Probabilistic security assessment of sustainable power grids

Multivariate analysis and dependence modeling for risk-based security assessment

Khuntia, Swasti

DOI
10.4233/uuid:08d75eb8-0fe2-455f-85a3-3966d97ff674

Publication date
2018

Document Version
Final published version

Citation (APA)
Khuntia, S. (2018). *Probabilistic security assessment of sustainable power grids: Multivariate analysis and dependence modeling for risk-based security assessment*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:08d75eb8-0fe2-455f-85a3-3966d97ff674

Probabilistic security assessment of sustainable power grids

Multivariate analysis and dependence modeling for risk-based security assessment

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 26 November 2018 at 10:00 o'clock

by

Swasti Ranjan KHUNTIA
Master of Science in Electrical Engineering
Illinois Institute of Technology, USA
born in Odisha, India

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof.ir. Mart A.M.M. van der Meijden	Delft University of Technology, promotor
Dr.Dipl.-Ing. José L. Rueda	Delft University of Technology, copromotor

Independent members:

Prof.dr. Gerd Kjølle	Norwegian University of Science and Technology, Norway
Prof.dr.-Ing. habil. Lutz Hofmann	Leibniz Universität Hannover, Germany
Prof.dr. Louis Wehenkel	University of Liege, Belgium
Prof.dr. Peter Palensky	Delft University of Technology
Prof.dr. Miro Zeman	Delft University of Technology

ISBN/EAN: 978-94-028-1283-1

Keywords: dependence modeling, load forecast, static security assessment, vine copula.

Cover design by Alexis Ierides
Printed in The Netherlands by Ipskamp Printing, Enschede

This research is financially supported by European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 608540 GARPUR [Generally Accepted Reliability Principle with Uncertainty modelling and through probabilistic Risk assessment] project.

DREAM, DREAM, DREAM.

DREAMS TRANSFORM INTO THOUGHTS
AND THOUGHTS RESULT IN ACTION."

DR. APJ ABDUL KALAM

Dedicated to my parents

ABSTRACT

Among the existing Renewable Energy Sources (RES), wind power has become significantly important to Transmission System Operators (TSOs) because of two reasons, namely:

i. large Wind Power Plants (WPPs) can be connected to bulk power system at transmission level, and
ii. large WPPs are being built or planned in regions with a high potential for the extraction of wind energy and TSOs must facilitate their integration.

This comes at a time when the electric power industry is undergoing an energy transition due to the increasing penetration of RES and decentralized generation while discarding fossil fuels to achieve a greener future in the form of a low-carbon power system. To accommodate the high penetration of wind power into the existing electrical grid infrastructure while TSOs are facing stranded expansion of transmission infrastructure, TSOs are investing knowledge and money into safe-guarding grid reliability and to meet the required security of supply. As the location of WPPs and demand sites are not always close by, transmission of energy has placed a burden on transmission links in the existing grid infrastructure. Complexity in terms of interspatial dependence and temporal correlation of load and wind power impose a challenging operational threat to TSOs. Thus, it is important to emphasize spatial and temporal dependency to assess system security as TSOs are paving the way for the transition from deterministic to probabilistic reliability management. It is to be noted that system security is one of the two aspects of power system reliability, with the other being system adequacy. The security level of a power system is determined by the likelihood and severity of violations.

Considering operational threats (line overloading, voltage instability, etc.) as a potential future challenge, the aim of this research is to assess the system security in terms of overload risk due to transmission line overloading by developing a reproducible statistical model which can account for the spatio-temporal dependency of load and wind power as joint probability distribution. Both load and wind power are exogenous in nature, meaning that they are determined by someone else than the TSO, and the TSO will have to adapt its behaviour accordingly. The reasons why load and wind power penetration have been chosen as modeling challenges for TSOs in terms of system security are:

i. *Uncertainty in electricity load*: Due to its nature, electricity load is uncertain and supplemented with the spatial distribution of load sites which are not always located close to generating sites.

ii. *Considerable wind power generation*: Wind power is characterized by variability and uncertain nature, and its generation capacity is dependent on geographic location. The European Wind Energy Association expects 50% of the electrical energy demand to be met by wind energy by 2050. In addition to wind on land, a substantial proportion of wind capacity from the North Sea is anticipated: 150 GW by 2030 and 350 GW by 2050.

To account for the above-mentioned challenges in the presence of exogenous variables, TSOs need to re-evaluate their system security. As a first step towards achieving the research objective, a critical review of current reliability assessment practices followed by TSOs on three time-horizons (short-, mid- and long-term) is performed to justify the need to adopt an interacted approach in operation and planning. It is followed by modeling and forecasting the electricity load in both the short-term and long-term horizons since the current practice among TSOs is to treat load growth as independent from wind power generation. From the time-horizons study, the yearly forecast with an hourly time-span is chosen for the short-term forecast based on operational planning while a 4 year-long forecast horizon with a monthly time-span is chosen for the long-term forecast based on grid development activities (as offshore wind farm construction takes approximately $3-4$ years). A novel neural network-based load forecasting model is developed for the short-term horizon with a focus on low forecast error and an innovative attempt at modeling forecast error distribution using truncated normal distribution. For the long-term horizon, a new forecasting model based on multiplicative error is developed with a focus on low forecast error and directional accuracy. Both the models account for temporal correlation and Gaussianity of distributions. However, the problem lies with spatio-temporal dependencies of load and wind power pointing out two critical features: non-Gaussian nature of data and the complexly dependent relationship between load and wind power.

In the case of continental Europe, the electricity grid is heavily interconnected and hence electricity produced in one place can meet the demand elsewhere. This transmission of energy comes at a cost for TSOs, where the transmission lines operate closer to their operational limits more and more frequently. As such, the risk of transmission line overloading and voltage instability cannot be avoided in the future. In the attempt to overcome the burden of transmission line overloading, this research will encourage TSOs to operate the grid within security limits while considering spatio-temporal dependency. A probabilistic approach in combination with high-dimensional spatio-temporal dependence modeling is proposed in this research. Modeling load and wind power as joint probability distribution for studying spatio-temporal dependence using vine copula is a novel attempt in this research. Use of vine copula facilitates building multi-dimensional copulas out of bivariate copulas as they are easy to estimate

and are well understood. This study will consider hourly resolution load and wind power data obtained from a U.S. utility spanning three years and spatially distributed in nineteen load and two wind power zones. Data collection, in terms of dimension, tend to increase in future and to tackle this high-dimensional data, a reproducible vine copula sampling algorithm is developed in this research. The developed sampling algorithm employs k-means, Gaussian Mixture Model and hierarchical linkage clustering techniques along with singular value decomposition technique to analyze high-dimensional data and ease the computational burden. This is deemed essential when the future operating condition will be data-centric and the selection of an appropriate clustering technique and copula family is realized by goodness of clustering and goodness of fit tests.

To assess the operational threat in terms of transmission line overloading, a risk-based security assessment is performed considering the spatio-temporal dependency of load and wind power using the developed vine copula modeling. A severity function is used to describe the transmission line overloading and a subsequent assessment is performed thereby. The overload risk index is treated as a security indicator in this research for risk-based security assessment. Probabilistic AC load flow with correlated parameters is performed on modified IEEE-39 test case (modified in terms of addition of WPPs) with significant wind penetration representing actual U.S. market zones and real-life load and wind power data from the same U.S. utility. The real-life data is mapped onto the test case based on the system data and in order to achieve realistic results. Two case studies representing future scenarios of massive wind power penetration and lower conventional generation are included to study the importance of spatio-temporal dependency. The impact of line overloading is described by the severity function and the probability of line overload. Simulation results prove the advantage of addressing spatio-temporal dependency to quantify the overload risk index, which is treated as a security indicator.

SAMENVATTING

Van de verschillende bronnen van duurzame energie heeft windenergie de speciale aandacht van netbeheerders, voornamelijk vanwege de volgende twee redenen:

i. grote windparken kunnen op het elektriciteitsnet aangesloten worden op transportnetniveau,
ii. grote windparken worden momenteel aangelegd, of zijn gepland, in gebieden met een hoog potentieel voor windenergie en netbeheerders moeten de aansluiting op het elektriciteitsnet faciliteren.

Dit gebeurt op een moment waarop de elektriciteitssector een energietransitie ondergaat, met name de toenemende integratie van duurzame energie en decentrale productie, terwijl de productie met fossiele brandstoffen wordt afgebouwd om een groenere toekomst in de vorm van een koolstofneutrale elektriciteitsvoorziening te creëren. Om de grootschalige integratie van windenergie in de bestaande elektrische infrastructuur mogelijk te maken, terwijl netbeheerders geconfronteerd worden met een gelimiteerde uitbreiding van het transportnet, investeren netbeheerders kennis en geld in het verbeteren van de betrouwbaarheid van het elektriciteitsnet om aan de leveringszekerheid te voldoen. Omdat de locaties van windparken en belastingsgebieden niet altijd bij elkaar in de buurt zijn, vormt het transport van windenergie een belasting voor de bestaande elektrische infrastructuur. De complexiteit van belasting en windproductie wat betreft ruimtelijke afhankelijkheid en tijdsafhankelijkheid, vormt een operationele uitdaging voor netbeheerders. Aandacht voor ruimtelijke afhankelijkheid en tijdsafhankelijkheid bij het vaststellen van de betrouwbaarheid van het elektriciteitsnet wordt daarom belangrijk geacht, aangezien netbeheerders de weg vrijmaken voor de overgang van deterministisch naar probabilistisch betrouwbaarheidsmanagement. Hierbij moet opgemerkt worden dat weerbaarheid tegen fouten in het elektriciteitsnet een van de twee aspecten van betrouwbaarheid van elektriciteitsvoorzieningssystemen is, waarbij de andere adequaatheid is. De mate van weerbaarheid wordt bepaald door de waarschijnlijkheid en het gevolg van verstoringen in het elektriciteitsnet.

Wat betreft de operationele bedreigingen (zoals overbelasting van lijnen en spanningsinstabiliteit) als toekomstige uitdaging, is het doel van dit onderzoek om de weerbaarheid te beoordelen in termen van het risico op overbelasting van het elektriciteitsnet ten gevolge van de overbelasting van transmissielijnen, door de ontwikkeling van een reproduceerbaar statistisch model dat rekening houdt met de ruimtelijke afhankelijkheid van belasting en windenergie als gezamenlijke kansverdeling. Zowel belasting als windenergie zijn exogeen van aard, d.w.z. ze worden niet door de netbeheerder bepaald, en deze zal zijn gedrag dienovereenkomstig

moeten aanpassen. De redenen om te kiezen voor de combinatie van belasting en windenergie als modelleringuitdaging voor netbeheerders op het gebied van weerbaarheid van het elektriciteitssysteem zijn:

i. *Onzekerheid in belasting*: De belasting is van nature onzeker en wordt aangevuld met de ruimtelijke verdeling van belastingslocaties die zich niet altijd in de buurt van de productielocaties bevinden.
ii. *Aanzienlijke productie van windenergie*: Windenergie wordt gekenmerkt door variabiliteit en onzekerheid en de productiecapaciteit is afhankelijk van de geografische locatie. De European Wind Energy Association verwacht dat in 2050 50% van de elektriciteitsbehoefte zal worden geleverd door windenergie. Naast wind op land wordt een aanzienlijk aandeel windcapaciteit verwacht op de Noordzee: 150 GW in 2030 en 350 GW in 2050.

Om rekening te houden met de bovengenoemde uitdagingen in aanwezigheid van exogene variabelen, moeten netbeheerders het begrip weerbaarheid van het elektriciteitssysteem opnieuw evalueren. Als eerste stap om de onderzoeksdoelstelling te bereiken, is een kritische evaluatie van de huidige betrouwbaarheidsanalyse door netbeheerders op drie tijdshorizonnen (korte, middellange en lange termijn) uitgevoerd om de noodzaak van een geïntegreerde aanpak in de planning en bedrijfsvoering te rechtvaardigen. Dit wordt gevolgd door het modelleren en voorspellen van de belasting, zowel op de korte als op de lange termijn, aangezien de huidige praktijk van de netbeheerders is om de groei van de belasting onafhankelijk van de windenergie te beschouwen. In dit onderzoek is gekozen voor een jaarlijkse prognose met een resolutie van een uur voor de kortetermijnprognose op basis van operationele planning, terwijl een prognosehorizon van 4 jaar met een maandelijkse tijdspanne is gekozen voor de langetermijnprognose op basis van netplanning (aangezien de constructie van windparken op zee ongeveer 3-4 jaar in beslag neemt). Een nieuw neuraal-netwerkgebaseerd voorspelmodel voor de belasting is ontwikkeld voor de kortetermijnhorizon met de focus op een kleine voorspellingsfout en de modellering van de foutdistributie door een afgekapte normale verdeling. Voor de langetermijnhorizon is een nieuw prognosemodel gebaseerd op de multiplicatieve fout ontwikkeld, met de focus op een kleine voorspellingsfout en goede richtingsnauwkeurigheid. Beide modellen houden rekening met tijdsafhankelijkheid en Gaussianiteit van de verdelingen. Het probleem zit echter in de ruimtelijke- en tijdsafhankelijkheden van de belasting en windenergie, wat op twee essentiële kenmerken duidt: de niet-Gaussiaanse aard en de complexe afhankelijkheid tussen de belasting en windenergie.

Wat betreft continentaal Europa, is het elektriciteitsnet sterk met elkaar verbonden en kan de elektriciteit die geproduceerd is op de ene locatie voldoen aan de vraag op een andere locatie. Dit energietransport brengt kosten met zich mee voor de

netbeheerders, terwijl transmissielijnen steeds vaker tegen de grens van hun operationele limiet werken. Het risico op overbelasting van transmissielijnen en op spanningsinstabiliteit kan in de toekomst niet altijd vermeden worden. Om het risico op overbelasting van hoogspanningslijnen te beperken, zal dit onderzoek netbeheerders helpen om het netwerk te gebruiken binnen de veiligheidslimieten, daarbij rekening houdend met ruimtelijke- en tijdsafhankelijkheid. Met andere woorden, een probabilistische benadering in combinatie met een hooggedimensioneerde ruimtelijke- en tijdsafhankelijkheid is vereist. De modellering van belasting en windenergie als gezamenlijke kansverdeling voor het bestuderen van de ruimtelijke- en tijdsafhankelijkheid met behulp van zgn. 'canonical vine copulas' is een eerste poging in dit onderzoek. Het gebruik van canonical vine copulas vergemakkelijkt het bouwen van multidimensionale copulas uit bivariate copulas, omdat deze gemakkelijk te schatten en goed te begrijpen zijn. Belastings- en windenergiewaarden afkomstig van een Amerikaans energiebedrijf met een resolutie van een uur voor een periode van drie jaar en ruimtelijk verdeeld in negentien belastings- en twee windenergielocaties zijn in dit onderzoek gebruikt. Het verzamelen van gegevens, in termen van omvang, zal in de toekomst toenemen en om dit aan te pakken, is in dit onderzoek een reproduceerbaar algoritme voor het combineren van vine copulas ontwikkeld. Het sampling algoritme maakt gebruik van een zgn. 'k-means, Gaussian Mixture Model' en 'hierarchical linkage clustering techniques', samen met de 'singular value decomposition technique', om hoogdimensionale gegevens te verwerken en de rekenlast te verlichten. Selectie van een geschikte clusteringstechniek en copula-familie is gerealiseerd door 'goodness of clustering and goodness of fit tests'.

Om de operationele dreiging in de vorm van overbelasting van transmissielijnen te beoordelen, is een risicogebaseerde weerbaarheidsbeoordeling uitgevoerd waarbij rekening is gehouden met de ruimtelijke- en tijdsafhankelijkheid van belasting en windenergie met behulp van de ontwikkelde vine-copula-modellering. Een weerbaarheidsindex is gebruikt om de overbelasting van transmissielijnen te beschrijven, waarmee de beoordeling uitgevoerd wordt. De overbelastingsrisico-index wordt in dit onderzoek als beveiligingsindicator voor de risicogebaseerde weerbaarheidsbeoordeling gehanteerd. Probabilistische wisselstroom load flow met gecorreleerde parameters is uitgevoerd op het aangepaste IEEE-39 netmodel (aangepast door toevoeging van windenergielocaties) met een significant aandeel windenergie die de werkelijke Amerikaanse marktgebieden vertegenwoordigt, met reële waarden voor de belasting en windenergie. Deze reële gegevens zijn op basis van de systeemgegevens overgebracht op het IEEE-39 netmodel om realistische resultaten te bereiken. Twee studies die toekomstige scenario's van grootschalige windenergie en een kleine hoeveelheid conventionele productie weergeven, zijn opgenomen om het belang van ruimtelijke- en tijdsafhankelijkheid te bestuderen. De impact van een lijnoverbelasting wordt weergegeven door de weerbaarheidsindex en de kans op een

overbelasting van de lijn. De resultaten van de simulaties onderbouwen de superioriteit van het beschouwen van ruimtelijke- en tijdsafhankelijkheid om de beschreven overbelastingsrisico-index als weerbaarheidsindicator te kwantificeren.

TABLE OF CONTENTS

ACKNOWLEDGMENT

First things first! More than four and a half years have already passed since I walked through the corridors of the IEPG Research Group for the first time. I could never have made it to this point on my own, I was able to succeed only with the support of countless people. I need to thank all who accompanied me on this Ph.D. journey full of adventures, challenges, incredible experience and lots of frustrating moments. The rich research experience gained would not have possible without the blessing, support and help of great people around me and some abroad. Some people have contributed directly to my research and others indirectly by supporting me in any form or way.

My deepest sincere appreciation and thanks go to my family who always supports me even from far away. Dear Mommy and Bapa, without you and your endless love and support that you have given me, I might not be the person I am today; Thanks for everything. My elder brother and my best friend, Bhai, you are the best! I made it just because of your inspiration. Many thanks go to my lovely sister-in-law Ananya and my cute nephew Dev.

I would like to thank all my teachers that lead me to the world of science. Sharing knowledge with young generations and preparing them a sound basis to create is why universities exist. The example set by many of my great teachers is one of the motivations that encouraged me to pursue this Ph.D. With this, I would like to express my sincere appreciation to my mentor team: Prof. Mart van der Meijden for his confidence to recruit me as a Ph.D. researcher for the ambitious GARPUR project and Dr. José Rueda for his supervision, inspiration, and advice to help me structure my work. I would like to express my sincerest gratitude to both of you for your guidance, understanding and supporting me towards the completion of my research. I am also grateful for your support for me to attend many conferences to broaden my professional network and meaningful summer schools to acquire valuable knowledge and experience. Thank you Dr. Michiel Pertijs for being a nice mentor and involved in discussions about my research. At CWI, I would like to thank Eric and Nanda for being great mentors.

My gratitude goes to the committee members as well for reading, evaluating my dissertation, and participating in my defense.

I consider myself very lucky that I was involved in the ambitious GARPUR project, a European Union collaborative project of nineteen partners and whose participants came from ten European countries. It was a great pleasure for me not only to communicate and work with experts in the field of power system reliability but also to have the opportunity to travel and get to know a more of those countries. I am very

glad to meet Prof. Wehenkel, Prof. Kjølle, Prof. Bell, Remy, Konstantin, Samuel, Camille, Oddbjorn and the rest of GARPUR team.

A big shout out to my gang at work. Bart: it was nice being a great colleague and a kind neighbor next to my desk in the office. Thanks a lot for your kind help in translating the summary and proposition of this thesis. Mario: thanks for being a great office mate and a friend. Matija, Hossein, Rishabh: I believe it was more fun being friends and not just colleagues. Matija and Tina: It was fun treating you Indian cuisine and thank you for being great friends. Hossein: Gent, it is fun having you as a good friend. Rishabh: Glad to meet you. Discussions on politics, TV series (you know it) are just interesting parts of our friendship. Andreas: I know you are having a great time in Sweden but I would like to thank you, starting from my interview to my first days at IEPG. This thankful message would be incomplete without appreciating our beautiful secretaries: Ellen and Sharmila. Many thanks to both of you for all kind of administrative and paper works, and talks; keep smiling always. I would like to thank other IEPGers as well for being awesome colleagues.

Away from work, I would like to extend my special thanks to Aditya, Geetha, and Gözde. My gratitude goes to Mihir, Kishore, Mohit, and Prashant for fun-time at Delft. Thank you, Alexis, for being a great flatmate, a friend, and my thesis cover designer. I would also thank people at IEEE Student Board, PromooD, and Young CIGRE Netherlands.

At last, I am really thankful to the unnamed people without whom this amazing journey was not possible.

Swasti Ranjan Khuntia
Delft, 1st November 2018

LIST OF PUBLICATIONS

JOURNAL PAPERS

1. **Khuntia, S.R.**, Rueda, J.L., & van der Meijden, M.A.M.M. (2018). Risk-based security assessment of transmission line overloading considering spatio-temporal modeling of load and wind power using vine copula. *Under review.*
2. **Khuntia, S.R.**, Rueda, J.L., & van der Meijden, M.A.M.M. (2018). Spatio-temporal modeling using vine copula: Application to electricity load and wind power. *Electric Power System Research*. Under review.
3. **Khuntia, S.R.**, Rueda, J.L., & van der Meijden, M.A.M.M. (2018). Long-term electricity load forecasting considering volatility using a multiplicative error model. *Forecasting*. Accepted.
4. **Khuntia, S.R.**, Rueda, J.L., & van der Meijden, M.A.M.M. (2016). Forecasting the load of electrical power systems in mid-and long-term horizons: a review. *IET Generation, Transmission & Distribution, 10(16)*, 3971-3977.
5. **Khuntia, S.R.**, Tuinema, B.W., Rueda, J.L., & van der Meijden, M.A.M.M. (2016). Time-horizons in the planning and operation of transmission networks: an overview. *IET Generation, Transmission & Distribution, 10(4)*, 841-848.
6. **Khuntia, S.R.**, Rueda, J.L., Bouwman, S., & van der Meijden, M.A.M.M. (2016). A literature survey on asset management in electrical power [transmission and distribution] system. *International Transactions on Electrical Energy Systems*, *26*(10), 2123-2133.

CONFERENCE PAPERS

1. **Khuntia, S.R.**, Rueda, J.L., & van der Meijden, M.A.M.M. (2017). Spatio-temporal study for modeling high dimensional future uncertainties: Univariate to multivariate model. In 2018 IEEE PES General Meeting, Portland
2. **Khuntia, S.R.**, Rueda, J.L., & van der Meijden, M.A.M.M. (2017). Smart asset management for electric utilities: Big data and future. In 2017 World Congress on Engineering Asset Management, Brisbane.
3. **Khuntia, S.R.**, Rueda, J.L., & van der Meijden, M.A.M.M. (2016). Volatility in electrical load forecasting for long-term horizon—An ARIMA-GARCH approach. In IEEE Probabilistic Methods Applied to Power Systems (PMAPS), Beijing.
4. **Khuntia, S.R.**, Rueda, J.L., & van der Meijden, M.A.M.M. (2016). Neural network-based load forecasting and error implication for short-term horizon. In International

Joint Conference on Neural Networks (IJCNN), Vancouver. Awarded "*IEEE CIS Outstanding Student-Paper Travel Grant*".

5. **Khuntia, S.R.**, Rueda, J.L., Bouwman, S., & van der Meijden, M.A.M.M. (2015). Classification, domains and risk assessment in asset management: A literature study. In IEEE International Universities Power Engineering Conference (UPEC), Staffordshire.
6. **Khuntia, S.R.**, Rueda, J.L., & van der Meijden, M.A.M.M. (2015). Mutual information based Bayesian analysis of power system reliability. In IEEE PowerTech, Eindhoven.

PREPRINTS

1. **Khuntia, S.R.**, Rueda, J.L., & van der Meijden, M.A.M.M. (2017). Pathway for multivariate dependence modeling in long-term horizon of electrical power system. *arXiv* preprint arXiv:1708.05404.

PROJECT DELIVERABLES

All public deliverables can be found at:
https://www.sintef.no/projectweb/garpur/deliverables/

1. Hamon, C., Perkin, S., Clement, R., Tournebise, P., Tabatabaeipour, S.M., Silber, P., Hering, P., **Khuntia, S.R.**, Rueda, J.L., Chaouachi, A., Kuwahata, R. (2016). How to upgrade reliability management for short-term decision making. GARPUR Consortium.
2. Wehenkel, L., Karangelos, E., Mannor, S., Dalal, G., Clément, R., Rueda, J.L., **Khuntia, S.R.**, Kjolle, G., Baldursson, F.M., Weber, C., Bellenbaum, J., Baldursdóttir, Í., Perkin, S., Marián, B., Campion, B. (2016). Guidelines for implementing the new reliability assessment and optimization methodology. GARPUR Consortium.
3. Vergnol, A., Campion, B., Sprooten, J., Gerasimov, K., Gamst, M., **Khuntia, S.R.**, Bell, K., Bukhsh, W., Martinez, E. (2016). Upgrading of the decision-making process for system development (2016). GARPUR Consortium.
4. Wehenkel, L., Karangelos, E., Marin, M., Mannor, S., Dalal, G., Giboa, E., Clément, R., Stevenin, P., Tournebise, P., Catrinu-Renstrøm, M.D., **Khuntia, S.R.**, Rueda, J.L., Janeček, P. (2016). Pathways for mid-term and long-term asset management. GARPUR Consortium.
5. Clement, R., Tournebise, P., Weynants, A., Perkin, S., Johansen, K., **Khuntia, S.R.**, Janeček, P., Catrinu-Renstrom, M.D., Kile, H. (2015). Functional analysis of Asset Management processes. GARPUR Consortium.

P.S.: *In addition to the public deliverables, I contributed to other **seven** internal deliverables which are accessible only by the GARPUR consortium members*.

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

In its attempt to achieve a greener future, the electric power industry has begun undergoing an energy transition towards the establishment of a low-carbon power system, especially since the increasing penetration of Renewable Energy Sources (RES) and decentralized energy generation have decreased reliance on fossil fuels. Since 2009, the trend of increasing penetration of RES in power systems, such as wind and solar energy, was accelerated by European directives aiming to reduce greenhouse gas emissions. This tendency will likely increase in order to meet the targets of the European Commission to establish a reliable, competitive and sustainable European interconnected power system [Zervos et al., 2011]. Complying to the 2016 Paris Agreement [Web2, 2016], such an energy transition radically alters the development and design context for the transmission grid, which results in further change or adoption of new or alternative technologies and practices in accordance with energy policies. By 2030, for instance, renewable energies will dominate conventional generation in the European electric power system as per the ENTSO-E Ten Year Network Development Plan (TYNDP) [Web3, 2016]. It aims to reduce domestic greenhouse gas emissions by 40% and increase RES penetration to at least 27% at the European level. In 2017, the total installed additional wind power capacity in Europe was $16.8GW$ ($15.6GW$ in the EU), an increase of 25% compared to the 2016 annual installations [EWEA, 2017]. To achieve the target, Transmission System Operators[1] (TSOs) need to *keep the lights on* while performing the fundamental tasks of system balancing and scheduling with substantial penetration of RES. In recent years, many efforts have been made in the field of research and development of alternative probabilistic reliability assessment methodology with European projects like iTesla [Web4, 2016], Umbrella [Web5, 2016] and GARPUR [Web8, 2017] as compared to the deterministic methods (e.g., N-1 criterion) that are still preferred by system operators for grid development

[1] A Transmission System Operator (TSO) is an entity entrusted with transporting energy in the form of natural gas or electrical power on a national or regional level, using fixed infrastructure which is defined by the European Commission. Similar organizational categories in the United States are: Independent System Operator (ISO) and Regional Transmission Organization (RTO).

and operation planning measures. The reason why deterministic methods are preferred is due to their robustness, ease-of-use and achievement of satisfactory results with respect to the accepted reliability level obtained so far by TSOs. Within such a deterministic approach, a relatively small set of selected, pre-specified credible contingencies (e.g. removal of generation units or transmission assets) are identified and checked so that the grid does not suffer from line overloading or voltage violation. Such a deterministic security assessment fails to check the probabilistic nature of system behavior and stochastic nature of exogenous variables.

As such, it is important to introduce exogenous variables that are explicitly modeled in a reliability management task. Exogenous variables are determined external to the TSOs, and the latter will have to adapt their behavior accordingly [GARPUR, 2016b]. Some examples of exogenous variables are electricity load, wind power, solar power, market output, forced outage rates and criticality of supply interruption. Among the well-known shortcomings of the current practices of TSOs is that, regardless of the time frame, the modeling of exogenous factors is not taken into account in a probabilistic manner. Such considerations are known to the operators and influence them, despite the fact that it is hardly possible to measure anything quantitatively. TSOs already work with the multi-scenario deterministic (power flow based) approach that is combined with matrix-based risk appraisal. However, the multi-scenario deterministic approach does not take into account the variability and uncertainty associated with exogenous variables. It is also imperative to note that these exogenous variables vary throughout the year and depend on the geographical area covered by the transmission system. For instance, wind and solar power generation are highly variable and uncertain in nature, resulting in a more locally distributed generation when compared to the traditional system with large centralized generation plants. Their integration into the power grid is a real challenge for TSOs, with respect to both infrastructure management and control of energy flows. In Germany, for example, wind power generation is concentrated in the northern part, solar power generation is concentrated in the southern part and load centers are mostly in the mid-western and southern parts of the country [Web1, 2013]. The variable nature of wind power with spatial diversity is destabilizing the electric grids (e.g., potential blackouts or weakening voltage). Hence, while renewable energy resources are significant, their location is non-uniformly distributed both in space and time, and often far from load centers. Regarding electricity load, energy consumption everywhere has increased tremendously in the last 150 years [ECF, 2010]. In 2015, ENTSO-E consumption reached $3{,}278TWh$, which represents a 1.4% increase compared to 2014. By 2016, ENTSO-E consumption had reached $3{,}322TWh$, which represented greater stabilization ($+0.6\%$) when compared to 2015 [Web4, 2016]. When observing the last 10-year period from 2005 to 2015, the electricity consumption

of households fell in the EU-28[2] by 0.9%. These figures on overall household electricity consumption are likely to be influenced, in part, by the average number of persons living in each household and the total number of households, both of which are linked to demographic events [Web7, 2017].

Uncertainty in wind power generation and load growth is bundled with the ageing power system's infrastructure. In terms of transmission infrastructure, the European electricity grid was built decades ago (nearly 30-40 years) and has provided highly reliable electricity to date. For example, TenneT[3] in the Netherlands has maintained an availability level of 99.99% in 2017 [Web9, 2018]. TSOs had a relatively constant activity level (one-way generation and consumption) with rather risk-averse and age-based replacement policies for system security. However, the practices are now being challenged with the massive integration of Wind Power Plants (WPPs) and feeding of load centers at distributed locations. However, trying to tap more wind power into the existing grid is challenging due to its irregular availability and variability. For instance, to increase the utilization of wind power, investments in wind farms are concentrated at locations with higher average wind speeds. TSOs must cautiously evaluate operation as well as future planning when power output fluctuations occur in such spatially distributed systems. In addition, it is well known that wind speed is temporally correlated at one location and both spatially and temporally correlated in different locations. Moreover, the location of WPPs and load centers are not so close by such that the non-dispatchable sources can be easily managed or curtailed. This spatial diversity imposes a burden on TSOs, which must operate the existing infrastructure despite uncertainty in terms of both demand (load centers) *and* generation (WPPs).

In principle, the participation of WPPs in the existing grid is different from conventional generators in regarding uncertain and variable output. Such uncertainty and variability introduce a level of risk that adversely affects day-ahead operational planning decisions. For example, variation in wind power hampers the power system's operation in real-time when WPPs are unable to deliver the required reserve capacities in real-time. The embedding of WPPs also raise concerns in terms of the planning and upgrading of existing infrastructure concerning the size, location and distribution of WPPs. Hence, it is vital to model the inter-spatial dependence and temporal correlation, but this should not only be done for wind power. A joint probability distribution, which is essentially a multivariate model considering electricity load and wind power, is additionally needed. The first step in this process is to obtain a tractable model that captures the uncertainties and correlations between both variables. To have a clearer understanding of the TSOs' perspective, Distributed Energy Resources (DERs) like solar or battery storage are not considered in this research since it is more pertinent at the

[2] This refers to the 28 member states of the European Union

[3] TenneT TSO is the transmission system operator in the Netherlands and in a large part of Germany.

distribution level. It is intended that the future power system will be data-centric and collecting data from the stochastic sources in terms of spatial and temporal resolution will result in a high dimensional database of varied features. This huge chunk of data, referred to as big data, is explored in electric power systems such as big data analytics. To tackle this high dimensional data, a suitable statistical approach is indeed required to mine useful information and ease the computational burden. Although the complexity of the electric grid tends to increase in terms of data, models and available tools, recent advances in the fields of mathematical programming, statistics, machine learning and power system simulation can be leveraged to construct suitable modeling tools to assist with system operation activities aiming for security of supply. A blend of a statistical model and power system simulation tool is needed, although the selection of a suitable method is constrained by the data availability and modeling approach.

To sum up, there is an immediate requirement for the development of spatio-temporal modeling of load and wind power as joint probability distribution for three reasons. *Firstly*, inter-spatial dependence and temporal correlation of load and wind power in any considered site is important. The literature review (detailed in chapter 4) reveals a consideration of load and wind power as independent variables and also some instances of temporal or spatial dependence. However, there are no significant findings that investigate the spatio-temporal dependence of these two exogenous variables. *Secondly*, a suitable spatio-temporal modeling approach will facilitate the improvement of both short-term operational planning and the long-term grid development of power grids. For instance, in short-term operational planning, accurate spatio-temporal modeling can help in assessing system security in terms of asset overloading or reducing operational costs by using forecast values for unit commitment or reducing wind curtailment. Similarly, in terms of long-term grid development, appropriate modeling can result in grid development plans that respond to load growth or massive integration of wind power. *Lastly*, a suitable spatio-temporal multivariate model can generate a rich synthetic database of normally distributed load and wind power data. Such a database will be of immense help to the research community and industry as well to assist in developing other statistical tools.

Due to the fact that the location of WPPs and demand sites are not always close by and that the transmission of energy places a burden on the existing grid infrastructure, this burden allows transmission lines to function more and more frequently close to their operating limits (both physical as well as power rating limits) and exerts unexpected stress on them (which can also vary in real-time). To date, for N no. of lines, TSOs check that the system is not overloaded for N-1 or N-2 contingencies by a certain percentage as specified by TSOs. The overloading percentage is country specific and varies from one to another. It can certainly be below or above the rated operating limit depending on the security margins (or risks) adopted. In addition, the voltage at all nodes is checked to ensure that it is within the required limits. As massive integration of

wind power is an important future plan, the risk of transmission line overloading and voltage instability cannot be neglected. The traditional approach for TSOs to make sure the network is sufficiently robust basically consists of stressing the average forecast and verifying that the security of supply is met. In the event of a transmission asset failure, system security is certainly under threat and it affects system dynamics which might increase the likelihood of line overload, low voltage or even voltage collapse. Understanding such challenges requires addressing the spatio-temporal modeling of load and wind power. Though achievements have been made in terms of efficient forecast of future demand and wind power generation, there are other vital concerns corresponding to wind power such as spatio-temporal correlations, variability, non-normality, non-stationarity, non-dispatchable nature of the energy source (unless there is adequate storage) and seasonal patterns to name a few. Some of the reasons leading to the transmission infrastructure operating under stress include:

i. Introduction of new technologies in terms of structural changes [more interconnections, variable RES, flexible AC transmission systems (FACTS), high voltage direct current (HVDC), and other power electronics-based devices] and uncertainty in electricity demand growth;
ii. Stranded network expansion due to public opposition to the construction of new transmission infrastructure, which forces the same aging grid to accommodate the massive penetration of RES;
iii. High variability of operating conditions and bidirectional power flows are introduced by stochastic power infeed and stochastic behavior of prosumers; and
iv. New operating policies like liberalization of the energy market and more intense trading, coupled with markets and higher demand-side participation.

The next section will elaborate on the scope of this research.

1.2 RESEARCH SCOPE

Prior to answering the question "*What is within the scope of this research?*" this research will attempt to bridge some scientific gaps in terms of:

- *Power system time-horizons*: This research starts by revisiting the *traditional concept* of time-horizons and activities performed under each time-horizon by TSOs. Traditionally, the three time-horizons and corresponding activities are: long-term horizon (system development), mid-term horizon (asset management) and short-term horizon (system operation). The *traditional concept* refers to the more established actions that are undertaken by TSOs, which are implemented more or less separately and can be described within a sequential approach. The transition from a sequential to an integrated (or interactive) approach is essential when

tackling exogenous variables. Within this research, such an interactive approach has been realized when TSOs need to make decisions for system development (in terms of constructing new transmission lines, etc.) and when they have to tackle the future uncertainty in terms of load growth and wind power generation at spatially distributed locations. If TSOs agree to undergo system development, they may face some challenges, such as those associated with:

  - Developing a new transmission corridor (plan and execution), which takes approximately 8-10 years and the construction of high voltage and extra high voltage transmission lines is very expensive;
  - Planning a cost-effective expansion plan to recover the capital expenditure of a new transmission corridor built through one zone when the corridor provides benefits to consumers in other zones;
  - Building long-distance transmission lines, such as interconnectors to WPPs, where high-quality wind is located with the rest of the grid, which also includes demand sites; and
  - Expanding the grid's infrastructure in the digital age means protecting the grid from both physical and cyber-attacks. Creating such a smart grid involves integrating the cyber system with the physical power system. Although adopting a cyber system has and will make the grid more energy-efficient and modernized, cyber-attacks can possibly threaten national infrastructure security and customer satisfaction.

- *Load modeling in short- and long-term horizons*: Due to the fact that the electricity load has a seasonal or temporal component, most current studies aim to model temporal correlation. While many models exist for short-term load forecasting, it is important to develop a model with high forecast accuracy and an efficient way of representing forecast error for reliability studies. When compared to the short-term, long-term forecasting is completely different and involves multiple factors (i.e., economics and so on) apart from historical load data. Volatility has been identified as a key factor that affects long-term forecasting and was used in developing the model. When considering the temporal correlation only and ignoring spatial correlation among load growth in different zones under a transmission system proved to be inefficient, a multivariate model was developed to solve this issue.
- *Spatio-temporal modeling of load and wind power as joint probability distribution*: Stochasticity of wind makes it difficult to predict accurate wind power output when only considering temporal wind behavior, although it is also affected by other geographical and technical factors like wind farm topology and wind turbine characteristics. Similar spatial patterns among wind power data favored spatial correlation studies for wind power and seasonal patterns of load favored temporal

correlation studies. It also should not be forgotten that meaningful correlation exists between load and wind power because both are significantly affected by weather. As the future sees uncertainty in both load and wind power, capturing the inherent dependence between load and wind power in different temporal and spatial contexts is lacking in the current state and achieved by adopting a multivariate modeling approach.

- *Use the developed model for a risk-based security assessment of transmission line overloading risk*: This research is innovatively attempting to assess transmission line overloading risk while addressing spatio-temporal dependence using vine copula. The reproducible sampling algorithm uses spatially distributed load and wind power data spanning a three years horizon to model joint normal distribution. The calculation of overload risk indices are accomplished with probabilistic AC load flow and real-life data mapped onto a modified IEEE 39-bus system. A severity function is employed to assess the consequences of overloading in terms of likelihood of its occurrence and its associated impact.

One key question that falls under this research scope and has been answered is: "*Is it a smart idea to operate the grid with the existing infrastructure based on correlation studies of demand growth and wind power generation since both are stochastic in nature*?" The next section describes the main research objective and research questions that ensue from the research scope.

1.3 RESEARCH QUESTIONS

The main research objective of this thesis is:

To develop a statistical model that can address the spatio-temporal dependency of multiple exogenous variables and, thereby, validate the same developed model to study the extent to which the grid can function closer to the operating limits by performing a risk-based security assessment in case of transmission line overload.

Overall, the research questions for this research can be summarized as:

Q.1. Power system time-horizons

- *What is the implication of different TSO actions taken in different time-horizons on power system reliability?*
- *Will the traditional concept of time-horizons be valid in future when there is uncertain load growth and high penetration of renewable energy into existing transmission grid infrastructure?*

Q.2. Load modeling and forecasting in short- and long-term horizons

- *How should uncertainty in load growth be addressed and what are the associated modeling challenges in the short-term and long-term horizons?*
- *How can forecast error be accounted for in terms of error distribution in the short-term horizon?*
- *What is the role of volatility in long-term forecasting and how does it impact the modeling framework?*

Q.3. Spatio-temporal modeling of load and wind power as joint probability distribution

- *How should load variability and wind power generation for spatially distributed locations in a large-scale system be modeled?*
- *How can both spatial as well as temporal correlations be effectively addressed?*
- *How can high dimensional data be accounted for when the future will be data-centric?*

Q.4. Use the developed model for a risk-based security assessment of transmission line overloading

- *Does the consideration of spatio-temporal dependence of load and wind power prove beneficial to quantify the risk of overloading transmission lines?*
- *How does the correlation impact the risk values of line overload?*
- *How do the risk values of individual lines or the entire system enable the system operator to assess system operation condition?*

1.4 ORIGINAL CONTRIBUTION

This research addresses and validates various novel content related to time-horizons, load modeling and forecasting in the short-term and long-term horizons, dependency modeling to address spatio-temporal correlation and a risk-based security assessment technique for transmission line overloading. The main contributions of this research, while answering to the research questions defined in section 1.3, can be summarized as:

- This research re-visited the definition of time-horizons for current and future operation and planning of the power system, explicitly describing the drawbacks of previous practices and the need for upgrading;
- This research built models to forecast load in the short-term and long-term horizons. A neural network-based load forecasting technique was developed for the short-term horizon and the truncated normal distribution for error modeling was proposed. For the long-term horizon, a multiplicative error model was proposed by addressing the volatility foreseen in long-term forecasting and focusing on both directional accuracy and minimal forecast error as well;

- This research modeled the inter-spatial dependency and temporal correlation between load and wind power as a joint probability distribution to obtain joint normal distribution using vine copula. A reproducible sampling algorithm has been developed for spatio-temporal modeling, which also tackles the high dimensionality of data using clustering and feature extraction techniques;
- In this research, a risk-based security assessment of transmission line overloading is performed taking spatio-temporal dependency of load and wind power into account. Using a severity function, the overloading risk is quantified by performing a probabilistic AC load flow on the modified IEEE 39-bus system with real load and wind power data. A comparative analysis of uncorrelated and correlated samples proves the advantage of considering spatio-temporal correlation.

1.5 THESIS STRUCTURE

This thesis consists of four technical chapters (chapters 2-5) that aim to address the above-mentioned research questions, followed by the conclusion and future recommendations (chapter 6). It is to be noted that the relevant literature reviews pertaining to each topic are included in their respective chapters. A pictorial overview of the structure of the thesis is shown in Fig. 1.1.

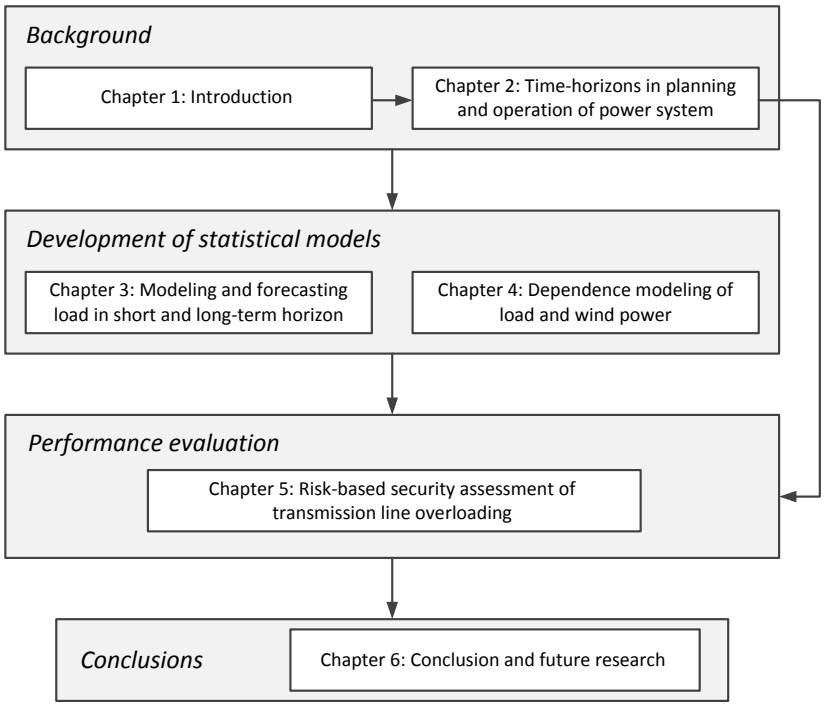


Fig. 1.1: Overview of the thesis

Chapter 2 lays the foundation of this research by re-visiting TSO actions undertaken during operation and planning in the three time-horizons. It also discusses the inter-dependency of activities in different time-horizons and the obligation to change utilities' current practices. This chapter further serves as an in-depth review of asset management.

