is nevertheless reasonable to assume that accurate prediction is feasible and that recorded flight parameters do somehow correlate with MQF.

Similar arguments can be developed for the other target loads listed in Table 4-2.

L.2 Principle Component Analysis

Principle Component Analysis (PCA), as introduced in section H.3, can be used to identify correlations between parameters, and thus to roughly identify important predictors. Here, the analysis is primarily executed to verify the assumption that the dataset from Load Classification Flights (LCF) to generate predictive DLDM and PLDM models is representative of feature data recorded on helicopters 1-3 in chapter 4.

Applying PCA for comparative analysis, however, indicates that the distribution and correlation of recorded flight parameters differ significantly between the LCF dataset and data from helicopters 1 and 2, as shown in Figure L.1. If the composition and correlation between parameters in the three compared databases are the same, then the principle components should overlap. However, Figure L.1 shows significant differences between the principle components of the three datasets, indicating that care must be taken when using statistically derived relations from one dataset for predictive purposes on another dataset. The differences in

the statistical properties between the datasets from LCF flights and helicopters 1 and 2 are analysed and confirmed by further analysis in sections L.3 and L.4 and may cause significant prediction errors if regression models generated by LCF data are applied to data from helicopters 1 and 2, as is the case in chapter 4.

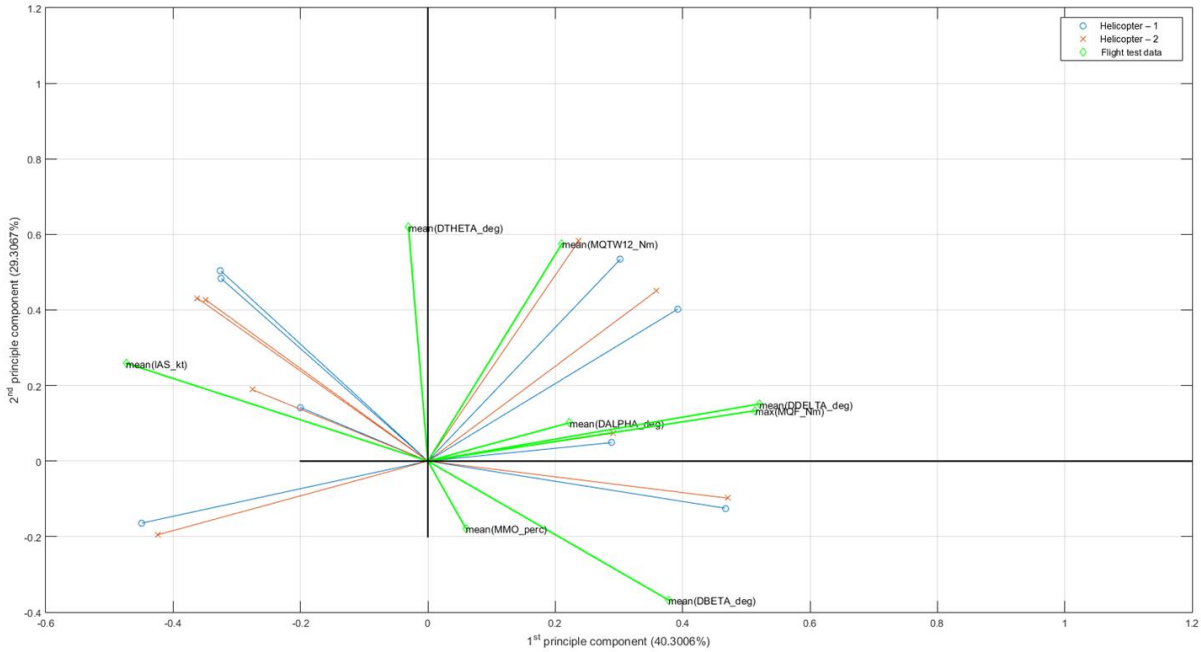


Figure L.1: Comparison between Principle Components from databases from flight tests, helicopter-1 and helicopter-2.

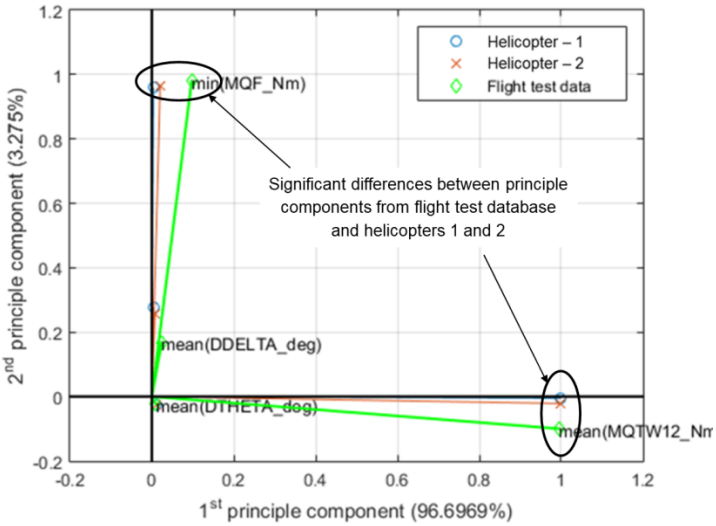


Figure L.2: Comparison between Principle Components from databases from flight tests, helicopter-1 and helicopter-2 - using a reduced feature set

Significant and consistent differences remain present in the principle components even if the dimensionality of the flight parameter dataset is reduced to only include three parameters expected to have a relatively high correlation, as shown in Figure L.2. Since the drive shafts of the main rotor and tail rotor are mechanically linked and are both driven by engine torque MQTW12, and since the main rotor speed is kept constant by an engine control law, it is expected that engine torque MQTW12, main rotor collective DTHETA and Fenestron

collective DDELTA are the primary predictors for Fenestron input torque MQF. The comparative principle component analysis in Figure L.2 indicates that engine torque and Fenestron torque are fairly independent and that Fenestron torque and collective have a high correlation.