Chapter 3 contributes to modeling and forecasting electricity load in the short- and long-term horizons. For the short-term horizon, a neural network-based model is developed to forecast hourly load and represent forecast error using a new distribution method called the truncated normal distribution. Challenges and requirements for forecasting in long-term horizon is also discussed in this chapter. For long-term horizon, a multiplicative error model is developed taking volatility into consideration. The methodology performance is checked during the Great Recession of 2008 to account for directional accuracy.

Chapter 4 describes the appropriate methodology to address spatio-temporal correlation using vine copula. The procedure to model the joint probability distribution using dependency modeling is elaborated in the form of a reproducible sampling algorithm. To tackle the high dimensional data, clustering and the feature extraction technique is employed and described in this chapter.

Chapter 5 presents the performance of the developed statistical models in the form of a risk-based security assessment. It describes the risk of transmission line overloading in the form of a severity function and aims at risk quantification. A probabilistic AC load flow is run on modified (with addition of WPPs) IEEE test cases and simulation results are described to acknowledge the consideration of spatio-temporal correlation for the assessment of transmission line overloading risk.

Chapter 6 is the final chapter and summarizes the key findings and scientific contributions from each of the previous chapters of this study. It also gives recommendations for future research in relation to forecasting and error modeling, extending the study to learn TSO-DSO interaction and the application of big data analytics.

1.6 CONTRIBUTION OF THIS THESIS TO GARPUR PROJECT

Power system reliability management aims to maintain power system performance at a desired level, while minimizing the socio-economic costs of keeping the power system at that performance level. Historically in Europe, network reliability management has been dependent on the so-called $N-1$ criterion: in case of fault of one relevant element (e.g. one transmission system element, one significant generation element or

one significant distribution network element), the elements remaining in operation must be capable of accommodating the new operational situation without violating the network's operational security limits. Not only today but keeping an eye on future, the increasing uncertainty of generation due to stochastic energy sources, combined with the opportunities provided e.g. by demand-side management and energy storage, there is a call for imagining new reliability criteria with a better balance between reliability and costs. The GARPUR project designed, developed, assessed and evaluated such new reliability criteria to be progressively implemented over the next decades at the pan-European level, while maximizing social welfare. GARPUR stands for Generally Accepted Reliability Principle with Uncertainty modeling and through probabilistic Risk assessment. It was a large collaborative R&D project co-funded by the European Commission 7th Framework Programme. Detailed information may be found on the website of the project [Web8, 2017]. Ensuring good reliability of the transmission assets comes at a cost, which motivated the introduction of asset management in the GARPUR project. In addition, variability and uncertainty of renewables penetrating into primary grid calls for re-evaluation of grid reliability. Major contribution of this research towards the project was addressing the gap in terms of data, models and tools for dealing with uncertainties for further upgrading reliability management. Fig. 1.2 shows the participation of this research in various work packages of project.

The contribution was distributed in Work Packages (WPs) 2, 4-7, 9. Within WP2, the main task involved ensuring the algorithmic feasibility, scalability, and sustainability while stating the main practical requirements for the Reliability Management Approach and Criterions (RMACs) [GARPUR, 2016a] that was developed in WP2. In an effort to analyze the structure of the mathematical formulations developed in WP2, the contribution focused on projecting the formulations in the form of feasible (i.e., tractable and scalable) algorithmic approximations. WP4 involved long-term planning (or grid development) studies, which looked into finding representative credible operating states. It aimed at defining how to manage the different uncertainties that occur during future operational activities, notably how to synthesize them while maintaining the credibility of the assumptions and physical meaning including relevant temporal and spatial correlations between them. From this research perspective, the focus was on tackling volatility in long-term demand forecasting and proposing a new methodology to predict demand growth in the long-term horizon.

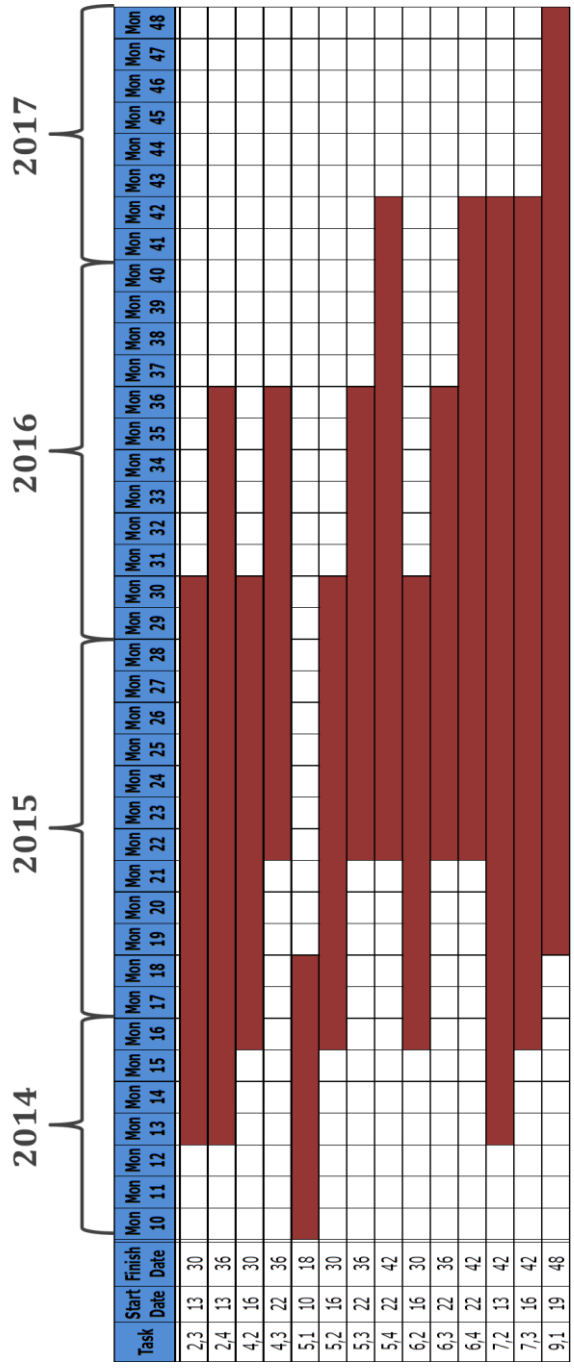


Fig. 1.2: Participation in GARPUR [Task x,y refers to work package x and sub-task y, Mon refers to Month number]

Work Package 5 involved mid-term planning studies which aimed at asset management decision making process, and it also addressed the long-term 'maintenance budgeting' problem. An overview of asset management with a more detailed description of the different input/output data, decisions, and sub-activities was carried out. It started with brief discussion on the two different notions of uncertainty, namely macro- and micro- uncertainties that need to be taken into account in the assessment of maintenance policies and/or outage schedules. Generative models of micro-scenarios were proposed in order to automatically generate the required input about uncertainties needed in the proposed algorithms. In order to produce on demand with the required flexibility the needed samples of micro-scenarios, it was decided to go for the development of a generative model that can be used both for the needs of mid-term and long-terms studies. In essence, such a generative model of micro-scenarios would be a software tool that can be called upon request with a certain number of input parameters, and that will generate efficiently a sample (of specified size) of micro-scenarios described by the relevant output-variables. By this, we mean that the generative model will use a probabilistic model from which it will sample a specified number of (say yearly) micro-scenarios independently. Though the work aimed to cover the specific needs and tools for uncertainty models in the context of the asset management tasks covered by WP5, the envisaged methods share many commonalities with similar ones envisaged in other WPs of GARPUR, e.g. on the one hand for long-term system expansion studies and on the other hand for short-term operation planning contexts. While we already have taken advantage of the two-fold contexts of asset management (namely long-term maintenance policy on the one hand, and mid-term outage schedule assessment on the other hand) in order to specify a common micro-scenario generation tool, it was believed that further work could be carried out to fully coordinate the various uncertainty models used in the different GARPUR contexts. The work in this WP concluded with future challenges on asset condition monitoring and use of big data techniques in asset management, which was envisioned for post-GARPUR research and development work.

Involvement in WP6 aimed at presenting the requirements for adapting available tools/models and identifying data needs for probabilistic reliability analysis and optimal decision-making in the short-term decision making process. In the GARPUR proposal, it was needed to generate sequences of realizations of the exogenous parameters (for example, realizations of nodal loads and RES production for each hour of a day). It is important to capture both the spatial as well as temporal correlation when generating sequences of realizations, and the proposed methodology was the use of vine copula models. The proposed method is adequate for generating such sequences or scenarios. Along with the specifications for uncertainty modeling, data requirements for realization of vine copula methodology was proved to be important. It was experienced that data availability is a deciding factor, whether to go for dependence modeling using

the joint normal transform or select available features from inadequate data and use dependency modeling using vine copula models. And, the selection of a suitable method is restricted by the data availability and the daily grid operation processes. Work package 7 designed the GARPUR Quantification Platform (GQP), that allowed comparison of different reliability management strategies via numerical simulations of their application in different applications contexts. Different use-cases of the platform were first defined, and the necessary inputs and outputs are defined according to these use-cases. Within WP6, a neural network based load forecasting methodology was proposed, which was also used in GQP in WP7.

REFERENCES

[ECF, 2010] Roadmap 2050, A practical guide to a prosperous, low-carbon Europe, European Climate Foundation, April 2010

[EWEA, 2017] European wind energy association. Wind in power: 2017 European statistics. [Online] https://windeurope.org/wp-content/uploads/files/about-wind/statistics/WindEurope-Annual-Statistics-2017.pdf

[GARPUR, 2016a] GARPUR Consortium. *D2.2: Guidelines for implementing the new reliability assessment and optimization methodology*. GARPUR, 2016.

[GARPUR, 2016b] GARPUR Consortium. *D5.2: Pathways for mid-term and long-term asset management*. GARPUR, 2016

[Web1, 2013] http://instituteforenergyresearch.org/analysis/germanys-green-energy-destabilizing-electric-grids/

[Web2, 2016] http://unfccc.int/paris_agreement/items/9485.php

[Web3, 2016] http://tyndp.entsoe.eu/2016/

[Web4, 2016] https://www.entsoe.eu/Documents/Publications/Statistics/electricity_in_europe/entsoe_electricity_in_europe_2016_web.pdf

[Web5, 2016] http://www.itesla-project.eu/

[Web6, 2016] http://www.e-umbrella.eu/

[Web7, 2017] http://ec.europa.eu/eurostat/statistics-explained/index.php/Electricity_production,_consumption_and_market_overview

[Web8, 2017] http://www.garpur-project.eu/

[Web9, 2018] Integrated Annual Report 2017, TenneT TSO B.V., 2018.

[Zervos et al., 2011] Zervos, A., Lins, C., Tesniere, L., & Smith, E. (2011). Mapping renewable energy pathways towards 2020. *European renewable energy council*, European Renewable Energy Council, Brussels.

CHAPTER 2

POWER SYSTEM TIME-HORIZONS: BACKGROUND AND FUTURE

2.1 INTRODUCTION

This chapter aims at answering the first research question Q.1., which deals with power system time-horizons,

- *What is the implication of different TSO actions taken in different time-horizons on power system reliability?*
- *Will the traditional concept of time-horizons be valid in the future when there is uncertain load growth and high penetration of renewable energy into the existing transmission grid infrastructure?;*

The content of this chapter is based on research papers [Khuntia *et al*., 2016a, Khuntia *et al*., 2016b] where the key point in answering the research question is by exploring the current status of the electrical power system. In the planning and operation of power system, actions are taken for different activities and in different time-horizons. The purpose of these actions is to secure a high reliability level and a persistent security of supply. In power systems context, reliability assessment can be divided into two categories; system adequacy and system security. System adequacy is generally considered to be the existence of sufficient facilities within the system to satisfy the function of the transmission system. System security is concerned with the ability of the system to respond acceptably to contingencies. Today's scenario places tremendous stress on transmission assets because of:

- *Development and uncertainty in electricity demand growth*.
- *Structural changes [more interconnections, variable RES, flexible AC transmission systems (FACTS), high voltage direct current (HVDC) systems, and other power electronics-based devices]*.
- *New operating policies like liberalization of the energy market, more intense trading, coupled markets, higher demand-side participation*.

This has led the power system reliability specialists to divide their activities into three main processes in which sets of decisions are taken [Wood & Wollenberg, 2012]:

i. Grid development (long-term horizon)
ii. Asset management (AM) (mid-term horizon)
iii. System operation (short-term horizon)

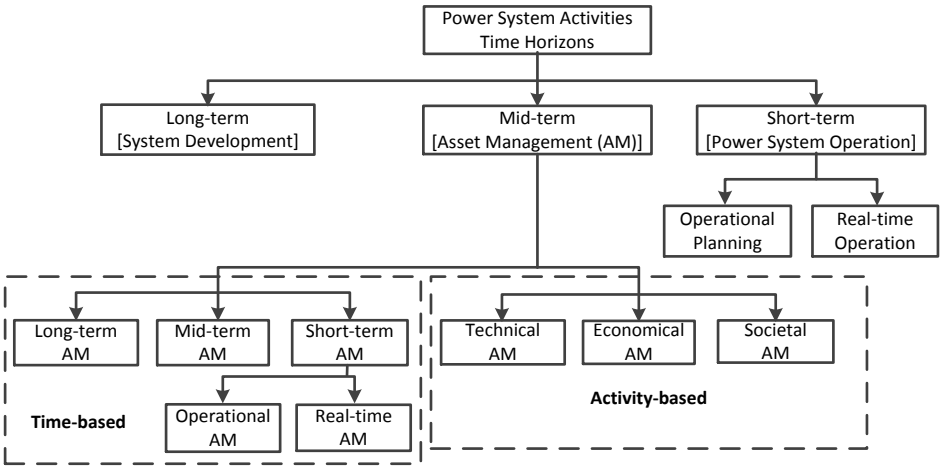


Fig. 2.1: Classification of the time-horizons according to literature

However, in reality under each of these three main processes, various sub-activities are performed on different time-horizons as illustrated in Fig. 2.1. Although the three main processes (grid development, asset management, and system operation) are described in the literature, there has been no explicit study on the time-horizons (long-, mid-, and short-term) and actual time scale (decades, years, months, etc.) that these processes focus on. This chapter aims at re-visiting the current state-of-the-art of the various activities performed by transmission system operators (TSOs) while reviewing the concept of each time-horizon and methodologies developed in the literature. As decisions taken in different time-horizons can influence each other, the interactions and overlapping are discussed which will help in future decision-making process. The actual time scale of these horizons can vary between different activities and has never been clearly mentioned in any published literature, which often leads to confusion in practice. For example, long-term grid development is performed on a time scale of decades, while long-term system operation has a time scale of weeks/months as shown in Fig. 2.2. In fact, Fig. 2.2 is first-of-its-kind to classify and distinguish among the time-horizons, processes and activities. It was developed to understand the concept of three processes and horizons. It gives an overview of the activities that are performed in each process, and time-horizon and shows the actual time scale of these activities.

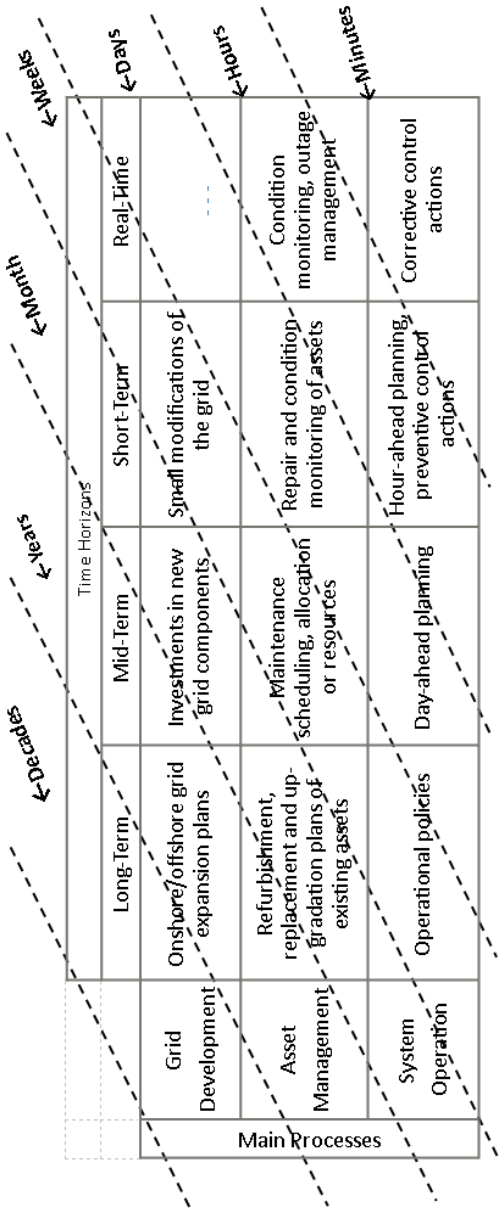


Fig. 2.2: Actions taken during different time-horizons [Khuntia *et al.*, 2016a]

Discussion on planning and operation brings 'reliability' into the picture. Power system reliability has been a subject of interest since the 1960s when [Billinton & Bollinger, 1968] published the first article in 1968. Since then, there has been growing interest in introducing various methods and theory to pursue reliable power system operation [Khuntia *et al.*, 2016a]. In the past three decades, there have been

developments in terms of new concepts, models, algorithms, software and its applications in assessing power system reliability. With the current pace of uncertainties getting introduced in power system operation and adoption of new technologies (RES, demand-side management, etc.), it is vital to reformulate the reliability management tasks.

In this chapter, section 2.2 focuses on describing the concept of different time-horizons and lists the different methodologies developed in various literature till date. Section 2.3 discusses the overlapping and interactions among the time-horizons, the possibilities for a combined reliability approach, and the challenges for the future. Finally, section 2.4 concludes the chapter.

2.2 REVIEW OF THREE MAIN PROCESSES

In this section, the three main processes and time-horizons are discussed supported by methods and tools developed till date.

2.2.1 GRID DEVELOPMENT

Grid development aims mainly at transmission system expansion planning. According to [Pereira et al., 1985], grid development can be divided into two parts:

i. Determining optimal investments in new system capacity.
ii. Determining system operating cost and supply reliability associated with the construction of this new capacity.

In short, developing the transmission infrastructure is one of the key priorities. An adequate transmission network is responsible for a safe, reliable, and efficient delivery of electrical energy to the consumers. Thus, grid development aims at providing solutions for the future. For an efficient planning, it is important to find the type, location, and timing of the network upgrades not only at a minimal cost but also considering socio-economic, environmental, legal, and political constraints. Since the fact that it generally covers the far future, grid development deals with a large number of uncertainties in various domains.

In practice, grid development is performed on longer and shorter time-horizons. Long-term grid development has a time scale of decades and includes the creation of grid expansion plans based on load/generation scenarios. In mid-term grid development, investments in the grid infrastructure (new connections, substations, etc.) are made. In short-term grid development, only small modifications of the network are made in the time scale of months. For instance, new protection systems, phase shifters to name a few can be installed in this time-horizon. Literature survey, in the last four decades, reveals that transmission system expansion planning has evolved due to the introduction of various mathematical models and techniques. An extensive list of

different models and techniques used in transmission expansion planning are enlisted in [Khuntia *et al*., 2016a]. To facilitate energy transition, there is an immense requirement of tools and knowledge-based schemes for decision making to integrate RES under market regulations and uncertainties [Milligan et al., 2012, Ugranli & Karatepe, 2016]. This is a challenging task because of specific properties of RES like stochastic behavior, non-linearities, and non-convexities. At the same time, electricity market also adds to the uncertainty [Munoz et al., 2014]. It can be deduced that risk and uncertainties have evolved due to advancement in technology, and will be evolving further. For instance, some of the novel emerging/prominent approaches for expansion planning are based on the least-effort criterion, maximum principle, minimizing the maximum regret or maximizing benefits, uncertainties, and security constraints. Introduction of stochastic studies are reported in [de la Torre et al., 2008] and use of optimization techniques date back to 1990s when [Hobbs, 1995] enlisted the use of optimization methods to tackle planning horizon in the electric utility. The reason was that mathematical modeling for grid development is a challenging task because of the presence of so many constraints and a high level of uncertainty. With this, we move to asset management.

2.2.2 ASSET MANAGEMENT

Asset Management (AM) is one of the key components in a transforming electric power industry. AM is closely related to grid development and system operation, hence forms a bridge between the long-term and short-term horizons. It is defined as the process of maximizing the return on investment of equipment over its entire life cycle by maximizing performance and minimizing costs (both capital expenditure and operational expenditure) at a given risk level [Tor & Shahidehpour, 2006]. The electric power industry is undergoing significant changes because of technical, socio-economic and environmental developments. The focus of TSOs has been on transmission assets that include transmission lines, power transformers, protection devices, substation equipment, and support structures. Transmission assets are capital-intensive and there is a requirement of utilizing them in the most efficient way.

CIGRE Joint Task Force JTF23.18 [Bartlett, 2002] describe AM as '*The Asset Management of Transmission and Distribution business operating in an electricity market involves the central key decision making for the network business to maximize long-term profits, whilst delivering high service levels to customers, with acceptable and manageable risks*.' Complying with the needs, TSOs are constantly striving to optimize the use of resources available for maintenance and new projects while ensuring system reliability is within satisfactory limits. As seen in Fig. 2.1, AM can be classified based on time domain and activity domain. The time-domain AM is categorized into long-, mid-, and short-term:

i. *Long-term asset management*: The time frame ranges from a year and beyond and it aims at replacement, refurbishment or up-gradation of existing assets like

phase-shifting transformers, reactive devices, and existing connections. This involves greater financial risks, and hence proper planning can avoid the risks involved in time delays, interest rates, and long-term load diversity.

ii. *Mid-term asset management*: The time frame of mid-term ranges from a few months to a year, and it involves optimal scheduling of asset maintenance and allocation of available resources. The primary aim is to extend the lifespan of existing assets through proper maintenance and optimally allocate the available generation resources through market modeling. Maintenance cost is the most crucial or driving factor and it can be greatly reduced when planned outages are scheduled according to the availability of resources during seasonal load distributions. So, an 'optimal' maintenance plan greatly reduces the possibility of unplanned outages. It is also the task of asset managers to check that maintenance schedule is planned based on system adequacy and fuel constraints on the non-maintainable system, like the availability of water in-flows for hydro plants. Reference [Tor & Shahidehpour, 2006] explain the mid-term asset management as:
    - minimizing corporate financial and physical risks based on planned and forced outages of assets
    - reducing operation costs for supplying customers in a competitive era
    - optimizing the allocation of volatile and limited natural resources for utilizing corporate assets
    - extending the lifespan of assets through proper operation and maintenance schedules
    - prolonging investment costs for the acquisition of new assets

iii. *Short-term asset management*: Short-term asset management is categorized into operational asset management (daily and weekly) and real-time asset management (outage management). Operational asset management aims at minimizing risks involved with assets, both physical and financial, due to load demand and hourly prices. Real-time asset management is also called asset outage management where contingency analysis forms a vital part. It helps in assessing the effect of unexpected outages due to change in weather conditions, any sudden breakdown or load fluctuations on the asset condition and performance. With technological advancements, real-time monitoring of assets is possible because of systems like Supervisory Control And Data Acquisition (SCADA) systems, remote terminal units and geographic information system (GIS). This has contributed significantly towards better management and decision-making process in short-term.

Based on the activity aspect, reference [Smit et al., 2006] categorize asset management into technical, economical and societal asset management, described below:

i. *Technical asset management*: Technical asset management refers to asset-related parameters such as the physical condition of assets, inventory, and maintenance. Aging of components is of primary concern that links to the physical condition of assets. Other areas in this aspect are the component condition, the failure probability of assets, inventory or spare parts and maintenance history and/or future planning.

ii. *Economical asset management*: Economical asset management evolved when technical asset management in many instances proved to be financially unstable. As the name suggests, economical asset management refers to financial aspects like maintenance costs and other costs related to the procurement of spare parts, maintaining the inventory and doing tests and assessment.

iii. *Societal asset management*: Societal asset management works closely with economical asset management. It refers to how the utilization of asset affects the society and the environment. The outage caused in high-priority buildings like the hospital is not acceptable. Also, any disturbances in other places like schools, government offices or convention centers will impact the status of distribution companies.

To classify the transmission assets falling under the scope of asset management, a survey was carried out as part of the GARPUR project and the results are shown in Fig. 2.3. It revealed that overhead lines, busbars, transformers, circuit breakers and protection systems are equally important by 9 TSOs out of 14 participating TSOs. It is evident that the grid components are naturally aging, especially when operated close to their technical limits, and most of them are subject to hostile external conditions in the open air. Asset management encompasses all the activities undertaken by a TSO to make sure that these devices are efficiently maintained, and thus can have an extended lifespan, or, once they have reached their end of life, are replaced by new ones. Inspection of assets and maintenance actions also aims to act as preventive measures which objective is to decrease the number of (unexpected) failures, therefore resulting in fewer service interruptions and a more efficient management of financial, human, and material resources.

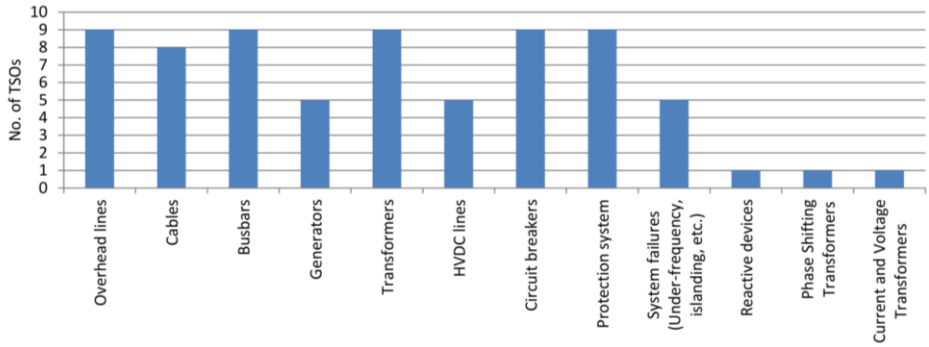


Fig. 2.3: Answers to question on components considered under asset management (Y-axis is number of TSOs) [Khuntia *et al.*, 2016b]

Aligning with the survey, literature study also reveals there have been explicit studies on maintenance of transmission assets, like power transformers, overhead lines, cables, and protection devices [Lindquist et al., 2008, Puglia et al., 2014, Velasquez-Contreras et al., 2011]. With the integration of RES, wind farms have been extensively studied from AM point of view in [Nilsson & Bertling, 2007, Puglia et al., 2014]. In general, the power transformer represents nearly 60% of the overall costs of the network and is ranked as one of the most important and expensive components [Jahromi et al., 2009]. Study reveals substantial research on power transformers about health monitoring, aging, and oil indicators [Ashkezari et al., 2014]. Similarly, studies have been carried out for overhead lines, underground cables, and circuit breakers which are more elaborated in [Khuntia *et al.*, 2016b]. With the advent of computational tools, information technology (IT), and human-machine interface in the last decade, [Kostic, 2003] studied on the application of IT in AM while focusing on energy management services. Various computational models and optimization techniques have been developed thereafter for maintenance planning, refurbishment, aging, and asset monitoring techniques like state diagram, fuzzy technique, ANN, PSO, linear programming, and other optimization techniques [Lindquist et al., 2005a, Lindquist et al., 2005b].

In the current state, one thing that binds together the physical and non-physical domains of asset management is data management. CIGRE WG D1.17 gives a clear picture of how asset management relies on asset data and information extracted from this data that is to be used in future planning [CIGRE, 2010b]. With the advent of computational tools and smart meters, a huge amount of data is collected by the utility companies that are used later for improving the performance of assets and/or maintenance policies. Data requirements for probabilistic concepts in asset management are huge and range from inspection rates and mean times to failure to probabilities of state transitions. For example, [Billinton & Allan, 1996] studied the effect of maintenance on the replacement time for transformers. The study used

reliability centered maintenance and a genetic algorithm to optimally schedule maintenance activities for the transformer. In the last decade, merging of data requirements with Information Technology (IT) and Human Machine Interface is shown by the studies of [Kostic, 2003]. It focused on the aspect of integrating IT in asset management by utilizing process data (e.g. SCADA systems, Energy Management Services (EMS)/Data Management Services (DMS)) in back-end tools such as enterprise resource planning, GIS, computerized maintenance management system and other analysis tools. The downside with the electric power industry is that utilities have used the existing models in an inefficient or wrong manner, and that paves way for data-mining process. A generic framework of data management that can be potentially used in asset management is shown in Fig. 2.4.

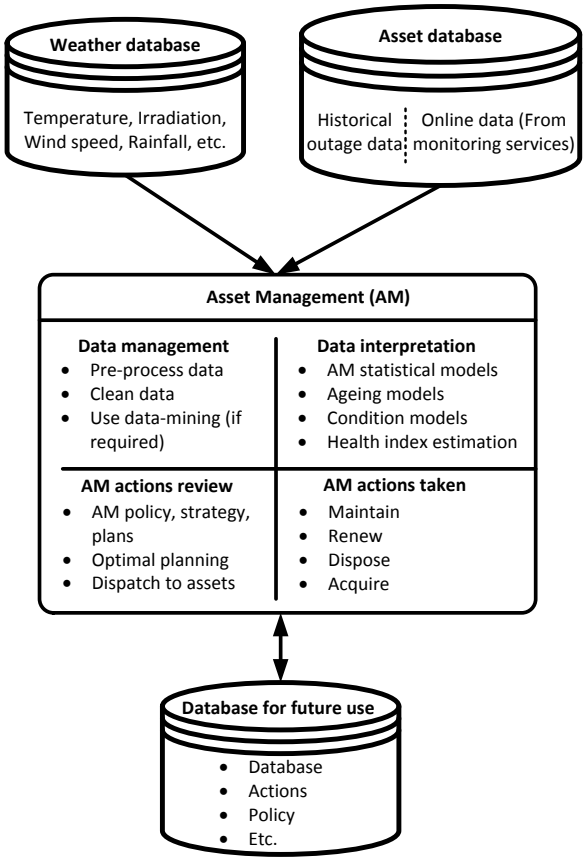


Fig. 2.4: A framework of data management used in asset management

Another survey on the usage of data in terms of data collection and the usefulness of data collection towards asset management is shown in Fig. 2.5.

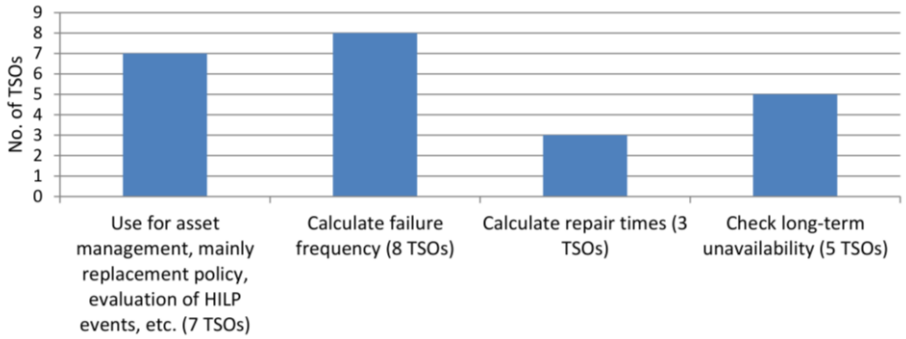


Fig. 2.5: Answers to question on usefulness of data collection for components (Y-axis is number of TSOs) [Khuntia *et al.*, 2016b]

2.2.3 SYSTEM OPERATION

System operation encompasses operational planning and real-time operation, which deals with activities ahead of real time. The duration of this time-horizon ranges from minutes/hours to several days ahead, though this can vary among different TSOs. Under system operation, reliability is also of primary concern and it is very important to maintain both security and adequacy levels at the acceptable levels with minimum socio-economic cost. System security level refers to the ability of the system to respond to failures [Billinton & Li, 1994, Billinton & Allan, 1996]. This ensures that the dynamics induced from any contingency or any operating conditions remain within an acceptable level. System adequacy indicates whether there are enough means in the system to fulfill its function, also during contingencies. The two sub-levels under system operation are:

i. *Operational planning*: Operational planning happens in several instances prior to the establishment of the system operating conditions. It constitutes the preparatory phases before the real-time operation. Also, operational planning ensures that right decisions are taken in advance such that reliability management is achievable within a prolonged future period of time, called the operational planning horizon. The horizon consists of a sequence of target real-time intervals. The operational planning time-horizon does not have a specific point in time, it can be week-ahead (W-1), two days (D-2) or one day (D-1) in advance as well as several (n) hours (H-n) before real time. Due to the unspecific points in time, operational planning brings significant uncertainties into consideration.

ii. *Real-time operation*: Real-time operation encompasses system operation for time intervals ranging $15 - 60$ min. During this time interval, it is assumed that the system operating conditions (scheduled generation, demand, inter-area exchange, and network configuration) are highly predictable. Fig. 2.6 shows the inter-dependency of real-time operation and operational planning. Real-time

operation is a series of activities, which are planned in a sequential manner. It starts with the preventive control, with a horizon of $1-2$ h, and aims operation at optimal cost under security constraints. Preventive action is always planned and covers failures or unexpected reactions from the system point of view. Taking preventive decisions such as switching equipment, rescheduling loads, is also part of the sequence. Furthermore, it oversees contingencies, and prepares or adjusts the system to take control decisions. Preventive control may be followed by two other control strategies, namely corrective control and emergency control. Corrective control is the first step taken following preventive control. The horizon is $0-15$ min and aims at maintaining the system intact. Emergency action is the control scheme of real-time operation. Both preventive, as well as corrective action, may end up in emergency action in the worst-case scenario. Emergency action is taken during any unplanned contingency or failure when the effect of a contingency is not sufficiently covered by means of preventive and corrective actions.

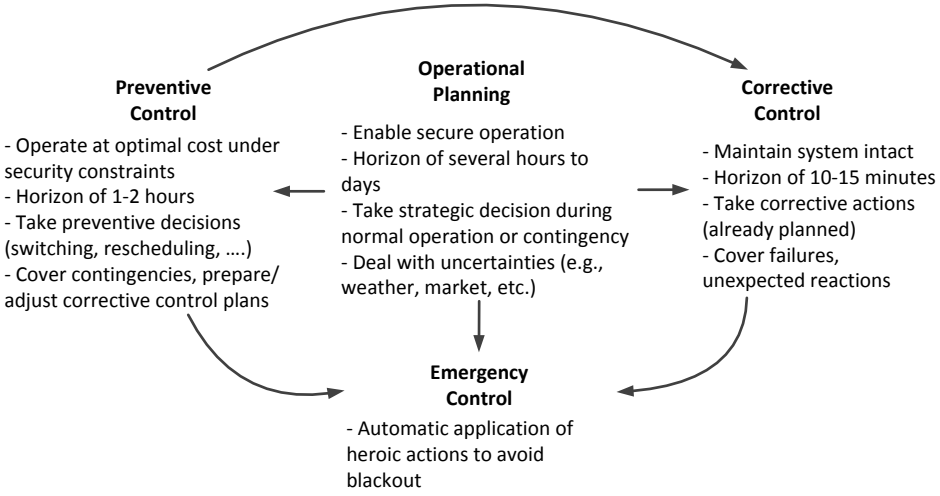


Fig. 2.6: Actions taken during short-term operational planning [Khuntia *et al.*, 2016a]

Literature study reveals comprehensive methodologies for preventive and corrective actions. Load shedding, considered as a corrective action, is studied in various literature [Echavarren et al., 2006, Otomega et al., 2014]. Other various mathematical modelling and optimization techniques include: PSO [Voumvoulakis & Hatziargyriou, 2010], decision [Krishnan et al., 2011, Liu et al., 2014, Xu et al., 2014a], model predictive control [Gong & Hiskens, 2008], ant colony system [Church et al., 2011, Ayan et al., 2015], GA and ANN [Kucuktezcan & Genc, 2012], differential evolution [Xu et al., 2014b], and various other optimization techniques [Arandian et al., 2014, Kucuktezcan

& Genc, 2015]. In system operation, decisions are taken within a limited time. Probabilistic risk analysis is already often used in grid development, but the application in system operation is relatively new. IEEE and CIGRE have developed task forces working on risk analysis and probabilistic techniques for planning and operation [McCalley et al., 2004]. Recently, various domains in risk-based planning have been studied, like power transfer limit, weather conditions, stability, and reserve generation. Reference [Preece & Milanović, 2015] combined probabilistic and fuzzy inference systems to categorize different degrees of risk, which facilitates the understanding of the planner. The paper focused on stability issues and the methodology was applied to a multi-area network, but the concept can also be applied to reliability problems. Reference [Ciapessoni et al., 2013] studied the advantages of integrating probabilistic and deterministic tools for enhancing security during short-term horizon. Reference [Reneses *et al*., 2006], for the first time, discussed the importance of coordination among the time-horizons, stating how long-term decisions impact short-term decisions.

2.3 ENERGY TRANSITION LEADING TO INTERACTED APPROACH

Energy transition towards low-carbon future with decentralized generation and massive integration of RES make it plausible for the TSOs to shift towards an interacted approach over the more conventional sequential approach. In power system reliability management, the sequential approach can be understood as long-term grid development, via mid-term asset management, towards short-term operation planning and real-time operation. The interacted approach resembles a future operating state where long-term grid development can interact with short-term operation planning and real-time operation. Such an approach entails exchange of information of different reliability management activities within the different time-horizons. For example, a candidate decision for grid development (long-term horizon) can be passed through the assessment of outage scheduling (mid-term horizon) as well as the assessment of future operational short-term operational planning (short-term horizon). Then the results of this assessment serves as a feedback to grid developers. Fig. 2.7 illustrates this notion and the concepts are explained in next sub-section. The aging of the power system infrastructure and the increasing penetration of renewable and dispersed generation presently induce new threats to the system security. Electricity generation from RES is both variable as well as uncertain, which makes their integration into existing traditional power systems a challenging task. TSOs are responsible to maintain the required level of security of supply with variations from both generations as well as the demand side. However, despite the urgency of further development, investment in grid development is stagnating. In Germany, for example, many of the projects planned are expecting delays [Web1, 2013] and one of the many reasons is public opposition at the local level (e.g., with respect to overhead transmission lines). In such situations, TSOs opt to plan

their operational planning decisions in an interacted manner based on grid development plans and asset management decision-making process. This section explains how such an interacted approach is viable in the future.

2.3.1 CONCEPT OF SEQUENTIAL AND INTERACTED APPROACH

This chapter reviewed the concept of different main processes and time-horizons in the planning and operation of power systems. Not only in terms of timescale but also the tasks involved in each of the time-horizons make them different from each other. In some cases, short-term planning may work out efficiently, but it may not be adequate in identifying the long-term needs of the system. For example, in the short term, a lower voltage and less expensive line addition may be adequate but may require an expensive upgrade within a decade. In contrast, an initially more expensive and higher capacity line might be less expensive in the long term [Milligan et al., 2012].

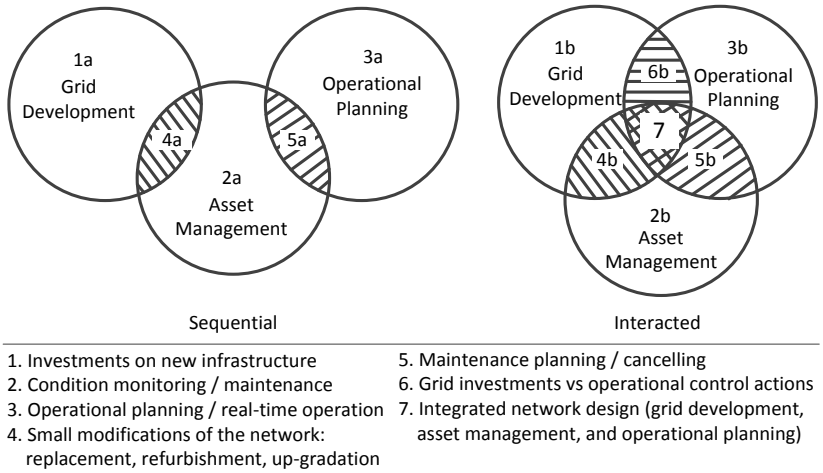


Fig. 2.7: Interactions among the three processes [Khuntia *et al.*, 2016a]

There is always an overlap among the different processes, as illustrated in Fig. 2.7. Without discussing this overlap, this work would be incomplete. Small modifications of the network can be required because of grid development or because of AM (area-4a in Fig. 2.7) and planned maintenance might be canceled during system operation because of a contingency (area-5a). In the past, the three processes consisted of more or less separate activities. This is illustrated as a sequential approach in Fig. 2.7. For example, the Dutch 380 kV ring was developed considering $N-2$ ($N-1$ during maintenance) redundancy (area-1a). As a result, this gave enough room to plan maintenance in AM (area-2a), and enough room for operational activities (area-3a). In the future, the overlap and interaction between the three main processes are expected to increase, because of the developments and energy transition. Earlier studies showed that

offshore network redundancy is mostly uneconomical. Onshore spinning generation reserve can serve as redundancy for the offshore network to maintain the required level of security of supply (area-6b in Fig. 2.7). If redundancy is not created in the offshore network (long-term activity), this has consequences for activities in other time-horizons.

Several (possible) challenges can be expected for the future as enlisted in Table 2.1. Uncertainty modeling is one of the primary challenges followed by big data. The management of a large amount of data poses a second challenge. Furthermore, the development of new risk tools and clear interpretation of results are of importance, as risk analysis is useless if the results cannot be translated into actions. Various projects on pan-European electric power system are working towards improving reliability or developing a new reliability criterion. Reference [Vefsnmo et al., 2015] in AFTER project developed a risk assessment tool to be used in the short-term horizon. The GARPUR project has developed new probabilistic reliability criteria taking into consideration the three time-horizons [Web8, 2017]. As transmission system planners face numerous challenges originated in load growth, increased penetration of RES, economic forces of deregulation, and development and integration of new technologies, this research presents recent literature to keep up with the advancement. Herewith, identifying load and wind power as exogenous variables, the next chapter focuses on modeling and forecasting of electricity load in short- and long-term horizons.

Table 2.1: Future [possible] challenges due to the interaction of different time-horizons

Challenge	**Description**
Uncertainty modeling	Modeling the variability of RES, market uncertainties, and variable demand, high-impact-low-probability (HILP) events, operating conditions, contingency modeling.
Data management	Handling large amounts of data (e.g. WAMS measurements, weather data, load data). Collecting suitable failure statistics of network components.
Tools	Develop complex methods in academia for easy understanding and use in the real world. Risk analysis of large-scale systems with a large amount of uncertainties within a reasonable computing time.
Result interpretation	Presentation of the results of probabilistic reliability analysis in clear, understandable and actionable indices.

2.3.2 GRID DEVELOPMENT AND OPERATIONAL PLANNING: CHALLENGES AHEAD

Within the transmission system framework, grid development deals with planning and decision making that alter transmission capacities either within a TSO's own network or towards other TSOs' network. An important element of the grid development process is to identify and analyze ideas for future expansion plans. To realize such plans, rich historical data and accurate forecasts of exogenous variables are crucial in devising

future development plans as we look into an uncertain future. This is accompanied by more realistic assumptions/expectations of future working conditions that facilitate optimal decision making.

Usually, TSOs employ a detailed model of their own grid and a specific set of scenarios (corresponding to load, generation, etc.) in order to simulate how an investment option would influence the operation of their grid or the facilitation of the electricity market. Decision making is quite uncertain in grid development when the time scale stretches to many months and years. This is in contrast to other processes in operational time scales such as outage scheduling and system operations when the uncertainty level is comparatively low. It is even more challenging when TSOs need to cope up with energy transition and aim at future low-carbon generation. Thus, TSOs need to make inferences and nearly accurate forecasts to generate more realistic market scenarios for the future. The scenarios can be in the temporal scale of generation mix (conventional and RES) and load growth, and also spatial distribution of generation sites (mostly RES) and load centers. Suitable modeling techniques are needed for large-scale integration of RES because of its stochastic nature. The current and future integration of large-scale RES is enabled by:

i. Connection of RES generation to the main network,
ii. Increasing the Grid Transfer Capability (GTC) between an area with excess RES generation and other areas, in order to facilitate a higher level of RES penetration.