L.3 Linear regression

Linear regression is a simple and easily-to-interpret method to build a predictive load model. It makes use of a linear combination of assigned features to best predict the value of a correlated target. Further details on linear regression are given in appendix H.1. It is expected that if the datasets from flight tests, helicopter-1 and helicopter-2 are comparable, then predictive models generated by these datasets should also be similar.

The available datasets from flight test data, helicopter-1 and helicopter-2 are each independently used to make a linear regression model to predict Fenstron input torque MQF by means of timeframe averages of the predictors listed in Table 4-1. Comparison of the resulting regression coefficients of the three linear regression models in Figure L.3 indeed shows comparable predictor weights. Since the generated prediction models are thus comparable, this provides confidence that a predictive model generated based on the flight test database, will yield accurate results when applied to helicopters 1 and 2. However, the differences between the regression models can nevertheless be observed clearly and confirm that care must be taken in assuming the applicability of a predictive model generated from one dataset to another dataset.

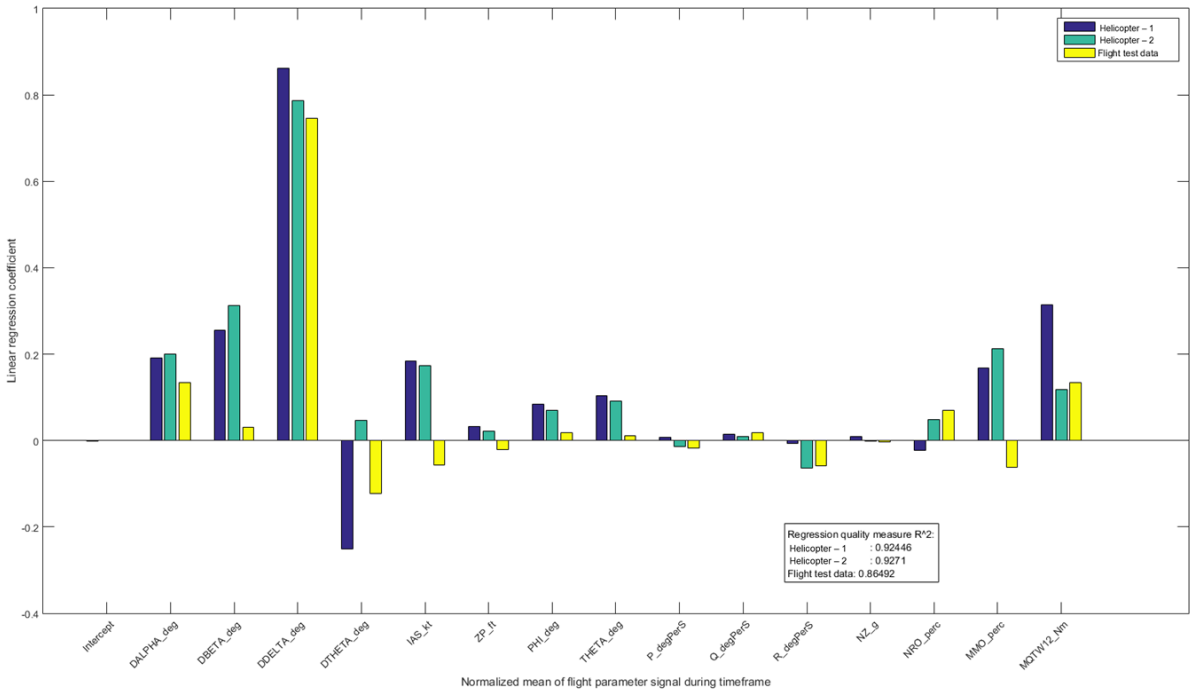


Figure L.3: Comparison between coefficients of linear regression models generated from data from flight tests, helicopter-1 and helicopter-2.

The estimated Fenestron blade angle DTHETA consistently has the largest influence on predicted MQF. Since the angle of attack of the Fenestron blades directly determines how much power is taken from the Fenestron driveshaft, this is a well-understood correlation. Other influential parameters are identified to include main rotor cyclic and pitch, which can be understood to control the amount of torque absorbed by the main rotor. Noting the presence of a constant main rotor speed control law and that the main rotor shaft and tail rotor shaft are mechanically linked and driven by engine output torque MQTW12, main rotor cyclic and collective thus indirectly indicate how much torque is taken by the tail rotor. Other parameters describing vehicle

attitude, the rate of attitude change, and atmospheric conditions can be seen to have a lower influence on Fenestron torque, at least on average.

The coefficients of determination R^2 shown in Figure L.3 for the regression models indicate a good correlation between the recorded flight parameters and target load. The consistent and significant difference in achieved regression quality between the dataset from flight tests on the one side and the datasets from helicopter-1 and helicopter-2 on the other may be explained by a difference in data complexity and variance. It is suspected that during test flights, the helicopter flies a large variety of maneuverers throughout the entire flight envelope and with a wide range of configuration. This makes the dataset more diverse, spanning a larger range of conditions and with a higher degree of data variance, thus making regression more difficult. The operational profile of commercially operated helicopters 1 and 2, however, is expected to be much more homogenous and to span a smaller range of operational conditions. For example, the relative time helicopter-1 and helicopter-2 spend in steady-state conditions such as level flight can be expected to be much higher than in the dataset from flight tests.

Comparison with the significantly higher R^2 values in section 4.2 obtained by more complex and non-linear regression models such as Artificial Neural Networks does indicate the need for the use of such complex models. The R^2 values of the linear regression models also compare well with values found in previous work, e.g. by Haas *et.al.* [86, 85].

L.4 Comparison of feature distribution and range

In order to confirm the presence of significant differences between the distribution properties of recorded flight data, Figure L.4 to Figure L.17 explicitly compare the range and variance of recorded flight parameters from the LCF dataset and helicopters 1-3. On the horizontal axis, the figures compare the range[41] of flight parameter data obtained from helicopters 1-3 (in blue, red, and orange respectively) and from LCF data (in green). The figures demonstrate that significant differences between the ranges of observed data exist. In most cases, but not all, does the CLF data cover the range of in-service data from helicopters 1-3. On the vertical axis, the figures compare the distribution of statistics, e.g. standard deviation, of the recorded flight parameters. From top to bottom, the figures show the quantiles of the distribution of the standard deviation for samples around the mean value on the horizontal axis. This comparison allows identifying differences in for example the signal noise that is present in the recorded parameters, and how these differences vary with mean parameter values. The figures include examples for many recorded flight parameters to demonstrate the generality of the observed differences. Similar results were obtained for all parameters listed in Table 4-1 and for more statistical features, such as kurtosis, extreme value, range, and skewness.

The analysis reveals that the range of LCF data does not fully cover the range of data recorded from helicopters 1 and 2. This is an important issue since it implies that the envelope of the LCF data does not provide full coverage for data recorded on other helicopters and that LCF-based regression models must thus incidentally perform extrapolation. In the case of the Artificial Neural Networks employed in chapter 4, this can be considered as problematic since this type of regression model can rapidly become unstable if extrapolation must be performed [85].

[41] Statistics have been calculated by the 2000 nearest samples to each grid point. The grid is linear between the dataset extremes and consists of 80 discretization points

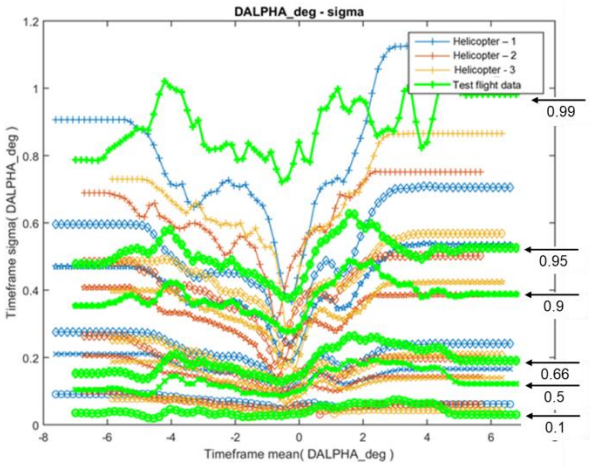


Figure L.4: Comparison of the range and distribution properties of timeframe features between helicopters 1-3 and LCF data. The data displays the following timeframe feature distribution quantiles from bottom to top: 0.1, 0.5, 0.66, 0.9, 0.95, and 0.99

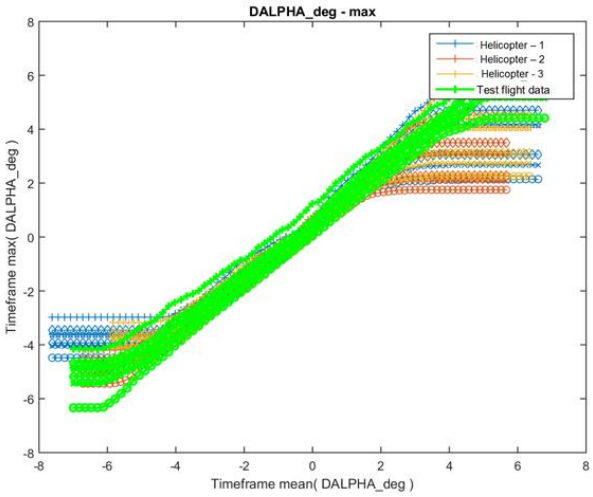


Figure L.5: Comparison of the range and distribution properties of timeframe features between helicopters 1-3 and LCF data.

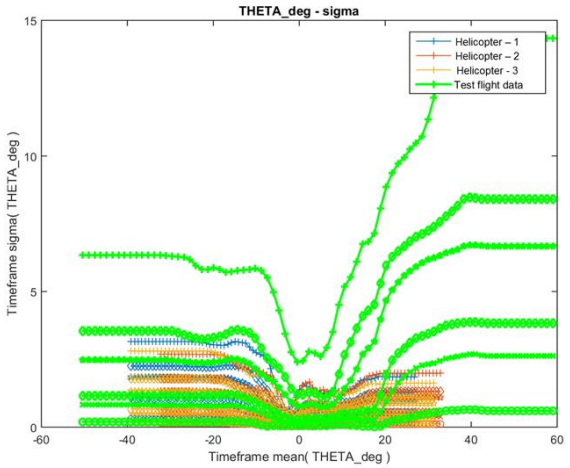


Figure L.6: Comparison of the range and distribution properties of timeframe features between helicopters 1-3 and LCF data.