Among the RES, wind power has gained significant attention among TSOs and is an accelerator in energy transition. By 2050, wind power is envisaged as the largest source of power in the EU in the 2050 roadmap [ECF, 2011]. The massive integration of wind power cannot be answered by fast grid development activities in terms of transmission system expansion. Such a need should be carefully assessed against a possible over-investment in expansion planning, resulting from an uncertain forecast of wind power generation and load respectively. The dilemma with system planners is that construction of WPP takes 2-3 years while it may take 8-10 years to plan, execute and construct a transmission corridor. Costs incurred in grid upgrade and new infrastructure is dependent on locations of wind generation and load centers. In Portugal, to increase the wind power penetration from 3% to 13% during the years 2004-2009, revenues worth €145 million is reported [Holttinen, 2012]. At the same time, system operation is increasingly challenging because with the existing transmission infrastructure, transmission assets operate more and more frequently at operating limits and line loading with higher currents leads to increased energy losses. Thus, apart from already used measures like redispatch, curtailment of wind power and load, this research looks into the future for possible solutions in terms of the development of new statistical

models for better forecast and use of big data analytics which results in the improved risk-based security assessment.

For operational planning, power grids are equipped with automatic generation control to handle any significant deviation and uncertainty in electricity load over time scales from seconds to hours. However, it is a challenge with massive penetration of distributed WPPs. A common practice among TSOs is to follow the change in net load rather than addressing to variation from a single generator or customer load. The same theory applies to consideration of aggregated wind power because the relative variability of wind power decreases with aggregation of more WPPs. Moreover, the number of *zero hour output* decreases with aggregated wind power over a large geographic area [Holttinen, 2012]. For instance, a single WPP can have zero output for more than 1000 hours in a year as compared to aggregated wind power in a large area which is always greater than zero. An added benefit of considering aggregated wind power is that the net installed wind capacity will not be reached at any given instant because of physical reasons suggesting same wind speed cannot be experienced over the large geographical area. Even if the same wind speed is attained, the technical availability is proven to be 95% of the time and the rest result with zero wind power.

A well-proven case of understanding the need for interacted approach is the case of Germany. Germany records a significant amount of onshore and offshore wind in the north close to the North Sea. Solar power is recorded in southern part of the country and load centers with high demand are located in the western and southern part of the country. With such spatial diversity, electrical energy has to be transmitted to regions from surplus north to deficit south. A map depicting a change in Germany's power balance from the year 2015 till the year 2030 is shown in Fig. 2.8. This comes at a point when energy transition aims at facilitating 40,000 MW of new generation plant in terms of RES by 2020 while slowly shutting down nuclear power plants and other fossil fuel generation plants. Energy transition is experienced at a time when the transmission system, built 30-40 years ago, have to facilitate the massive integration of variable and uncertain RES. In particular, transporting the wind energy from north to meeting demand in the south. Thus, the system needs to be optimally operated to maintain the required security of supply. At the present state, the long transmission corridor linking the north and south was never intended to carry such massive energy. The power production in the north has to be curbed because it does not have anywhere to go. This *feed-in management* [Web1, 2018] incurs further costs because wind power producers must be compensated when their turbines are switched off. It results in cheap electricity prices in the north and when the energy cannot be transported there is re-dispatch of clean energy.

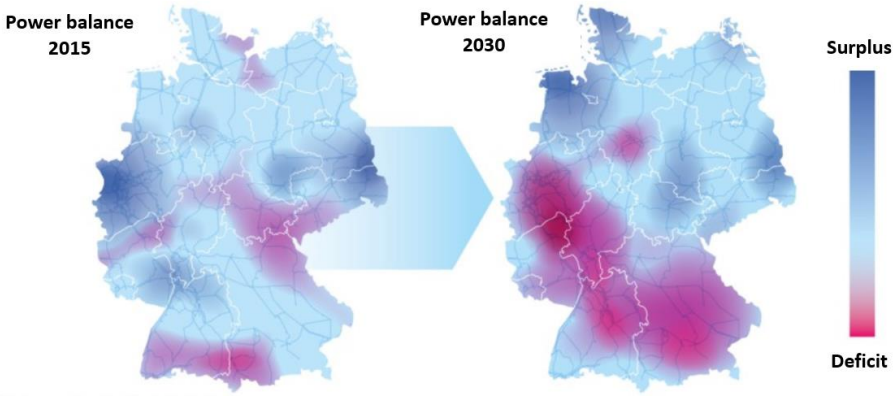


Fig. 2.8: Change in power balance recorded by Amprion [Web2, 2018]

2.4 CONCLUSION

Operating the current and designing the future power system with massive penetration of wind power must ensure that the system is both secured and adequate at all time scales. As different independent actions are taken by TSOs in different time-horizons, the nature of these actions vary from one TSO to another as concluded from this research. Among the current practices of TSOs, it was learned that the old concept of time-horizons, which refers to the sequential approach adopted by TSOs, is challenged by exogenous variables as the electric power system undergoes energy transition in the form of massive integration of RES and introduction of new technologies. While this research aims at studying the interaction of operational planning with grid development, the following attributes of exogenous variables need consideration:

- Electricity load is uncertain by nature. In addition, adaptation to new technologies, changing economy and addition of flexible load and energy autonomy of households makes it more challenging to forecast load in short- and long-term horizons. Suitable models to forecast load in both the time-horizons is realized in this thesis in order to generate more realistic market scenarios for the future.
- Variability of wind power tends to decrease with either *increase of wind generators within a WPP* or *consideration of aggregated wind power*. There may be a situation of *saturation wind power potential* which is a condition explained as the maximum wind power that can be extracted upon increasing the number of wind turbines over a large geographic region, independent of societal, environmental, climatic, or economic considerations. It affects the operational planning of TSOs when accurate wind power forecast is needed. And as the electrical grid was not planned for massive integration of stochastic generation,

this saturation effect might overshadow TSO's grid development plans (e.g., constructing new transmission corridor) to enable the integration.

As TSOs are collecting a huge amount of chronological load and wind power data, this research identifies the key parameters to harness the information from spatially distributed chronological data towards assessment of system security. They are:

- *Exogenous variables: Load and wind power*
- *Time-horizons: Short-term (and long-term horizon scenarios)*
- *Target: Assessment of system security by calculating overloading risk of transmission lines*

In the next chapter, modeling and forecasting electricity load in short and long-term is explained using two newly developed models.

REFERENCES

[Arandian et al., 2014] Arandian, B., Hooshmand, R. A., & Gholipour, E. (2014). Decreasing activity cost of a distribution system company by reconfiguration and power generation control of DGs based on shuffled frog leaping algorithm. *International Journal of Electrical Power & Energy Systems*, *61*, 48-55.

[Ashkezari et al., 2014] Ashkezari, A. D., Ma, H., Saha, T. K., & Cui, Y. (2014). Investigation of feature selection techniques for improving efficiency of power transformer condition assessment. *IEEE Transactions on Dielectrics and Electrical Insulation*, *21*(2), 836-844.

[Ayan et al., 2015] Ayan, K., Kılıç, U., & Baraklı, B. (2015). Chaotic artificial bee colony algorithm based solution of security and transient stability constrained optimal power flow. *International Journal of Electrical Power & Energy Systems*, *64*, 136-147.

[Bartlett, 2002] Bartlett S. Asset management in a de-regulated environment. On behalf of CIGRE Joint Task Force 23.18 and Australian Working Groups 2002, CIGRE Paris

[Billinton & Bollinger, 1968] Billinton, R., & Bollinger, K. E. (1968). Transmission system reliability evaluation using Markov processes. *IEEE Transactions on Power Apparatus and Systems*, (2), 538-547.

[Billinton & Li, 1994] Billinton, R., Li, W.: 'Reliability assessment of electric power systems using Monte Carlo methods' (Plenum Press, New York, 1994)

[Billinton & Allan, 1996] Billinton, R., Allan, R.N.: 'Reliability evaluation of power systems' (Plenum Press, New York, 1996)

[Church et al., 2011] Church, C., Morsi, W. G., El-Hawary, M. E., Diduch, C. P., & Chang, L. C. (2011). Voltage collapse detection using Ant Colony Optimization for smart grid applications. *Electric Power Systems Research*, *81*(8), 1723-1730.

[Ciapessoni et al., 2013] Ciapessoni, E., Cirio, D., Grillo, S., Massucco, S., Pitto, A., & Silvestro, F. (2013). An integrated platform for power system security assessment implementing probabilistic and deterministic methodologies. *IEEE Systems Journal*, *7*(4), 845-

853.

[CIGRE, 2010b] CIGRE WG D1.17. Generic guidelines for life time condition assessment of HV assets and related knowledge rules. CIGRE, 2010

[de la Torre et al., 2008] de la Torre, S., Conejo, A. J., & Contreras, J. (2008). Transmission expansion planning in electricity markets. *IEEE transactions on power systems*, *23*(1), 238-248.

[ECF, 2011] Energy Roadmap 2050. (2011). *European Commission*, Brussels.

[Echavarren et al., 2006] Echavarren, F. M., Lobato, E., & Rouco, L. (2006). A corrective load shedding scheme to mitigate voltage collapse. *International Journal of Electrical Power & Energy Systems*, *28*(1), 58-64.

[Gong & Hiskens, 2008] Gong, B., & Hiskens, I. A. (2008, September). Two-stage model predictive control for voltage collapse prevention. In *Power Symposium, 2008. NAPS'08. 40th North American* (pp. 1-7). IEEE.

[Hobbs, 1995] Hobbs, B. F. (1995). Optimization methods for electric utility resource planning. *European Journal of Operational Research*, *83*(1), 1-20.

[Holttinen, 2012] Holttinen, H. (2012). Wind integration: experience, issues, and challenges. *Wiley Interdisciplinary Reviews: Energy and Environment*, *1*(3), 243-255.

[Jahromi et al., 2009] Jahromi, A., Piercy, R., Cress, S., Service, J., & Fan, W. (2009). An approach to power transformer asset management using health index. *IEEE Electrical Insulation Magazine*, *25*(2), 20-34.

[Khuntia *et al*., 2016a] Khuntia, S. R., Tuinema, B. W., Rueda, J. L., & van der Meijden, M. A. (2016). Time-horizons in the planning and operation of transmission networks: an overview. *IET Generation, Transmission & Distribution*, *10*(4), 841-848.

[Khuntia *et al*., 2016b] Khuntia, S. R., Rueda, J. L., Bouwman, S., & Meijden, M. A. (2016). A literature survey on asset management in electrical power [transmission and distribution] system. *International Transactions on Electrical Energy Systems*, *26*(10), 2123-2133.

[Kostic, 2003] Kostic, T. (2003, July). Asset management in electrical utilities: how many facets it actually has. In *Power Engineering Society General Meeting, 2003, IEEE* (Vol. 1, pp. 275-281). IEEE.

[Krishnan et al., 2011] Krishnan, V., McCalley, J. D., Henry, S., & Issad, S. (2011). Efficient database generation for decision tree based power system security assessment. *IEEE Transactions on Power systems*, *26*(4), 2319-2327.

[Kucuktezcan & Genc, 2012] Kucuktezcan, C. F., & Genc, V. I. (2012). A new dynamic security enhancement method via genetic algorithms integrated with neural network based tools. *Electric Power Systems Research*, *83*(1), 1-8.

[Kucuktezcan & Genc, 2015] Kucuktezcan, C. F., & Genc, V. I. (2015). Preventive and corrective control applications in power systems via big bang–big crunch optimization. *International Journal of Electrical Power & Energy Systems*, *67*, 114-124.

[Lindquist et Lindquist, T. M., Bertling, L., & Eriksson, R. (2005, June). Estimation of disconnector contact condition for modelling the effect of maintenance and

al., 2005a] ageing. In *Power Tech, 2005 IEEE Russia* (pp. 1-7). IEEE.

[Lindquist et al., 2005b] Lindquist, T., Bertling, L., & Eriksson, R. (2005, January). A method for age modeling of power system components based on experiences from the design process with the purpose of maintenance optimization. In *Reliability and Maintainability Symposium, 2005. Proceedings. Annual* (pp. 82-88). IEEE.

[Lindquist et al., 2008] Lindquist, T. M., Bertling, L., & Eriksson, R. (2008). Circuit breaker failure data and reliability modelling. *IET generation, transmission & distribution*, *2*(6), 813-820.

[Liu et al., 2014] Liu, C., Sun, K., Rather, Z. H., Chen, Z., Bak, C. L., Thøgersen, P., & Lund, P. (2014). A systematic approach for dynamic security assessment and the corresponding preventive control scheme based on decision trees. *IEEE Transactions on Power Systems*, *29*(2), 717-730.

[McCalley et al., 2004] McCalley, J., Asgarpoor, S., Bertling, L., Billinion, R., Chao, H., Chen, J., Endrenyi, J., Fletcher, R., Ford, A., Grigg, C. and Hamoud, G.,Hamoud, G. (2004, June). Probabilistic security assessment for power system operations. In *Power Engineering Society General Meeting, 2004. IEEE* (pp. 212-220). IEEE.

[Milligan et al., 2012] Milligan, M., Donohoo, P., O'Malley, P.: 'Stochastic methods for planning and operating power system with large amounts of wind and solar power'. NREL/ CP-5500-56208, Golden, CO, 2012

[Munoz et al., 2014] Munoz, F. D., Hobbs, B. F., Ho, J. L., & Kasina, S. (2014). An engineering-economic approach to transmission planning under market and regulatory uncertainties: WECC case study. *IEEE Transactions on Power Systems*, *29*(1), 307-317.

[Nilsson & Bertling, 2007] Nilsson, J., & Bertling, L. (2007). Maintenance management of wind power systems using condition monitoring systems—life cycle cost analysis for two case studies. *IEEE Transactions on energy conversion*, *22*(1), 223-229.

[Otomega et al., 2014] Otomega, B., Glavic, M., & Van Cutsem, T. (2014). A two-level emergency control scheme against power system voltage instability. *Control Engineering Practice*, *30*, 93-104.

[Pereira et al., 1985] Pereira, M. V. F., Pinto, L. M. V. G., Cunha, S. H. F., & Oliveira, G. C. (1985). A decomposition approach to automated generation/transmission expansion planning. *IEEE Transactions on Power Apparatus and Systems*, (11), 3074-3083.

[Preece & Milanović, 2015] Preece, R., & Milanović, J. V. (2015). Probabilistic risk assessment of rotor angle instability using fuzzy inference systems. *IEEE Transactions on Power Systems*, *30*(4), 1747-1757.

[Puglia et al., 2014] Puglia, G., Bangalore, P., & Tjernberg, L. B. (2014, July). Cost efficient maintenance strategies for wind power systems using LCC. In *Probabilistic Methods Applied to Power Systems (PMAPS), 2014 International Conference on* (pp. 1-6). IEEE.

[Reneses *et al.*, 2006] Reneses, J., Centeno, E., & Barquin, J. (2006). Coordination between medium-term generation planning and short-term operation in electricity markets. *IEEE Transactions on Power Systems*, *21*(1), 43-52.

[Smit et al., 2006] Smit, J. J., Quak, B., & Gulski, E. (2006, October). Integral decision support for asset management of electrical infrastructures. In *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on* (Vol. 3, pp. 2622-2628). IEEE.

[Tor & Shahidehpour, 2006] Tor, O., & Shahidehpour, M. (2006). Power distribution asset management. In *Power Engineering Society General Meeting, 2006. IEEE* (pp. 7-pp). IEEE.

[Ugranli & Karatepe, 2016] Ugranli, F., & Karatepe, E. (2016). Transmission expansion planning for wind turbine integrated power systems considering contingency. *IEEE Transactions on Power Systems*, *31*(2), 1476-1485.

[Vefsnmo et al., 2015] Vefsnmo, H., Kjølle, G., Jakobsen, S. H., Ciapessoni, E., Cirio, D., & Pitto, A. (2015, June). Risk assessment tool for operation: from threat models to risk indicators. In *PowerTech, 2015 IEEE Eindhoven* (pp. 1-6). IEEE.

[Velasquez-Contreras et al., 2011] Velasquez-Contreras, J. L., Sanz-Bobi, M. A., & Arellano, S. G. (2011). General asset management model in the context of an electric utility: application to power transformers. *Electric Power Systems Research*, *81*(11), 2015-2037.

[Voumvoulakis & Hatziargyriou, 2010] Voumvoulakis, E. M., & Hatziargyriou, N. D. (2010). A particle swarm optimization method for power system dynamic security control. *IEEE Transactions on Power Systems*, *25*(2), 1032-1041.

[Web1, 2018] Weblink: https://www.tennet.eu/electricity-market/german-market/eeg-kwkg/feed-in-management/

[Web2, 2018] Weblink: https://www.cleanenergywire.org/dossiers/energy-transition-and-germanys-power-grid

[Wood & Wollenberg, 2012] Wood, A. J., & Wollenberg, B. F. (2012). *Power generation, operation, and control*. John Wiley & Sons.

[Xu et al., 2014a] Xu, Y., Dong, Z. Y., Zhang, R., & Po Wong, K. (2014). A decision tree-based on-line preventive control strategy for power system transient instability prevention. *International Journal of Systems Science*, *45*(2), 176-186.

[Xu et al., 2014b] Xu, Y., Dong, Z. Y., Luo, F., Zhang, R., & Wong, K. P. (2013). Parallel-differential evolution approach for optimal event-driven load shedding against voltage collapse in power systems. *IET Generation, Transmission & Distribution*, *8*(4), 651-660.

CHAPTER 3

MODELING AND FORECASTING ELECTRICITY LOAD IN SHORT- AND LONG-TERM HORIZONS

3.1 INTRODUCTION

The need for moving from sequential to integrated interactions among the three time-horizons was discussed in chapter 2. In addition, it was identified that electricity load and wind power are the exogenous variables that need immediate attention in terms of modeling for TSOs. This chapter aims at answering the second research question Q.2., which deals with load modeling and forecasting in short- and long-term horizons,

- *How should uncertainty in load growth be addressed and what are the associated modeling challenges in the short-term and long-term horizons?*
- *How can forecast error be accounted for in terms of error distribution in the short-term horizon?*
- *What is the role of volatility in long-term forecasting and how does it impact the modeling framework?;*

The content of this chapter is based on research papers [Khuntia et al., 2016c, Khuntia et al., 2016d, Khuntia et al., 2016e, Khuntia et al., 2018d]. Load forecasting is essential at the present state when deregulation introduced competitiveness among different active market players. This is different from earlier days when electricity sectors were regulated and utility monopoly considered short-term load forecast for ensuring the security of supply and long-term load forecast for grid development. Deregulation along with new players like RES and demand-side management being introduced, it is vital to model and forecast load in a more accurate manner. Depending on the forecast horizon, modeling of the monthly or hourly timestamps is adopted for the model. For the sake of clarity as defined in Fig. 1.2, yearly forecast with hourly time-resolution is chosen for the short-term forecast based on operational planning while a 4-year forecast horizon with monthly time-resolution is chosen for the long-term

forecast based on grid development activities as offshore wind farm construction takes approximately $3-4$ years. For both the cases, the elemental quantity of interest is net system load.

Load forecasting has always been an important part of the planning and operation of electric utilities, i.e., both transmission and distribution companies. With technological advancement, change in economic condition and many other factors (to be discussed in this chapter), load forecasting is becoming more important. Forecast accuracy is largely affected by load impacting factors and actions taken in different time-horizons. However, due to its stochastic and uncertainty characteristics, it has been one challenging problem for electrical utilities to accurately forecast future load demand. In the past as well as in today's date, load forecasting is an integral part of planning for more than just utilities; system operators (transmission and distribution), generation companies, energy regulators, and financial institutions have a vested interest in load forecast accuracy. For instance, the short-term load forecasting accuracy plays a pivotal role in evaluating the long-term planning needs. In the long-term horizon, it is more vital to predicting what will eventually be needed than to know exactly when it will be needed. Based on time-scale, load forecast can be broadly classified into three main categories [Willis & Northcote-Green, 1983]:

- *Short-term load forecast (STLF)*: The time-period of STLF lasts for a few minutes, hours to one-day ahead or a week. STLF aims at economic dispatch and optimal generator unit commitment while addressing real-time control and security assessment.
- *Mid-term load forecast (MTLF)*: The time-period of MTLF is a month to a year or two. MTLF aims at maintenance scheduling, coordination of load dispatch and price settlement so that demand and generation are balanced.
- *Long-term load forecast (LTLF)*: The time-period of LTLF is few years (>1 year) to 10–20 years ahead. LTLF aims at system expansion planning, i.e. generation, transmission, and distribution. In some cases, it also affects the purchase of new generating units.

Each of the three categories is equally important for the smooth operation of the power system, and any error/uncertainty in forecast affects the economy and control aspect of the power system. Especially in the mid- and long-term horizons, since load forecasting is highly related to the system development, attention has been paid to the impact of load forecasting on system design [Willis, 1983] and economics [Ranaweera et al., 1997]. An accurate forecast leads to better maintenance plan during mid-term, and generation and expansion planning during a long-term horizon. The preciseness of long-term forecast significantly affects the development of future generation systems. For example, construction of a new generation plant takes approximately $5-10$ years or offshore wind farm construction takes $\sim 3-4$ years and involves a huge amount of

capital investment. In order to meet the demand and make the economic growth continuous, load forecasting is required for the related electricity utilities. Utilities do not want a huge investment going in vain. Both an overestimation as well as underestimation of the forecast will result in discontent among utilities and substantial investment for the construction of new generation units. So, accurate forecasting helps in assessing the needs in relation to planning, designing, environmental admitting to constructing step of power plants, and subsequent planning of transmission and distribution systems.

There are dozens of different load forecasting methods that have been used and documented during the last 50 years though the majority of them fall into the category of STLF. MTLF and LTLF are much less popular as research topics as compared with STLF; dozens of papers on STLF are published every year for each paper on MTLF or LTLF. The reason, of course, is that forecasting for the mid-term and especially for the long-term is a whole different problem from forecasting for the short term. It cannot be done by simply fitting a model (either statistical or computational) over a dataset, and then extrapolating from it. It is evident from [Box et al., 2008, Chatfield, 2001], that MTLF or LTLF is usually ignored because of the complications. Xia et al. [Xia et al., 2010] reported the difficulty in accurate forecasting since the factors are not *stable random*, but rather *unstable random* factors like governance within a country. Leahy and Foley [Leahy & Foley, 2012] discussed the impact of the long-term weather forecast and wind penetration on electrical load in Ireland. The work showcased the importance to consider the combined potential impacts of prolonged cold weather and periods of low winds under future projected generation scenarios. Reference [Makridakis et al., 2008] clearly stated that long-term forecasting *requires a different approach*, and suggested that these forecasts should be based on (a) identifying and extrapolating mega-trends going back in time as far as necessary (e.g. they discuss the variations in the price of copper, since the year 1800); (b) analogy and (c) constructing scenarios to consider future possibilities. The influence of economic factors on load in the long-term horizon becomes only visible on longer time scales or in extreme situations such as the economic crisis of 2009 [Troccoli, 2009]. Effect of weather (mostly, temperature) is extensively discussed in the work also. It reported that during winter, a drop of temperature by 1°C causes an additional power request of about 1.8 GW in France. Weather forecast, itself, is difficult in longer horizon. So, it can be concluded how complicated load forecasting for mid-/long-term horizon is. The problem of robust MTLF/LTLF can be foreseen as principal part of strategies design for substitutable development and optimal equipment renovation of energy systems under energy-saving technical progress. One of the feasible ways here is to design such strategies using integral dynamical models employment, as suggested in [Hritonenko & Yatsenko, 2012]. Here readers may refer to the extensive bibliography in these manuscripts on the use of integral dynamic models. Reference [Hong, 2014] performed a study on past,

current and future trends in energy forecasting. This paper showcased the trend in spatial, STLF, LTLF and energy price forecasting in a lucid manner. It quoted '*When you flick that switch, you expect the lights to go on – but the business of keeping them on is not nearly as straightforward*'. For example, it is feasible to forecast the next day load of a certain zone with less error percentage but quite challenging or even impossible to forecast the next summer peak load with a similar accuracy. In such case, it is possible to forecast the weather-normalized summer peak load based on average peak weather condition for that particular zone.

In this chapter, section 3.2 focuses on the development of a neural network model for short-term load forecast and modeling forecast error using *truncated normal distribution*. Section 3.3 describes the importance of volatility in the long-term load forecast and focuses on development of a multiplicative error model for a long-term forecast. Finally, section 3.4 concludes the chapter.

3.2 SHORT-TERM LOAD FORECASTING USING NEURAL NETWORK

In recent years, the significance of short-term load forecast (STLF) has increased and it will continue because over- or under-estimate of the future load has a significant impact on the efficiency of operation of any electrical utility. Various operational decisions such as economic scheduling of the generating capacity, scheduling of fuel purchase and system security assessment are based on such forecasts [De Felice & Yao, 2011]. Literature study reveals that short-term load forecasting has been extensively studied, and many load modeling and forecasting techniques have been developed [Gross & Galiana, 1987, Moghram & Rahman, 1989, Taylor & McSharry, 2007]. These methodologies can be broadly classified into two categories:

i. Traditional or statistical STLF methods like time series, regression analysis and gray model, which is based on load patterns. As the name suggests, statistical STLF methods forecast load using a mathematical combination (additive or multiplicative) of historical load values and other impacting factors like weather. These models are easy to interpret because they explain the relationship between load and impact factors. But often criticized because of their limitations in modeling the non-linear relationship between them.
ii. Non-traditional or artificial intelligence STLF methods like fuzzy, neural networks and other intelligent load forecasting methods. Compared to traditional methods, these methods are flexible and can model both non-linearity as well as complexity. Their popularity rose with the advent of computers in the early 1990s. In addition, they are widely accepted because of the fact that they do not need any prior modeling experiences and work as a black box. The black box uses customizable algorithms that automatically classify the input data and later correlate the data with the relevant output values.

The scope of this research is forecasting electricity load in the short-term horizon using a feedforward neural network. Neural network-based load forecasting is one of the most widely used non-traditional load forecasting methods and it is evident from literature study [Rui & El-Keib, 1995, Hippert et al., 2001, Baliyan et al., 2015]. Reference [Baliyan et al., 2015] can be accessed for a complete review of different neural network-based methods developed lately. In the short-term horizon, the challenge for planners is not only to forecast load with high accuracy but also to form an accurate picture of the day-ahead load profile. The day-ahead load includes point forecasts of the load in each hour and also acknowledges the precision, or lack thereof, associated with the forecasted value. This research addresses the uncertainty associated with day-ahead forecasting and error implication associated with this uncertainty. An important question is: *What are the need for error and uncertainty analysis in short-term load forecasting*? One of the vital inputs in the short-term load forecast is weather data. Thus, it is evident that for day-ahead load forecast, use of weather forecast is the most viable option. There are high chances that online operation of weather forecast model introduces the associated error into load forecasting model. Reference [Chen et al., 2005] performed a comprehensive study showing the significant effect of weather on load forecast. It can also impact the training of the neural network as discussed by [Yoo & Pimmel, 1999]. For our research, weather forecast error is not considered; rather a historical database of weather information (dry bulb and dew point temperature) is used to train the network. In addition, statistical properties of forecast errors are studied and performance analysis of neural network based model is carried out by making comparisons to the normal distribution[4].

The rest of the section is organized as follows: sub-section 3.2.1 discusses the neural network-based load forecasting methodology. Sub-section 3.2.2 explains the forecast steps. Forecast results are demonstrated in sub-section 3.2.3. Sub-section 3.2.4 discusses the implication of truncated normal distribution to represent forecast error. Finally, sub-section 3.2.5 concludes this section on short-term load forecasting using neural networks.

3.2.1 NEURAL NETWORK-BASED SHORT-TERM LOAD FORECAST

Load forecasting has always gained attention and till today it is still difficult to accurately predict load. This is because load time series exhibit seasonality (daily, weekly, monthly) and the number of impacting factors increasing with the adoption of new technologies. Thus, choosing the right amount of data and a suitable model is equally important. For instance, if the inputs to the forecasting model are insufficient, it will be difficult or even impossible to come up with an accurate forecast no matter how

[4] In this research, the terms *normal distribution* and *Gaussian distribution* are synonyms.

good the chosen model is. The use of neural networks has been a widely studied electricity load forecasting technique since the 1990s [Peng et al., 1992]. This is understandable from the fact that neural networks are able to solve non-linear and complex problems and hence make them suitable for forecasting problems. Again, this is due to the fact that instead of relying on explicit rules or mathematical functions, neural networks draw a link between input and output data. Thus, in comparison to other traditional and non-traditional models, neural networks hold a good promise for the purpose of load forecasting. For our research, historical data for years 2010-2013 was used to forecast day-ahead load for 2014. The various input parameters for the training data is illustrated in Fig. 3.1. An exception in this study is that random disturbances, consumer class, and demand-side management are not considered. This is because the primary aim of this research is to analyze error and uncertainty in forecasted value. It is vital because the associated financial penalties for high error values are high considering almost all power system studies (both operational and expansion planning) are based on forecast values.

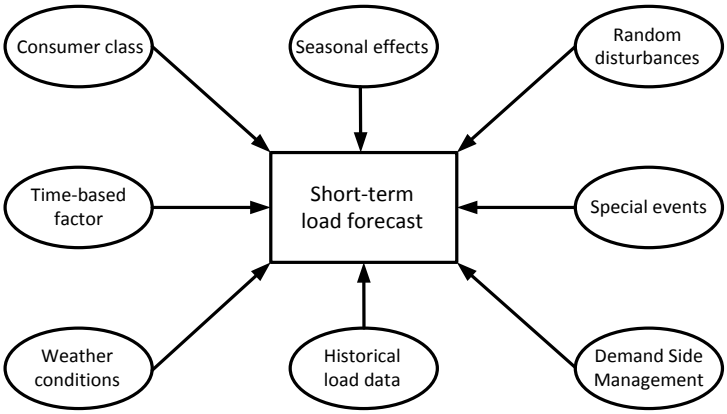


Fig. 3.1: Input parameters for training data [*Note*: *Consumer class, random disturbances, and demand-side management are excluded in this study*]

A neural network model is described by three distinct features [Weron, 2007]:

i. *Architecture*: The architecture describes the neural connections. Typically, network elements are arranged in a relatively small number of connected layers of elements between network inputs and outputs. The outputs are linear or non-linear functions of its inputs. The inputs may be the outputs of other network elements as well as actual network inputs. The three common neural network family models are feedforward, radial based, and recurrent types. One of the popular architecture for STLF is feedforward architecture with backpropagation,

which uses continuously valued functions and supervised learning. The same is employed in this research.

ii. *Processing*: This feature describes how a neural network produces output for every input and weight based on the training algorithm. It also describes how the neural network adapts its weight for every training vector.

iii. *Training*: The actual weights assigned to element inputs are determined by matching historical time and weather data to historical electricity load data in a pre-operational training period.

For model parameters selection, a systematic approach with regards to the number of hidden layers, number of nodes, batch size, epochs, network performance, desired activation functions, and training period has to be unambiguously formulated. In this research, the gradient descent based Levenberg-Marquardt optimization technique is used as it has one of the best learning rates when compared to other available functions in forecasting problems [Saini & Soni, 2002]. It is also the fastest backpropagation algorithm available with the MATLAB (version 2016a) and supervised algorithm as well, although it does require more memory than other algorithms. Neural network modeling can be categorized into two groups: the first group corresponds to single output node with next hour or next day peak or total forecast load and second group corresponds to multiple output nodes corresponding to sequential hourly loads for next day (24 hours). The developed model in this research is a three-layered feedforward backpropagation network, as shown in Fig. 3.2 with non-linear activation function in the hidden layer and linear function for the output layer. More details on the forecasting steps are explained in the next section.

3.2.2 UNDERSTANDING THE FORECAST STEPS

The neural network dataset is divided into two sets. The first set of data is used to train the neural network and called the training dataset. The second set of data is used to test the trained neural network-based forecast model. For this research, the training dataset comprises four years (2010-2013) load and weather data with hourly resolution. Weather data is represented by dry bulb temperature and dew point temperature. An additional parameter that is used in training dataset is the type of day (weekday, weekend or holiday). It acts like a flag indicating if it is a holiday or a weekend. The model is tested on completely out-of-sample data from 2014.

The inputs are fed into the input layer and multiplied by interconnection weights, and then passed through an activation function before passed on to the next layer. The mathematical model of the artificial neuron is modeled in a similar manner of the biological architectural set-up. Axons and synapses of the neuron are modeled as inputs and weights respectively. The strength of the connection between an input and a neuron is denoted by the value of the weight. The weighted inputs are added together

and passed through a nonlinear activation transfer function in the hidden layer. Finally, the activation function controls the amplitude of the output of the neuron. For activations, a sigmoid transfer function was used for the hidden layer and pure linear transfer function for the output layer. An advantage of sigmoid activation function is, unlike the linear function, the output of the activation function is always going to be in the range $[0,1]$ compared to $[-\infty, \infty]$ of the linear function. So the activations are bound in a range. The output layer is 24-hour or day-ahead load forecast. Choosing the hidden layer is tricky. Though there is no pre-specified rule for ideal selection of network layers and neurons in a neural network architecture, it is aimed for a simple three-layered architecture for better forecast results following the study [Peng et al., 1992]. Networks with more than one hidden layer are generally more complex. Again, the number of neurons in the hidden layer is chosen based on the fact that overspecializing might occur. Overspecializing is similar to overfitting when too many neurons lead to loss of generalizing capacity of the architecture. On the other hand, due to lack of enough hidden layer neurons, the network may find it difficult to learn the behavior of the series. For this research, the model was tested with varying number of hidden layer neurons, ranging from ten to fifteen based on [Adepoju et al., 2007]. Fourteen neurons were finally used because it offered a better model characteristic and this was achieved by training of network.

Training of the network is evaluated by epochs. An epoch is a single step in training a neural network, i.e., one forward pass and one backward pass of all the training dataset. In order to train the network for our training dataset, one epoch will be too big to feed to the network. Hence, the training dataset is divided into smaller batches. Batch size refers to the total number of training examples present in a single batch. Each batch consists of 73 training examples, i.e., type of day (1), previous day load (24), dry bulb temperature (24), dew point temperature (24). It is often observed that error reduction is inversely proportional to the increase in epoch training, but there might be instances of data overfitting which will result in an increase in error propagation. To infer, the best performance is observed from the epoch with the lowest validation error. In this study, best validation performance was observed at epoch 199, as seen in Fig. 3.3., and it can be argued that it attained in less number of epochs.

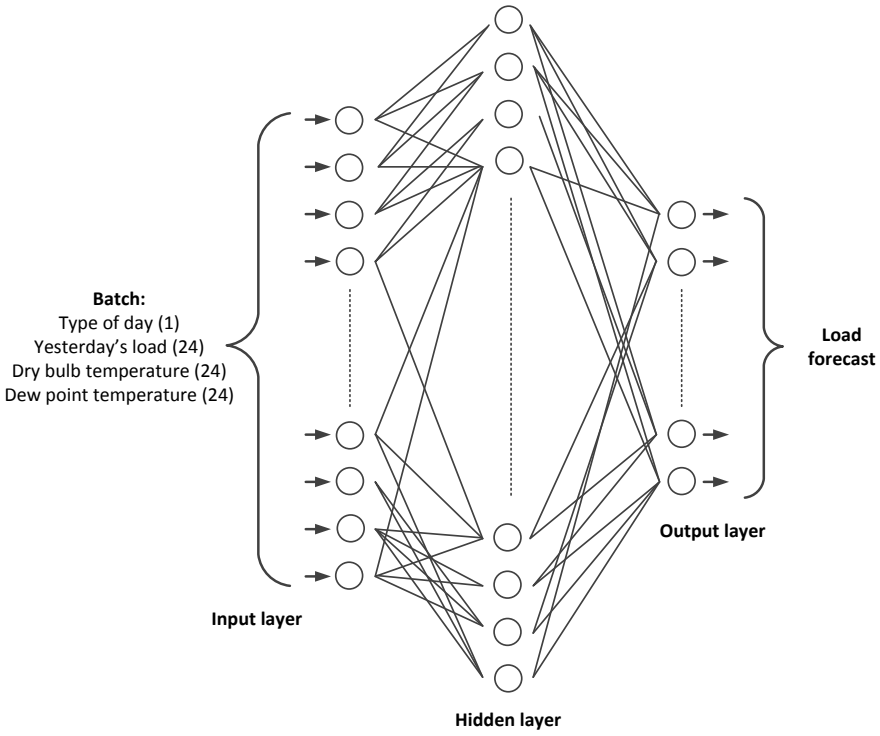


Fig. 3.2: Neural network architecture for short-term load forecast

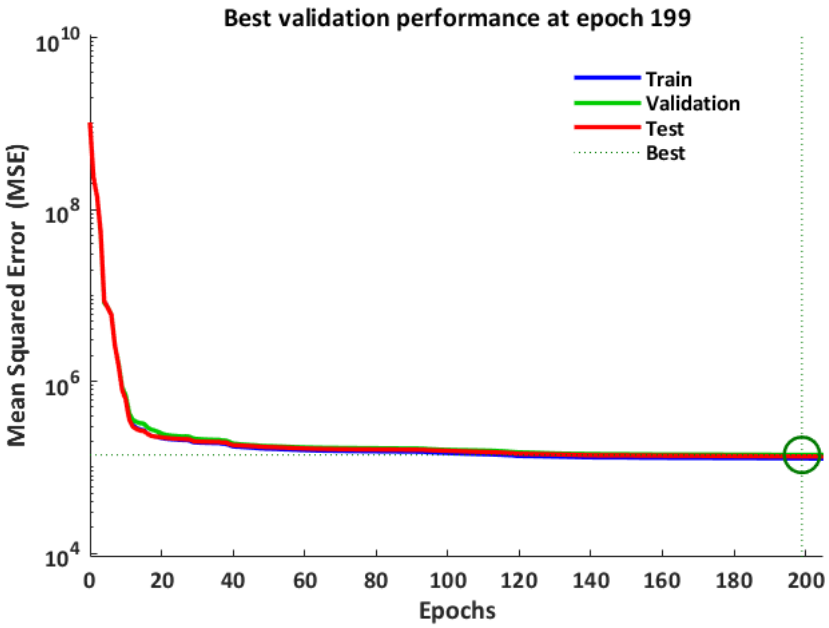


Fig. 3.3: Epoch training of the neural network

3.2.3 FORECAST RESULTS

To assess the effectiveness of the proposed forecast methodology, the Neural Network toolbox in MATLAB (version 2016a) [MATLAB, 2017] is used. As discussed in the previous section, forecasting electricity load using neural network can be observed as a non-linear function influenced by a number of factors, such as historical load, weather-related factors, time factors, etc.. These factors are modeled as network inputs. The flowchart in Fig. 3.4 illustrate the essential steps to forecast load using the neural network. The first one involves data preparation. The proposed forecasting model is trained off-line using publicly available data from ISO-New England (ISO-NE) [Web7, 2016]. Historical load, dry bulb temperature, dew point temperature and holiday list from state *ME* under ISO-NE is used. The values in historical load correspond to actual system load as determined by metering. Dry bulb and dew point temperature data are used as weather-related data. The latter step covers details with regard to network architecture and suitable model parameters, i.e., defining type and size of network, learning paradigm, network performance function, etc.. The parameters are as given below:

- *No. of layers: 3 (Input layer, Hidden layer, Output layer)*
- *No of neurons in hidden layer: 10 to 15*
- *No of neurons in output layer: 24*
- *Activation function of hidden layer: Sigmoid*
- *Activation function of output layer: Linear*
- *Training algorithm: Backpropagation*
- *Training dataset: years 2010-2013 with historical load and weather data of hourly resolution*
- *Test dataset: Year 2014*
- *Momentum factor (γ): 0.98*
- *No of data sets in each batch: 73*
- *No. of epochs for training: 1000*
- *Learning rate (α): 0.1*

The forecast results are compared with historical data, as shown in Fig. 3.5. The solid red curve indicates the historical load over the date range (19/06/2014 – 14/07/2014), while the blue curve indicates the forecast results. Weekly load profile corresponding to forecast values for the same date range is shown in Fig. 3.6. It illustrates the load profile of selected dates and depicts that the load demand is much higher during weekdays than the weekends. The load curve reflects the activity of a population with respect to electrical power consumption over a given period of time. It gives an insight into the consumption pattern of an area. This weekly pattern is repeated more or less throughout the year.

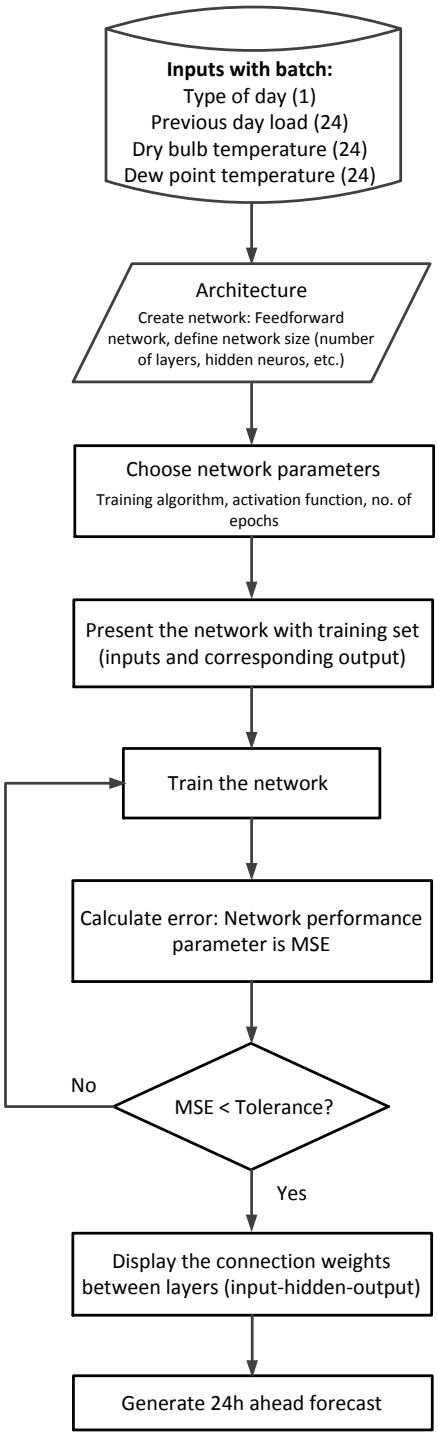


Fig. 3.4: Flowchart for the development of a supervised neural network-based load forecast model

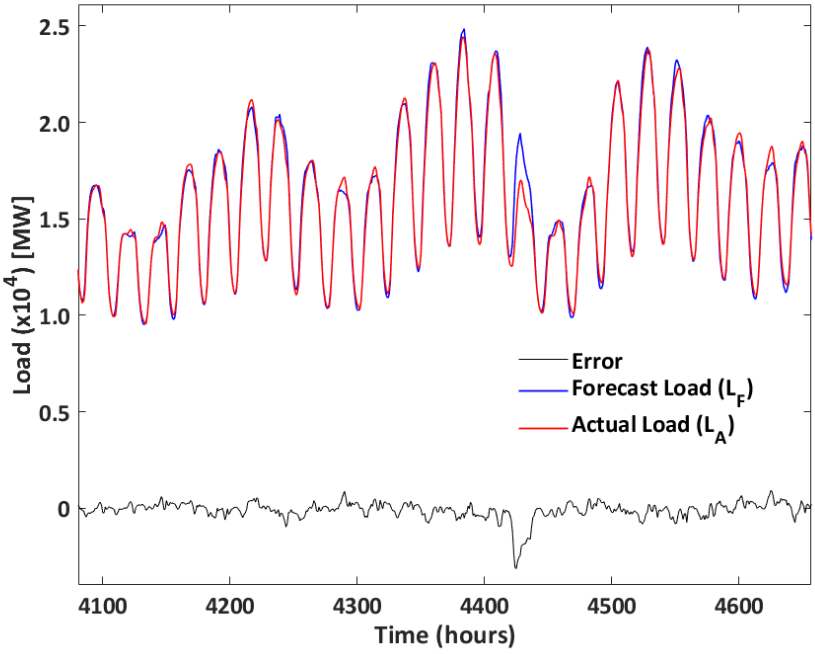


Fig. 3.5: Forecast load (L_F), actual load (L_A) and error for the date range (19/06/2014 – 14/07/2014)

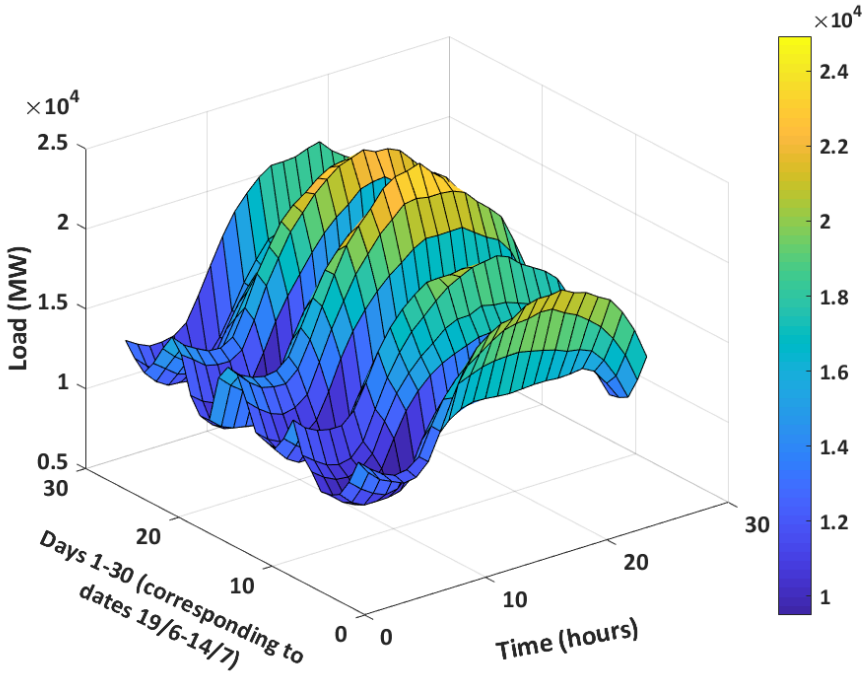


Fig. 3.6: Hourly load forecast for date range (19/06/2014 – 14/07/2014)

In order to model the statistical uncertainty information, large volumes of historical and real-time data are needed. As illustrated in Fig. 3.7, a sliding window is used for acquiring continuous statistical information on load influencing factors (i.e., timestamp, forecast horizon). Such a framework is adapted from [Makarov et al., 2010]. The time-frame can be tuned accordingly depending on forecast requirements. The forecast resolution is the time interval between two consecutive data records. The time-horizon is the length of the look-ahead interval, and the forecast update interval is the time interval for updating the forecast. The structure is supported by a table of data requirements for STLF as shown in Table 3.1.