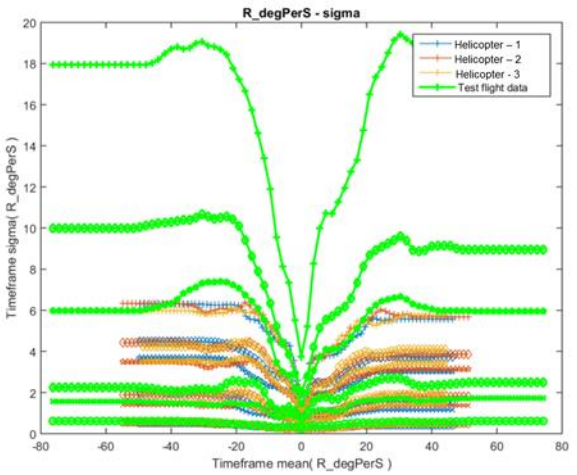


Figure L.7: Comparison of the range and distribution properties of timeframe features between helicopters 1-3 and LCF data.

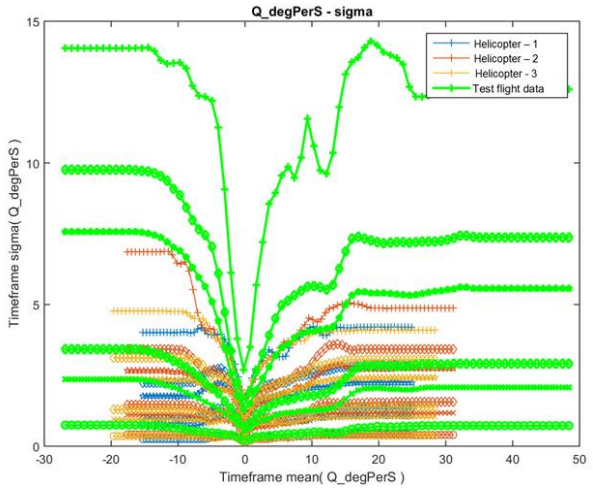


Figure L.8: Comparison of the range and distribution properties of timeframe features between helicopters 1-3 and LCF data.

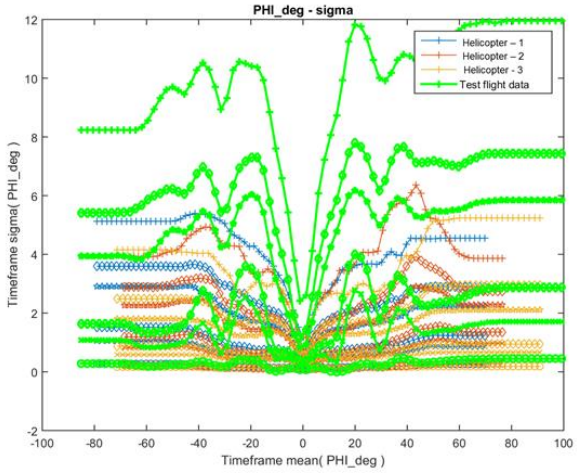


Figure L.9: Comparison of the range and distribution properties of timeframe features between helicopters 1-3 and LCF data.

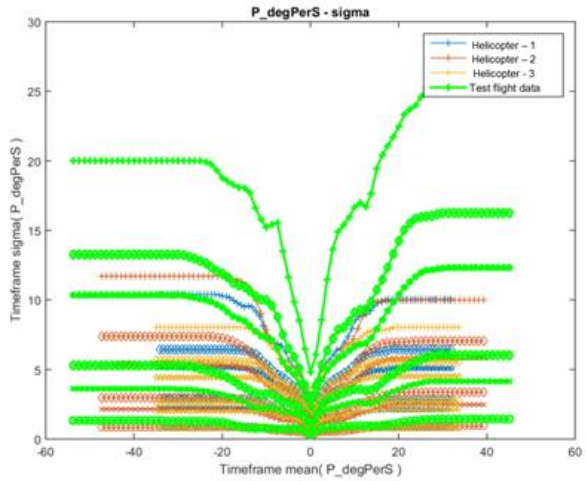


Figure L.10: Comparison of the range and distribution properties of timeframe features between helicopters 1-3 and LCF data.

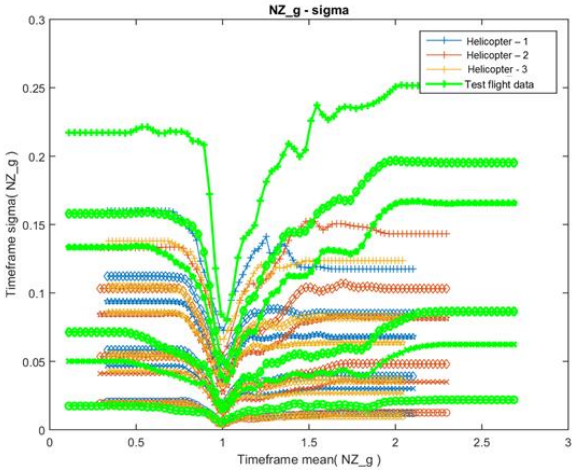


Figure L.11: Comparison of the range and distribution properties of timeframe features between helicopters 1-3 and LCF data.