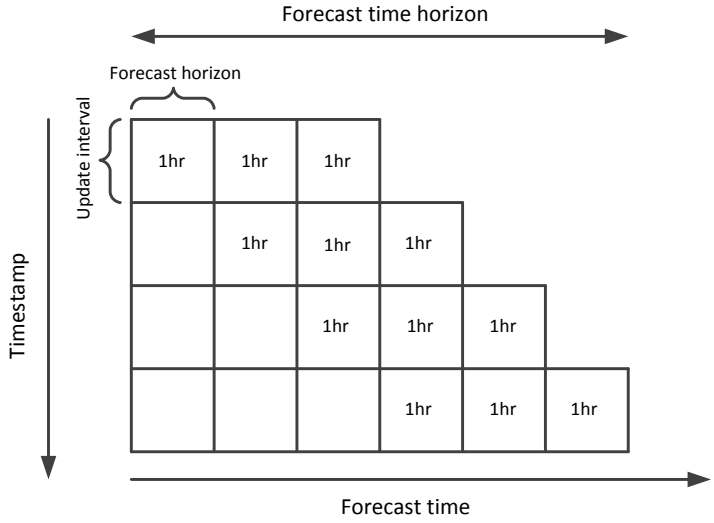


Fig. 3.7: Day-ahead load forecast structure

Table 3.1: Data Requirements for STLF

Data	Resolution	Forecast horizon
Day-ahead load forecast	1hr	24h
Hour-ahead load forecast	30mins.	24h
Real-time load forecast	5mins.	1h

3.2.4 ERROR IMPLICATION

In STLF, accurate load forecasts are very important because they determine the scheduling of generation units for next day or maybe few hours ahead. A slight deviation in forecast accuracy results in the suboptimal commitment of generation unit in the day-ahead market, which is avoided by utilities. Forecast error on the test set, as seen in Fig. 3.8, is defined as the difference between the actual load (L_A) and the

forecast load (L_F). It is evident from the figure that day-ahead load forecast error varies within the ±7% range. Few random overshoots above the range of +10% and -20% can be observed in the figure, which is excluded for error analysis. Such exclusions are wisely chosen because the uncertainty associated with the forecast error plays a vital role in influencing the resulting uncertainty obtained from the forecast model.

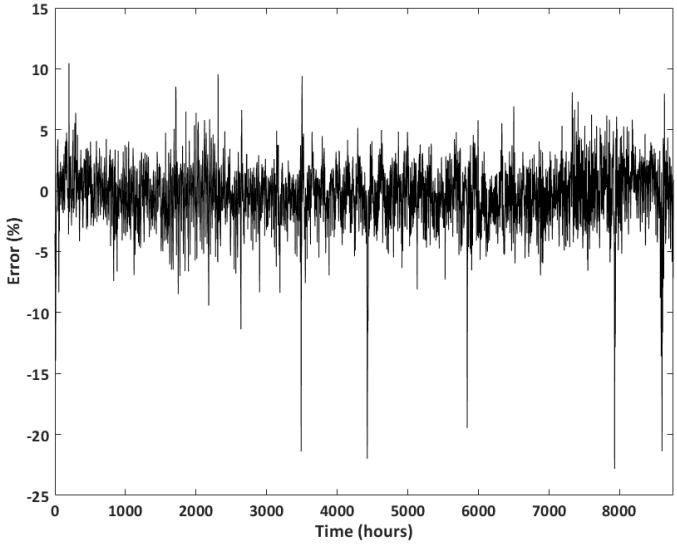


Fig. 3.8: Load forecast error (%)

The normal distribution is amongst the most common method to describe load forecast error. It can be described by the first two statistical moments, namely, mean (μ) and variance (σ^2). The third and fourth moments, namely skewness and kurtosis, are often close to zero if the observed distribution is well represented by the normal distribution. For error analysis and to check the efficiency of the neural network model, two widely used performance metrics called Mean Absolute Percentage Error ($MAPE$) and Coefficient of Variation (CV) are used. Considering two time series, actual load $L_A(t)$ and forecasted load $L_F(t)$ for time $t = 1,2 \ldots T$, the $MAPE$ is defined as:

$$MAPE = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{L_A(t) - L_F(t)}{L_A(t)}\right| \times 100\% \tag{3.1}$$

The coefficient of variation (CV), also called relative standard deviation, measures the ratio of the forecast error standard deviation (σ) to the error mean (μ). For forecast error (L_{FE}), CV is defined as:

$$CV = \frac{\sigma}{\mu} = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(L_{FE}(t) - \frac{1}{T}\sum_{t=1}^{T} L_{FE}(t)\right)^2}}{\frac{1}{T}\sum_{t=1}^{T} L_{FE}(t)} \tag{3.2}$$

For the neural network-based load forecasting model, the following results are obtained:

$$\mu = -61.89MW;\ \sigma = 351.35MW$$

$$MAPE\ (\%) = 1.74; CV = -5.67$$

Both, $MAPE$ and CV are traditional relative error metrics traditionally reported in the forecast-related literature although $MAPE$ is more commonly reported. Following the traditional method of examining distributions, an error histogram is plotted as shown in Fig. 3.9. The histogram is plotted with bin size of 100 so as to exclude the unwanted distribution and focus on large forecast errors. Few observations from the histogram are:

- The dotted line shows a normal distribution with the same mean and standard deviation as the forecast errors.
- The observed error distribution, in Fig. 3.9, is more peaked with narrower shoulders and larger tails than the normal distribution assumption would suggest. Q-Q plot of forecast error is shown in Fig. 3.10 to support the non-normality assumption.
- One of the most critical features of the observed distribution is the negative mean bias, represented by a mean value of $-61.89MW$.
- The histogram shows that load forecast errors does not fit a normal distribution. Also, the Q-Q plot, in Fig. 3.10, confirms that error distribution do not follow a normal distribution.
- The actual load forecast error distribution has more mass around zero than what a normal probability density function (PDF) would predict. A solution to this anomaly is choosing the logistic distribution as it proved to fit the data better. The actual distribution used to model load forecast uncertainty is not crucial as long as it accurately represents the data.

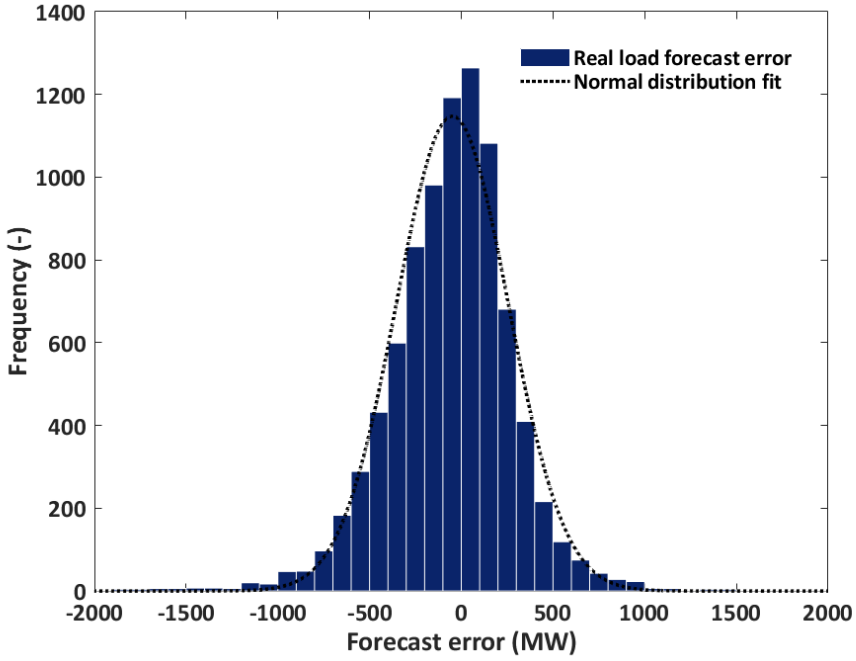


Fig. 3.9: Histogram and normal distribution fit of forecasted error with μ=-61.89 and σ=351.35

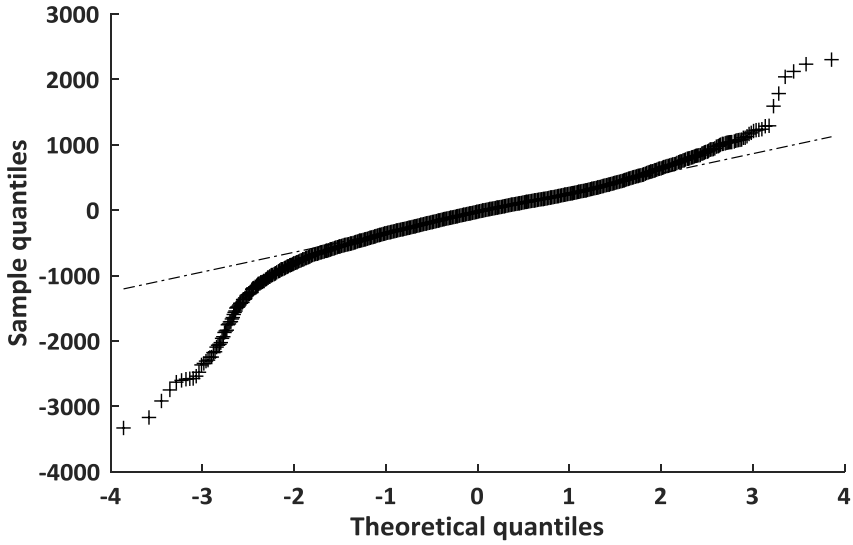


Fig. 3.10: Q-Q plot of forecast error distribution

In this error and uncertainty analysis, load forecast errors are summed for each dispatch interval in the past within a sliding window as shown in Fig. 3.7. The sliding

window size is selected to collect sufficient statistical information regarding the forecast errors. The information can be accumulated separately for each forecast horizon; for instance, for the hour-ahead forecast, two hours ahead forecast, and so on. Based on the collected statistics, the approach evaluates the percentile intervals (also called confidence intervals or uncertainty ranges) for each forecast horizon and different level of confidence. These intervals are assumed to be the same in the future dispatch interval; that is, for the next hour, the hour after that, and so on. It is used, in later part, to analyze the truncated normal distribution of load forecast error.

Based on the results obtained, intensive approaches for the uncertainty analysis of forecast errors is carried out in this study following the work on wind forecast uncertainty [Makarov et al., 2010]. The two approaches are described below:

i. *Distribution Fitting Approach*: Distribution fitting approach is fitting of probability distributions and is based on assumptions about a specific standard form of random variables; for example, normal, uniform or Poisson distributions. This approach assigns a probability to an event when the random variable takes on a specific, discrete value, or falls within a quantified range of continuous values. It is solely based on the standard distributions and selected set of its parameters. From Fig. 3.9, the forecast error distribution is more peaked with narrower shoulders and larger tails than the normal distribution assumption would suggest. Hence, selecting a distribution model means choosing a standard probability distribution and then adjusting its parameters to fit the data. A proposed solution to the above problem is fitting the error distribution dataset with truncated normal distribution (TND). A mathematically defensible way to preserve the main features of the normal distribution while avoiding extreme values involves the truncated normal distribution, in which the range of definition is made finite at one or both ends of the interval. The PDF of such a truncated normal distribution is written as:

$$PDF(x;\mu,\sigma,a,b) = \begin{cases} 0 \; if \; x \le a \\ \dfrac{\frac{1}{\sigma} PDF_N\left(\frac{x-\mu}{\sigma}\right)}{CDF_N\left(\frac{b-\mu}{\sigma}\right) - CDF_N\left(\frac{a-\mu}{\sigma}\right)} \; if \; a < x < b \\ 0 \; if \; b \le x \end{cases} \tag{3.3}$$

where, μ is the mean value of non-TND

σ is the standard deviation of the non-TND

a, b are upper and lower limits of the non-TND

$x\epsilon(a,b), -\infty \le a < b \le \infty$

$PDF_N(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$ is the PDF of standard normal distribution

$CDF_N(\cdot)$ is the CDF of standard normal distribution

The use of truncated normal distribution in wind forecast error has been reported in refs. [Xie et al., 2011, Lu et al., 2013]. One of the many reasons to use truncated normal distribution is because though normal distribution is a good fit for our forecasted error data, for physical reasons it is known that data can never be negative. Also, the values of a normal distribution can, in theory, assume any value over the range from $-\infty$ to $+\infty$, which may lead to significant computational errors in situations where the distribution's outcomes are constrained. For error analysis, it is desirable to consider data within a particular range of interest as per the planner, which we might symbolize as $[A, B]$, $[A, +\infty)$ or $(-\infty, B]$, depending on the truncation we apply. Following the study of [Billinton & Allan, 1996], the range of the truncated normal distribution can be extended, say 99.95% confidence interval of $[\mu - 3.5\sigma, \mu + 3.5\sigma]$, to represent a large forecasting error distribution. Accordingly, the modified load forecast error, shown in Fig. 3.11, is represented by a truncated normal distribution, in which the mean is the hourly power forecast and the standard deviation is 5% of the mean:

$$f(x) = \begin{cases} 0, & x < \mu - 3.5\sigma \\ & or\ x > \mu + 3.5\mu \\ \dfrac{1}{\alpha\sqrt{2\pi\sigma}} . e^{-(x-\mu)^2 / 2\sigma^2}, & \mu - 3.5\sigma \le x \le \mu + 3.5\sigma \end{cases} \tag{3.4}$$

where, $\alpha = \int_{\mu-3.5\sigma}^{\mu+3.5\sigma} \left(1/\sqrt{2\pi\sigma}\right) . e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$

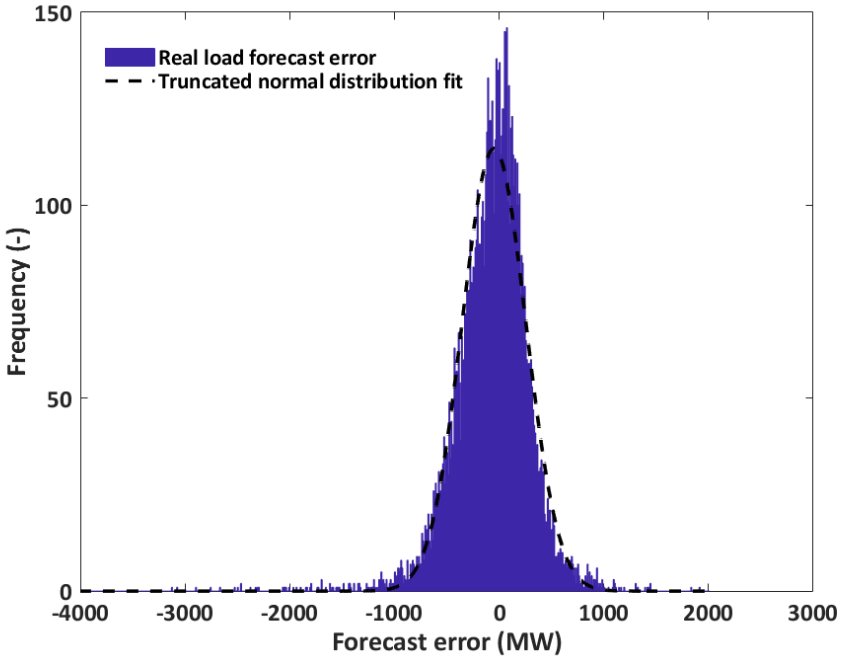


Fig. 3.11: Histogram of modified load forecast error

ii. *Empirical Probability Approach*: Statistical analysis of load forecast error distribution using empirical probability approach is not studied extensively. Empirical probability is a type of non-parametric distribution that does not follow any standard probability distribution. In this approach, the models make no assumptions about the form of the underlying distribution, so no parameter estimates are needed [Murphy, 2012].

Compared to real-life with physical data, like load forecast error distribution, the empirical PDF is simply the histogram. Integrating the PDF using cumulative sum produces the empirical CDF. If a sample comes from a parametric distribution (such as a normal distribution), its empirical CDF will resemble the parametric distribution as in the case of load forecast error shown in Fig. 3.12. The empirical CDF assigns probability 1 (*y*-axis) over *n* to each of *n* observations in the analyzed dataset. If not, the empirical distribution still gives an estimate of the CDF for the distribution. As seen in the left-hand figure of Fig. 3.12, the continuous and stairs empirical CDF overlap on each other, thus proving that the distribution is not very diverse. The confidence bounds are spaced evenly from empirical CDF.

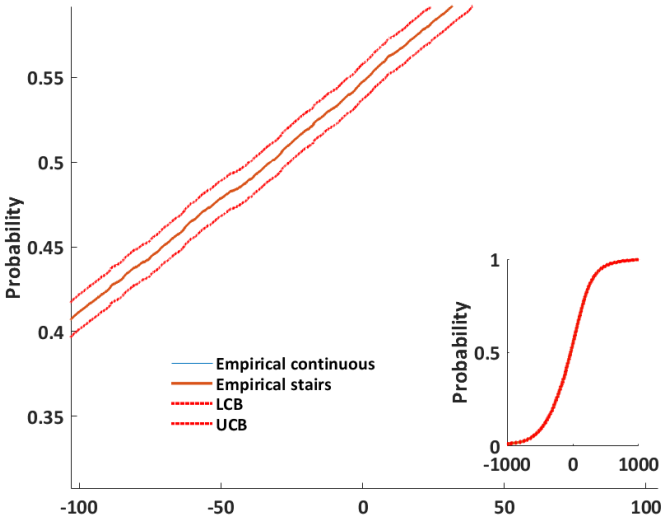


Fig. 3.12: Empirical CDF of forecasted error with μ=-61.89 and σ=351.35 [*Sub-image is truncated to ±1000. LCB: Lower Confidence Bound and UCB: Upper Confidence Bound*]

3.2.5 DISCUSSIONS

A neural network-based load forecasting model was designed and implemented. Real data from U.S. utility is used to train the neural network architecture and results are obtained with a high degree of accuracy ($MAPE$ of 1.74%). A set of optimized weights and the associated biases was obtained after training the network with real load and weather data. Forecasting was followed by uncertainty analysis to represent forecast error distribution. It was observed that the error distribution does not follow a normal distribution. Performance analysis was performed to show that truncated normal distribution is a more accurate means of modeling the error distribution.

3.3 LONG-TERM LOAD FORECASTING CONSIDERING VOLATILITY USING MULTIPLICATIVE ERROR MODEL

Long-term load forecasting plays a vital role for utilities and planners in terms of grid development and expansion planning. An overestimate of long-term electricity load will result in substantial wasted investment on the construction of excess power facilities, while an underestimate of the future load will result in an insufficient generation and inadequate demand. As a first-of-its-kind, this research proposes the use of multiplicative error model (MEM) in forecasting electricity load for the long-term horizon. MEM originates from the structure of autoregressive conditional heteroscedasticity (ARCH) model where conditional variance is dynamically

parameterized and it multiplicatively interacts with an innovation term of time series. Historical load data, accessed from a U.S. regional transmission operator, and recession data, accessed from National Bureau of Economic Research, for years 1993-2016 is used in this study. The advantage of considering volatility is proven by out-of-sample forecast results as well as directional accuracy during the great economic recession of 2008. To incorporate future volatility, backtesting of MEM is performed. Two performance indicators used to assess the proposed model are mean absolute percentage error (for both the in-sample model fit and out-of-sample forecasts) and directional accuracy.

Load forecasting in the long-term horizon is important for electric utilities and planners in terms of grid expansion planning, future investments and revenue analysis for long-term decision-making process. Moreover, it plays a vital role in the economic and social development of a country (or specific region in case of some utilities). A more realistic range of future generation scenarios can be modeled when the electricity consumption is increasing at a faster rate in this globalizing world. For instance, annual load forecasting is favored among utilities and is one of the common long-term load forecasting problems. It can alleviate the disparity between demand and generation, thereby maintaining the required level of security of supply. Choosing a right horizon for long-term varies from one utility to another based on their policies. Usually a monthly or yearly time-step for one to ten years ahead in long-term load forecasting is helpful in inter-tie tariff setting and long-term grid investment return problems.

It is often difficult to forecast load over a such a long planning horizon and it is due to the stochastic nature of demand growth and the influential parameters. Most of these parameters are, by nature, unpredictable and uncontrollable. Examples are socio-economic developments, the occurrence of special events and/or climatic conditions, and regulatory requirements. Any considerable deviation in forecast results in over expenditure on generation/transmission infrastructure or energy resource waste. Hence, in order to improve the forecast accuracy in the long-term horizon, attention is needed either in terms of *improvement of existing employed techniques* or *development of a new technique to consider all the aforementioned factors*. Forecast accuracy influences the decision making of generation and transmission companies on their plans to address future load growth and market volatilities. Based on the forecast, electric utilities coordinate their resources to meet the actual demand using a cost-effective plan. As we look into the future, in the energy consumption scenario, it is expected that electricity will play a major role as we move towards the declined use of de-carbonized heat pumps in the residential sector and the addition of electric vehicles (EVs) and other hybrid vehicles in the transportation sector. Such transition will have a significant impact on the overall load profile. To visualize, Fig. 3.13 depicts the future complexity in the context of load forecast, various players in action and the inter-dependency that needs attention too. Stochasticity in future scenarios, such as economy (GDP: Gross Domestic Product), demographics (change in energy usage of end-users), energy

generation mix (renewable energy generation depends on forecasting of a resource like wind, solar irradiation) and technology advancement, influences the forecast accuracy to a large extent. For utilities, the evolution of prosumer complicates the forecasting methodology. A prosumer (combination of words *producer* and *consumer*) is a consumer that has its own production facility (e.g., rooftop solar panel). Because of decentralized power generation, prosumers have evolved recently and played an active role in electricity generation and the provision of grid services. Not to be forgotten is the spatial complexity as the area expands from the DSO level to multi-TSO level. Decisions about grid development are mostly based on the accuracy of predictions of both the scale and occurring geographical locations of energy consumption in the long-term horizon. Any change in the electric consumption patterns is compensated by financial incentives and/or electricity tariffs. In such cases, accurate forecasting will certainly provide support for the utility (TSO/DSO) to estimate the amount of investment needed.

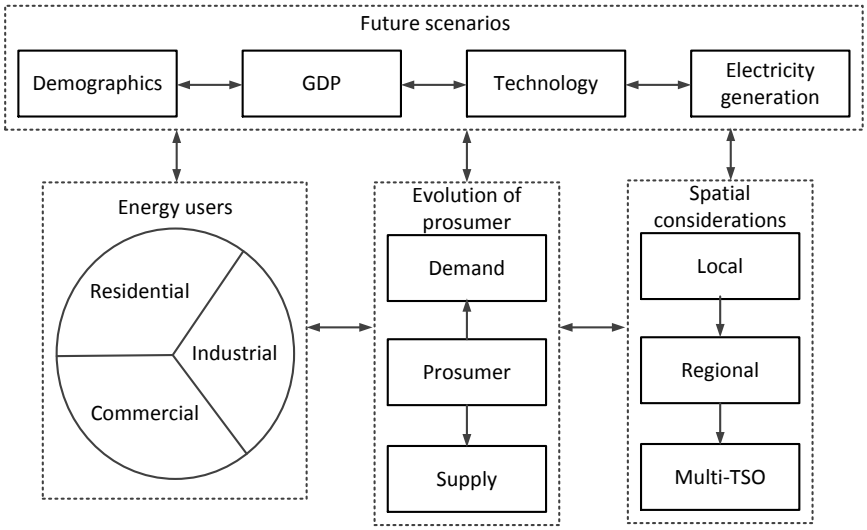


Fig. 3.13: Complexity in long-term load forecast [GDP: Gross Domestic Product]

This research sums up the need for accurate LTLF as,

- Firstly, moving towards greener future is accredited with development in new technology and integration of renewable energy into the primary grid while discarding fossil fuels is becoming important. In the Paris Agreement 2016 [Web2, 2016], it was agreed upon to move towards renewable energy from the more conventional energy. Such a move is realized with an accurate and reliable forecast of the electrical energy demand. Despite advancements in battery

technology, storing energy for the long-term purpose is not the viable option. Thus, accurate and reliable forecasting is required for planning the right tools.

- Secondly, considerable changes in weather factors like temperature, rainfall and hot/cold days. Any change in climatic variables will have a direct impact on the demand pattern. Erratic weather events posed due to climate change pose some serious burden on forecasters to accurately model load growth considering long-term horizon.
- Thirdly, maintaining the security of supply during the energy transition. In today's date, existing grids are performing under stress to deliver the growing demand in the presence of variable stochastic renewable energy sources.
- Lastly, the occurrence of events like the great economic recession of 2008 that jolted the economic backbone of many countries. Such an event is termed as black swan event. A black swan event is an incident that occurs randomly and unexpectedly and has wide-spread ramifications. The effect of the great recession was widespread and energy investment worldwide plunged into tougher financing environment and weakening final demand for energy [Web1, 2009]. This reminds the importance to study the financial aspects of long-term forecasting by energy forecasters in electric utilities.

Based on the needs, the key contributions of this research can be listed as:

- Reviews recent advancement in long-term load forecasting in terms of techniques and models developed.
- Provides a comprehensive and critical evaluation of long-term load forecasting considering historical volatility.
- Proposes the use of multiplicative error model (MEM) to model conditional mean and to forecast aggregated zonal monthly load. In this research, we consider a forecast horizon of 4 years as a solution for electric utilities and planners based on the fact that construction of offshore wind farms takes approximately 3-4 years depending on the capacity [Khuntia *et al*., 2016a].

The rest of the section is organized as follows: sub-section 3.3.1 gives a background on LTLF and volatility, sub-section 3.3.2 introduces MEM and is followed by forecast methodology in sub-section 3.3.3. Sub-section 3.3.4 analyses the forecast results based on real data. Finally, sub-section 3.3.5 concludes the section.

3.3.1 BACKGROUND ON LONG-TERM LOAD FORECAST

Electricity load forecasting in the long-term horizon is an important part of the transformation of electric power systems and it has appealed more and more attention from both academics and industry. By principle, a load forecasting model aims at a

mathematical representation of the relationship between load and influential parameters. Such a model is identified with coefficients that are used to forecast values by extrapolating the relationship to the desired lead time. Eventually, the accuracy of the model depends on both the selected model as well as the accuracy of the estimated parameters. Literature study reveals that long-term load forecasting received less attention compared to short-term load forecasting. This is because of the complexity involved in achieving an accurate forecast. Long-term load forecasting is based on the integration of concepts from theoretical foundations of economic theory with knowledge on financial, statistical, probability and applied mathematics to make inferences about the load growth/fall and technology evolution. Reference [Khuntia *et al*., 2016e] illustrates rationally the concept of long-term load forecasting and also presents recent development within the electric power industry. Reference [Hong, 2014] performed a study on past, current, and future trends in energy forecasting while stating the trend in spatial, short-term and long-term load forecasting, and energy price forecasting in a lucid manner. Reference [Feinberg & Genethliou, 2005] proposed three methods suitable for long-term forecasting as time series approach, econometric approach, and end-use approach. For long-term forecasting, all approaches require historical data and they are broadly categorized into traditional (or statistical) and non-traditional (or artificial intelligence AI) based methodologies.

Traditional methods include regression-based models and time series methods. Reference [Saab et al., 2001] proposed univariate autoregressive models to forecast load with monthly time-step in Lebanon. Multiple linear regression models were proposed in [Mohamed & Bodger, 2005]. Reference [Kandil et al., 2001] implemented a knowledge-based expert system to support the choice of the most suitable load forecasting model with practical application. However, traditional methods are criticized for their weakness of non-linear fitting capability. In AI-based techniques, artificial neural network (ANN) is one of the most popular models. Its application in forecasting Greek long-term energy consumption for the years ahead is reported in [Ekonomou, 2010]. Reference [El-Ela et al., 2009] used ANN on the Egyptian electrical network for long-term peak load forecasting. Reference [Xia et al., 2010] reported the superiority of ANN for medium and long-term load forecasting in terms of accuracy and robustness. Hybrid of fuzzy and ANN are reported in [Chen, 2012] for forecasting Taiwan's annual electricity load and in [Padmakumari et al., 1999] for long-term electrical energy consumption in India. Other AI techniques include support vector regression models (SVR) [Hong, 2009, Jianjun et al., 2016] and SVR with simulated annealing algorithms [Pai & Hong, 2005].

Use of metaheuristic methods such as genetic programming [Karabulut et al., 2008], fruit-fly algorithm [Li et al., 2013], gravitational search algorithm [Abdi & Beigvand, 2016] and particle swarm optimization (PSO) [Ünler, 2008, AlRashidi & El-Naggar, 2010] are also reported. Other methods include long-term forecasting based on partial least

squares method [Meng & Niu, 2011] and complete decomposition method [Sun, 2001]. Recent study includes forecasting for country-specific such as Spain [Moral-Carcedo & Pérez-García, 2017], Greece [Ekonomou, 2010, Angelopoulos et al., 2017], Lebanon [Saab et al., 2001], Turkey [Yumurtaci & Asmaz, 2004, Dilaver & Hunt, 2011, Hamzacebi & Es, 2014]. More recently, reference [Kaboli et al., 2017] used gene expression programming for long-term prediction of electrical energy consumption in ASEAN-5 countries and projected up to 2030 according to rolling-based forecasting procedure. The results are compared with those obtained from ANN, SVR, adaptive neuro-fuzzy inference system (ANFIS), rule-based data mining algorithm, gene expression programming (GEP), linear and quadratic models optimized by PSO, cuckoo search algorithm (CSA) and backtracking search algorithm (BSA).

It was learned that the developed models aim at predicting accurate peak load or electrical energy consumption while comparing with any traditional model. However, one aspect that has received less attention in long-term load forecasting, when the whole energy scenario is growing in terms of complexity and dynamics, is volatility. The concept of volatility is prevalent in financial markets and it refers to the degree of erratic variations of a process over time. It is used as a criterion to study the risk associated with a financial asset. Reference [Zareipour et al., 2007] showed that power markets have greater volatility levels than other financial markets like crude oil, natural gas or stock prices. Literature study reveals volatility studies on various electricity markets: Spanish, Californian, UK and PJM electricity markets [Benini et al., 2002], Ontario and some of its neighboring markets [Zareipour et al., 2007], German market [Auer, 2016], Australian electricity market [Boland et al., 2016] to name a few. Reference [Li & Flynn, 2004] examined and compared the volatility of 14 deregulated markets through the “price velocity” metric. The Nordic pool was studied in [Simonsen, 2005] considering volatility clustering, log-normal distribution, and long-range correlations. In time series forecasting of electrical load, volatility is defined as a deviation from the mean which corresponds to risk. An advantage of such an approach is that once the time series model is understood, it is possible to simulate the data generation for any lead time in the future. Reference [Khuntia *et al*., 2016d] explained the importance of volatility in long-term load forecast, which no work reported earlier. Extending the concept of volatility forecast to load forecast in the long-term horizon is adopted in this research.

Volatility is a fundamental issue in financial and econometrics domain, and virtually present in all financial decision making. The concept of volatility in financial markets refers to the degree of unpredictable fluctuations of a process over time. Volatility can be broadly classified into five major types: price volatility, stock volatility, historical volatility, implied volatility, and market volatility [Web3, 2018]. In this study, historical volatility is used to account for implied volatility. In the financial market, historical volatility is understood as how much volatility a stock has had over the past time-frame

(say, 12 months). If the stock price varied widely in the past 12 months, it is considered more volatile and riskier. Implied volatility is understood as how much volatility the stock will have in the future. In fact, volatility is forecastable because of a number of persistent properties: (i) it appears in clusters, (ii) it changes over time and has unusual jumps, (iii) it does not grow to infinity and is persistent in specific time-span, and (iv) it reacts different for an increase or decrease of the considered entity. For instance, load forecast in the long-term horizon takes into account socio-economic factors like population growth and gross domestic product (GDP) along with explicit factors like historical load and weather data. Presence of economic factors induces volatility, or what is called as implied volatility. In fact, implied volatility is generally treated to be the best available forecast as it has certain characteristics that can increase the accuracy of forecast values. Likewise load, future volatility prediction is an extremely difficult task because the actual realization of the future process volatility will be influenced by events that happen in the future. Thus, it is important to develop a model that can fit the sequence of calm and turbulent periods. Studies reveal that ARIMA technique, one of the widely used forecasting techniques, is inadequate in long-term forecasting because it suffers from the mean convergence problem [Shumway & Stoffer, 2011]. It means that ARIMA forecast *converges to the mean* of the observations as the forecast horizon grows. To address the short-coming and treating volatility as an influential parameter, the next section introduces the concept of MEM and its application to load forecast in the long-term horizon.

3.3.2 MULTIPLICATIVE ERROR MODEL FOR LONG-TERM LOAD FORECAST

Multiplicative error model (MEM) was introduced by Engle in 2002 [Engle, 2002] as an adaptation of autoregressive conditional duration model [Engle & Russell, 1998] to be used for time series that always receive positive values. Literature study on MEM reveals its application in financial risk and volatility forecasting [Lanne, 2006, Han et al., 2015, Caporin et al., 2017]. A search about the application of MEM in load forecasting reveals no information, not even for short-term forecasting which is common among forecasters. Hence, the proposed model is first-of-its-kind to introduce MEM for load forecasting. As the electricity load is always represented as a non-negative time series, MEM is presumed to be a good fit to forecast. The MEM for a non-negative time series (y_t) on $[0, +\infty)$ and considering $\mathcal{F}_{t-1}$ as information available for forecasting y_t is written as [Engle, 2002]:

$$y_t = \mu_t \varepsilon_t \tag{3.5}$$

where, the range of the disturbance ε_t is non-negative on $[0, +\infty)$, unit mean and unknown constant variance given as $\varepsilon_t|\mathcal{F}_{t-1} \sim D(1, \psi)$ for positive distribution D. μ_t is conditional on $\mathcal{F}_{t-1}$ and positive, described on a parameter vector θ as:

$$\mu_t = \mu(\theta, \mathcal{F}_{t-1}) \tag{3.6}$$

When $\mathcal{F}_{t-1}$ includes only historical values of the series, μ_t can be generalized as:

$$\mu_t = \delta + \sum_{i=1}^{p} \alpha_i \mu_{t-i} + \sum_{j=1}^{q} \beta_j y_{t-j} \tag{3.7}$$

where, δ is constant, term $\sum_{i=1}^{p} \alpha_i \mu_{t-i}$ represent an inertial component, and term $\sum_{j=1}^{q} \beta_j y_{t-j}$ represent more recent observation. Equation 3.7 is referred to as referenced MEM of order (p, q). Model specifications can be modified to adapt to the needs of the load forecast. For instance, residuals at t-th observation denoted as $\vartheta_t = y_t - \mu_t$ and $\alpha_1^* = \alpha_1 + \beta_1$, equation 3.7 can be written as:

$$y_t = \delta + \alpha_1^* y_{t-1} + \vartheta_t - \alpha_1 \vartheta_{t-1} \tag{3.8}$$

Equation 3.8 represents an ARMA model with heteroskedastic errors and is the cornerstone of this modeling approach. The procedure of finding and validating a suitable MEM for a given dataset is discussed in the next section.

3.3.3 FORECAST METHODOLOGY CONSIDERING REAL DATA

In order to realize a suitable long-term forecasting model, one must start with a rich historical database, construct the model, identify the appropriate model order and finally evaluate the forecast results. Fig. 3.14 shows the steps to forecast load using MEM. Since MEM falls under time series models, we follow the Box-Jenkins methodology of building a model with certain adaptations [Box et al., 2008]. Starting from data preparation, the first part involves stationarity checking, data fitting, and model identification while checking various statistical properties of the time series. Identifying the right model, estimating parameters and checking the model adequacy falls under this part. In the second part, MEM is validated for forecasting both as in-sample fit and out-of-sample prediction. Modeling of MEM starts with the identification of autoregressive and moving average parameters of non-negative time series that has predictive power regarding the directional change, and later added by persistent error specifications that eventually improves forecast. MEM differs from linear regression models in the sense that the mean equation, which is a scalar factor, is multiplied with the independent and identically distributed ($i.i.d.$) error term. The scalar factor evolves in a conditionally autoregressive manner, hence, favorable for forecasting. The assumption of $i.i.d.$ means that the error terms behave randomly with constant mean

and variance over a considered time-horizon. However, in reality, both the original time series as well as error time series are highly correlated and do not behave as an $i.i.d.$ process.

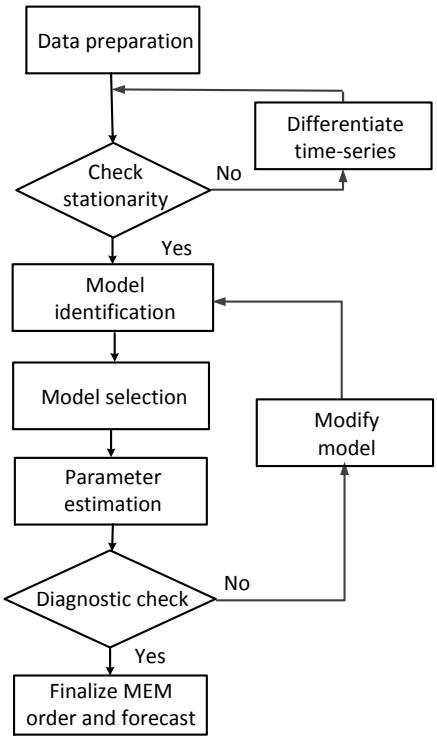


Fig. 3.14: Flowchart for model identification, selection, and forecast

3.3.3.1Database generation

Forecast accuracy strongly depends on the quality of available historical data. A poor history, composed only by anomalous or average events, may polarize the analysis and affect the quality of the forecast values. For this study, historical data of a specific load zone region under a U.S. regional transmission operator [Web4, 2018] and data describing economic recession as extracted from National Bureau of Economic Research (NBER) [Web5, 2018] is considered. Hourly load data for years 1993-2016 is extracted and sampled to monthly aggregated load as shown in Fig. 3.15. The use of monthly time-stamp enables in understanding the monthly energy consumption. Recession data for years 1993-2012 is used to build the predictor. From Fig. 3.15, it is evident that the load growth is on a fairly positive side apart from a few incidents where a downturn in demand is observed. Such an incident is the year span of 2006-2009, where the year 2007-2008 is identified by large variability in demand value because of spikes and negative demand growth coincide with the great recession of 2008. The de-seasonalized data, shown in blue color in Fig. 3.15, helps in obtaining a goodness-of-fit

measure. It is achieved by dividing the original data by the seasonal factors (12 months as it is monthly aggregated data) to get something that looks more like a straight line.

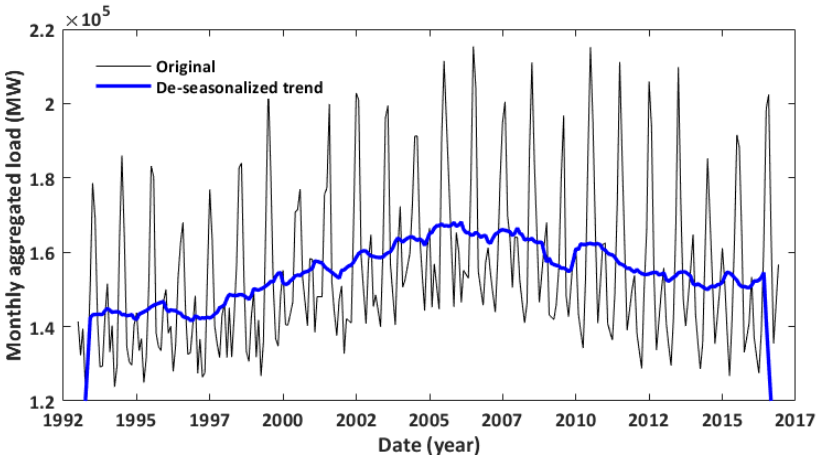


Fig. 3.15: Monthly aggregated load data and de-seasonalized trend for years 1993-2016

3.3.3.2Stationarity and autocorrelation test

A visual inspection of Fig. 3.15 suggests non-stationary time series with a linear trend and seasonal periodicity. Tests reveal that non-stationarity is apparent as both mean and variance increase with time. The class of MEM requires time series to be stationary so that its statistical (up to the second order moment) properties do not depend on time. This is coherent with any time series forecasting because non-stationary time series are erratic and unpredictable. Phillips-Perron (PP) test is used for stationarity check [Elder & Kennedy, 2001]. For any time series $x_t = ax_{t-1} + e_t$ where e_t is the residual, PP test checks for the null hypothesis ($H_0: a = 0$ vs. $H_1: a \neq 0$). Use of PP test is preferred over the widely used Augmented Dickey-Fuller (ADF) because of its non-parametric nature. In addition to the steps from ADF test, PP test corrects the statistics to account for autocorrelations and heteroscedasticity. The time series is checked for 0 lags and both the tests reject the null hypothesis with a p-value of 0.001. Thus, the time series is differenced to obtain a stationary time series and next step is to determine the presence of autoregressive or moving average terms to correct any autocorrelation that exists in the differenced time series.

To check the presence of correlation, two tests used to check the null hypothesis ($H_0: no\ autocorrelation$ vs. $H_1: correlation$) are Ljung-Box Q-test (Q) and Durbin-Watson (D) test [Johnson Jr et al., 1987]. The Q-test statistic for R residuals, L lags is written as,

$$Q = R(R+2)\sum_{l=1}^{L}\left(\frac{\rho(l)^2}{(R-l)}\right) \tag{3.9}$$

where, $\rho(l)$ is the autocorrelation coefficient at lag l. The Durbin-Watson statistic (D) is conditioned on the order of the observations (rows) or in this case, is the number of months. The D-test statistic for n-observations is written as:

$$D = \frac{\sum_{i=2}^{n}(R_i - R_{i-1})^2}{\sum_{i=1}^{n} R_i^2} \tag{3.10}$$

Presence of autocorrelation in a time series indicates that the values of adjacent observations are correlated. Fig. 3.16 shows the autocorrelation function (ACF) and partial autocorrelation function (PACF) plot giving evidence of the presence of autoregressive and moving average parameters. The ACF plot reveals the presence of significantly large autocorrelations, particularly at every 12th lag. Presence of autocorrelation suggests the data is dependent and correlated and needs modification. Table 3.2 displays the detailed statistics of original and differenced time series. Taking a look at the third and fourth moments of distribution (skewness and kurtosis), it is realized that the differenced time series is close to normal. It is slightly left-skewed which means that the left tail is longer. Skewness involves the third moment of the distribution and kurtosis involves the fourth moment. The outliers in the distribution, therefore, have even more effect on the kurtosis than they do on the skewness. In a symmetric distribution, both tails increase the kurtosis, unlike skewness where they offset each other. The mean and standard deviation have the same units as the original data, and the variance has the square of those units. However, the kurtosis, like skewness, has no units: it is a pure number. The standard value of kurtosis of a normal distribution is 3.