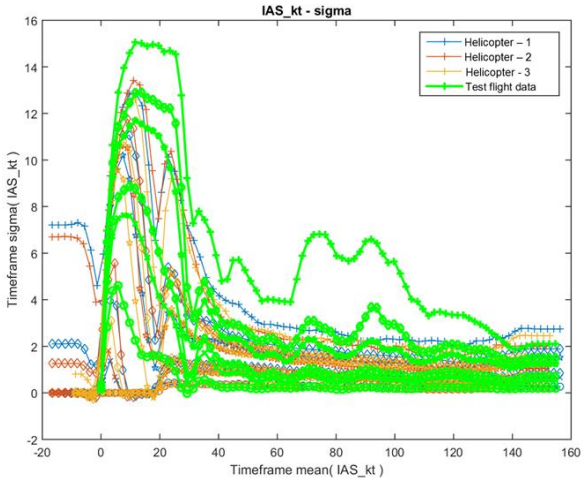


Figure L.12: Comparison of the range and distribution properties of timeframe features between helicopters 1-3 and LCF data.

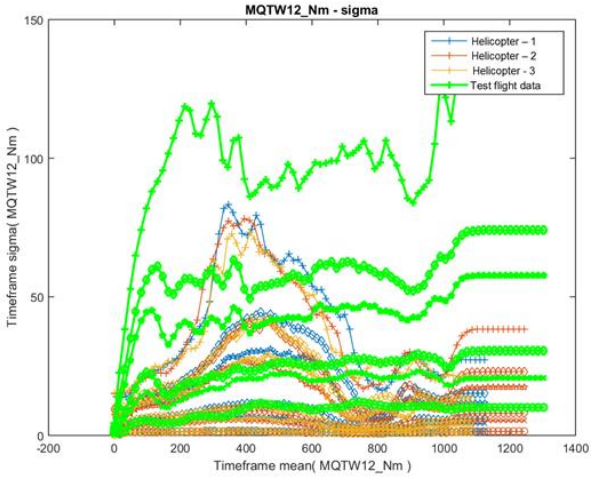


Figure L.13: Comparison of the range and distribution properties of timeframe features between helicopters 1-3 and LCF data.

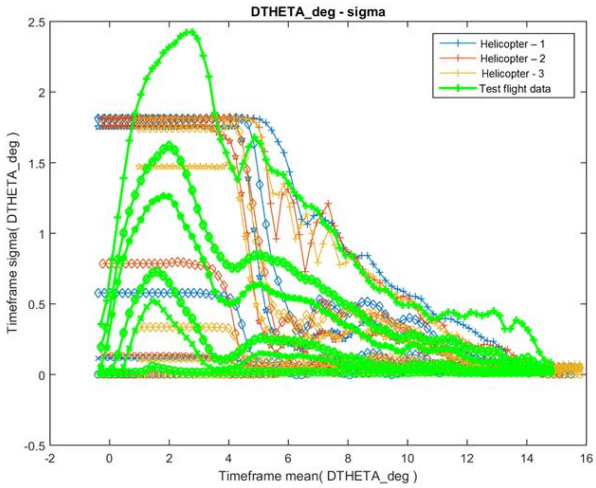


Figure L.14: Comparison of the range and distribution properties of timeframe features between helicopters 1-3 and LCF data.

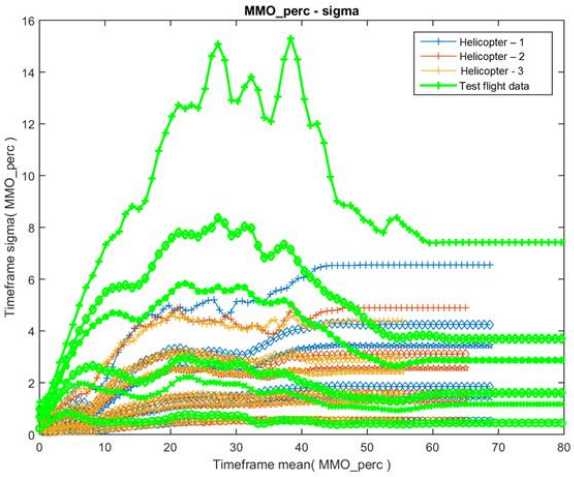


Figure L.15: Comparison of the range and distribution properties of timeframe features between helicopters 1-3 and LCF data.

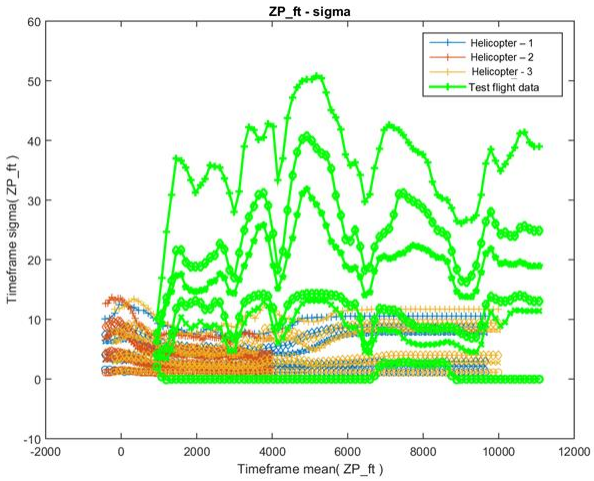


Figure L.16: Comparison of the range and distribution properties of timeframe features between helicopters 1-3 and LCF data.