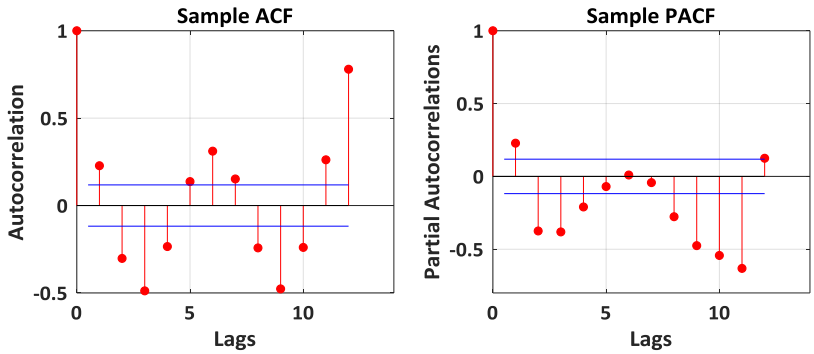


Fig. 3.16: Sample autocorrelation function (ACF) and partial autocorrelation function (PACF) plot of differenced time series with significance limit of 20% indicated by blue line

Table 3.2: Detailed statistics of load time series

	Original	Differenced
Mean	154581.1	53.40888
Max	215379.9	45842.8
Min	123786.5	-49286.9
Median	149928.9	433.1667
Standard deviation	20514.4	17879.6
Skewness	1.015012	-0.10097
Excess kurtosis	3.449577	3.005507

3.3.3.3Volatility check and multiplicative error modeling

Next step in modeling is to check if the differenced time series shows any cluster of volatility and satisfy the homoscedastic assumption of constant variance or heteroskedastic behavior. It may happen that squared values of the differenced time series exhibit significant serial correlation. It means that values are again dependent but serially uncorrelated. So, the sample autocorrelation and partial autocorrelation test is repeated for squared residual followed by Q-test and DW-test. The tests re-confirm our model selection [Ljung & Box, 1978], and the corresponding plot of autocorrelation and partial autocorrelation function is shown in Fig. 3.17.

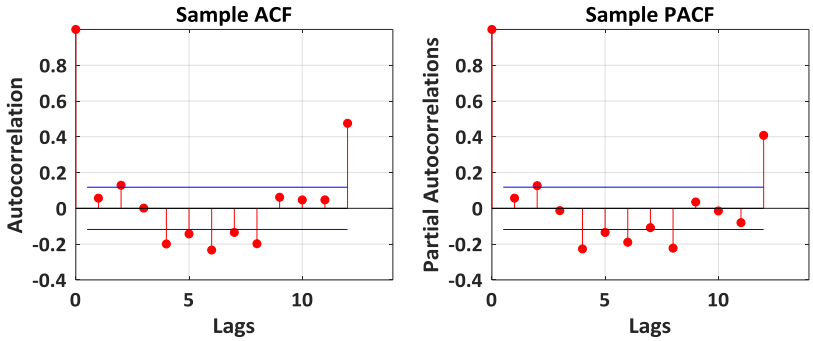


Fig. 3.17: Sample autocorrelation function (ACF) and partial autocorrelation function (PACF) plot of squared residuals with significance limit of 20% indicated by blue line

The ACF and PACF plots in Fig. 3.17 verifies the presence of conditional heteroscedasticity and also facilitates in identifying the appropriate model order for de-seasonalized differenced time series. As stated in [Engle, 2002], generalized autoregressive conditional heteroscedasticity (GARCH) models are a form of MEM and form the basis of the proposed model. With reference to equation 3.5, if μ_t is the conditional expectation of y_t w.r.t available information (or historical values), its parameters can be estimated by specifying a GARCH for the conditional second moment of $\sqrt{y_t}$ while imposing its conditional mean to be zero. Reference [Ahoniemi, 2006] augmented the regression model with GARCH error modeling, and the same concept is

adapted for this study. The standard model common to both the processes and its square while rewriting equation 3.5 is:

$$Z_t = \sqrt{h_t} e_t$$
$$Z_t^2 = h_t e_t^2 \tag{3.11}$$

In the squared equation, the dependent variable (Z_t) is non-negative with conditional mean h and a non-negative multiplicative error term $e_t \sim i.i.d.(0,1)$ with unit mean. This can be estimated by taking the load residual as the dependent variable of a GARCH model. The GARCH model is an extension of the ARCH model, in the way that it allows current volatility to be dependent on its lagged values directly. For more information on ARCH and GARCH, reference [Bollerslev, 1986] is recommended. The model can be estimated by taking Z_t as the dependent variable, with specifications of zero mean and an error process. In such case, the *conditional variance* is then the *conditional mean* of Z_t^2 [Brooks & Oozeer, 2002]. Rewriting equation 3.5, the GARCH model with order $p \geq 0$ and $q \geq 0$ is defined as [Bollerslev, 1986]:

$$Z_t = \sqrt{h_t} e_t \tag{3.12}$$

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Z_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}^2 \tag{3.13}$$

for the square root of duration, and where $\alpha_0 > 1$, $\alpha_i \geq 0$ and $\beta_j \geq 0$ are constants with

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1 \tag{3.14}$$

and $e(t)$ is independent of $Z_{t-k}, k \geq 1$.

Selecting the right order (p, q) is achieved by following one of the many order selection tests. Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) tests are chosen in this study. The reasons for choosing the two criterions are that both the tests assess the fit between model predicted and original values and penalize models with a larger number of parameters. Tests confirmed the use of order (1,1) multiplicative error model. Fig. 3.18 shows the innovation plot for a sample size of 101 (0-80 range shown in Fig. 3.18), and it can be concluded that clusters of volatility appear

in a periodic manner. The innovation is the difference between the observed value of a variable at time *t* and the optimal forecast of that value based on information available prior to time *t*. Thus, the movement of non-linearity is not only dependent on the previous values but for the whole time series, it is uncorrelated. Volatility tends to cluster into periods with higher and lower volatility. This effect proves that volatility at some time must be dependent on its historical values, say with some degree of dependence.

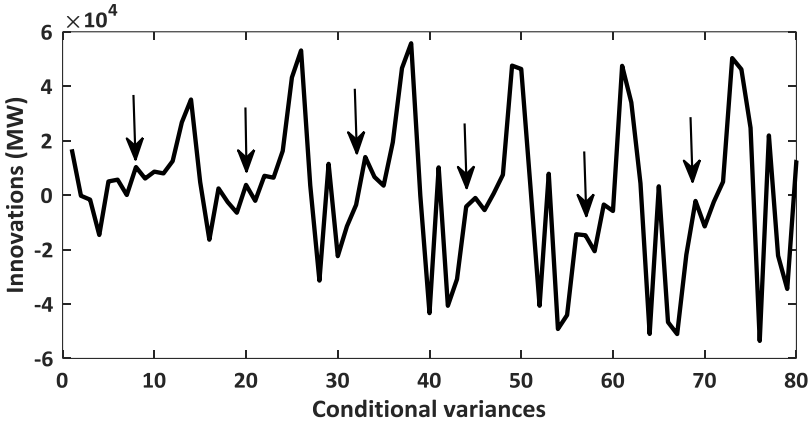


Fig. 3.18: Innovation plot showing clusters of volatility (pointers show the clusters)

After the model order is identified, maximum likelihood estimator is used in estimating the parameters. Regardless of the low standard errors, parameter estimation is still feasible. As the sample size runs from $N \to \infty$, the probability that the value of the estimators shows a large divergence from the true (which is unknown) parameter values goes to 0, making it a consistent estimator. Estimation is achieved with conditional variance $h_t \sim i.i.d.(0,1)$, and with an assumption that error distribution follows student t-distribution, a version of the generalized error distribution, whose density is given as,

$$f(x) = \frac{ve^{\left(-\frac{1}{2}\left|\frac{x}{\lambda\mu}\right|^{v}\right)\frac{1}{\mu}}}{\lambda 2^{\left(1+\frac{1}{v}\right)}\Gamma\left(\frac{1}{v}\right)} \tag{3.15}$$

where v is the positive measure of fat tail, $\lambda = \sqrt{2^{-(2/v)}\,\Gamma(1/v)/\Gamma(3/v)}$, and $\Gamma(.)$ is the gamma function defined as $\Gamma(x) = \int_0^\infty y^{x-1}e^{-y}dy$. This assumption helps in the better modeling of excess kurtosis (in Table 3.2). It also approximates the normal distribution

as the degrees of freedom grow to infinity. Presence of fat tail is evident from the Q-Q plot in Fig. 3.19.

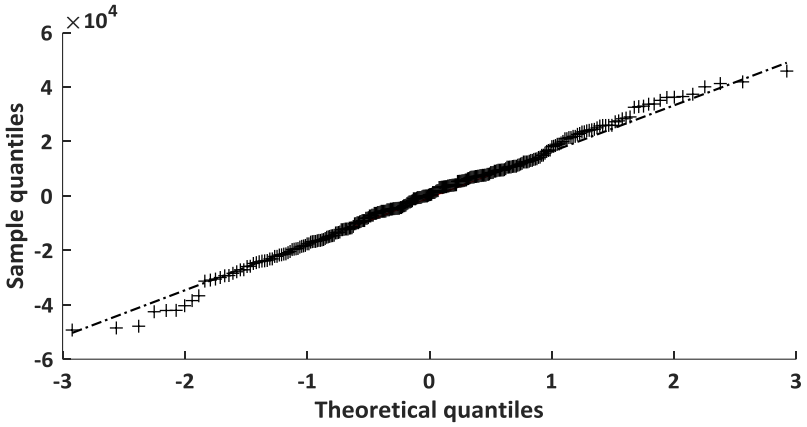


Fig. 3.19: Q-Q plot of residuals

3.3.4 RESULTS AND ANALYSIS

Result analysis consists of two parts: first part consist of in-sample model fit using load and economic data followed by out-of-sample forecast, and the second part is checking directional accuracy by forecasting for the year 2008 during the great economic recession.

3.3.4.1In-sample model fit and out-of-sample forecast

A forecast horizon of 4 years is chosen in this study for both in-sample model fit as well as out-of-sample forecast. It is based on the assumption that off-shore wind-farm plant needs at least 3-4 years for completion, which is itself a long-term grid development action [Khuntia *et al*., 2016a]. For the in-sample model fit, the study embodies fitting the MEM using load data and recession data of years 1993-2008 and then evaluating its performance on the data set for years 2009-2012. When assessing point forecasts with mean square errors, it appears to be useful to use a longer in-sample period for model estimation as followed in this study. Fig. 3.20 shows the in-sample model fit for the years 2009-2012. It is clear from Fig. 3.20 that the in-sample model fit is below the original values for the last three years (2010-2012) but the year 2009. For the year 2009, the original values show a lower peak aggregated load as compared to model fit prediction. Hereby, it is essential to note that the economy was just reviving after the great economic recession of 2008. Thus, the year 2009 shows a lower peak while the in-sample model fit does not recognize and is higher. However, the model learns and consequently the peak is lower. To check the model accuracy, apart from visual inspection, an expected loss function is required to assess the model performance and check the model accuracy. Use of appropriate loss function also aims at summarizing the accuracy of the point estimate and future distribution. The two loss functions used

in this research are Mean Absolute Percentage Error ($MAPE$) and Mean Squared Error (MSE), both being unit-free measures. While the optimal point forecast under mean absolute error (without percentage) is the median, MSE represents the (conditional) mean. For two sets of n-observations ($x_{i,\dots,n}, y_{i,\dots,n}$), $MAPE$ is defined as

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - x_i}{y_i}\right| \tag{3.16}$$

where y_i is the original monthly aggregated load and x_i is the predicted monthly aggregated load. In-sample model fit accuracy is achieved with MAPE of 4.98%. MSE is defined as,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2 \tag{3.17}$$

The MSE, a quadratic and symmetric function, is a measure of how close a fitted line is to data points. The smaller the MSE, the closer the fit is to the data. However, studies reveal that median error measures are not sensitive [Armstrong and Collopy, 1992]. Thus, we employ Root Mean Squared Error ($RMSE$), which is just the square root of the MSE and it measures the deviation in terms of MW. The $RMSE$ describes the magnitude of the error in terms that would be relatively more useful to decision makers. It can be argued that both $MAPE$ and MSE are less appealing because percentages do not have obvious implications for decision making. While $MAPE$ is scale independent, MSE is more sensitive to a few large errors than to many small errors. In addition, squared error terms may be more difficult for decision makers to understand. The $RMSE$ is calculated as the distance, on average, of a data point from the fitted line, measured along a vertical line. For in-sample model fit, calculated $RMSE$ is $7.7 \times 10^3 MW$.

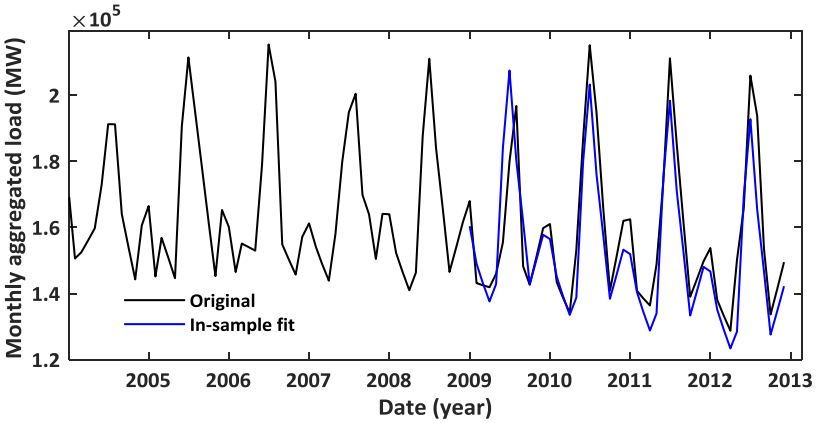


Fig. 3.20: In-sample fit for years 2013-2016 with a MAPE of 4.98%

No forecasting analysis is complete without performing out-of-sample forecasts. For better out-of-sample forecasts, the most crucial choice is splitting the series between training and test periods. Unfortunately, no study exists so far that discusses how to choose the decision point [Hansen & Timmermann, 2012]. In this study, training dataset of 1993-2012 is chosen to forecast the next four years (2013-2016). The accuracy of the MEM is improved with backtesting technique where the aim is to achieve a dynamic model that can address the future volatilities. With 48 months as forecast horizon and monthly timestamp, the MEM is built every month and forecasts ahead 48 months. The forecast result compares with original values and averages the error. In such a manner, the out-of-sample result improves as the model learns and adapts from past results. Fig. 3.21 shows the out-of-sample forecast results with MAPE of 7.09% and $RMSE$ of $1.09 \times 10^4 MW$. A high error percentage as compared to the in-sample model fit is understood from the long forecast horizon.

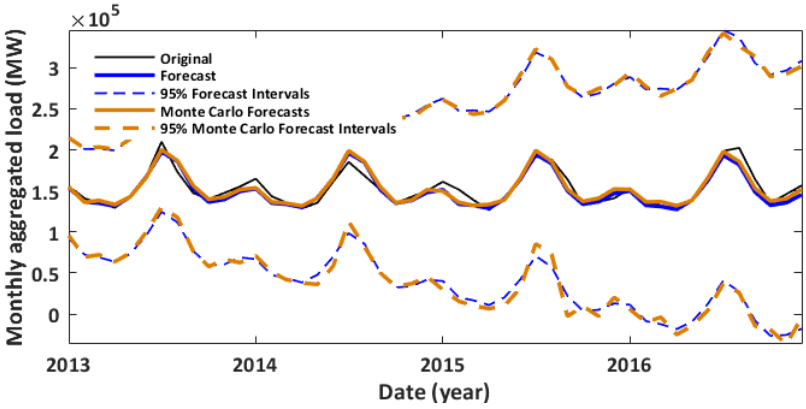


Fig. 3.21: Out-of-sample forecast for years 2013-2016 with 95% confidence interval both upper and lower and a MAPE of 7.09%

To better evaluate the model accuracy, Monte Carlo simulation is run for 500 sample paths by choosing a confidence interval of 95%. The motivation behind calculating range forecasts this way is to evaluate the likelihood that a particular forecast will be accurate within a specified confidence bound. In this way, the values within the confidence interval of the conditional mean describe the considerable range of values of the point on the line. Thus, the conditional mean for all values of time series indicates how much the entire MEM prediction can considerably move from sample to sample. It eases in predicting the range of likelihood values that an observation in the next time step may take. The confidence interval of the out-of-sample forecast presents a range for the mean rather than the distribution of individual data points. Fig. 3.21 shows a comparative analysis of out-of-sample forecast and the Monte Carlo simulation results. Both the forecast as well as confidence intervals from the two outputs are virtually indistinguishable. To understand the intervals, a value of 0.05 corresponds to

predicted upper and lower intervals where there is a 5% chance that original values will not be in that range.

It is implicit that forecast error measure increases with forecast horizon. Thus, to check the forecasts, both in-sample model fit as well as the out-of-sample forecasts are evaluated by means of their MSE-values. A comparison of MSE for the 48 months horizon for both in-sample model fit and out-of-sample forecast is shown in Fig. 3.22. The error difference is quite large at the beginning of the horizon and increases in the middle, evident in 13-36 months. An exceptional performance of out-of-sample forecast is observed when it outperforms the in-sample model fit results. The error comparison graph reveals that out-of-sample forecasts better reflect the information available to the forecaster in real time.

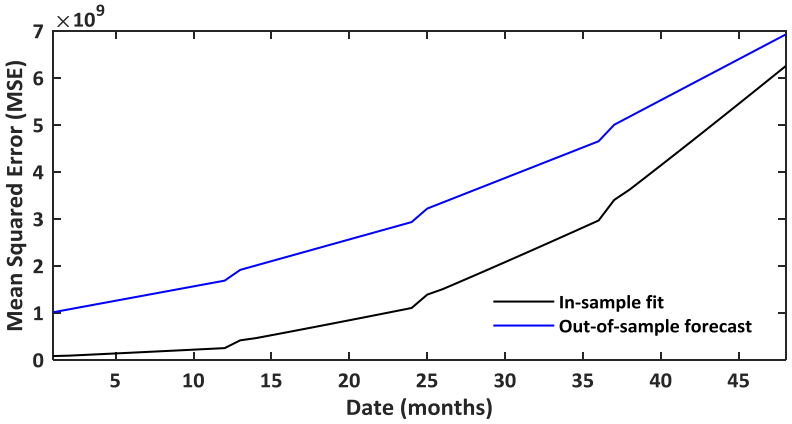


Fig. 3.22: Comparison of Mean Squared Error (MSE) values for in-sample model fit and out-of-sample forecast

While evaluating the forecast results, we take a glimpse back at Fig. 3.19 and observe that the data is skewed to the right. The Q-Q plot also displays sizeable excess kurtosis or fat tails. Also referring to Table 3.2, the high skewness and kurtosis value is an indicator of non-normal time series. To verify the claim, Jarque–Bera (JB) test is considered in our study. It is usually used for large datasets because other normality tests are not reliable for large datasets. The JB-test verifies the null hypothesis ($H_0: normal$ vs. $H_1: non-normal\ distribution$). The JB-test statistic is written as [Jarque & Bera, 1987]:

$$JB = N\left(\frac{s^2}{6} + \frac{(k-3)^2}{24}\right) \tag{3.18}$$

where, N is the sample size, s is the skewness coefficient and k is the kurtosis coefficient. A value of 1 from JB-test indicates the data is non-normally distributed. The residual distribution is fitted with Student's t-distribution, which has a thicker tail and is thus more tolerant of extremes. The study is repeated by including both fat-tails and

volatility to verify if the forecast improves and the result is significant. Fig. 3.23 shows the forecast for years 2013-2016 with and without accounting for fat-tails. The inclusion of fat-tail is significant because it represents a greater likelihood of extreme events occurring similar to the financial crisis, also called the black swan event [Taleb, 2007]. Some notable features of volatility that should be clearly mentioned are: volatility appears in clusters apparent from Fig. 3.19, volatility changes over time and that jumps in the volatility are unusual, volatility does not grow to infinity; it rather stay within some spans, and the fourth characteristic is that the volatility reacts differently on a drop in the demand than it does for an increase in the demand. The estimated MEM parameters are shown in Table 3.3. To support the range for in-sample model fit, one of the assumptions in the study is that a *t-statistic > 2* in magnitude correspond to approximately a 95% confidence level. The t-statistic column is the parameter value divided by the standard error and is normally distributed for large samples. It measures the number of standard deviations the parameter estimate is away from zero.

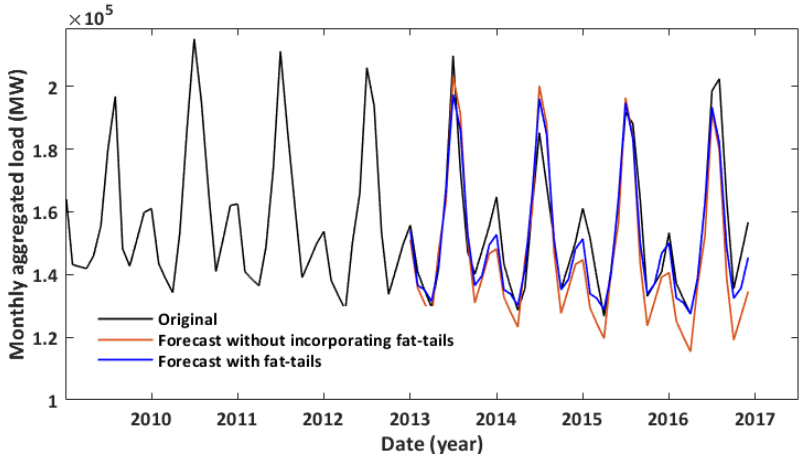


Fig. 3.23: Forecast for years 2013-2016 with and without incorporating fat-tails

Table 3.3: Multiplicative Error Model parameters

Parameter	Multiplicative error model values	Standard errors	t-statistic
α	1.368271e+07	0.000153268	7.22314e+10
α_1	0.703561	0.012491	40.1238
β_1	0.0241376	0.0201932	1.5713

3.3.4.2Directional accuracy of forecast methodology

The second part of result analysis is checking directional accuracy during the year 2008 when the great economic recession hit the whole world and the U.S. was largely affected. Since the data is from U.S. utility, it was decided to check the robustness of

the proposed model during that period. The loss functions were used to evaluate the accuracy of the proposed methodology. However, their usage does not distinguish the direction of errors. In other words, the positive forecast errors (i.e., when historical values more than forecasted values) and negative forecast errors (i.e., when historical values less than forecasted values) are counted equally in the error metrics. It can be agreed that ignoring the direction of errors simplifies the efforts of evaluating forecasting accuracy. Nevertheless, it should be noted that the directions of errors often have an economic impact in long-term forecast applications. For example, in power system planning, positive forecast errors (i.e. historical values more than forecast values) can result in planning inadequate capacity and in turn loss of load service. On the other hand, negative forecast errors (historical values less than forecast values) can result in wasting of resources by deploying more capacity than necessary. Note that economic loss corresponding to losing load (due to positive errors) is often different from that corresponding to resource wasting (due to negative errors).

An out-of-sample forecast is performed for the year 2008 with a training dataset of years 1993-2007. When economic factors play a pivotal role, the need to study market movements is important. Not many forecast studies include the significance of directional forecasting and how its accuracy supports the statistical parameters. Fig. 3.24 shows the two overlapped time series. A long period of uniform load growth was interrupted in the early 2000s till mid-2000s. In fact, the 2000s show two distinct jumps in historical load data (seen in Fig. 3.15): one was triggered by energy crisis because of fluctuating oil prices, and one was prompted by the great recession of 2008. Since then, load growth has regularly displayed volatility relative to the pre-2000s. As the real load growth has not changed much over time, still large fluctuations tend to be concentrated over somewhat short periods, thus embodying directional accuracy along with improved and accurate forecast result is preferred.

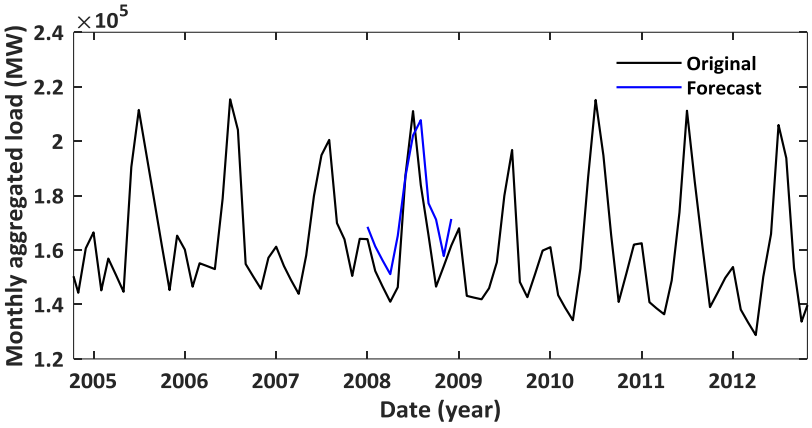


Fig. 3.24: Forecast during the great recession of 2008

3.3.5 DISCUSSIONS

This chapter proposes a novel implementation of MEM to forecast electricity load in the long-term horizon. The proposed forecasting model aims at accurate forecasting of the monthly aggregated load for a horizon of 4 years (or 48 months) and the error metrics used in validating the model are mean squared error and mean average percentage error. To account for future volatility, the proposed MEM is built upon recession data accessed from NBER. MEM is able to model the volatility in time series by accounting for conditional variance. The term conditional variance in MEM denotes the dependency on past sequence of events and is quite contrasting to unconditional which implies long-term behavior assuming null knowledge on past events. Moreover, the discovered non-linearity could be well handled using MEM. The two performance indicators for this research are: point forecast with a low error percentage as proved by mean error metrics for both in-sample and out-of-sample forecasts, and the directional accuracy during the great recession. The inclusion of heteroscedastic errors improves forecast performance and also shows that it is possible to predict the direction of change of residuals in the presence of conditional heteroscedasticity, even if the residuals themselves cannot be predicted.

3.4 CONCLUSIONS

Till date, research has focused on univariate distributions to model load or wind power, and they are univariate because they show the probability density function of only one variable. The choice for univariate models is credited with features like easiness to ensure robustness. In addition, they are considered to be sufficient for short lead times because weather variables like wind and irradiation that affect wind and solar forecast tend to change in a smooth fashion over short time frames. The same was witnessed for load forecasting in the short-term horizon using a feedforward neural network architecture. And to describe the forecast error distribution, section 3.2 proposed truncated normal distribution as a suitable option. However, a shortcoming of using univariate distribution, keeping future uncertainties in mind, is the fact that one cannot catch the dependencies and correlations when the exogenous variables evolve in both space and time. Hence, a multivariate modeling approach is required. In simplest words, a multivariate model is able to describe the multiple variables (or distributions) which can be regarded as a set of univariate distributions. In such cases, univariate distribution is called marginal distribution while the multivariate distribution is named as joint probability distribution because it reveals the joint probability density function. In the next chapter, marginal distributions comprising of load and wind power are modeled as joint probability distribution using vine copula.

It was learned that distributions of load are usually fat-tailed and skewed and the same follows for wind power which will be discussed in the next chapter. In building a multivariate dataset of load and wind power, it can be well understood that the marginal distributions do not conform to normality assumption and the dependencies between the variables are misleading too. Thus, the normal assumption of correlation coefficient to calculate the dependence between two variables is challenged. Moreover, learning dependencies over correlation is important when modeling multiple variables that evolve in both space and time. Few limitations of correlation coefficient are identified as:

i. The correlation coefficient measures the linear association between random variables. It fails to capture the non-linear dependencies that exist between load and wind power.
ii. As correlation is described as a scaled covariance, the correlation coefficient is always influenced by the distribution of marginals.
iii. Correlation expresses the dependence numerically as a number rather than describing it as some function to illustrate the non-linear association.
iv. Lastly, a correlation coefficient of one does not indicate a positive dependence and a zero correlation does not indicate independence.

Based on the limitations, the next chapter introduces the spatio-temporal dependence modeling of load and wind power as a multivariate model using vine copula.

REFERENCES

[Abdi & Beigvand, 2016] Abdi, H., & Beigvand, S. D. (2016). Long-term load forecasting based on gravitational search algorithm. *Journal of Intelligent & Fuzzy Systems*, *30*(6), 3633-3643.

[Abu-El-Magd & Sinha, 1982] Abu-El-Magd, M. A., & Sinha, N. K. (1982). Short-term load demand modeling and forecasting: a review. *IEEE transactions on systems, man, and cybernetics*, *12*(3), 370-382.

[Adepoju et al., 2007] Adepoju, G. A., Ogunjuyigbe, S. O. A., & Alawode, K. O. (2007). Application of neural network to load forecasting in Nigerian electrical power system. *The Pacific Journal of Science and Technology*, *8*(1), 68-72.

[Ahoniemi, 2006] Ahoniemi, K. (2006). Modeling and forecasting implied volatility-an econometric analysis of the VIX index. *HECER Discussion Paper, 129*.

[Alfares & Nazeeruddin, 2002] Alfares, H. K., & Nazeeruddin, M. (2002). Electric load forecasting: literature survey and classification of methods. *International Journal of Systems Science*, *33*(1), 23-34.

[AlRashidi & El-Naggar, 2010] AlRashidi, M. R., & El-Naggar, K. M. (2010). Long term electric load forecasting based on particle swarm optimization. *Applied Energy*, *87*(1), 320-326.

[Angelopoulos Angelopoulos, D., Psarras, J., & Siskos, Y. (2017, June). Long-term electricity

et al., 2017] demand forecasting via ordinal regression analysis: The case of Greece. In *PowerTech, 2017 IEEE Manchester* (pp. 1-6). IEEE.

[Armstrong and Collopy, 1992] Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International journal of forecasting*, *8*(1), 69-80.

[Auer, 2016] Auer, B. R. (2016). How does Germany's green energy policy affect electricity market volatility? An application of conditional autoregressive range models. *Energy Policy*, *98*, 621-628.

[Baliyan et al., 2015] Baliyan, A., Gaurav, K., & Mishra, S. K. (2015). A review of short term load forecasting using artificial neural network models. *Procedia Computer Science*, *48*, 121-125.

[Benini et al., 2002] Benini, M., Marracci, M., Pelacchi, P., & Venturini, A. (2002, July). Day-ahead market price volatility analysis in deregulated electricity markets. In *Power engineering Society Summer Meeting, 2002 IEEE* (Vol. 3, pp. 1354-1359). IEEE.

[Boland et al., 2016] Boland, J., Filar, J. A., Mohammadian, G., & Nazari, A. (2016). Australian electricity market and price volatility. *Annals of Operations Research*, *241*(1-2), 357-372.

[Bollerslev, 1986] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, *31*(3), 307-327.

[Box et al., 2008] Box, G.E., Jenkins, G.M., Reinsel, G.C.: 'Time series analysis: forecasting and control' (Wiley, Hoboken, 2008)

[Brooks & Oozeer, 2002] Brooks, C., & Oozeer, M. C. (2002). Modelling the implied volatility of options on long gilt futures. *Journal of Business Finance & Accounting*, *29*(1-2), 111-137.

[Bunnoon et al., 2010] Bunnoon, P., Chalermyanont, K., & Limsakul, C. (2010). A Computing Model of Artificial Intelligent Approaches to Mid-term Load Forecasting: a state-of-the-art-survey for the researcher. *International Journal of Engineering and Technology*, *2*(1), 94.

[Caporin et al., 2017] Caporin, M., Rossi, E., & de Magistris, P. S. (2017). Chasing volatility: A persistent multiplicative error model with jumps. *Journal of Econometrics*, *198*(1), 122-145.

[Chatfield, 2001] Chatfield, C. (2001). *Time-series forecasting*. Chapman & Hall, New York.

[Chen et al., 2005] Chen, H., Du, Y., & Jiang, J. N. (2005, June). Weather sensitive short-term load forecasting using knowledge-based ARX models. In *Power Engineering Society General Meeting, 2005. IEEE* (pp. 190-196). IEEE.

[Chen, 2012] Chen, T. (2012). A collaborative fuzzy-neural approach for long-term load forecasting in Taiwan. *Computers & Industrial Engineering*, *63*(3), 663-670.

[De Felice & Yao, 2011] De Felice, M., & Yao, X. (2011). Short-term load forecasting with neural network ensembles: A comparative study [application notes]. *IEEE Computational Intelligence Magazine*, *6*(3), 47-56.

[Ekonomou, 2010] Ekonomou, L. (2010). Greek long-term energy consumption prediction using artificial neural networks. *Energy*, *35*(2), 512-517.

[El-Ela et al., El-Ela, A. A., El-Zeftawy, A. A., Allam, S. M., & Atta, G. M. (2009). Long-term

2009] load forecasting and economical operation of wind farms for Egyptian electrical network. *Electric Power Systems Research*, *79*(7), 1032-1037.

[Dilaver & Hunt, 2011] Dilaver, Z., & Hunt, L. C. (2011). Turkish aggregate electricity demand: an outlook to 2020. *Energy*, *36*(11), 6686-6696.

[Elder & Kennedy, 2001] Elder, J., & Kennedy, P. E. (2001). Testing for unit roots: what should students be taught?. *The Journal of Economic Education*, *32*(2), 137-146.

[Engle & Russell, 1998] Engle, R. F., & Russell, J. R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, 1127-1162.

[Engle, 2002] Engle, R. (2002). New frontiers for ARCH models. *Journal of Applied Econometrics*, *17*(5), 425-446.

[Feinberg & Genethliou, 2005] Feinberg, E. A., & Genethliou, D. (2005). Load forecasting. *Applied mathematics for restructured electric power systems*, 269-285.

[Ghiassi et al., 2006] Ghiassi, M. D. K. Z., Zimbra, D. K., & Saidane, H. (2006). Medium term system load forecasting with a dynamic artificial neural network model. *Electric Power Systems Research*, *76*(5), 302-316.

[Ghods & Kalantar, 2011] Ghods, L., & Kalantar, M. (2011). Different methods of long-term electric load demand forecasting; a comprehensive review. *Iranian Journal of Electrical & Electronic Engineering*, *7*(4), 249-259.

[Gross & Galiana, 1987] Gross, G., & Galiana, F. D. (1987). Short-term load forecasting. *Proceedings of the IEEE*, *75*(12), 1558-1573.

[Hamzacebi & Es, 2014] Hamzacebi, C., & Es, H. A. (2014). Forecasting the annual electricity consumption of Turkey using an optimized grey model. *Energy*, *70*, 165-171.

[Han et al., 2015] Han, H., Park, M. D., & Zhang, S. (2015). A multiplicative error model with heterogeneous components for forecasting realized volatility. *Journal of Forecasting*, *34*(3), 209-219.

[Hansen & Timmermann, 2012] Hansen, P. R., & Timmermann, A. (2012). Choice of sample split in out-of-sample forecast evaluation. Working paper, Stanford University.

[Hippert et al., 2001] Hippert, H. S., Pedreira, C. E., & Souza, R. C. (2001). Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on power systems*, *16*(1), 44-55.

[Hong, 2009] Hong, W. C. (2009). Electric load forecasting by support vector model. *Applied Mathematical Modelling*, *33*(5), 2444-2454.

[Hong, 2014] Hong, T. (2014). Energy forecasting: Past, present, and future. *Foresight: The International Journal of Applied Forecasting*, (32), 43-48.

[Hritonenko & Yatsenko, 2012] Hritonenko, N., & Yatsenko, Y. (2012). Energy substitutability and modernization of energy-consuming technologies. *Energy Economics*, *34*(5), 1548-1556.

[Jarque & Bera, 1987] Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, 163-172.

[Jianjun et al., 2016] Jianjun, W., Li, L., & Ding, L. (2016). Application of SVR with backtracking search algorithm for long-term load forecasting. *Journal of Intelligent & Fuzzy*

Systems, *31*(4), 2341-2347.

[Johnson Jr et al., 1987] Johnson Jr, A. C., Johnson, M. B., & Buse, R. C. (1987). *Econometrics: Basic and applied* (p. 90). New York.

[Kaboli et al., 2017] Kaboli, S. H. A., Fallahpour, A., Selvaraj, J., & Rahim, N. A. (2017). Long-term electrical energy consumption formulating and forecasting via optimized gene expression programming. *Energy*, *126*, 144-164.

[Kandil et al., 2001] Kandil, M. S., El-Debeiky, S. M., & Hasanien, N. E. (2001). The implementation of long-term forecasting strategies using a knowledge-based expert system: part-II. *Electric Power Systems Research*, *58*(1), 19-25.

[Karabulut et al., 2008] Karabulut, K., Alkan, A., & Yilmaz, A. S. (2008). Long term energy consumption forecasting using genetic programming. *Mathematical and Computational Applications*, *13*(2), 71-80.

[Khuntia *et al*., 2016c] Khuntia, S. R., Rueda, J. L., & van der Meijden, M. A. (2016, July). Neural network-based load forecasting and error implication for short-term horizon. In *Neural Networks (IJCNN), 2016 International Joint Conference on* (pp. 4970-4975). IEEE.

[Khuntia *et al*., 2016d] Khuntia, S. R., Rueda, J. L., & van der Meijden, M. A. M. M. (2016, October). Volatility in electrical load forecasting for long-term horizon—An ARIMA-GARCH approach. In *Probabilistic Methods Applied to Power Systems (PMAPS), 2016 International Conference on* (pp. 1-6). IEEE.

[Khuntia *et al*., 2016e] Khuntia, S. R., Rueda, J. L., & van der Meijden, M. A. (2016). Forecasting the load of electrical power systems in mid-and long-term horizons: a review. *IET Generation, Transmission & Distribution*, *10*(16), 3971-3977.

[Khuntia et al., 2018d] Khuntia, S. R., Rueda, J. L., & van der Meijden, M. A. (2018). Long-term electricity load forecasting considering volatility using multiplicative error model. *Forecasting*, Accepted.

[Lanne, 2006] Lanne, M. (2006). A mixture multiplicative error model for realized volatility. *Journal of Financial Econometrics*, *4*(4), 594-616.

[Leahy & Foley, 2012] Leahy, P. G., & Foley, A. M. (2012). Wind generation output during cold weather-driven electricity demand peaks in Ireland. *Energy*, *39*(1), 48-53.

[Li & Flynn, 2004] Li, Y., & Flynn, P. C. (2004). Deregulated power prices: comparison of volatility. *Energy Policy*, *32*(14), 1591-1601.

[Li et al., 2013] Li, H. Z., Guo, S., Li, C. J., & Sun, J. Q. (2013). A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm. *Knowledge-Based Systems*, *37*, 378-387.

[Ljung & Box, 1978] Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297-303.

[Lu et al., 2013] Lu, N., Diao, R., Hafen, R. P., Samaan, N., & Makarov, Y. V. (2013, July). A comparison of forecast error generators for modeling wind and load uncertainty. In *Power and Energy Society General Meeting (PES), 2013 IEEE* (pp. 1-5). IEEE.

[MATLAB, 2017] MATLAB. Natick, Massachusetts: The MathWorks Inc., 2017.

[Makarov et al., 2010] Makarov, Y. V., Guttromson, R. T., Huang, Z., Subbarao, K., Etingov, P. V., Chakrabarti, B. B., & Ma, J. (2010). Incorporating wind generation and load

forecast uncertainties into power grid operations. *Report PNNL-19189, PNNL*.

[Makridakis et al., 2008] Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). *Forecasting methods and applications*. John wiley & sons, New York.

[Meng & Niu, 2011] Meng, M., & Niu, D. (2011). Annual electricity consumption analysis and forecasting of China based on few observations methods. *Energy Conversion and Management*, *52*(2), 953-957.

[Moghram & Rahman, 1989] Moghram, I., & Rahman, S. (1989). Analysis and evaluation of five short-term load forecasting techniques. *IEEE Transactions on power systems*, *4*(4), 1484-1491.

[Mohamed & Bodger, 2005] Mohamed, Z., & Bodger, P. (2005). Forecasting electricity consumption in New Zealand using economic and demographic variables. *Energy*, *30*(10), 1833-1843.

[Moral-Carcedo & Pérez-García, 2017] Moral-Carcedo, J., & Pérez-García, J. (2017). Integrating long-term economic scenarios into peak load forecasting: An application to Spain. *Energy*, *140*, 682-695.

[Murphy, 2012] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

[Nezzar et al., 2016] Nezzar, R. M., Farah, N., Khadir, M. T., & Chouireb, L. (2016). Mid-Long Term Load Forecasting using Multi-Model Artificial Neural Networks. *International Journal on Electrical Engineering and Informatics*, *8*(2), 389.

[Padmakumari et al., 1999] Padmakumari, K., Mohandas, K. P., & Thiruvengadam, S. (1999). Long term distribution demand forecasting using neuro fuzzy computations. *International Journal of Electrical Power & Energy Systems*, *21*(5), 315-322.

[Pai & Hong, 2005] Pai, P. F., & Hong, W. C. (2005). Support vector machines with simulated annealing algorithms in electricity load forecasting. *Energy Conversion and Management*, *46*(17), 2669-2688.

[Peng et al., 1992] Peng, T. M., Hubele, N. F., & Karady, G. G. (1992). Advancement in the application of neural networks for short-term load forecasting. *IEEE Transactions on Power Systems*, *7*(1), 250-257.

[Ranaweera et al., 1997] Ranaweera, D. K., Karady, G. G., & Farmer, R. G. (1997). Economic impact analysis of load forecasting. *IEEE Transactions on Power Systems*, *12*(3), 1388-1392.

[Rui & El-Keib, 1995] Rui, Y., & El-Keib, A. A. (1995, March). A review of ANN-based short-term load forecasting models. In *System Theory, 1995., Proceedings of the Twenty-Seventh Southeastern Symposium on* (pp. 78-82). IEEE.

[Saab et al., 2001] Saab, S., Badr, E., & Nasr, G. (2001). Univariate modeling and forecasting of energy consumption: the case of electricity in Lebanon. *Energy*, *26*(1), 1-14.

[Saini & Soni, 2002] Saini, L. M., & Soni, M. K. (2002). Artificial neural network based peak load forecasting using Levenberg–Marquardt and quasi-Newton methods. *IEE Proceedings-Generation, Transmission and Distribution*, *149*(5), 578-584.

[Shumway & Stoffer, 2011] Shumway, R. H., & Stoffer, D. S. (2011). Time series regression and exploratory data analysis. In *Time series analysis and its applications* (pp. 47-82). Springer New York.

[Simonsen, Simonsen, I. (2005). Volatility of power markets. *Physica A: Statistical*

2005] *Mechanics and its Applications*, *355*(1), 10-20.

[Srinivasan & Lee, 1995] Srinivasan, D., & Lee, M. A. (1995, October). Survey of hybrid fuzzy neural approaches to electric load forecasting. In *Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century., IEEE International Conference on* (Vol. 5, pp. 4004-4008). IEEE.

[Sun, 2001] Sun, J. (2001). Energy demand in the fifteen European Union countries by 2010— A forecasting model based on the decomposition approach. *Energy*, *26*(6), 549-560.

[Taleb, 2007] Taleb, N. N. (2007). Black swans and the domains of statistics. *The American Statistician*, *61*(3), 198-200.

[Taylor & McSharry, 2007] Taylor, J. W., & McSharry, P. E. (2007). Short-term load forecasting methods: An evaluation based on european data. *IEEE Transactions on Power Systems*, *22*(4), 2213-2219.

[Troccoli, 2009] Troccoli, A. (Ed.). (2009). *Management of weather and climate risk in the energy industry*. Springer.

[Ünler, 2008] Ünler, A. (2008). Improvement of energy demand forecasts using swarm intelligence: The case of Turkey with projections to 2025. *Energy Policy*, *36*(6), 1937-1944.

[Web1, 2009] https://www.iea.org/publications/freepublications/publication/impact.pdf

[Web3, 2018] https://www.thebalance.com/volatility-definition-and-types-3305968

[Web4, 2018] https://www.pjm.com/markets-and-operations/ops-analysis/

[Web5, 2018] http://www.nber.org/cycles/cyclesmain.html

[Web7, 2016] http://iso-ne.com/markets/hstdata/znl_info/hourly/index.html

[Weron, 2007] Weron, R. (2007). *Modeling and forecasting electricity loads and prices: A statistical approach* (Vol. 403). John Wiley & Sons.

[Willis & Northcote-Green, 1983] Willis, H. L., & Northcote-Green, J. E. (1983). Spatial electric load forecasting: a tutorial review. *Proceedings of the IEEE*, *71*(2), 232-253.

[Willis, 1983] Willis, H. L. (1983). Load forecasting for distribution planning-error and impact on design. *IEEE Transactions on Power Apparatus and Systems*, (3), 675-686.

[Xia et al., 2010] Xia, C., Wang, J., & McMenemy, K. (2010). Short, medium and long term load forecasting model and virtual load forecaster based on radial basis function neural networks. *International Journal of Electrical Power & Energy Systems*, *32*(7), 743-750.

[Xie et al., 2011] Xie, L., Gu, Y., Zhu, X., & Genton, M. G. (2011, May). Power system economic dispatch with spatio-temporal wind forecasts. In *Energytech, 2011 IEEE* (pp. 1-6). IEEE.

[Yoo & Pimmel, 1999] Yoo, H., & Pimmel, R. L. (1999). Short term load forecasting using a self-supervised adaptive neural network. *IEEE transactions on Power Systems*, *14*(2), 779-784.

[Yumurtaci & Asmaz, 2004] Yumurtaci, Z., & Asmaz, E. (2004). Electric energy demand of Turkey for the year 2050. *Energy Sources*, *26*(12), 1157-1164.

[Zareipour et al., 2007] Zareipour, H., Bhattacharya, K., & Cañizares, C. A. (2007). Electricity market price volatility: The case of Ontario. *Energy Policy*, *35*(9), 4739-4748.

CHAPTER 4

SPATIO-TEMPORAL MODELING OF LOAD AND WIND POWER

4.1 INTRODUCTION

This chapter aims at answering the third research question Q.3., which deals with spatio-temporal modeling of load and wind power as a joint probability distribution for the short-term horizon,

- *How should load variability and wind power generation for spatially distributed locations in a large-scale system be modeled?*
- *How can both spatial as well as temporal correlations be effectively addressed?*
- *How can high dimensional data be accounted for when the future will be data-centric?*

The content of this chapter is based on research papers [Khuntia et al., 2018a, Khuntia et al., 2018b]. Spatial distribution of WPPs and load centers make it plausible to study the inter-spatial dependence and temporal correlation for the effective working of the power system. Both load and wind power, are characterized by uncertainty and variability that are also identified as two future bottlenecks. In this chapter, the concept of vine copula is introduced to model spatio-temporal dependence efficiently. Use of vine copula facilitates building multi-dimensional copulas out of bivariate copulas as they are easy to estimate and are well understood. Hourly resolution of load and wind power data obtained from a U.S. regional transmission operator spanning three years and spatially distributed in nineteen load and two wind power zones are considered in this research. Data collection, in terms of dimension, tend to increase in future and to tackle this high dimensional data, a reproducible sampling algorithm using vine copula is developed. The sampling algorithm employs k-means, Gaussian Mixture Model (GMM) and hierarchical linkage clustering techniques along with singular value decomposition technique to ease the computational burden. Selection of appropriate clustering technique and copula family is realized by the Goodness of Clustering (GoC) and Goodness of Fit (GoF) tests. The chapter concludes with a discussion on the importance

of spatio-temporal modeling of load and wind power and the advantage of the proposed sampling algorithm using vine copula.

The rest of this chapter is organized as follows: sub-section 4.2 presents a background on spatio-temporal modeling of load and wind power till date and why there is a need for a joint probability distribution. Sub-section 4.3 introduces copulas and vine copulas for spatio-temporal modeling. Sub-section 4.4 presents the modeling framework using vine copulas and explains each step lucidly with simulation results. Finally, sub-section 4.5 concludes the chapter.

4.2 BACKGROUND

The future sees tremendous renewable energy in-feed to the existing electric grid network at both transmission and distribution level. This is accompanied by uncertainty in load growth for which TSOs have to prepare the system in a secure and adequate manner where the key performance indicator is the security of supply. In accordance to the Paris Agreement 2016 [Web2, 2016], the move from conventional energy (fossil fuel and nuclear power plants) towards renewable energy is adopted globally. Trying to tap more of renewable energy into the existing grid is however challenging, pertaining to its irregular availability, variability, and link to varying atmospheric factors. For instance, to increase the utilization of renewable energy, it is understood that investments in wind farms are concentrated at locations with higher average wind speeds and solar farms at locations with higher average solar insolation. TSOs have to cautiously evaluate operation as well as future planning when power output fluctuations occur for such spatially distributed systems. An accurate knowledge of time and spatial features is then beneficial to model the behavior of the power system under different RES penetration.

Wind generation is driven by wind patterns, which tend to follow certain geographical spatial correlations. Till date, modeling of wind power focused on a single wind farm (or aggregation of wind power in single WPP). In this manner, they do not account for potential information from neighboring sites, for example, other WPPs or meteorological stations. Spatial modeling is vital when wind power error in a WPP might propagate to WPPs in other locations during the following period when they are affected by same meteorological factors. As a massive integration of wind power is witnessed in Europe and U.S., considering inter-spatial dependence along with temporal correlation is important for wind power modeling. Load modeling follows the same path and to understand we consider aggregated load forecast of a load zone. For load modeling with historical data for a specific temporal scale, it is likely to obtain the forecasts between the maximum and minimum values of the training set. Between the two extrema, there might be some hot or cold days which went unnoticed in the training set. But, if we include the historical data from the closest load zone, there is a high possibility that the hot and cold days are taken into consideration. This is because

electricity load is affected by weather parameters and, hence, the energy usage of adjacent load zone is also taken into the new training set.

It is evident that incorporating inter-spatial dependence can help improve the modeling accuracy and thus the trend has been towards exploiting all of the available data in modeling. The first step is to obtain a tractable model that captures the uncertainties and correlations among the exogenous variables.

Before we dive into details, it is to be noted that,

- This research is performed at the transmission level and assumed that WPPs are connected to the transmission system. Hence, aggregated zonal load and wind power are our exogenous variables.
- The spatio-temporal model presented in this research is aimed at short-term power system application such as the static security assessment.
- It is intended that solar power and other DERs are concentrated at the distribution level and, thus, excluded in this research.
- Interaction of transmission and distribution system operators is excluded from this research.
- The terms dependence and correlation are used frequently. Correlation is a natural dependence measure for a multivariate dataset. Specifically, in this research, dependence refers to a non-linear association that exists between load and wind power as both spatial and temporal aspect is considered. However, correlation is used to describe the dependence and for further reading, reference [Embrechts et al., 2002] is recommended.

4.2.1 Brief Background on Spatio-Temporal Study of Load

The spatio-temporal features of electricity load can be explained by two underlying spatio-temporal processes, namely weather and human activities [Shi et al., 2017]. The weather of adjacent neighborhoods or cities tends to be more alike than those far apart. Similarly, human activities in adjacent neighborhoods tend to be highly correlated. Electricity load has a long-anticipated factor due to its very strong seasonality feature, i.e., daily, weekly, and monthly, along with weather-based variation although the weather-based variables that affect load can differ according to location. Accounting for temporal correlation is explained by relatively regular load profiles which is an outcome of aggregation of a large number of loads. This is referred to as seasonal or temporal component. Due to the fact that electricity load has seasonal or temporal component, most previous studies aimed at modeling temporal correlation while overlooking spatial correlation among load variation in different zones. [Melo et al., 2014] performed spatial load forecasting to determine spatial resolution. In a previous study, [Melo et al., 2012] study the distribution of load variability in a city and the relationships among different areas into account. Considering spatio-temporal aspect of

electricity load will result in more accurate modeling. It will be possible to realize extreme values beyond a fixed target load from the training spatio-temporal dataset which results from different load profiles in different zones. Most of spatio-temporal studies focused at distribution level where the target is residential homes and data is collected from advanced metering infrastructure (AMI) [Tascikaraoglu & Sanandaji, 2016] and without the use of AMI [Shin et al., 2011]. A dynamic spatio-temporal model was developed in [Shi et al., 2017] taking Southern California's electricity load time series. However, this research aims at presenting a viewpoint and criticality of spatio-temporal modeling from the transmission level.

4.2.2 BRIEF BACKGROUND ON SPATIO-TEMPORAL STUDY OF WIND POWER

Stochasticity of wind makes it difficult to predict accurate wind power output by only considering temporal wind behavior when it is affected by other geographical and technical factors like wind farm topology and wind turbine characteristics [Lenzi et al., 2017]. As such, it is a common belief among researchers that for the spatial pattern, the dependence is relatively stronger for elements that are closer to each other [Osborn et al., 2011]. The stochastic and variable nature of wind power has its implications on power system reliability, spinning reserve, and to some extent on operating cost. Literature study reveals that inter-spatial dependence and temporal correlation was studied separately until recently [Louie, 2014a, Haghi & Lotfifard, 2015, Malvaldi et al., 2017, Wei et al., 2017]. A spatial study of wind power in different zones in UK was studied in [Miranda & Dunn, 2007] using a multivariate regression model. The study showed a multivariate time series model for real wind speed data from multiple sites is complicated in nature, particularly the presence of a large number of wind sites. Reference [Maisonneuve & Gross, 2011] proposed a wind regime model for planning studies. The study aimed at modeling both seasonal and diurnal trends of wind power and its correlation to the same trends of electricity load. In the temporal aspect, transformation and standardization of non-Gaussian and non-stationary characteristics of wind power are studied in [Brown et al., 1984Torres et al., 2005] by application of regression models. Use of copula for spatio-temporal scenario studies is reported in [Tastu et al., 2013] and more discussion is followed in section 4.3. Reference [Papavasiliou et al., 2015] address spatial correlation for wind power modeling using a noise vector based regression model in an attempt where a single multivariate time series model is decoupled into distinct univariate time series models. In a large clustered WPP, spatio-temporal correlations of wind have significant impacts on the overall uncertainty of wind power outputs [Wan et al., 2003]. Such erratic characteristic has significant impact on power system planning and operation [Miettinen & Holttinen, 2017].

4.2.3 BRIEF BACKGROUND ON SPATIO-TEMPORAL STUDY OF LOAD AND WIND POWER

It was learned that load data exhibits strong seasonality that favored temporal correlation studies. In reality, an easy and effective load prediction for tomorrow can be realized by using historical load data from the same day of week or month. This is based on studying large autocorrelations between the same days of the week or month. Such a dependence characteristic is also observed in spatial data. A similar spatial pattern among wind power data favored spatial correlation studies for wind power. Not to be forgotten that meaningful correlation exists between load and wind power because both are significantly affected by weather. A suitable spatio-temporal multivariate model should capture both inter-spatial dependence as well as temporal correlation embedded in the multivariate dataset. Capturing the inherent dependence between load and wind power in different temporal and spatial context is achieved by adopting a multivariate modeling approach. There is an immediate need for the development of spatio-temporal modeling of load and wind power as joint normal distribution for three reasons. *Firstly*, inter-spatial dependence and temporal correlation of load and wind power in any considered site are important. From sub-sections 4.2.1 and 4.2.2, the literature study revealed consideration of load and wind power as independent variables and also some instances of temporal or spatial correlation. However, there are no significant findings that investigate the spatio-temporal dependence of these two exogenous variables. *Secondly*, a suitable spatio-temporal modeling approach will facilitate improving both short-term operational planning and long-term grid development of power grids. For instance, in short-term operational planning, an accurate spatio-temporal modeling can help the TSO in assessing system security in terms of asset overloading or reducing operational costs by using forecast values for unit commitment or reducing wind curtailment. Similarly, in terms of long-term planning, accurate modeling can result in assessing grid development plans to answer the load growth or massive generation of wind power. For example, in this research, control area of U.S. regional transmission operator is considered (more details about the dataset follow later in section 4.4). As of 2016, 1GW of installed wind power along with other generation sources serve large load centers along the east coast and mid-western region [Web6, 2018]. Under the renewable portfolio standard, 11.3% (32GW) and 13.9% (42GW) of the total load are expected in the years 2021 and 2026 in the form of massive integration [Web7, 2018]. A map of future wind farm projects is shown in Fig. 4.1 [Web8, 2018]. *Lastly*, a suitable spatio-temporal multivariate model can generate a rich synthetic database of normally distributed load and wind power data. Such a database will be of immense help to research community and industry as well to develop other statistical tools. Examples of online tools to visualize the most promising areas where wind farms can be profitably installed are IRENA [Web10, 2018] global atlas and NREL's [Web11, 2018] wind prospector. However, the data extracted from

these wind atlases lack the clarity when they are used to learn the behavior of power system operation in shorter time-horizons (15 mins or 1 hour).

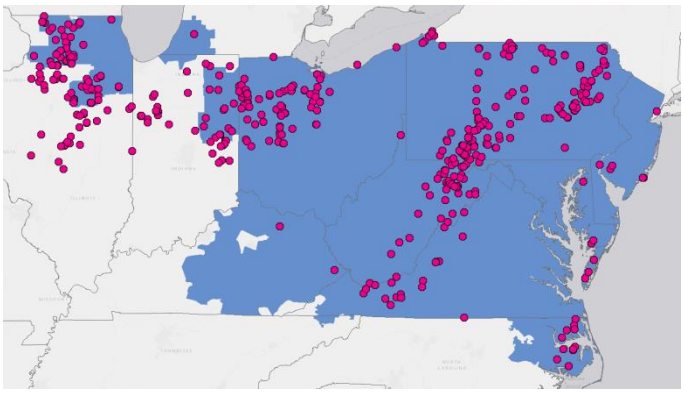


Fig. 4.1: Map of future wind farm projects under one of the US regional transmission operator (blue shade is the control area and red dots are the planned wind farms)

The key contributions of this chapter can be listed as:

- The chapter discusses the work performed till date on spatio-temporal modeling. Also, it reviews the challenges associated with multivariate modeling while serving as a short literature study for future research work.
- A detailed framework of copula and vine copula is presented. A first-hand application of vine copula to address spatio-temporal correlation is described in this work.
- It is intended that future power system will be data-centric and collecting data from stochastic sources in terms of spatial and temporal resolution will result in a high dimensional database of varied features. This huge chunk of data, referred to big data, is explored in electric power systems in terms of big data analytics. To address the need for such analytics, this chapter presents a suitable statistical approach, which is indeed required to study the dependence among the multivariate variables. As a solution, this research developed a reproducible vine copula-based sampling algorithm using clustering and feature extraction technique to tackle the high dimensional data.
- A reproducible sampling algorithm to model spatio-temporal correlation using canonical vine is developed. A spatio-temporal vine copula is employed to model a spatio-temporal dataset of size, say $n + 1$, which is composed of one central location and its n-neighbors in space and time. The first tree of the vine is realized by spatio-temporal bivariate copulas, reflecting the fact that the dependence structure changes over space and time. And the selection of

bivariate copulas is an added benefit of spatio-temporal vine copulas. Bivariate spatio-temporal copulas are a convex combination of different copula families that are parameterized by spatial and temporal distance (also called marginal distributions or conditional bivariate distributions). The remainder of the vine, i.e. the vine of the variables conditioned under the value of the central location, is modeled as a n-dimensional vine.

4.3 SPATIO-TEMPORAL MODELING USING COPULA AND VINE COPULA

A suitable spatio-temporal model should be developed such as to capture both inter-spatial dependence as well as temporal correlation embedded in the multivariate dataset. To realize such a model, the modeling framework can be divided into four major tasks:

- *Understanding spatio-temporal covariance and correlation*
- *Modeling one-dimensional marginal distributions*
- *Modeling stochastic dependence using copula*
- *And, spatio-temporal modeling using vine copula*

While the first task involves understanding the concept of spatio-temporal covariance and correlation, the rest three tasks focus on modeling. Apparently, it is important to capture these three modeling aspects. For example, if the marginal distributions remain the same, the joint probability distribution can change due to changes in the dependence structure. Obtaining the right joint distribution given the marginals is, however, a non-trivial problem, since there exist an infinite number of joint distributions with the same marginals, corresponding to an infinite number of stochastic dependence structures between the random variables, i.e., load and wind power. The next sub-sections elaborate on the four major tasks.

4.3.1 UNDERSTANDING SPATIO-TEMPORAL COVARIANCE AND CORRELATION

It is important to describe and understand the concept of spatio-temporal covariance and correlation structure for dependence modeling of multivariate time series. Covariance is a measure to indicate the extent to which two random variables change in tandem while correlation is a special case of covariance and defined as a measure to represent how strongly two random variables are related. Correlation matrix is calculated when the considered variables are not comparable. Otherwise, the input data must be normalized, and then the normalized covariance matrix can be used. However, choosing covariance over correlation is advantageous as it remains unaffected by the change in space and time. Such a characteristic is useful in spatio-temporal studies.

The spatio-temporal dependencies at different lags can be computed by the empirical spatio-temporal covariance function as explained in [Cressie & Wikle, 2015]. For any spatial lag h and time lag τ, the estimated empirical spatio-temporal covariance Cov_{emp} is written as,

$$Cov_{emp}(h;\tau) = \frac{1}{|N_s(h)|}\frac{1}{|N_t(\tau)|}\sum_{s_i,s_j\in N_s(h)}\sum_{t,r\in N_t(\tau)}\left(y_{s_i}(t)-\hat{\mu}_{s_i}\right)\left(y_{s_j}(r)-\hat{\mu}_{s_j}\right) \quad (4.1)$$

where,

$$\hat{\mu}_s = \frac{1}{T}\sum_{t=1}^{T} y_s(t)$$

$N_s(h)$ refers to the pairs of spatial locations with spatial lag h, $N_t(\tau)$ refers to the pairs of timestamp pairs with time lag τ, and $|N(\cdot)|$ refers to the cardinality of the set $N(\cdot)$. $y_s(t)$ is the seasonality differenced random variable at any two spatial locations s_i and s_j at timestamp τ and T is the total number of timestamps. In this study, Equation 4.1 is valid for time series that are stationary both in space and time. Hence, an adequate differencing operation is required to make the original time series stationary. The empirical (*also stationary*) spatio-temporal correlation ρ_{emp} at spatial lag h and time lag τ is then written as,

$$\rho_{emp}(h;\tau) = \frac{Cov_{emp}(h;\tau)}{Cov_{emp}(0;0)} \quad (4.2)$$

In equation 4.1, the spatial lag h between any two spatial locations s_i and s_j is calculated as [Cressie & Wikle, 2015],

$$h = \left\lceil \frac{d(s_i,s_j)}{MaxDist} \times M \right\rceil \quad (4.3)$$

where, $d(s_i,s_j)$ is the geographical distance between any two spatial locations s_i and s_j, $MaxDist$ is the maximum geographical distance between the two spatial locations and M is the maximum spatial lag.

4.3.2 Modeling One-Dimensional Marginal Distributions

Within the framework of multivariate distribution, the univariate distribution is called marginal distribution while the multivariate distribution is named as joint probability distribution because it reveals the joint probability density function. The one-

dimensional marginal distributions, also called the marginals, capture the stochastic behavior of the individual random variable X. The marginal cumulative density function (CDF) of X is defined as:

$$F_X(x) = P(X \leq x) \tag{4.4}$$

An important property of CDF is that the cumulative distribution function of X applied to X itself yields a uniformly distributed random variable. Mathematically, it can be written as [Kurowicka& Cooke, 2006]:

$$For\ x\epsilon[0,1]: P(F_X(X) \leq x) = P\left(X \leq F_X^{-1}(x)\right) = F_X[F_X^{-1}(x)] = x \tag{4.5}$$

It is to be noted that equation 4.5 forms the base of Monte-Carlo sampling method. And, a numerical simulation technique like Monte-Carlo approach is generally applied in order to explore and exploit the impacts of wind power uncertainty on the power grid whilst considering load. However, this creates a heavy computational burden due to the necessarily large sample size and hence it is not preferred in this research. Now, using equation 4.5, sampling a random variable X with F_X can be performed by first sampling a random realization y from a uniform random variable Y in $[0,1]$, and then applying the transformation $x = F_X^{-1}(y)$. In such case, the samples x follow the distribution F_X. It can be extended for sampling from real measured data by using the empirical CDF. The sampling method in equation 4.5 is effective for sampling any single random variable whose CDF is known. When multiple correlated random variables need to be sampled, this methodology is inadequate since it does not capture any measure of the dependence between the random variables. However, the marginal $CDFs$ are still important as will be seen in the next section when copula is introduced.

4.3.3 Modeling Stochastic Dependence Using Copula

In multivariate analysis, modeling stochastic dependence is a challenging task because the variables within the multivariate dataset do not always have standardized marginal distributions. To answer such challenge, one solution is that the dependence between multiple correlated random variables can be captured by different measures of dependence. For general multivariate random variables, Spearman's rank coefficient can be used to study the non-linear, monotonic relationship between two random variables [Embrechts et al., 2001]. Spearman's rank coefficient helps in defining the dependence structure based on rank with specific functions, called copula functions. Copulas are functions that couple the marginal distribution functions of the random variables into their joint distribution function and therefore describe the dependence structure between these random variables [Sklar, 1959]. Using copula functions, it is possible to simulate two random variables that are correlated according to rank

correlation by first simulating a copula and later transforming the obtained ranks into respective marginals.

The name copula has its origin from Latin word which means “bond“ or “couple“. Thus, copulas are functions that couple multivariate distribution functions to their one-dimensional uniform marginal distributions functions [Nelsen, 2007]. And the advantage is that the joint distribution function is built based on two independent tasks comprising the modeling of the dependence and the modeling of the marginal distribution functions. Use of copula is not new in the field of electric power systems. Literature study reveals the use of Gaussian copulas to evaluate short-term scenarios for wind power generation [Pinson & Girard, 2012, Hagspiel et al. 2012], wind power forecasting error [Wei et al., 2017], transmission network planning [Park et al., 2015] and empirical copulas for modeling the dependence structure between the wind speed and the wind power output [Gill et al., 2012].

Copulas are used to describe the dependence between random variables and they need to be both estimated as well as calibrated. By definition, two random variables X and Y with F_X, F_Y are joint by copula C if their joint distribution can be written as [Morales et al., 2008]:

$$F_{XY}(x, y) = C(F_X(x), F_Y(y)) \tag{4.6}$$

The function C is therefore defined on uniformly random variables, and the CDFs can be used to map the uniformly random variables to X and Y. A copula is a probability distribution function, which is used to join or combine several marginal distributions into a joint distribution. The foundation theory of copula is based upon Sklar’s theorem [Sklar, 1959]. Let $X = \{X_1, X_2, X_3 \dots X_n\}$ be a n-dimensional random vector with a continuous marginal CDF $\{F_1, F_2, F_3 \dots F_n\}$. The relationship between multivariate CDF H of X is written as:

$$H(X) = C\big(F_1(X_1), F_2(X_2), \dots, F_n(X_n)\big) \quad X \in R^n \tag{4.7}$$

where unique function $C: [0,1]^d \rightarrow [0,1]$ is called the copula. A function $C: [0,1]^n \rightarrow [0,1]$ is called an n-dimensional copula if it satisfies the following conditions:

- $C(u_1, \dots, u_n)$ is increasing in each component u_i.
- $C(u_1, \dots, u_{k-1}, 0, u_{k+1}, \dots, u_n) = 0$ for all $u_i \epsilon [0,1], i \neq k, k = 1, \dots, n$.
- $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$ for all $u_i \epsilon [0,1], i = 1, \dots, n$.
- For all $(a_1, \dots, a_n)$, $(b_1, \dots, b_n) \epsilon [0,1]^n$ with $a_i < b_i$,

$$\sum_{i_1=1}^{2} \dots \sum_{i_n=1}^{2} (-1)^{i_1+\dots+i_n} C\left(x_{1i_1}, \dots, x_{1i_n}\right) \geq 0 \tag{4.8}$$

where $x_{j1} = a_j$ and $x_{j2} = b_j$ for all $j \epsilon \{1, \dots, n\}$.

For a thorough understanding, it is advised to follow [Patton, 2009] which reviews the use of copulas in econometric modeling and Genest et al. [Genest et al., 2009] for an elaborate bibliometric overview of copulas. Algorithm 4.1 describes the steps to sample two random variables using copula.

Algorithm 4.1: Sampling of two correlated random variables using copula

Inputs: Two uniform independent random variables U_{r1}, U_{r2} (say load and wind power), correlation coefficient, copula function.

Outputs: Sampled distributions

Sample two uniform random variables U_{r1}, U_{r2}, and obtain the realizations u_{r1}, u_{r2}.

$u_1 = u_{r1}$: presents sampling of the rank distribution U_1

Calculate copula function $C_{12|u_1,\rho_{r12}}$, i.e., the conditional distribution of U_2 for rank correlation ρ_{r12} and given u_1.

Sample the rank distribution U_2: $u_2 = C^{-1}_{12|u_1,\rho_{r12}}(u_{r2})$, using the inverse copula function. Outcome can be either 0 or constant, depending on the value of independent sample u_{r2}

Transform the rank distributions according to the marginals: $x_1 = F_1^{-1}(u_1)$ and $x_2 = F_2^{-1}(u_2)$

End

Note that for a given rank coefficient, different copula functions can be used. A particular instance of Algorithm 4.1 is the use of a Gaussian copula. When using the Gaussian copula, the overall method in Algorithm 4.1 is called the joint normal transform [Papaefthymiou & Kurowicka, 2009]. The mean of the Gaussian copula is zero and the covariance matrix is:

$$R = \begin{pmatrix} 1 & Cov \\ Cov & 1 \end{pmatrix} \tag{4.9}$$

A property of the Gaussian copula is that the covariance Cov can be computed from the rank correlation of X and Y as follows:

$$Cov = 2sin\left(\frac{\pi}{6}\rho_r(X,Y)\right) \tag{4.10}$$

Note that Cov is not the covariance between X and Y but the covariance used in the Gaussian copula. The equation above links this covariance with the rank coefficient of X and Y. The above procedure describes the joint normal transform for two random

variables. The procedure can be generalized to *n* random variables by applying the above to all pairs of random variables. The resulting Gaussian copula will be *n*-dimensional. The application of joint normal transform can be used for any system with following data availability:

- *Marginal distributions*: The system load distribution, wind speed distributions and solar power generation at each generation node of the system.
- *Dependence structure*: The product moment correlation matrix is calculated from the rank correlation matrix, between all pairs of exogenous parameters.

For spatio-temporal modeling, we first discuss the possibilities of using copula and its limitations for multivariate correlated variables. Algorithm 4.1 works adequately when all required data is available because it is needed to calculate the correlation and then later use copula to model the stochastic dependence. In reality, not all required data is available. The reason can be anything from bad measurement devices to confidential information or just 'stochasticity'. When the problem involves high uncertainty, mutual correlations between the stochastic inputs is not possible and hence the joint normal transform is not recommended. However, in such cases, we typically try to capture the most important dependence relations and leave others unspecified.

For temporal studies, statistical dependence is revealed through correlations. The linear correlation coefficient measures the linear association in the interval $[-1,1]$ [Embrechts et al., 2001], and it is a measure of linear association which is invalid for spatio-temporal studies. Reference [Le et al., 2015] showed that for wind power, linear correlation coefficients fail to provide sufficient information on the temporal scale about the total generation whether it will be at par the threshold or not. This is because dependence relations are non-linear even when coupled with data on the marginal distributions of wind power at each wind farm sites. Therefore, being able to calculate the dependence in wind power independent of marginal distributions is of great advantage for system planners as it allows modeling wind power generation more accurately. Such bottlenecks associated with the linear correlation coefficient are answered with copulas.

A big advantage of copulas over correlation is modeling fat-tailed distributions, as it is not possible in terms of correlations because the variance of such distributions can be infinite. The fat-tailed character of marginal distributions of load time series is established and discussed in chapter 3. The stronger tail dependence, as well as skewed behavior of time series, can be easily justified with copulas. Tail dependence measures how large the distribution is when multiple time series have extremely large values. Copulas can be used to exploit this kind of distribution, which is basically the dependence between two random variables in the upper-right and lower-left quadrants of their domains [Nelsen, 2007].

Copula estimation is tricky in higher dimensions. The selection of an appropriate copula function is very important, as inappropriate selection can lead to unacceptable errors. Of all copulas, the Gaussian copula is the most commonly used copula due to its computational convenience. However, in this study, a more comprehensive approach is adopted by first testing a number of standard copulas on multivariate datasets as presented in [Louie, 2014b]. A bottleneck encountered with copulas is that they perform better for bivariate distributions and that the individual pattern of chosen random variables must be described by the same parametric family of univariate distributions. Moreover, multivariate copulas are neither good; hence, vine copulas are preferred as they allow a more flexible dependence structure.

4.3.4 SPATIO-TEMPORAL MODELING USING VINE COPULA

Vines are a representation of high dimensional copulas that are constructed from a sequence of nested bivariate copula components called 'pair-copulas'. They are flexible because any combination of bivariate copulas can be used for the pair-copulas. Vine copula models decompose a multivariate copula into a set of bivariate copulas and each bivariate copula can be described as a branch of a graph connecting two consecutive marginal distributions or their conditional bivariate distributions. The multivariate distribution is a combination of univariate marginal distributions and the distribution of the copula. This is a more practical way to represent high dimensional copula problems. Vine copulas allow to flexibly combine bivariate copulas to form multivariate copulas leading to distributions of higher dimensions, thus, allowing to build vine copulas that are aware of separating distances across space and time. To achieve this, the building blocks of vine copula are composed out of convex combinations of bivariate copulas. The weights of the convex combination as well as the copulas' parameters are defined by the distance over space and time, thus modeling spatial and temporal correlation. This explains the motivation to use vine copula for spatio-temporal modeling of non-Gaussian datasets, where the non-Gaussianity not only refers to marginal distributions but also to the dependence structure between locations.

Vine copula follows a nested tree structure with edges and nodes. By definition, a vine copula on n variables is a nested set of trees T_j where the edges of the j^{th} tree become the nodes of the $(j+1)^{st}$ tree for $j = 1, \dots, n$. In general, vine decompositions are referred to as regular vines (R-vines). A regular vine on n variables is defined as a vine in which two edges in tree j are joined by an edge in tree $j+1$ only if these edges share a common node. Each edge in the regular vine may be associated with a conditional rank correlation and a copula, and each node with a marginal distribution. All assignments of rank correlations to edges of a vine are consistent and each one of these correlations may be realized by a copula. Based on the bivariate and conditional bivariate distributions, the joint distribution can be constructed. Use of vine copulas to

tackle power system uncertainty is reported in [Sun et al., 2016, Sun et al., 2017, Wang et al., 2018a] and probabilistic forecast for multiple wind farms in [Wang et al., 2018b].

A regular vine can be decomposed to either,

i. D (drawable) – vine where each node in T_j has a degree of at most 2 and conditioning is done sequentially, or;

ii. C (canonical) – vine in which each tree T_j has a unique node of degree $n - i$ where the first variable is used as a conditioning variable for the following ones.

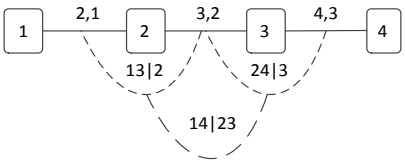


Fig. 4.2: D-vine on four variables

In D-vine, conditioning is performed sequentially whereas in C-vine the first variable is used as conditioning variable for the following ones. Fig. 4.2 shows the D-vine on four uniform variables labeled X_1, X_2, X_3, X_4. Distributions specified by conditional rank correlation on a D-vine can be sampled, and an algorithm to execute it is presented in Algorithm 4.2, which can be expanded from 4 to n variables. The algorithm involves sampling four independent uniform $[0,1]$ variables U_1, U_2, U_3, U_4. The conditional correlation between variables (i, j) given k is given as $\rho_{r_{ij|k}}$. The CDF for X_j given U_i under the conditional copula with correlation $r_{ij|k}$ is given as $F_{\rho_{rij|k;U_i}}(X_j)$.

Algorithm 4.2: Sampling of 4 uniform random variables using D-vine

Inputs: Four uniform random variables $\{X_1, X_2, X_3, X_4\}$, conditional rank correlations and corresponding CDFs ($\{F_1, F_2, F_3, F_4\}$)

Outputs: Sampled distributions

$x_1 = u_1$.

$x_2 = F^{-1}_{\rho_{r12;x_1}}(u_2)$.

$x_3 = F^{-1}_{\rho_{r23;x_2}}(F^{-1}_{\rho_{r13|2;F_{\rho_{r12;x_2}}(x_1)}}(u_3))$.

$x_4 = F^{-1}_{\rho_{r34;x_3}}(F^{-1}_{\rho_{r24|3;F_{\rho_{r23;x_3}}(x_2)}}(F^{-1}_{\rho_{r14|23;F_{\rho_{r13|2;F_{23;x_2}(x_3)}}(F_{\rho_{r12;x_2}}(x_1))}}(u_4)))$

End

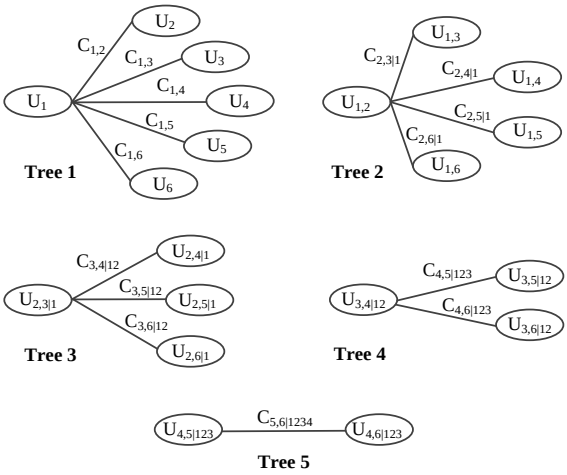


Fig. 4.3: C-vine of CDF ($u \in [0,1]$) with five trees

Fig. 4.3 illustrates a graphical representation to represent the joint density decomposition using C-vine. In general, a C-vine copula selects a root node in each tree, and all pair-wise copulas connecting with this node are modeled and conditioned on all of the previous root nodes. The nodes of Tree 1 correspond to marginal density functions while each edge corresponds to pair-copula density given as $C_{2,3|1}$ in Tree 2. The notation means that the copula model between variable 2 and 3 is conditional on 1. The full density of this C-vine copula is given by:

$$
\begin{aligned}
&C(U_1, \dots, U_6) = C_{5,6|1234}\left(U_{4,5|123}, U_{4,6|123}\right) \\
&. C_{4,5|123}\left(U_{3,4|12}, U_{3,5|12}\right). C_{4,6|123}\left(U_{3,4|12}, U_{3,6|12}\right) \\
&. C_{3,4|12}\left(U_{2,3|1}, U_{2,4|1}\right). C_{3,5|12}\left(U_{2,3|1}, U_{2,5|1}\right). C_{3,6|12}\left(U_{2,3|1}, U_{26|1}\right) \\
&. \dots \\
&\dots \\
&\dots \\
&\quad C_{1,2}(U_1, U_2). C_{1,3}(U_1, U_3). C_{1,4}(U_1, U_4). C_{1,5}(U_1, U_5). C_{1,6}(U_1, U_6)
\end{aligned}
\tag{4.11}
$$

The conditioned variables $U_{i|v}$, $i \in \{2, . . ,6\}$ and $v \in \{\{1\}, \{1, \dots, 4\}\}$, are derived through the copulas in the preceding tree (e.g. from tree 1):

$$
u_{i|1} \coloneqq F_{i|1}(u_i|u_1) = \frac{\partial C_{1i}(u_1, u_i)}{\partial u_1} \text{ at } u_1, \quad i \in \{2, \dots, 6\} \tag{4.12}
$$

One of the bottlenecks often encountered is modeling dependence structure of variables with the mixed type of dependencies (e.g., tail dependencies, asymmetries). However, adopting such a modeling framework facilitates in addressing the bottleneck by modeling each pair-copula as different parametric copula function to model the complex dependence structure.

To understand the spatio-temporal modeling using vine copula model, a spatio-temporal random field H is considered such that

$$H: S \text{ x } T \text{ x } \Phi \longrightarrow \mathbb{R} \tag{4.13}$$

where S corresponds to the spatial domain, T corresponds to the temporal domain and both with an underlying probability space Φ. For a section of the spatio-temporal random field defined as $H = (h(s_0, t_0), h(s_1, t_1) \dots h(s_n, t_n))$ of size $n + 1$, the section consists of one pivotal location and its n-neighbors in distinct spatio-temporal locations $(s_0, t_0), (s_1, t_1) \dots (s_n, t_n) \in S \text{ x } T$. Normally some spatial locations would be sampled at multiple time instances. And as the dependence structure changes over space and time, the first tree of the vine is realized by spatio-temporal bivariate copulas. The rest of the vine, i.e. the vine of the variables conditioned under the value of the central location, is modeled as some n-dimensional C-vine. To understand the functional capability of C-vine, Fig. 4.4 shows an example of spatio-temporal n-dimensional C-vine copula. The temporal extension of the spatial copula at different time lags for 3 spatial locations with Euclidean distance defined as $h_E \coloneqq \left\| s_i - s_j \right\|$, $s_i \; \forall \; i, j \in \{0,1,2,3\}$ & $t_C = 1 \dots n$. This should not be confused with spatial lag h explained in sub-section 4.3.1. Every curved connection is modeled by the same spatio-temporal copula C_{h_E, t_C} but with different spatial and temporal distances, h_E and t_C deduced from the indicated spatio-temporal locations. It is already assumed that marginals are stationary and combining them with multivariate copula results in a multivariate distribution of the spatio-temporal random field. And this multivariate distribution is later used for application studies like simulation or prediction.

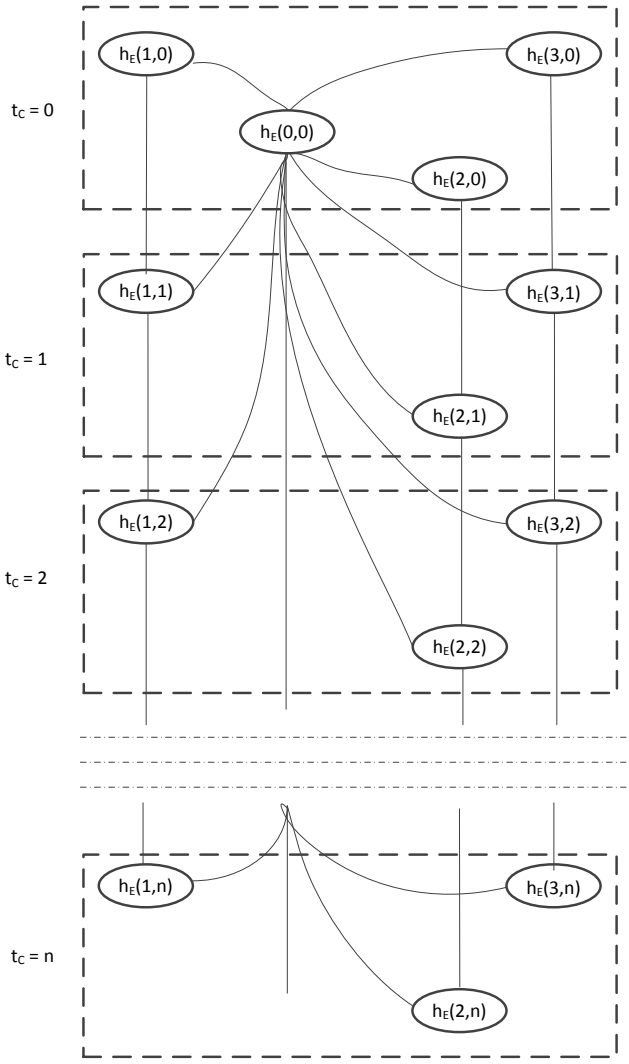


Fig. 4.4: Spatio-temporal n-dimensional C-vine copula

4.4 MODELING FRAMEWORK AND ASSESSMENT BASED ON REAL DATA

A numerical analysis is essential to understand the modeling framework. This section describes the real data used as input, the developed sampling algorithm used for vine copula construction and the output sampled dataset. Algorithm 4.3 explains the developed algorithm step-by-step and each step of the algorithm is further explained in sub-sections. All computation is performed in MATLAB (version 2017b) environment on an Intel Core i7 with 8 cores and 8GB RAM.

Algorithm 4.3: High dimensional spatio-temporal dataset modeling using vine copula

Inputs: High dimensional dataset of size (d x N), representing M data points and N features.

Outputs: S x N dimensional sampled dataset

1. Perform clustering to partition the high dimensional data. k number of clusters are selected after performing GoC test on sample size S
2. Feature extraction of k clusters using singular value decomposition (SVD)
3. Calculate copula function and construct vine copula models for k clusters. Choice is different copula functions can be tested and GoF test is performed to select the best copula function
4. Simulate the copula function for k clusters using cluster weight obtained in Step 1
5. Reconstruct dataset from low to high dimension with all features using eigenvectors from Step 2
6. Output as S x N dimensional sampled dataset
7. **End**

4.4.1 INPUTS

For this research, publicly available load and wind power data are taken from one of the U.S. regional transmission operator [Web4, 2018]. Aggregated zonal load data (nineteen numbers) and wind power data (two numbers) spanning three years with hourly resolution is used in this study and is shown in Fig. 4.5. The load and wind power from each zone is described by a distinct time series corresponding to a distinct position in space defined as the *weighted centroid*. Such a *weighted centroid* is required to calculate the spatial correlation using the geographical coordinates of zones. Since the exact coordinates are treated confidentially by utilities, an approximated *weighted centroid* is defined in this research to locate an approximated 'center' of load zones and wind power generation zones. A detailed explanation of the latitudes and longitudes to calculate the approximated *weighted centroid* is available in Appendix A1. The *weighted centroid* approach is able to describe the approximate dependencies between zones and a more realistic relationship is determined by the actual size of the zone. It can be argued that the resulting dependency will be affected by such aggregated data and approximated *weighted centroid* approach. However, such an approach still provides valuable information about dependencies. The three market regions are *MIDATL, WEST* and *SOUTH* and a detailed composition of these regions with load and wind power zones is shown in Table 4.1.

To visualize the complexity, scatter plot with marginal histograms of four load zones (*AP*, *CE*, *DAY* and *DUQ*) and one wind power zone (*WEST*) under the *WEST* zone is shown in Fig. 4.6. The marginal histograms (in the diagonal) reveal non-Gaussian nature while the scattered plots reveal the non-linear dependencies and also suggest a weak correlation. This does not mean lack of relationship, but rather a lack of linear relationship. In such cases, the marginal distributions do not conform to normality assumption and the dependencies between the variables are misleading too. This is valid for dependence studies between individual load and wind power, between different load zones and even between the output of two wind power zones.