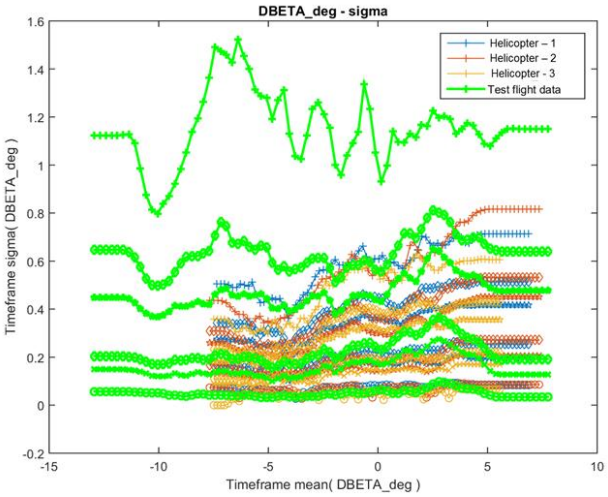


Figure L.17: Comparison of the range and distribution properties of timeframe features between helicopters 1-3 and LCF data.

Comparison of the normalized distribution of variance over the range of feature parameters exposes detailed differences in the apparent noise of recorded flight parameters. These differences are indicators that the quality and precision of recorded data can differ significantly between the LCF data and commercially operated helicopters, but also to a lesser extent between serial helicopters themselves. This observation is important as it lowers confidence in the representativeness of probabilistic prediction error models generated from LCF data. As it can be expected that random prediction errors are partially caused by random noise in the recorded flight parameters, the expected error distributions may be non-conservatively biased if noise in LCF recordings is systematically lower than for in-service helicopters. However, in the analysed case, LCF noise rather seems to systematically overestimate in-service noise and should thus not result in non-conservative estimation biases. The raised noise level in LCF data may be explained by the use of noisy flight test instrumentation or legacy sensors, or legacy recording data processing equipment.

L.5 Discussion

The analysis conducted in sections L.1 to L.4 demonstrates considerable differences between datasets from LCF data and from commercially operated in-service helicopters. As the LCF data is used to generate predictive models for in-service helicopters and is assumed to be representative for these, these difference may cause unexpected prediction errors. Specifically, the validity of the following two conditions has been tested and found to be at least partially invalid:

- **Domain coverage:** especially in the case of non-linear regression models, such as Artificial Neural Networks, it is important to make sure that in-service predictions are the result of regression interpolations. If predictions are made for data points outside of the envelope spanned by the data used to generate the predictive models, then this must effectively be done by extrapolation. Non-linear regression model models can become unstable and ill-defined when performing extrapolation. It is therefore recommended to develop methods that can be used to systematically identify cases where the regression models perform extrapolation and that can provide confidence that such extrapolations are robust and accurate.
- **Variance coverage:** as a derived requirement, if insufficient data is gathered to sufficiently sample and characterize noise data variance and regression variance, then there is an increased risk that error distributions will be predicted incorrectly. In addition, the risk of inadvertent-over-specialization of the predictive models on the training dataset, i.e. LCF data, is increased. This therefore limits the applicability and validity of LCF-based statistical prediction models for data from actual in-service helicopters, such as helicopters 1-3.

Despite the limited representativeness of the LCF data for data from helicopters 1-3, the reliability and accuracy of PLDM predictions could still be demonstrated in sections 4.2.5 and 4.3.4. It is therefore concluded that effect of the limited representativeness of the LCF data for actual in-service situation upon PLDM prediction accuracy and reliability are small. Nevertheless, for future work it is recommended to more systematically analyse up to what extend the coverage and features of LCF can be allowed to vary and deviate from actual in-service data without significantly reducing the accuracy and reliability of PLDM predictions.

Appendix M. Minimum remaining reliability

In order to adapt an individual Service Life Limit (SLL) according to actual usage by following AC-27-1B MG-15 [58], the complete end-to-end process and toolchain for Virtual Fatigue Life Monitoring (VFLM) must be analysed and all risks must be mitigated accordingly. Such an analysis and a discussion of their requirements are beyond the scope of the work presented here. However, some considerations on minimum remaining reliability in case of VFLM failure are nevertheless included to aid such analysis in the future.

All the components studied in this work and listed in Table 4-2, are critical components. Therefore, it is common practise to initially impose a Design Assurance Level (DAL) A to the VFLM system and the VFLM algorithms. DAL-A by ARP-4761 [33] and DO-178B [143] implies the most stringent set of software and hardware requirements. Upon closer inspection, a less restrictive classification may, however, be possible. Failure of a DAL-A software implies that "critical functionality to safely fly or land the aircraft is lost". However, a failure of the VFLM system or VFLM algorithm does not imply the onset of a direct and immediate catastrophic failure. The worst case VFLM failure condition does not correspond to any direct and immediate impact on flight safety, but merely a reduction in safety margins. In the worst case, the probability of a fatigue failure simply rises with the same rate as it would have in the scenario that is being used to set a legacy SLL, i.e. as in chapter2. Only once the legacy SLL is surpassed, can the probability of failure rise above the allowable limit.

Using the numerical reliability simulation methods introduced in chapter 2 to 4 it can be computed how the probability of failure of a component accumulating fatigue damage according to the DMP increases over time. These simulations have been carried out for all the components listed in Table 4-2 in order to demonstrate some generality of the argument. The results of the simulation are shown in Figure M.1 to Figure M.7. There graphs have been generated using a similar simulation technique as introduced in section 4.2.7 but where damage is assumed to accumulate according to the DMP. The graphs show that the confidence level with which a target probability of failure of a SLL can be substantiated decreases with increasing component life (i.e. SLL).