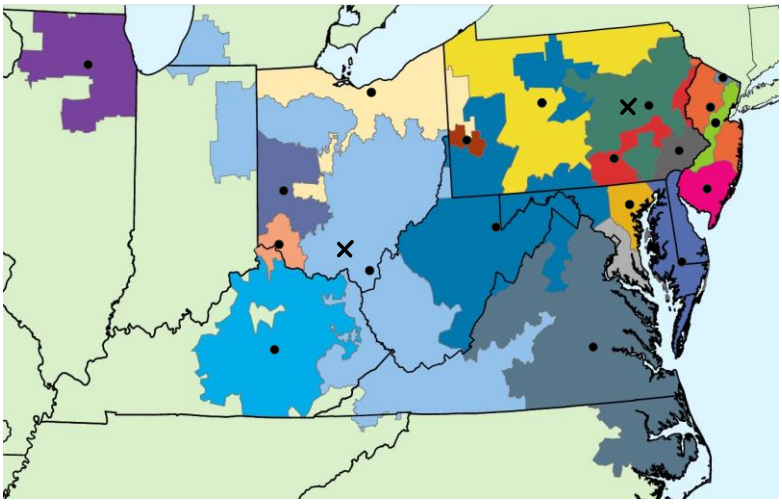


Fig. 4.5: Map showing the control areas of PJM with load and wind power zones' approximated weighted load centroid (●) and wind power centroid (**X**)

Table 4.1: Load and wind power zones as seen in Fig. 4.5

Market region	**Load zone**	**Wind power zone**
MIDATL	AE	MIDATL
	BC	
	DPL	
	JC	
	ME	
	PE	
	PL	
	PN	
	PS	
	RECO	
WEST	AEP	WEST
	AP	
	ATSI	
	CE	
	DAY	
	DEOK	
	DUQ	
	EKPC	
SOUTH	DOM	-

The m load zones (ten *MIDATL*, eight *WEST* and one *SOUTH*) and n wind power zones (one *MIDATL* and one *WEST*) for *t* time length (hourly resolution with a horizon of three-years) are written as:

$$X = \begin{bmatrix} L_1(t_1) & \cdots & L_m(t_1) & W_1(t_1) & \cdots & W_n(t_1) \\ L_1(t_2) & \cdots & L_m(t_2) & W_1(t_2) & \cdots & W_n(t_2) \\ \cdots & \ddots & \cdots & \cdots & \ddots & \cdots \\ L_1(t_t) & \cdots & L_m(t_t) & W_1(t_t) & \cdots & W_n(t_t) \end{bmatrix} \in \mathbb{R}^{t \times N} \quad (4.14)$$

where N refers to total measurements ($m + n$). For spatio-temporal modeling, we first visualize spatial correlation for the original data. Equation 4.3 is used to calculate the spatial lags. In this research, $MaxDist$ between any two zones (including load and wind power zones) is 1235 kilometers and M is 34 based on the minimum distance between any two zones. The spatial correlation plot, shown in Fig. 4.19, correspond to correlation coefficient between load zones in *MIDATL* and *WEST*. The correlation calculation uses hourly interval aggregated load on individual load zone for the years 2014 to 2016. The correlation coefficient for A and B is calculated using covariance, written as:

$$\rho_{AB} = \frac{Cov(A, B)}{\sigma_A \sigma_B} \quad (4.15)$$

where, σ_A, σ_B are the standard deviation and $Cov(A, B)$ is the covariance of A and B. Correlation plot reveals a weak correlation between zones. This lack of correlation between load zones is important to determine potential investment deferment in generation and planning interconnector capacity to utilize the lack of correlation.

To account for temporal correlation, the time series is checked for seasonality. It is to be noted that the same procedure was followed in chapter 3 while forecasting load in the long-term horizon. Presence of seasonality is evident in load time series and as an example, the original load time series from zone *AE* of *MIDATL* for the year 2014 with hourly resolution is shown in Fig. 4.7. And, Fig. 4.8 shows the peaks corresponding to the weekly trend. It is understood that electricity load data shows daily and weekly periodicity. Thus, the data needs to be differenced at both 24 and 168 lags, and this is checked and repeated for all load data. Backward differencing is normally used, and the 24th and 168th difference address the periodicity. For a time series y_t, the transformation is written as:

$$\Delta_{24}\Delta_{168}\, y_t = (1 - L^{24})(1 - L^{168}) y_t \quad (4.16)$$

where, Δ is the difference operator and L is the lag operator. After the lag operator polynomials ($(1 - L^{24})(1 - L^{168})$) are created, both are multiplied to get the desired lag operator polynomial. The differenced time series for zone *AE* of *MIDATL* is shown in Fig. 4.9 which is deseasonalized. Similarly, to account for temporal correlation in wind power data, all the wind data is analyzed. Fig. 4.10 shows the original wind power time

series. Wind power data also show distinctive seasonal and diurnal patterns as seen in Fig. 4.11. However, after checking for lags, seasonal differencing is performed for wind power data for 24 lags. So, equation 4.16 is modified for wind power as:

$$\Delta_{24}\, y_t = (1 - L^{24}) y_t \tag{4.17}$$

As the multivariate dataset is pre-processed with seasonal differencing to remove the periodicity, the second step in preparing the data is performing normalization. Normalization serves the purpose of bringing the multivariate variables into the same scale. For the load data, Z-score scaling is introduced to standardize the data for each zone as represented in [Shi et al., 2017]. Z-scores is the most commonly used method and it converts all indicators to a common scale with an average of zero and standard deviation of one. However, the wind power was normalized with respect to the installed wind capacity by comparing each of the datasets from different zones. Thus, the normalized wind power (W_{norm_i}) for each zone and hour i is calculated as,

$$W_{norm_i} = \frac{W_i}{Cap_{inst}} \tag{4.18}$$

where W_i is the actual wind power produced for hour i and Cap_{inst} is the installed wind capacity of the zone.

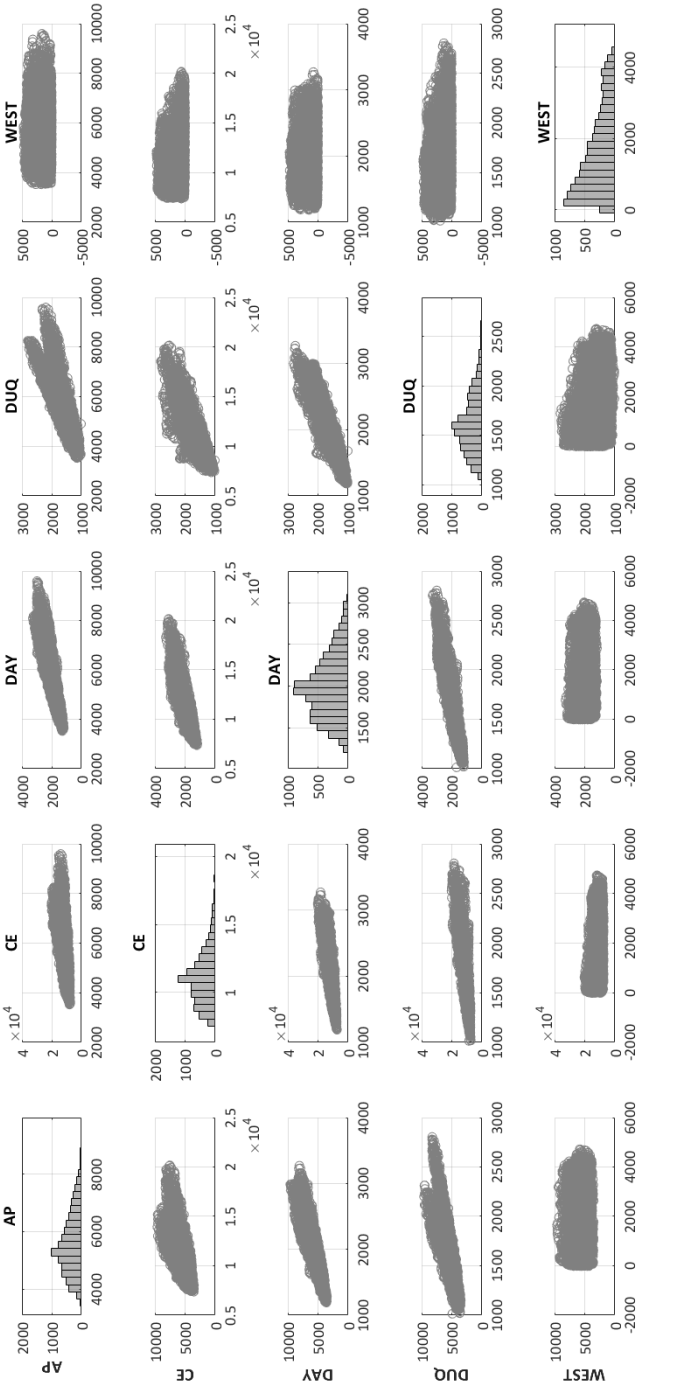


Fig. 4.6: Scatter plot with marginal histograms of original data of four load zones (AP, CE, DAY, DUQ) and one wind power zone (WEST)

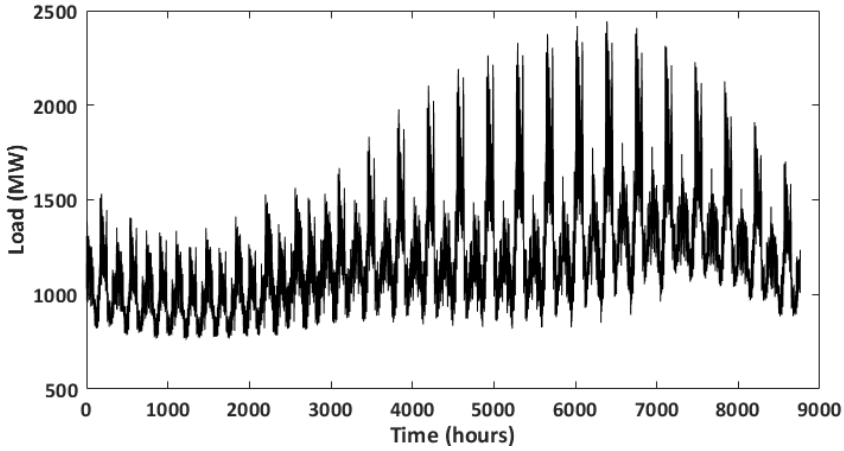


Fig. 4.7: Original load time series with hourly resolution (Year 2014 and zone *AE*)

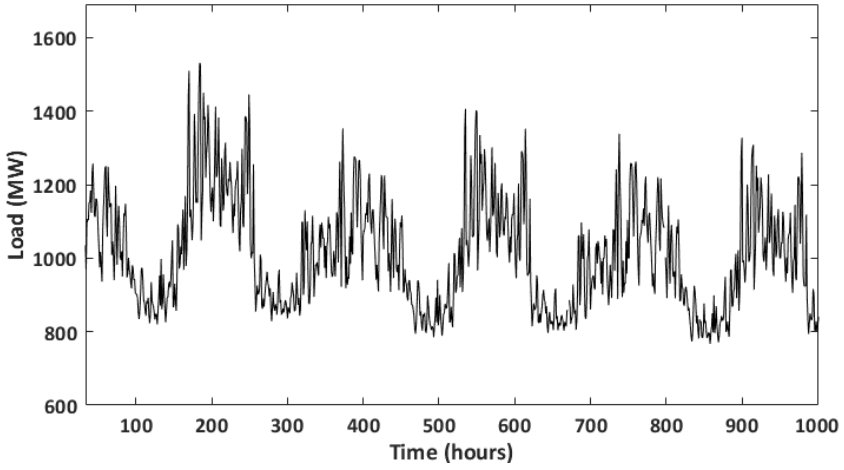


Fig. 4.8: Original load time series showing the weekly periodicity (Year 2014 and zone *AE*)

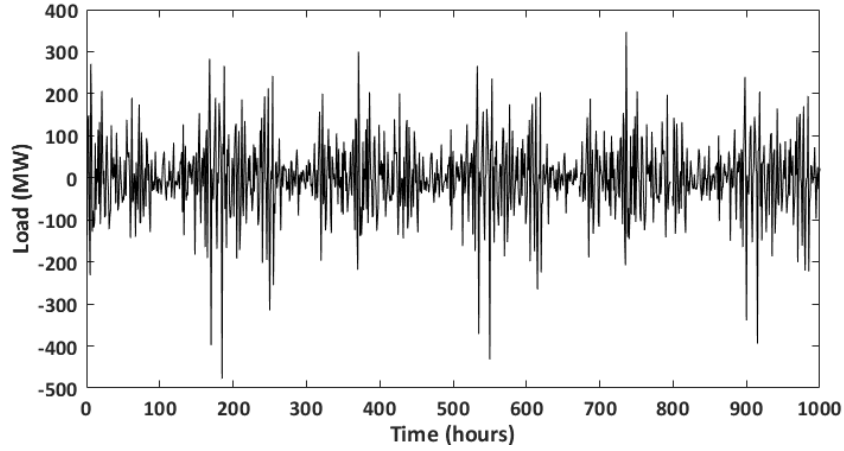


Fig. 4.9: Load time series of zone *AE* after seasonal differencing at lags 24 and 168

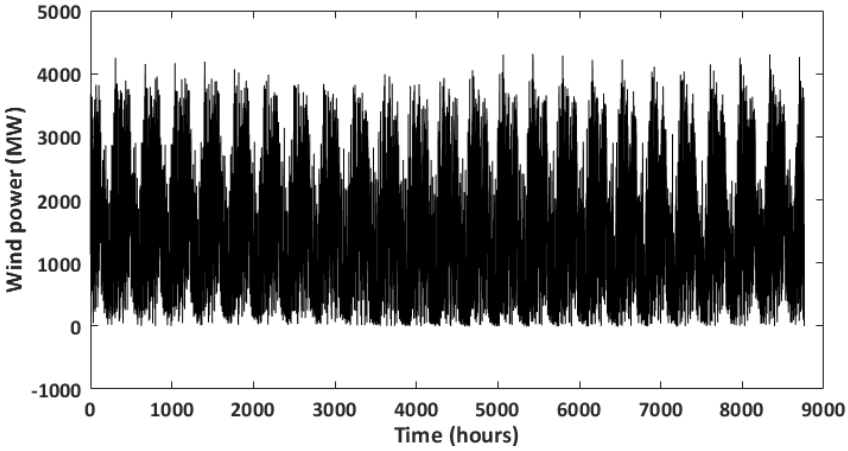


Fig. 4.10: Original wind power time series with hourly resolution (Year 2014 and zone *WEST*)

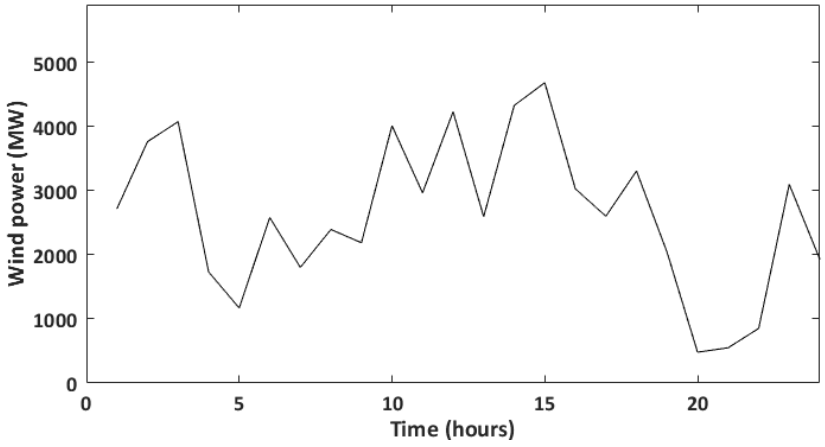


Fig. 4.11: Original wind power time series for 24-hour duration (Year 2014 and zone *WEST*)

Following the pre-processing steps, the multivariate dataset is normalized and is free of any trend and seasonality. To account for the large-sized dataset and to ease computational burden, the sampling procedure starts with performing clustering and followed by feature extraction. The problem addressed here is extracting the required features when the number of clusters is unknown. With such an approach, the method also reduces the dimensionality of the dataset.

4.4.2 STEP 1 (DATA CLUSTERING)

The power system network operates with a wide range of operating conditions and deals with non-uniform multivariate data from demand and energy injection centers. Both electricity load and wind power generation patterns are determined by different drivers depending on time and location as both vary with respect to time and space.

The data collected, corresponding to the peak value, does not always come from a single collection model and rather multiple models. Such high dimensional data often come with many highly correlated variables and succeeding in selecting all variables in a group of correlated variables can be very difficult. Clustering helps in partitioning the M-data points into groups of similar statistical characteristics or k clusters. The aim of data clustering is to discover the "natural" group(s) of a set of patterns in a multivariate dataset. Use of cluster analysis is widespread in any discipline that involves a study of multivariate data. Till date, many clustering algorithms have been proposed based on different application scenarios. And the associated literature on clustering is vast as hundreds of clustering algorithms have been recommended in the literature. For an overview of different clustering algorithms, readers can refer to surveys [Berkhin, 2006, Law, 2006] where the different clustering algorithms are discussed. Fig. 4.12 shows a classification of clustering algorithms. Literature study reveals that the most important way to classify clustering algorithms is partitional versus hierarchical clustering [Law, 2006]. Partitional clustering aims at creating a *flat* partition of the set of objects with each object belonging to one and only one group. On the other hand, hierarchical clustering aims at creating a tree of objects, where branches merging at the lower levels correspond to higher similarity. In this research, we will examine three widely used clustering algorithms used in analysis multivariate datasets: the *k*-means, Gaussian Mixture Model (GMM) and Hierarchical Linkage (HL) algorithms. Out of the three, GMM is a special case as it is a mixture-based clustering and its statistical nature gives us a solid foundation for analyzing its behavior.

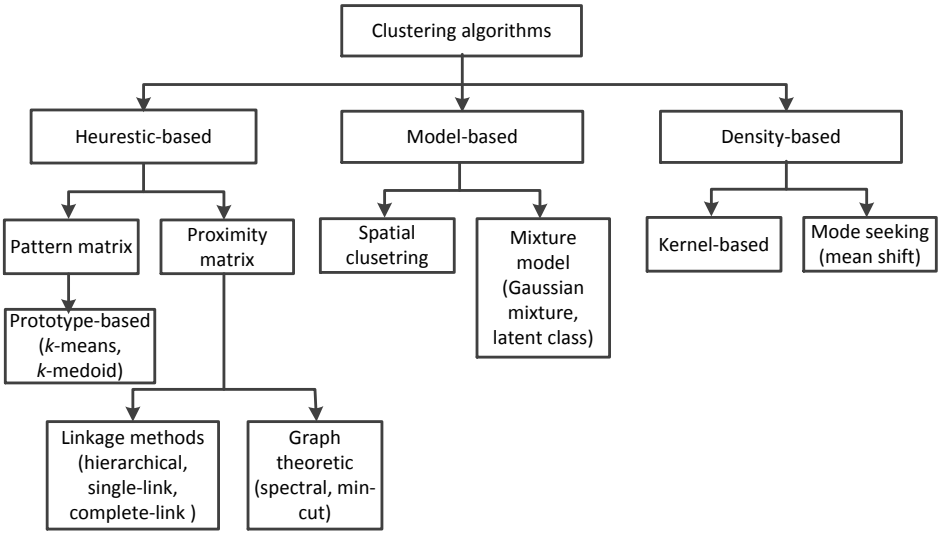


Fig. 4.12: Classification of clustering algorithms [Law, 2006]

k-means is the most widely used unsupervised clustering technique as it is easy to understand and implement [Hartigan et al., 1979]. k-means clustering aims at partitioning M-data points into k-clusters, where each data point belongs to the cluster with its nearest mean. The k-means clustering works as an objective function where the aim is to minimize a squared error function,

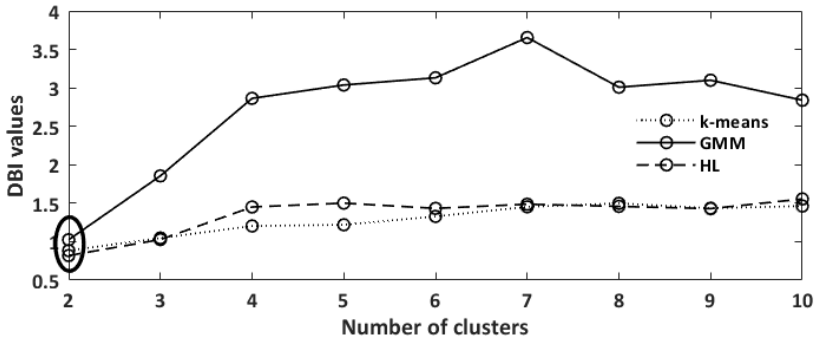
$$min \sum_{j=1}^{k} \sum_{i=1}^{M} \left\| x_i^{(j)} - c_j \right\|^2 \tag{4.19}$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center c_j. In contrast to simplicity, k-means has problems in discovering clusters that are not spherical in shape. It also encounters some difficulties when different clusters have a significantly different number of points. Since it is a minimization function, k-means also requires a good initialization to avoid getting trapped in a poor local minimum. It makes a number of assumptions about the data and it does not search through every possible partitioning of the data, hence, it was opted to test GMM and HL techniques. Gaussian mixture model (GMM) is a clustering algorithm to estimate probability function by using a finite linear combination of Gaussian model in which the weights of each Gaussian component is defined as a prior probability of each component. A step-by-step explanation of GMM clustering is available online [Web12, 2018]. GMM clustering technique uses the probability of a sample to determine the feasibility of it belonging to a cluster. And, to obtain the optimal model parameters, Expectation-Maximum (EM) optimization approach is used rather than maximum likelihood estimation because it avoids infinitely possibility problem for some sparsely distributed single points.

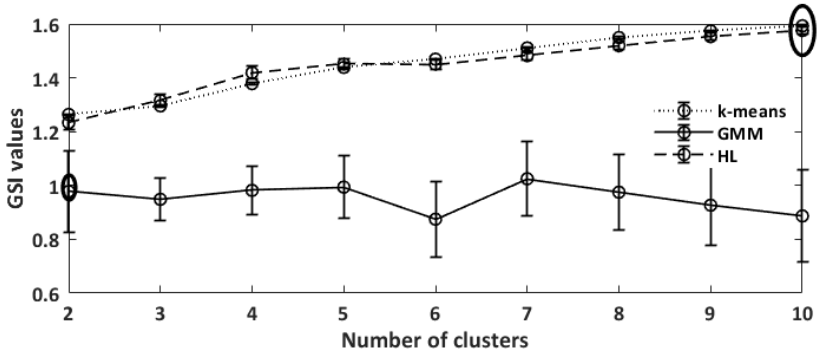
Compared to both k-means and GMM, hierarchical linkage (HL) clustering technique builds clusters incrementally. The clustering technique begins by assigning each sample to its own cluster (top level), and at each step, the two clusters that are most similar are merged. It continues until all of the clusters have been merged. In comparison to k-means, there is no need to specify a k parameter as one can navigate the layers of hierarchy to see which number of clusters is optimum. In addition, k-means clustering is usually more efficient run-time wise compared to GMM and HL clustering since k value is usually specified.

After deciding the clustering algorithms, choosing the right number of clusters is important to validate the chosen clustering algorithm. To select the right number of clusters, GoC test is performed using Davies-Bouldin index (DBI) and gap statistics index (GSI). A detailed explanation about the GoC tests is included in Appendix A2. k values ranging $\{2, \dots, 10\}$ were assessed using DBI and GSI statistics to choose the number of clusters. DBI quantifies the average similarity between the chosen number of clusters

[Davies et al., 1979]. In theory, it is desirable for the clusters to be as distinct from each other as possible and hence, the clustering technique which minimizes the DBI value is the ideal one for GoC test. Lower values of DBI correspond to better clustering validity. The results from GoC test for DBI is shown in Fig. 4.13(a). From the figure, values of $\{k = 2,2,2\}$ are obtained for the three clustering techniques. The second GoC test is GSI, which compares the within-cluster dispersion to its expectation under an appropriate null reference distribution [Tibshirani et al., 2001]. Fig. 4.13(b) shows the GoC test and corresponding GSI values. Gap statistics is maximized at $k = 10$ for k-means and HL clustering. For GMM, gap statistics is maximized at $k = 2$.



(a)



(b)

Fig. 4.13: GoC test and corresponding (a) DBI values (b) GSI values

GoC test is required to carefully select the number of clusters k. If k is too small, the GMM is unable to well capture the distribution of data, (especially when $k = 1$, GMM will degenerate to Gaussian maximum likelihood). On the other hand, if k is too large, except for computation complexity, a severe overfitting problem will result. Based on the GoC tests, $k = 2$ is chosen in this study to generate a large sample size of $S = 30000$ from the pre-processed historical dataset. Such a large sample size is chosen to guarantee a good accuracy of the estimated value. Clustering results in sorting the multivariate dataset (X) into homogeneous clusters ($X^k \subset X$, for $k = 1,2$)

which are strongly related to each other and thus provide similar information. Next step is feature extraction from the clusters which selects a small subset of actual features and remove redundant features.

4.4.3 STEP 2 (FEATURE EXTRACTION)

In spatio-temporal modeling, feature extraction in the form of dimension reduction is reasonable given that the true spatio-temporal feature often exists on a lower dimensional structure [Cressie & Wikle, 2015]. As the name suggests, dimensionality reduction is the process to transform a high dimensional dataset into a low dimensional space, while retaining most of the useful information from the original data. The principle behind such a transformation is that the useful information in the original high dimensional dataset can be represented by a small number of features. Generally, in a high dimensional space, the data points do not spread-out randomly but, rather, in a certain structure that can be easily exploited. Thus, dimensionality reduction can circumvent this problem by reducing the number of features in the data set before the training process. Doing this reduces the computation time and the resulting features in low dimension take less space to store. Other advantages of dimensionality reduction are easy interpretation and visualization, because of the low dimensional space. However, if not performed correctly, there are high chances that dimensionality reduction will result in information loss. And there is no way to extract the lost information from low dimension to high dimensional space. Reference [Law, 2006] classify dimensionality reduction into three types:

i. *Feature selection and feature weighting*: Feature selection, also known as variable selection, deals with the selection of a subset of features that are most appropriate for the task at hand. A feature is either selected (because it is relevant) or discarded (because it is irrelevant). Feature weighting, on the other hand, assigns weights (usually between zero and one) to different features to indicate the saliencies of the individual features.

ii. *Feature extraction*: In feature extraction, a small set of new features is constructed by a general mapping from the high dimensional data. The mapping often involves all the available features.

iii. *Feature grouping*: In feature grouping, new features are constructed by combining several existing features. Feature grouping can be useful in scenarios where it can be more meaningful to combine features due to the characteristics of the domain.

In this research, feature extraction is chosen. Before performing feature extraction, the clustered observations using k-means from the original domain are transformed to the rank-uniform domain via the empirical cumulative distribution function ($ECDF$).

Thus, each clustered dataset X^k is transformed to Y^k in the $[0,1]^N$ domain. Such a transformations help in reducing the sensitivity of feature extraction techniques.

Feature extraction is a standard statistical method for simplification of high dimensional multivariate datasets. It helps in filtering out irrelevant data pertaining to datasets in return for more pertinent and informative data. In this research, singular value decomposition (SVD) is employed to perform principal component analysis (PCA) since it is efficient and numerically robust [Wall et al., 2003]. Given an arbitrary $p \times q$ matrix $X \in \mathbb{R}^{p \times q}$, then there exists matrices U and V (both with orthogonal columns), and positive numbers $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ (where $r = \min(p, q)$), such that:

$$X = \sum_{k=1}^{r} \sigma_k U_k V_k^T = UDV^T \tag{4.20}$$

with U_k and V_k denoting the k^{th} column of U and V, respectively, and D is a $p \times q$ matrix for which the numbers σ_k (the singular values) are placed on the main diagonal. For a proof, see e.g. [Wall et al., 2003]. In this research, it is assumed that the singular values are arranged in descending order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$. For a given matrix X we use the notation $\sigma_i(X)$ or $\lambda_i(X)$ to denote the i-th (ordered) singular or eigenvalue, respectively. If there is no danger of confusion, the explicit reference to the matrix will be suppressed. Recall that there is a useful relationship between the singular values of a matrix $X \in \mathbb{R}^{p \times q}$ and the eigenvalues of the related matrices XX^T and X^TX:

$$\sigma_i(X) = \sqrt{\lambda_i(XX^T)} = \sqrt{\lambda_i(X^TX)} \tag{4.21}$$

where $i = 1{:}\min(p, q)$. This connection will be used extensively in the analysis below.

SVD is a computational method often employed to calculate principal components for a dataset. The set of principal components for each cluster will represent the original data, and they are determined by computing the eigenvalues and eigenvectors of the corresponding correlation matrix. It determines an optimal linear transformation: $y = Ax$, for p-dimensional data x into another q-dimensional transformed vector y for each cluster k. The linear transformation matrix A is optimal from maximal information retention criterion (IRC) viewpoint, given as:

$$IRC = \frac{\sum_{j=m+1}^{p} \lambda_j}{\sum_{i=1}^{p} \lambda_i} \tag{4.22}$$

This q-dimensional transformed vector ($q < t$) defines the reduced number of variables. For each cluster k obtained in Step 1 (sub-section 4.4.2), PCA is performed in three steps, (i) *centralize the data and compute the mean*, (ii) *then generate scatter matrix and compute eigenvalues (λ) and eigenvectors (m)*, and (iii) *project the data comprising principal components*. The eigenvectors, arranged in descending order

according to decreasing information content, represent the principal components of the clustered dataset, whereas the eigenvalues indicate to total variance accounted for by each principal component. Eigenvalues are obtained with respect to the covariance matrix to obtain relative importance of principal components. And the descending order of q-dimensional transformed vector y allows for straightforward feature extraction as well as dimensional reduction by discarding elements with lowest information content. Thus, feature extraction on each Y^k results in low dimensional dataset $\mathcal{H}^k \in \mathbb{R}^{t^k \times q}$, where t^k represents the number of observations in cluster k. It is to be noted that eigenvectors are later used in Step 5 in the resampling step.

4.4.4 STEP 3 (VINE COPULA CONSTRUCTION)

Till now, for each cluster, we have extracted the features based on which vine copula will be constructed for spatio-temporal modeling. For high-dimensional distributions, there are a significant number of possible pair-copulae constructions. Thus, the next step is describing a multivariate dependence structure with vine copula and selecting a bivariate copula family for each edge in the selected vine as well as estimating its parameters. For vine copula modeling, the obtained observations in Step 2 should be fitted in the uniform domain. Such a uniform transformation is achieved by $ECDF$s of $\mathcal{H}^k$ to obtain $\mathbb{C}^k$. In this research, families of bivariate copulas, namely, Gaussian copulas, Student's-t copulas, asymmetric Clayton and its corresponding 90°, 180° and 270° rotational copula types are implemented in C-vine. Construction of C-vine tree starts with selecting a root node, which is achieved by generating Kendall rank correlation matrix and adding the correlations across each location with respect to other locations [Genest & Favre, 2007]. The location with highest value of Kendall rank correlation coefficient is chosen as root node followed by other nodes in the tree.

After selecting the root node, estimating the conditional copulas in the tree is performed. To select the appropriate copula, GoF test is performed to check the copulas that can be rejected. Goodness of Fit (GoF) techniques are used for assessment whether a distribution is suitable to describe a data set or not. The null hypothesis for the GoF test states that the data is sampled from a normal distribution. When the p-value is greater than the predetermined critical value, the null hypothesis is accepted and thus we conclude that the data fits well or is normally distributed. Two GoF tests used in this study are Kolmogorov-Smirnov (K-S) and Cramer-von Mises (CvM) test. Both the tests are based on estimated $ECDF$. The idea of using $ECDF$ in testing normality of data is to compare the $ECDF$, based on the data with the CDF of the normal distribution, to see if there is a good agreement between them. The K-S test is a non-parametric test and is used to check if a sample comes from a hypothesized continuous distribution. The K-S statistic (D) is written as [Stephens, 1986]:

$$D = sup_x|F_h(x) - F_o(x)| \tag{4.23}$$

where, $F_h(x)$ is the $ECDF$ of hypothesized distribution and $F_o(x)$ is the $ECDF$ of observed distribution. A powerful and refined version of K-S test, called the Cramer-von Mises (CvM) test is also used in this study. The CvM statistic (ω^2) is written as [Stephens, 1986]:

$$\omega^2 = \int_{-\infty}^{\infty} [F_h(x) - F_o(x)]^2 dF_o(x) \tag{4.24}$$

Following the GoF test, $ECDF$s of different copula selection for each cluster at one branch of tree is shown in Fig. 4.14. For $k = 1$, p-value is 0.8950 (K-S test) and 0.8765 (CvM test) and 270° rotated Clayton copula is selected. Similarly, for $k = 2$, p-value is 0.2441 (K-S test) and 0.2422 (CvM test) and 180° rotated Clayton copula is selected. A high p-value indicates that it is a good fit since the acceptable level is ≥ 0.05. This is repeated for each bivariate copula in rest of the tree.

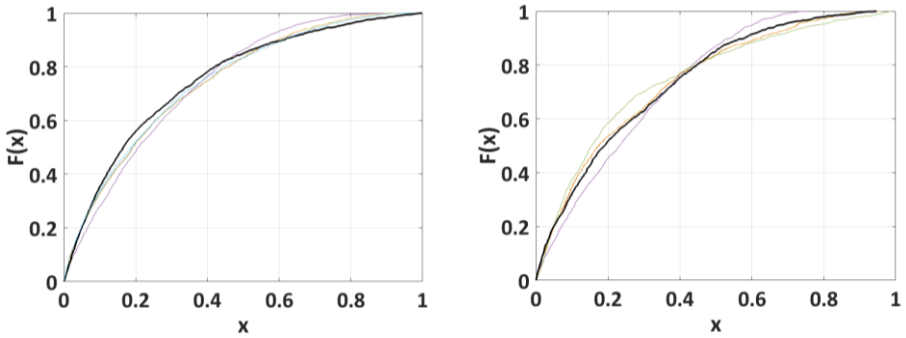


Fig. 4.14: $ECDF$ for estimated copulas (black line represents the selected copula)

4.4.5 Step 4 (Vine Copula Simulation)

Step 4 operates in coordination with step 3 in terms of vine copula simulation. For the chosen sample S, each parametric cluster-model generates samples $\mathbb{C}^k$ of size $t_S^k \times q$, where $t_S^k = S \times W^k$ and W^k is the weight of cluster k. Construction of C-vine is based on W^k obtained in *Step 1* (sub-section 4.4.2). Once the model has been stated and estimated, a key question is to check whether the initial model assumptions are realistic. Again, a GoF test is performed on C-vine for each cluster. The chi-square (χ^2) test is used to graphically represent GoF. Likewise K-S and CvM tests, the chi-square test is used to test if a sample of data came from a population with a specific distribution. At one branch of the tree, for $k = 1$, p-value is 0.9290 (K-S test) and 0.9122 (CvM test),

and for $k = 2$, p-value is 0.1145 (K-S test) and 0.1095 (CvM test) is calculated. A graphical comparison of GoF for K-S test and chi-square test is shown in Fig. 4.15.

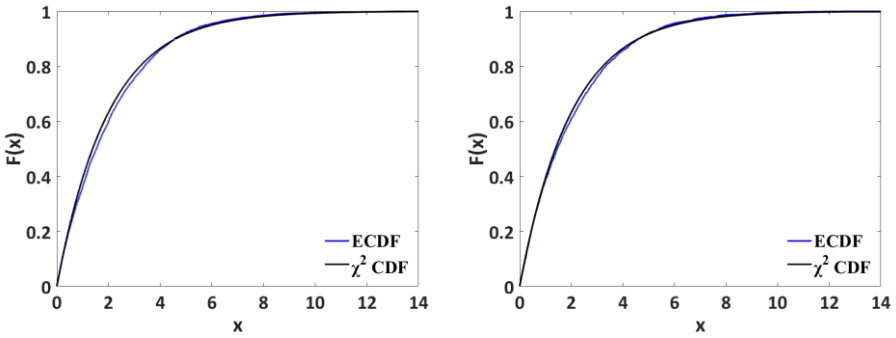


Fig. 4.15: $ECDF$ of K-S test and chi-square (χ^2) CDF plot for (a) $k = 1$, and (b) $k = 2$ showing the GoF in vine copula simulation

4.4.6 Step 5 (Resampling)

Performing inverse $ECDF$ transformation ($ECDF^{-1}$) on the sampled output to retrieve high dimensional data is achieved in this step. In the first step, samples of each cluster $\mathbb{C}^k$ are transformed back to the domain of $\mathcal{H}^k$ of size $t_S^k \times q$ by transforming $\mathbb{C}^k$ through $ECDF^{-1}$ of original dataset $\mathbb{C}^k$. The second step is transforming $\mathcal{H}^k$ to high dimensional space $\mathbb{R}^{t_S^k \times N}$, denoted by the dataset Y^k. And the last step involves the transformation of Y^k in the $[0,1]^N$ domain to original dataset X^k through $ECDF^{-1}$. In the end, the high dimensional sampled dataset is $\bar{X} = \bar{X}^1 \cup \bar{X}^2 \cup \ldots \cup \bar{X}^k \in \mathbb{R}^{S \times N}$, where $\bar{X}^k$ corresponds to the sampled dataset for cluster k.

In order to evaluate the performance of C-vine sampling, two-sample Kolmogorov-Smirnov (K-S) test is employed for the historical and simulated dataset for each variable in the multivariate dataset to check the null hypothesis that they are drawn from the same marginal distribution. It is easy to confuse the two-sample K-S test (which compares two groups) with the one sample K-S test used in Step 3, also called the K-S GoF test, which tests whether one distribution differs substantially from theoretical expectations. The two-sample K-S test statistic quantifies a distance between the empirical distribution functions of two samples. Theoretical description of the two-sample K-S test is included in Appendix A2.

A resampling method is employed to randomly generate comparison samples from the historical and sampled datasets based on [Sun et al., 2016]. For each of the 21 variables, 200 data points from the historical dataset and 400 data points from the sampled dataset were drawn in random. And the process was iterated for 500 times. Thus, the total number of times the K-S test performed is (500x21) times. Fig. 4.16 shows the ECDF plot against the p-values calculated from K-S test. In the test, reference

dataset is the historical dataset plotted against itself, which has a uniform distribution. And the sampled dataset from the sampling procedure is also perfectly uniform, aligning to reference dataset p-values. Thus, we can conclude that the sampled dataset does not suffer from information loss.

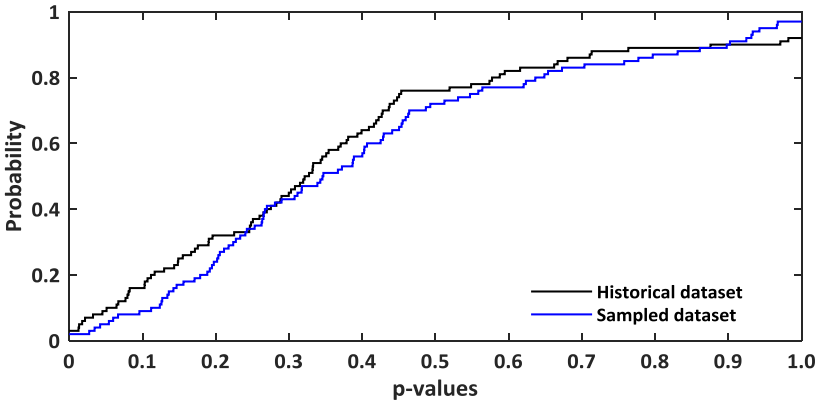


Fig. 4.16: Two-sample K-S test for historical and sampled dataset

4.4.7 OUTPUT

Following all the steps, the output obtained is a high dimensional sampled dataset with hourly time steps. When handling high dimensional dataset, computational efficiency is a vital concern. For this research, a mix of k-means clustering and C-vine sampling accounted for a computation time of 25.88 seconds. Fig. 4.17 and Fig. 4.18 shows the sampled load and wind power time series of zone *AE* and *WEST* respectively. A visual inspection of Fig. 4.18 shows some negative data points in the sampled dataset. The presence of negative values in the sampled wind power dataset is because wind power is treated as a negative load in the sampling procedure. Pair-wise comparison of histograms and scatter plots for the marginal distributions represented by four load zones and one wind farm zone under *WEST* market region, same as in Fig. 4.6, is shown in Fig. 4.19. The histograms in diagonal of the figure show the normally distributed marginals with heavy tail characteristic while the scattered plots reveal non-linear dependence between the variables. Now we compare the spatial correlation plot for the load zones in the sampled dataset with respect to the original data for zones *MIDATL* and *WEST*. The spatial correlation plot for the original dataset is shown in Fig. 4.20(a) and Fig. 4.20(b). Similarly, for the sake of visual comparison, the spatial correlation plot for the sampled dataset comparing the load zones for *MIDATL* is shown in Fig. 4.20(c) and for *WEST* is shown in Fig. 4.20(d). The correlation color map is adjusted to the same scale ($\rho = [0,1]$) While Fig. 4.20(c) accounts for more positive correlation, Fig. 4.20(d) displays some very strong correlation as well as some zero correlation. Understanding the variety in spatial correlation is *WEST* is understood by the long separation (distances) between the load zones. To further check the

dependency in the multivariate sampled output, the correlation coefficient is calculated for the dataset using equation 4.15. The detailed table of correlation coefficients is shown in Table 4.2. The values represent a non-linear dependence among the variables in the multivariate dataset calculated from bivariate copula. And, it should not be confused with the well-known linear correlation coefficient. Fig. 4.21 shows the overall spatial correlation including all load and wind power zones. Again, for visual comparison, Fig. 4.21(a) shows the correlation for original dataset and Fig. 4.21(b) shows the correlation for the sampled dataset.

Inferencing the correlation coefficients suggests positive, zero and negative correlation as well. Minimum variance is obtained for maximum negative correlation ($\rho = -0.2675$) and maximum variance in case of maximum positive correlation ($\rho = 1$). A negative correlation means that high values of the one dataset are matched with low values of the other, while positive correlation means that high values of the one dataset are always matched with high values of the other. This matching has a direct impact on the behavior of the sum: negative correlation prevents extreme values from happening at the same time, while positive correlation urges coincidence of extreme events. In terms of physical significance, a positive correlation between load and wind power explains the fact that both tend to increase and decrease at the same time, thereby facilitating load following task of the power system. On contrary, a negative correlation suggests that increase in load demand is identified by a decrease in wind power generation (and vice versa), thereby asking for more production from load following plants in the power system. The presence of a negative correlation between load and wind power is of physical significance. It explains the need to balance out the wind power fluctuations in different zones with corresponding load fluctuations to maintain a steady supply. In context of this study, it is fairly understandable that more wind farms will be integrated under the control area as shown in Fig. 4.1. In fact, the two wind power zones show a weak correlation of *0.54* which is understandable from their location. The reason for the negative correlation between load and wind power is explained by the fact that wind often blows during a period of low electricity demand. To suffice the negative correlation, the zones need to have adequate transmission connection so as to utilize the generated wind power at the location with high demand.