As discussed in section 3.5.2, airworthiness regulations do not explicitly specify the confidence level with which the reliability of a SLL must be substantiated. Allowing a reduced confidence level with which a reliability level of 0.999999 is substantiated enables to set a SLL beyond 20.000 flight hours for almost all components. In practise, this means that even if VFLM fails to register any fatigue damage whereas the component is actually flown according to the worst-case DMP usage, the probability of a fatigue failure can still be regarded as limited. For example, setting a limit to the fatigue life extension that can be enabled by VFLM can make sure that even in the case of worst-case VFLM failure, the probability of a fatigue failure does not fall below an acceptable, though reduced, level. This enables to classify VFLM failure as causing a "reduction of safety margins" and not immediate catastrophic failure. The safety classification and corresponding DAL level of the VFLM system may thus be brought down significantly and may assist in staying clear from unnecessarily demanding and expensive hardware and software requirements.

CAUTION: Service life limits presented in Figure M.1 to Figure M.7 are computed for academic purposes only and are not approved by any OEM or airworthiness authority.

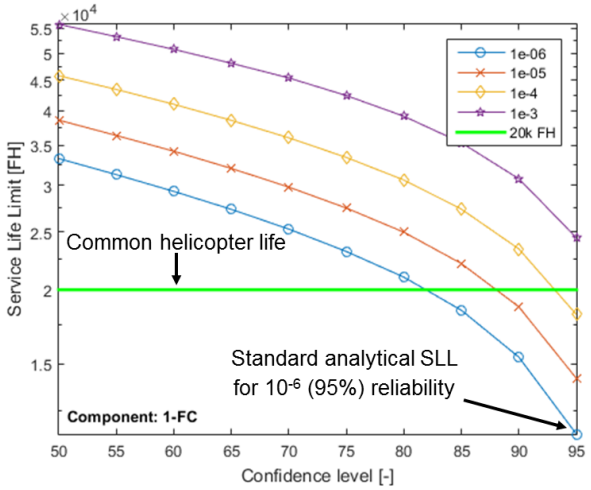


Figure M.1: SLL for the hydraulic actuator housing (component 1) for several reliability requirements and assuming design usage. The legend contains several levels of probability of failure and a reference line corresponding to the regular end-of-life of a helicopter. The outer lower right point, for example, corresponds to an SLL with 10^{-6} (95%) reliability.

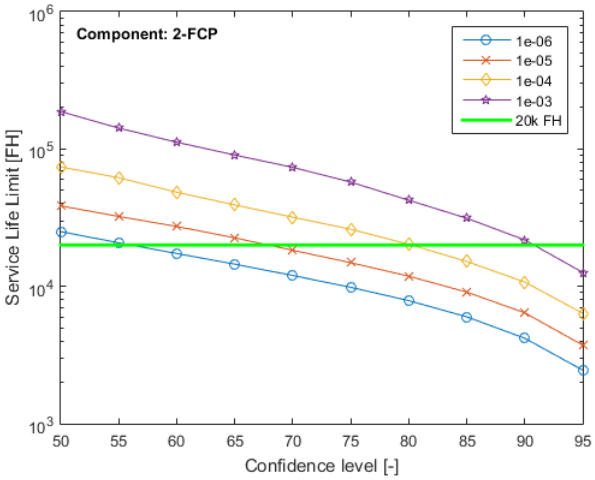


Figure M.2: SLL for the collective control rod (component 2) for several reliability requirements and assuming design usage.

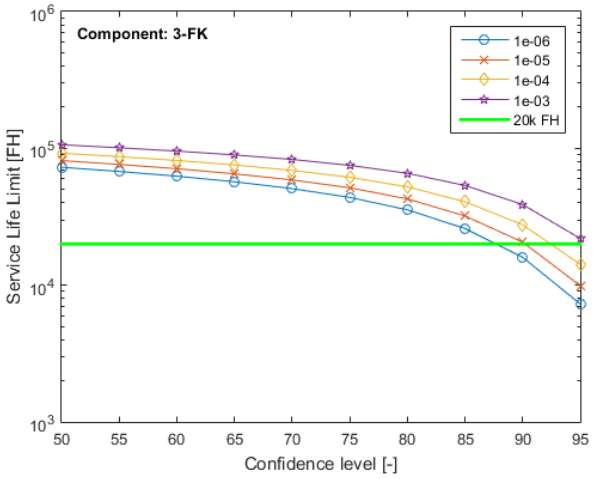


Figure M.3: SLL for the gimbal (component 3) for several reliability requirements and assuming design usage.

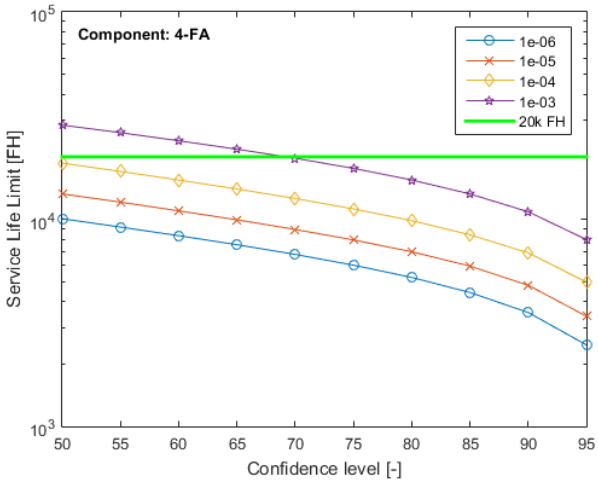


Figure M.4: SLL for the forked level (component 4) for several reliability requirements and assuming design usage.

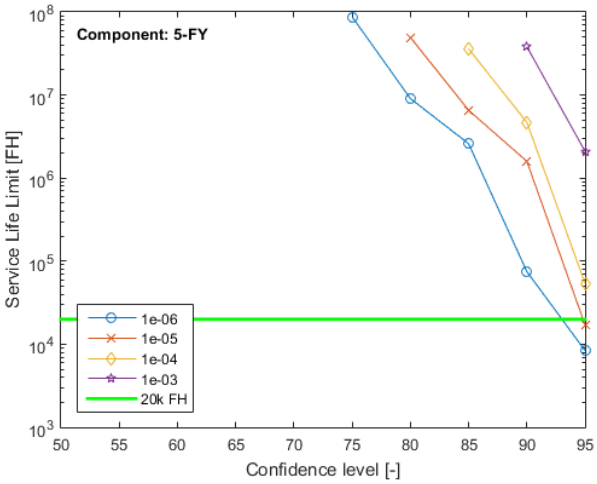


Figure M.5: SLL for the upper gearbox housing (component 5) for several reliability requirements and assuming design usage. Only data points with a finite SLL are displayed.

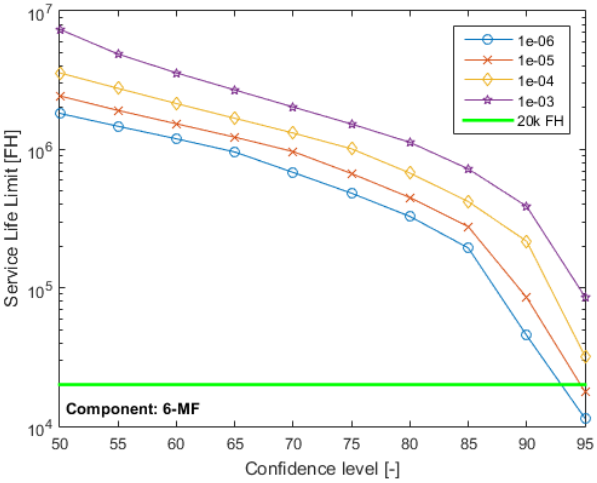


Figure M.6: SLL for the lower gearbox housing torque output (component 6) for several reliability requirements and assuming design usage.

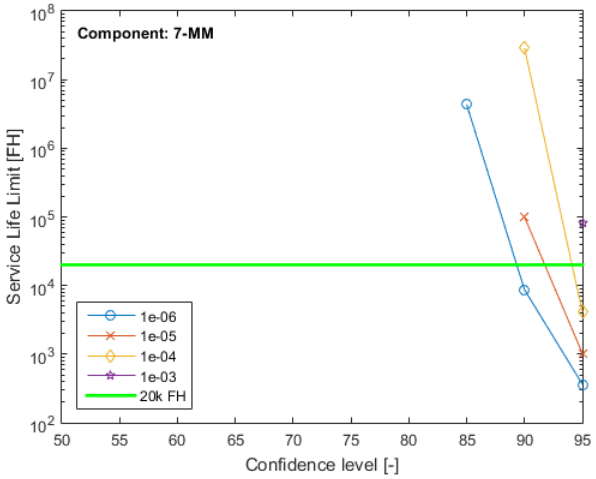


Figure M.7: SLL for the main rotor mast (component 7) for several reliability requirements and assuming design usage. Only data points with a finite SLL are displayed.

Appendix N. Curiculum Vitae

Personal Information

Name:	Sam Hiawatha Dekker
Birth:	September 5th 1988, Amsterdam
Nationality:	Dutch

Professional

January 2016 – current	**Project leader Condition Based Maintenance: Marenco Swisshelicopter**
November 2012 - December 2015	**PhD Researcher, Health and Usage Monitoring Systems & Flight Data: Airbus Helicopters Germany & Delft University of Technology Graduate School**
April 2012 – October 2012	**Master thesis, Health and Usage Monitoring Systems & Flight Data: Eurocopter Germany**
November 2011 – March 2012	**Internship, Health Usage Monitoring Systems & Flight Data: Eurocopter Germany**

Education

September 2010 – October 2012	**Master of Science in Aerospace Engineering, Delft University of Technology**
September 2007 – August 2010	**Bachelor of Science in Aerospace Engineering, Delft University of Technology**
September 2001 – June 2007	**Gymnasium, Vossius Gymnasium Amsterdam**

Professional memberships

January 2017 – present	**Swiss Alliance for Data Intensive Services** Working group: Data-Driven Predictive Maintenance of Industrial Assets

Publications

September 2017	Safety for Condition Based Maintenance of flight-critical parts, Smart Maintenance Conference 2017, Winterthur
July 2016	A Bayesian tolerance interval estimation method for fatigue strength substantiation of rotorcraft dynamic components, International Journal of Fatigue
July 2016	Reliability modelling for rotorcraft component fatigue life prediction wit assumed usage, The Aeronautical Journal, also presented during the 2014 European Rotorcraft Forum
November 2013	Helicopter Fatigue Monitoring with Direct Load & Damage Modelling, 5th International HELI World Conference

Patents

April 2016	Probabilistic load and damage modelling for fatigue life management, US 20170293712 A1
October 2012	Fatigue management system and method of operating such a fatigue management system, P2725337 A1