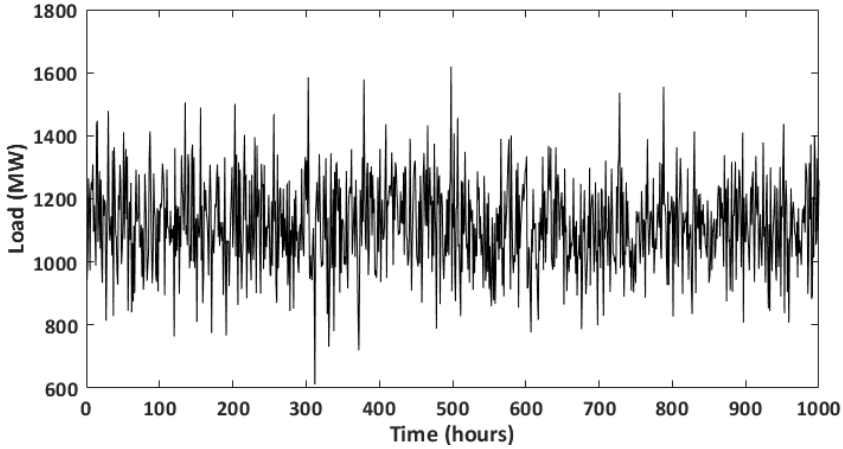


Fig. 4.17: Sampled load time series with hourly resolution of zone *AE*

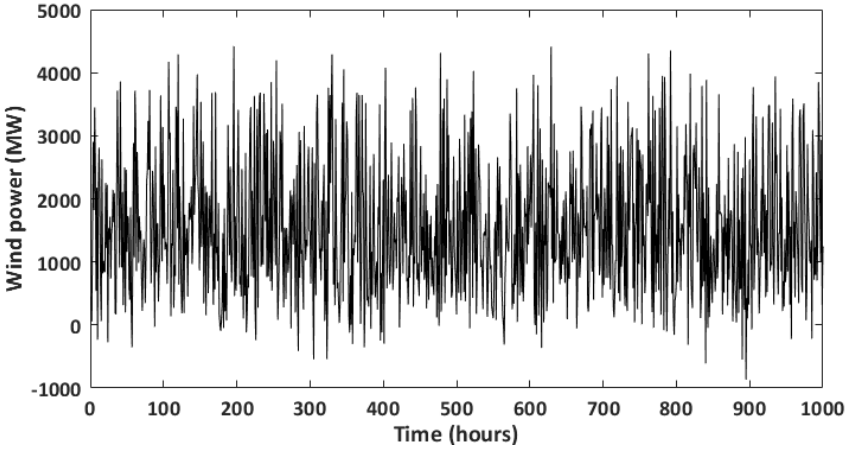


Fig. 4.18: Sampled wind power time series with hourly resolution of zone *WEST*

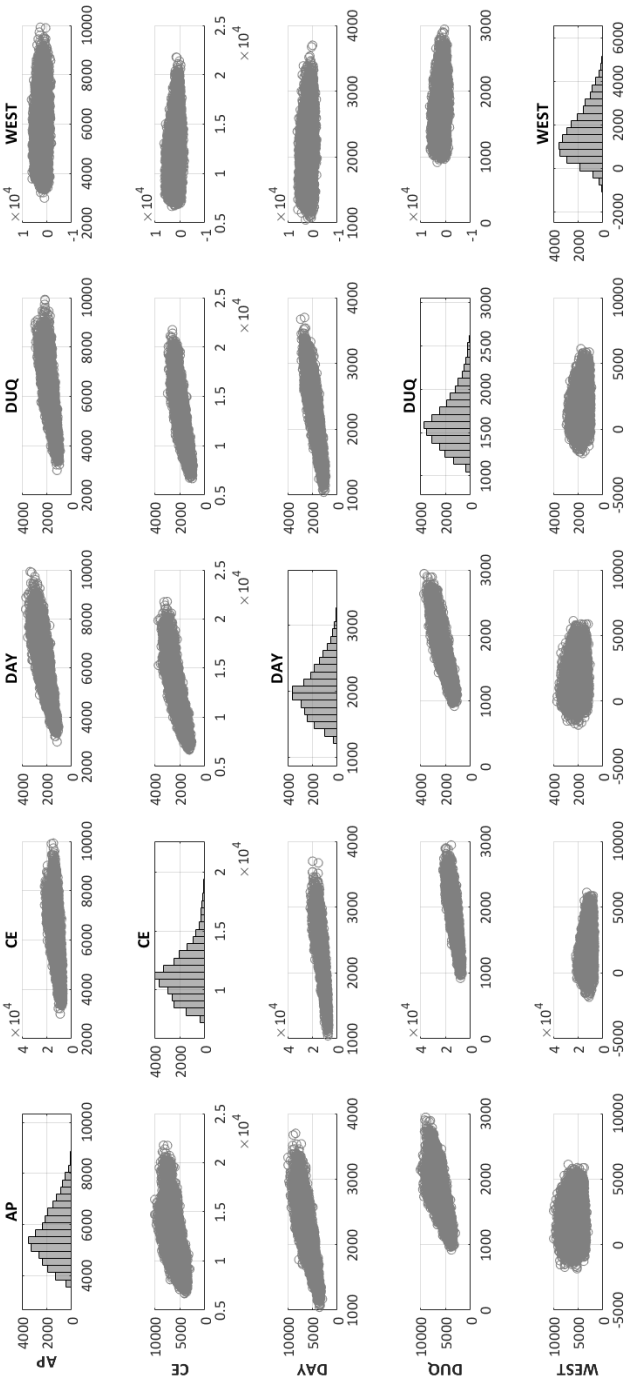


Fig. 4.19: Scatter plot with marginal histograms of sampled output of four load zones and one wind power zone

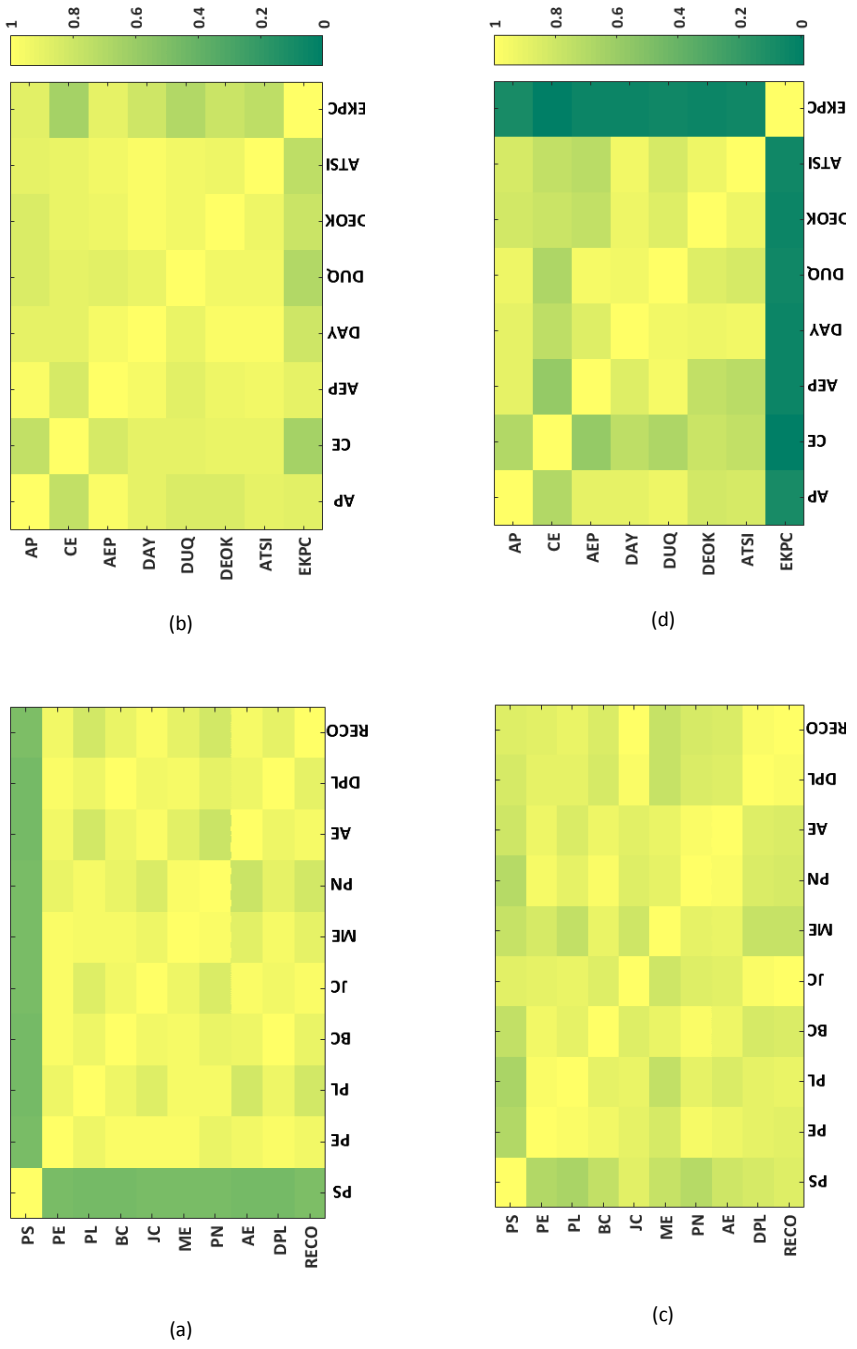
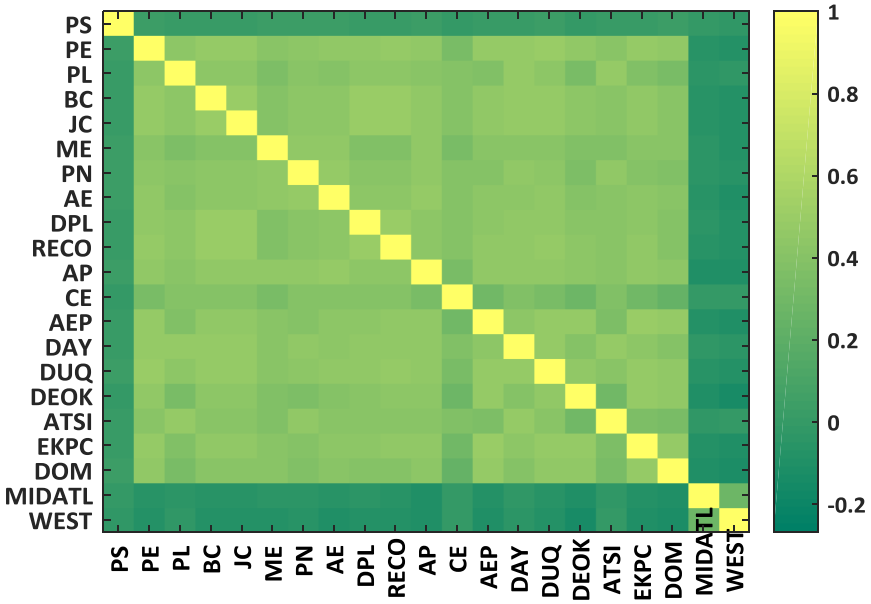
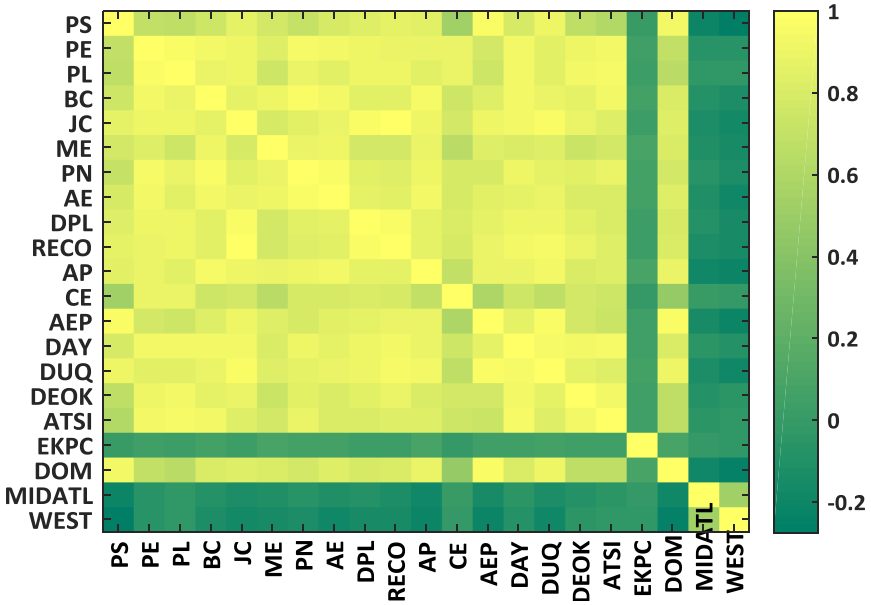


Fig. 4.20: Spatial correlation plot of different load zones (a) original dataset *MIDATL* zone, (b) original dataset *WEST* zone, (c) sampled dataset *MIDATL* zone, and (d) sampled dataset *WEST* zone



(a)



(b)

Fig. 4.21:Spatial correlation plot of load and wind power zones of (a) original dataset, (b) sampled dataset

Table 4.2: Correlation table for load and wind zones after sampling

	PS	PE	PL	BC	JC	ME	PN	AE	DPL	RECO	AP	CE	AEP	DAY	DUQ	DEOK	ATSI	EKPC	DOM	MIDATL	WEST
PS	1	0,69	0,67	0,75	0,88	0,77	0,72	0,8	0,83	0,87	0,85	0,53	0,96	0,78	0,91	0,67	0,61	0,02	0,93	-0,21	-0,28
PE	0,69	1	0,97	0,94	0,9	0,83	0,96	0,92	0,9	0,89	0,89	0,9	0,77	0,92	0,85	0,91	0,93	0,04	0,7	-0,07	-0,07
PL	0,67	0,97	1	0,9	0,91	0,76	0,89	0,85	0,9	0,91	0,85	0,88	0,76	0,94	0,85	0,94	0,94	0,03	0,66	-0,03	-0,03
BC	0,75	0,94	0,9	1	0,87	0,91	0,97	0,94	0,84	0,85	0,94	0,74	0,83	0,93	0,89	0,87	0,92	0,08	0,81	-0,09	-0,12
JC	0,88	0,9	0,91	0,87	1	0,8	0,86	0,89	0,97	0,98	0,9	0,78	0,91	0,94	0,96	0,88	0,84	0,03	0,83	-0,13	-0,16
ME	0,77	0,83	0,76	0,91	0,8	1	0,9	0,91	0,77	0,77	0,89	0,65	0,82	0,81	0,84	0,73	0,76	0,09	0,82	-0,1	-0,16
PN	0,72	0,96	0,89	0,97	0,86	0,9	1	0,97	0,84	0,83	0,91	0,8	0,8	0,91	0,86	0,85	0,9	0,07	0,77	-0,07	-0,12
AE	0,8	0,92	0,85	0,94	0,89	0,91	0,97	1	0,87	0,86	0,94	0,79	0,85	0,88	0,89	0,8	0,82	0,07	0,82	-0,11	-0,19
DPL	0,83	0,9	0,9	0,84	0,97	0,77	0,84	0,87	1	0,97	0,86	0,8	0,86	0,9	0,92	0,86	0,81	0,03	0,79	-0,09	-0,15
RECO	0,87	0,89	0,91	0,85	0,98	0,77	0,83	0,86	0,97	1	0,87	0,78	0,9	0,92	0,95	0,88	0,82	0,02	0,81	-0,12	-0,15
AP	0,85	0,89	0,85	0,94	0,9	0,89	0,91	0,94	0,86	0,87	1	0,69	0,9	0,9	0,92	0,82	0,84	0,09	0,88	-0,18	-0,2
CE	0,53	0,9	0,88	0,74	0,78	0,65	0,8	0,79	0,8	0,78	0,69	1	0,59	0,75	0,67	0,78	0,76	-0,01	0,47	0,01	-0,01
AEP	0,96	0,77	0,76	0,83	0,91	0,82	0,8	0,85	0,86	0,9	0,9	0,59	1	0,87	0,96	0,76	0,73	0,05	0,97	-0,15	-0,21
DAY	0,78	0,92	0,94	0,93	0,94	0,81	0,91	0,88	0,9	0,92	0,9	0,75	0,87	1	0,94	0,94	0,94	0,05	0,81	-0,06	-0,09
DUQ	0,91	0,85	0,85	0,89	0,96	0,84	0,86	0,89	0,92	0,95	0,92	0,67	0,96	0,94	1	0,86	0,83	0,06	0,92	-0,12	-0,19
DEOK	0,67	0,91	0,94	0,87	0,88	0,73	0,85	0,8	0,86	0,88	0,82	0,78	0,76	0,94	0,86	1	0,93	0,04	0,68	-0,09	-0,05
ATSI	0,61	0,93	0,94	0,92	0,84	0,76	0,9	0,82	0,81	0,82	0,84	0,76	0,73	0,94	0,83	0,93	1	0,05	0,67	-0,05	-0,02
EKPC	0,02	0,04	0,03	0,08	0,03	0,09	0,07	0,07	0,03	0,02	0,09	-0,01	0,05	0,05	0,06	0,04	0,05	1	0,09	-0,01	-0,03
DOM	0,93	0,7	0,66	0,81	0,83	0,82	0,77	0,82	0,79	0,81	0,88	0,47	0,97	0,81	0,92	0,68	0,67	0,09	1	-0,19	-0,25
MIDATL	-0,21	-0,07	-0,03	-0,09	-0,13	-0,1	-0,07	-0,11	-0,09	-0,12	-0,18	0,01	-0,15	-0,06	-0,12	-0,09	-0,05	-0,01	-0,19	1	0,54
WEST	-0,28	-0,07	-0,03	-0,12	-0,16	-0,16	-0,12	-0,19	-0,15	-0,15	-0,2	-0,01	-0,21	-0,09	-0,19	-0,05	-0,02	-0,03	-0,25	0,54	1

On the interaction of load and wind power, discussion on the correlation between them is vital as it ascertains the capability of wind power to equalize the changes in load fluctuation. Due to the location constraint of wind power, WPPs are usually far away from load centers. Fig. 4.22 shows the overall correlation between each zone pair (both load and wind power) as a function of their approximate distance. The zone pair corresponds to Table 4.2 and the location details are provided in Appendix A1.

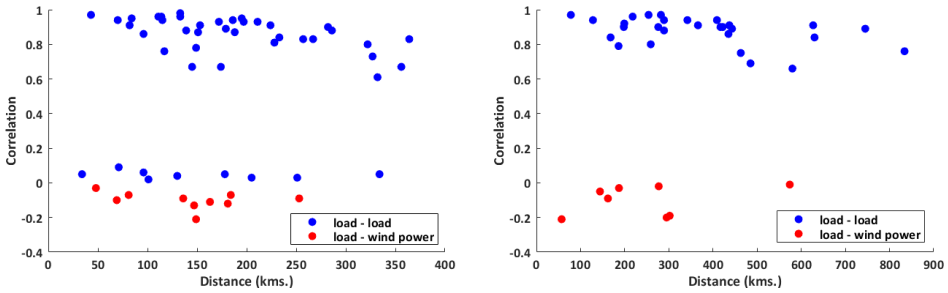


Fig. 4.22: Spatial correlation of load and wind power zone pairs under (a) *MIDATL* and (b) *WEST*

The correlation between zone pairs could vary because of changes in their installed wind power capacity. Since an approximated centroid approach is employed in this research, the exact relationship will also be influenced by the actual size of the zones. Although the correlation patterns will be affected, these patterns still provide useful information. The load-wind power zone pairs in MIDATL show significant weak and negative correlation when compared to *WEST*. Thus, in order to meet the increased demand in one load zone within the *MIDATL* market zone, a potential interconnector will benefit the system operator to supply the increased demand by wind power generated in another zone. The notion of weak correlation with the increase in distance does not hold valid for the sampled dataset. While in MIDATL, there are instances of zero correlation for the load-load pair, the zero correlation is independent of the distance between the zones. Surprisingly, WEST has no instances of zero correlation between the load zones even when the separation is ~800kms. The variations will sometime occur in the same directions and help the system, and on other times in opposite directions making load following more difficult. Although understanding the correlation of wind power output is important for the incorporation of wind, it will be more critical for determining the regulation reserves necessary as the correlation of changes in wind power and load. In such case, system operators will seek to supplement the peak demand from conventional generation sources or energy storage if available.

4.5 CONCLUSIONS

This research can be concluded with five distinct sections:

Addressing spatio-temporal correlation of load and wind power: In this research, a canonical vine based sampling methodology is developed to address spatio-temporal correlation for the high-dimensional multivariate dataset. With increased penetration of wind power into the existing grid, it was deemed vital to model the complex interdependencies introduced by wind power along with electricity load. The study revealed that chronological simulation of the multivariate dataset (load and wind power in this case) using vine copula is possible with conditional distribution calculated by multivariate copula to model the inter-spatial dependence and temporal correlation simultaneously. Such a modeling technique is realized with the developed reproducible sampling algorithm and it can be employed for power system studies (such as security assessment studies, generation and transmission expansion planning, optimal outage scheduling, stochastic unit commitment) involving a massive integration of stochastic generation.

Advantage of vine copula modeling: In stochastic optimization for multi-temporal problems, a set of scenarios is defined to describe the uncertainty associated to demand and generation at each temporal scale. Use of vine copulas facilitates in generating such scenarios for multi-temporal optimization simulations. And, the application ranges to scenarios necessary in stochastic programming, which is a critical decision tool in power systems analysis, economic dispatch, and planning problems.

Ease of computational burden: The developed sampling algorithm introduces a systematic way of reducing the original high dimensional dataset to low dimensional space while maintaining essential properties of the original dataset. With TSOs collecting a large amount of data, the future can be seen as data-centric in terms of large sized high-dimensional data with various features that can surely challenge computational efficiency. To tackle the high dimensionality and variability of data sets, the proposed model is able to ease the computational burden by employing clustering and feature extraction techniques.

Reproducing the methodology to include solar power: This research aimed at addressing spatio-temporal correlation from TSO point of view. And, that is the reason for choosing aggregated zonal load and wind power. To include solar power, this sampling methodology can be extended and reproduced if distribution feeder data is available for load, wind and solar power.

Future challenge in terms of microscale modeling: Advantage of the developed model is that wind farms with poor historical measurements or future planned wind-farms, parameters can be estimated with the spatio-temporal vine copula according to their geographical locations. However, one of the future challenges is modeling

microscale dependence using vine copula since it cannot explain the dependence within the grouped data.

REFERENCES

[Aslan & Zech, 2005] Aslan, B., & Zech, G. (2005). New test for the multivariate two-sample problem based on the concept of minimum energy. *Journal of Statistical Computation and Simulation*, *75*(2), 109-119.

[Berkhin, 2006] Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25-71). Springer, Berlin, Heidelberg.

[Brown et al., 1984] Brown, B. G., Katz, R. W., & Murphy, A. H. (1984). Time series models to simulate and forecast wind speed and wind power. *Journal of climate and applied meteorology*, *23*(8), 1184-1195.

[Cressie & Wikle, 2015] Cressie, N., & Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

[Davies et al., 1979] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224-227.

[Embrechts et al., 2001] Embrechts, P., Lindskog, F., & McNeil, A. (2001). Modelling dependence with copulas. *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*.

[Embrechts et al., 2002] Embrechts, P., McNeil, A., & Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 176-223.

[Genest & Favre, 2007] Genest, C., & Favre, A. C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, *12*(4), 347-368.

[Genest et al., 2009] Genest, C., Gendron, M., & Bourdeau-Brien, M. (2009). The advent of copulas in finance. *The European Journal of Finance*, *15*(7-8), 609-618.

[Gill et al., 2012] Gill, S., Stephen, B., & Galloway, S. (2012). Wind turbine condition assessment through power curve copula modeling. *IEEE Transactions on Sustainable Energy*, *3*(1), 94-101.

[Haghi & Lotfifard, 2015] Haghi, H. V., & Lotfifard, S. (2015). Spatiotemporal modeling of wind generation for optimal energy storage sizing. *IEEE Transactions on Sustainable Energy*, *6*(1), 113-121.

[Hagspiel et al. 2012] Hagspiel, S., Papaemannouil, A., Schmid, M., & Andersson, G. (2012). Copula-based modeling of stochastic wind power in Europe and implications for the Swiss power grid. *Applied Energy*, *96*, 33-44.

[Hartigan et al., 1979] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100-108.

[Khuntia et al., 2017] Khuntia, S. R., Rueda, J. L., & van der Meijden, M. A. (2017, August). Smart asset management for electric utilities: Big data and future. In *12th World Congress on Engineering Asset Management (WCEAM),* Brisbane.

[Khuntia et Khuntia, S. R., Rueda, J. L., & van der Meijden, M. A. (2018, July). Spatio-

al., 2018a]	temporal study for modeling high dimensional future uncertainties: Univariate to multivariate model. In *2018 IEEE PES General Meeting, Portland,* IEEE.
[Khuntia et al., 2018b]	Khuntia, S. R., Rueda, J. L., & van der Meijden, M. A. (2018). Spatio-temporal modeling using vine copula: Application to electricity load and wind power. *Electric Power Systems Research*, under review.
[Kurowicka& Cooke, 2006]	Kurowicka, D., & Cooke, R. M. (2006). *Uncertainty analysis with high dimensional dependence modelling*. John Wiley & Sons.
[Law, 2006]	Law, H. C. (2006). *Clustering, dimensionality reduction, and side information*. PhD Thesis. Michigan State University.
[Le et al., 2015]	Le, D. D., Gross, G., & Berizzi, A. (2015). Probabilistic modeling of multisite wind farm production for scenario-based applications. *IEEE Transactions on Sustainable Energy*, *6*(3), 748-758.
[Lenzi et al., 2017]	Lenzi, A., Pinson, P., Clemmensen, L. H., & Guillot, G. (2017). Spatial models for probabilistic prediction of wind power with application to annual-average and high temporal resolution data. *Stochastic Environmental Research and Risk Assessment*, *31*(7), 1615-1631.
[Louie, 2014a]	Louie, H. (2014). Correlation and statistical characteristics of aggregate wind power in large transcontinental systems. *Wind Energy*, *17*(6), 793-810.
[Louie, 2014b]	Louie, H. (2014). Evaluation of bivariate Archimedean and elliptical copulas to model wind power dependency structures. *Wind Energy*, *17*(2), 225-240.
[Maisonneuve & Gross, 2011]	Maisonneuve, N., & Gross, G. (2011). A production simulation tool for systems with integrated wind energy resources. *IEEE Transactions on Power Systems*, *26*(4), 2285-2292.
[Malvaldi et al., 2017]	Malvaldi, A., Weiss, S., Infield, D., Browell, J., Leahy, P., & Foley, A. M. (2017). A spatial and temporal correlation analysis of aggregate wind power in an ideally interconnected Europe. *Wind Energy*, *20*(8), 1315-1329.
[Melo et al., 2012]	Melo, J. D., Carreno, E. M., & Padilha-Feltrin, A. (2012). Multi-agent simulation of urban social dynamics for spatial load forecasting. *IEEE Transactions on Power Systems*, *27*(4), 1870-1878.
[Melo et al., 2014]	Melo, J. D., Carreno, E. M., Calvino, A., & Padilha-Feltrin, A. (2014). Determining spatial resolution in spatial load forecasting using a grid-based model. *Electric Power Systems Research*, *111*, 177-184
[Miettinen & Holttinen, 2017]	Miettinen, J. J., & Holttinen, H. (2017). Characteristics of day-ahead wind power forecast errors in Nordic countries and benefits of aggregation. *Wind Energy*, *20*(6), 959-972.
[Miranda & Dunn, 2007]	Miranda, M. S., & Dunn, R. W. (2007, June). Spatially correlated wind speed modelling for generation adequacy studies in the UK. In *Power Engineering Society General Meeting, 2007. IEEE* (pp. 1-6). IEEE.
[Morales et al., 2008]	Morales, O., Kurowicka, D., & Roelen, A. (2008). Eliciting conditional and unconditional rank correlations from conditional probabilities. *Reliability Engineering & System Safety*, *93*(5), 699-710.
[Nelsen, 2007]	Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

[Osborn et al., 2011] Osborn, D., Henderson, M. I., Nickell, B. M., Lasher, W., Liebold, C., Adams, J., & Caspary, J. (2011). Driving forces behind wind. *IEEE Power and Energy Magazine*, *9*(6), 60-74.

[Papaefthymiou & Kurowicka, 2009] Papaefthymiou, G., & Kurowicka, D. (2009). Using copulas for modeling stochastic dependence in power system uncertainty analysis. *IEEE Transactions on Power Systems*, *24*(1), 40-49.

[Papavasiliou et al., 2015] Papavasiliou, A., Oren, S. S., & Aravena, I. (2015, January). Stochastic modeling of multi-area wind power production. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on* (pp. 2616-2626). IEEE.

[Park et al., 2015] Park, H., Baldick, R., & Morton, D. P. (2015). A stochastic transmission planning model with dependent load and wind forecasts. *IEEE Transactions on Power Systems*, *30*(6), 3003-3011.

[Patton, 2009] Patton, A. J. (2009). Copula–based models for financial time series. In *Handbook of financial time series* (pp. 767-785). Springer, Berlin, Heidelberg.

[Pinson & Girard, 2012] Pinson, P., & Girard, R. (2012). Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, *96*, 12-20.

[Sklar, 1959] Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, *8*, 229-231.

[Shi et al., 2017] *Shi, J., Liu, Y., & Yu, N. (2017, September). Spatio-temporal modeling of electric loads. In Power Symposium (NAPS), 2017 North American (pp. 1-6). IEEE*.

[Shin et al., 2011] Shin, J. H., Yi, B. J., Kim, Y. I., Lee, H. G., & Ryu, K. H. (2011). Spatiotemporal load-analysis model for electric power distribution facilities using consumer meter-reading data. *IEEE Transactions on Power Delivery*, *26*(2), 736-743.

[Stephens, 1986] Stephens, M. A. (1986). Tests based on EDF statistics. *Goodness-of-fit Techniques*, *68*, 97-193.

[Sun et al., 2016] Sun, M., Konstantelos, I., Tindemans, S., & Strbac, G. (2016, June). Evaluating composite approaches to modelling high-dimensional stochastic variables in power systems. In *Power Systems Computation Conference (PSCC), 2016* (pp. 1-8). IEEE.

[Sun et al., 2017] Sun, M., Konstantelos, I., & Strbac, G. (2017). C-Vine copula mixture model for clustering of residential electrical load pattern data. *IEEE Transactions on Power Systems*, *32*(3), 2382-2393.

[Tascikaraoglu & Sanandaji, 2016] Tascikaraoglu, A., & Sanandaji, B. M. (2016). Short-term residential electric load forecasting: A compressive spatio-temporal approach. *Energy and Buildings*, *111*, 380-392.

[Tastu et al., 2013] Tastu, J., Pinson, P., & Madsen, H. (2013). Space-time scenarios of wind power generation produced using a Gaussian copula with parametrized precision matrix. *Tech. Univ. Denmark, Tech. Rep.*

[Tibshirani et al., 2001] Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411-423.

[Torres et al., 2005] Torres, J. L., Garcia, A., De Blas, M., & De Francisco, A. (2005). Forecast of hourly average wind speed with ARMA models in Navarre (Spain). *Solar Energy*, *79*(1), 65-77.

[Wall et al., 2003] Wall, M. E., Rechtsteiner, A., & Rocha, L. M. (2003). Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis* (pp. 91-109). Springer, Boston, MA.

[Wan et al., 2003] Wan, Y. H., Milligan, M., & Parsons, B. (2003). Output power correlation between adjacent wind power plants. *Journal of Solar Energy Engineering*, *125*(4), 551-555.

[Wang et al., 2018a] Wang, Y., Zhang, N., Kang, C., Miao, M., Shi, R., & Xia, Q. (2018). An efficient approach to power system uncertainty analysis with high-dimensional dependencies. *IEEE Transactions on Power Systems*, *33*(3), 2984-2994.

[Wang et al., 2018b] Wang, Z., Wang, W., Liu, C., Wang, Z., & Hou, Y. (2018). Probabilistic Forecast for Multiple Wind Farms Based on Regular Vine Copulas. *IEEE Transactions on Power Systems*, *33*(1), 578-589.

[Web6, 2018] https://www.pjm.com/library/reports-notices/rtep-documents/2016-rtep.aspx

[Web7, 2018] https://www.pjm.com/committees-and-groups/subcommittees/irs/pris.aspx

[Web8, 2018] https://www.pjm.com/renewables/default.html

[Web10, 2018] https://irena.masdar.ac.ae/gallery/#gallery

[Web11, 2018] https://maps.nrel.gov/wind-prospector/

[Web12, 2018] https://people.csail.mit.edu/rameshvs/content/gmm-em.pdf

[Wei et al., 2017] Wei, H. U., Yong, M. I. N., Yifan, Z. H. O. U., & Qiuyu, L. U. (2017). Wind power forecasting errors modelling approach considering temporal and spatial dependence. *Journal of Modern Power Systems and Clean Energy*, *5*(3), 489-498.

CHAPTER 5

SPATIO-TEMPORAL MODELING FOR STATIC SECURITY ASSESSMENT

5.1 INTRODUCTION

This chapter aims at answering the fourth research question Q.4., which deals with the use of the developed model in chapter 4 for a risk-based security assessment of transmission line overloading risk,

- *Does the consideration of spatio-temporal dependence of load and wind power prove beneficial to quantify the risk of overloading transmission lines?*
- *How does the correlation impact the risk values of line overload?*
- *How do the risk values of individual lines or the entire system enable the system operator to assess system operation conditions?*

The content of this chapter is based on research paper [Khuntia et al., 2018c] and it aims at validating the developed spatio-temporal model for static security assessment. The security level of a power system is determined by the likelihood and severity of security violation. In this chapter, a risk-based security assessment (RBSA) is performed on transmission line overloading considering spatio-temporal dependency of load and wind power penetration using vine copula. The research work is inspired by growing WPPs and uncertainty in load growth as we look into a low-carbon future power system. Location of WPPs and load sites are not always close by, hence, the transmission of energy puts a burden on existing grid infrastructure. This unwanted burden necessitates transmission lines to operate more and more frequently close to their operating limits. As the future sees a massive integration of wind power, the risk of transmission line overloading cannot be avoided. Complexity in terms of inter-spatial dependence and temporal correlation of wind power and load impose challenging operational threat in terms of transmission line overloading to system operators and grid planners.

The aim is to adopt a probabilistic approach to extract information on operating conditions and use appropriate metrics to identify the suitable model to give sufficient

confidence with respect to assessment of the future operation of the system for a given generation capacity, load or market scenario. Therefore, this chapter addresses the advantages of modeling load and wind power as a joint probability distribution using the canonical vine to address the spatio-temporal dependence. Probabilistic AC optimal power flow is performed on a modified IEEE 39-bus system with significant wind penetration and real load and wind power data from a U.S. utility. The data is mapped onto the test-case to achieve realistic results. Load flow calculation can help in performing steady-state voltage and overload evaluations for post-disturbance system conditions. In this research, the probability of line overload is calculated from load flow and the severity function describes the risk of line overloading. Two case studies depicting future operating conditions of massive wind power penetration with reduced fossil fuel and nuclear power generation are considered. Simulation results prove the advantage of addressing spatio-temporal dependency to quantify the overload risk index, which is treated as a security indicator.

The rest of this chapter is organized as follows: sub-section 5.2 presents a background on spatio-temporal modeling in risk-based security studies and an overview of probabilistic approach in transmission line overloading studies. Sub-section 5.3 discusses database generation and preparation for RBSA studies. Sub-section 5.4 discusses the two case studies along with result analysis. Finally, sub-section 5.5 concludes this chapter.

5.2 BACKGROUND

The electric power system infrastructure is subjected to increasing stress due to fundamental changes in both *generation* and *demand* side. A larger picture of such stress is encountered in the form of transmission loading pattern.

- On the *generation* side, integration of stochastic generation sources in the form of variable RES is a challenging task for TSOs. Among the RES, wind power has gained significant attention among TSOs because of three reasons, namely,
  - large WPPs can be connected to bulk power system at the transmission level,
  - large size WPPs are being built/planned in regions with high potential for wind energy and TSOs have to facilitate their integration, and,
  - TSOs are more and more limited by local constraints to build new transmission infrastructure in time.

  Over the last decade, electricity production from WPPs has reached considerable levels in various parts of Europe and the U.S. [TechRep, 2010, TechRep, 2015, Milligan et al., 2015]. Expansion of WPPs in terms of farm-size and unit capacity is significant at transmission level and its integration possesses challenges for TSOs both in terms of operation and security [TechRep, 2016]. This study

assumes massive wind power penetration occurs at transmission level and is deemed quite challenging for TSOs in planning and decision making. To have a clear view from TSO perspective, DERs like solar or storage is not considered in this study since it is more prominent at the distribution level.

- On the *demand* side, the advent of new technologies and growing number of variable generation sources at the distribution level is challenging itself. The traditional power system network was designed for the passive load without any plans for communication and digitalization.

In such cases, the system is being asked to perform in ways and in a context for which it was not designed, eventually resulting in performance under increasing stress. The European electricity grid, for example, a meshed and complex network, was built decades ago and has provided highly reliable electricity till date. In an event of a transmission asset failure, system security is certainly under threat and it affects system dynamics which might increase the likelihood of line overload, low voltage or even voltage collapse. As the security of supply is at risk, one of the ways to address the issue is to perform an overload analysis. Literature study reveals some commonly used indices like overload, cascading overload, low voltage and voltage instability [Ni et al., 2003a]. To account for the mentioned developments in reliability management, TSOs need to re-evaluate the system security [Khuntia *et al*., 2016a]. One such way is identified in this research, which constitutes dependence studies between load and wind power as a necessary modeling aspect. A discussion on the need for spatio-temporal modeling of load and wind power is presented followed by a probabilistic model for line overload and calculation of system risk index.

5.2.1 NEED FOR SPATIO-TEMPORAL MODELING OF LOAD AND WIND POWER

The location of WPPs and load sites are not always close-by such that the non-dispatchable sources can be easily managed or curtailed. This spatial diversity imposes a burden on the TSOs who have to operate the existing infrastructure with uncertainty from both ends, i.e., load and wind power. The participation of WPPs into existing grid is different from conventional generators in terms of location and output generation, which is uncertain and variable. It adversely affects day-ahead operational planning decisions that introduce a level of risk for TSOs. For example, variation in wind power hampers power system operation in real-time when WPPs are unable to deliver the required reserve capacities in real-time. Embedding of WPPs raise concerns in terms of planning and upgrading of existing infrastructure in terms of size, location and distribution of WPPs [Xie et al., 2011]. Though achievements have been made in terms of the accurate forecast of future load and wind power generation, there are other vital concerns corresponding to wind power such as spatio-temporal dependency, variability, non-normality, non-stationarity, non-dispatchable (unless there is adequate storage)

and seasonal patterns to name a few. Wind speed is temporally correlated at one location and for different locations wind speed is both spatially as well as temporally correlated [Osborn et al., 2011]. A statistical space-time model considering terrain, wind speed and direction was proposed in [Xie et al., 2014]. Spatial dependence of wind for transmission line overloading is found in [Li et al., 2015]. It is to be noted that spatial dependency within multiple wind-farms as studied in [Morales et al., 2010a] is out of the scope of this research. Addressing temporal correlation for both load and wind power can be found in [Usaola, 2010, Morales et al., 2010b, Abdullah et al., 2013]. To the best of our knowledge, no study has reported the importance of addressing spatio-temporal dependency of load and wind power till date. It is important to address the spatio-temporal dependency from the transmission point of view when TSOs are facing stagnant expansion planning because (i) weak transmission capacity causes reduced integration capacity of WPPs (ii) redundant transmission capacity results in resource waste.

5.2.2 BACKGROUND ON POWER SYSTEM RISK ASSESSMENT STUDIES

Bulk interconnected power systems with distributed and geographically isolated generators and demand centers constitute a majority of the power network. With increasing RES and other DERs, the present day power systems are dynamic in nature with network topology changing more and more frequently with the change in demand. As nuclear power plants and fossil fuel plants are phased out to include more RES in form of massive wind power penetration, uncertainty in load demand actuates the power network to operate at loading limits; thus making it susceptible to blackout under minor/major disturbances. In order to operate the power system economically, the state of the system has to be identified as secure/insecure. Security assessment studies aim to balance the system security as well as the economy for power system operation. Power system security can be divided into two, namely, static and dynamic security. Static security analysis targets steady-state post-disturbance conditions, namely it is assumed that the system reaches operating equilibrium after a disturbance and it is checked whether system limits are violated. Dynamic security analysis targets system stability after a disturbance, and therefore it is investigated whether the system can reach a new state of equilibrium after a disturbance. Sometimes static security reliability assessment can be referred by literature as adequacy assessment, and dynamic security reliability assessment can be met simply as security reliability assessment. A classification hierarchy is shown in Fig. 5.1. Another way of classification is based on assessment techniques. The three schemes of security assessment are, namely, (i) deterministic security assessment that is more traditional and considers a set of most credible contingencies resulting in high operating costs, (ii) probabilistic security assessment that considers probabilistic indices LOLP (Loss of load probability for likelihood of events) and EENS (Expected energy not supplied for both likelihood as well

as severity of events), and (iii) risk-based security assessment (RBSA) that considers both likelihood and severity of events allowing the power system to operate closer to or beyond its limits. Further, RBSA is categorized as; (i) static RBSA that considers the risk of overload and voltage violations and (ii) dynamic RBSA that considers the risk of instability in terms of voltage and swing transient. The deterministic security assessment methods get usually many conservative results in overload analysis. With recent advancement and more adoption of risk theory in power system, the risk assessment method is gradually evolving and acknowledged.

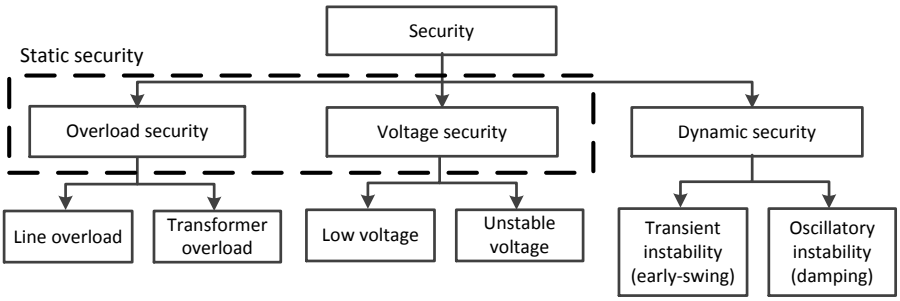


Fig. 5.1: Classification of power system security [McCalley, 2005]

This research focuses on static RBSA in terms of transmission line overload violation identified as a risk by developing a spatio-temporal modeling technique using vine copula. The risk assessment method is proposed for operational planning. Such a modeling framework will facilitate TSOs to improve short-term operational planning decisions while coordinating long-term grid development plans (e.g., integrating more WPPs) and increase the security margins and decrease the idle capacity reserves and related costs. In regard to transmission line overload, literature study reveals the study of static RBSA with 'N-1' contingency [McCalley et al., 1999], risk visualization using Poisson distribution [Ni et al., 2003a, Ni et al., 2003b], online static RBSA with forecasted operating condition [Arya et al., 2006], possibility and severity of risk occurrence [Dai et al., 2001].

Risk assessment by computing risk indices based on over-limit probability and severity to recognize system weakness more realistically is entailed in this study. In power system studies, risk assessment is performed in four steps [Ni et al., 2003b]:

i. describe an index that represents system risk,
ii. select a system state and calculate its probability,
iii. evaluate the outcome of the system state, and
iv. calculate the risk index.

For any operating condition, the risk associated with ith state S_i at time t is calculated for all possible values of probability and severity associated with it can be written as [Ni et al., 2003b],

$$Risk_t = \sum_{i=1}^{n} Prob_t(S_i) Sev_t(S_i) \tag{5.1}$$

where $Prob_t(S_i)$ is the state probability, $Sev_t(S_i)$ is the associated severity of state i, and n is the total number of system states. The line overload probability can be measured by the probability mass function of line flows. Reference [Ni et al., 2003b] categorized severity function into three types, namely, discrete, continuous and percentage of rating violation severity function. The concept of severity functions has been used in recent studies [Xie et al., 2014] to investigate transmission line overloading in power systems with wind and load-power generation correlation.

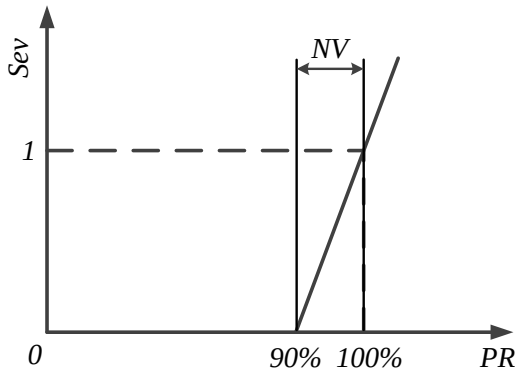


Fig. 5.2: Linear severity function of line overload (PR: Percentage of Rating, NV: Near Violation)

Fig. 5.2 shows the severity function as a continuous function of real line power as a percentage of rating (PR) and severity (Sev), written as a function of j-th branch apparent power flow F_j and rated apparent power flow F_j^{max}

$$PR = \frac{F_j}{F_j^{max}} x100\%$$
$$Sev_j = \begin{cases} 0, & (PR/100) < 0.9 \\ 10(PR/100) - 9, & (PR/100) \geq 0.9 \end{cases} \tag{5.2}$$

When the line flow exceeds 90% of its rating, near violation for overload takes place which increases linearly as power flow exceeds the limit. Severity function signifies the extent of security violation and helps in quantifying the severity. In the case of transmission line overload, severity function is defined for each line and the apparent power flow in the line determines the associated risk. For each line, the severity function defined in equation 5.2 evaluates to 1 at the deterministic limits, i.e., 100% of

line apparent power flow rating. It is to be noted that higher risk values do not necessarily indicate a larger interruption of security of supply and vice-versa. For example, a 130% overload of a transmission line might have higher risk value than a 130% overload of another transmission line but it does not necessarily lead to increasing the monetary penalty incurred due to unavailability of the security of supply.

5.2.3 SCOPE OF THIS RESEARCH

It was learned that despite the sustainable features of wind power, the spatial distribution of WPPs and load sites contribute to the change in operating conditions of the power system. The associated risk of change in operating conditions must be quantified and fully explored. In this study, calculation of risk indices is accomplished with probabilistic AC optimal power flow, which is used for the analysis of joint probability distributions resulting from the vine copula sampling algorithm. The choice for optimal power flow (OPF) is to determine a steady-state operating point that optimizes the use of system resources while maintaining operational constraints such as voltage regulation, generator limits and line flow limits. This proves helpful in static security assessment studies. OPF ensures that for every load scenario there is an optimal amount of power generated and restricted into a set of constraints to ensure its operational success. OPF is formulated as a nonconvex, nonlinear constrained optimization problem and a variety of algorithms have been used to solve it. In this research, one of the widely used optimization technique “interior-point method” is used for solving the OPF problem. The interior-point method is famous among power flow researchers due to their robustness and convergence characteristics. Another advantage of using the interior point method is that it is able to solve both linear and nonlinear problems. The stopping criterion is generally based on the coefficient of variation when it is less than a certain value.

As such, the accuracy of computed risk indices depends on the accuracy of power flow results. Such a power flow analysis provides a suitable tool to study the relationship between the bus injection fluctuations in terms of near violation and overload in system operation state. Presence of WPPs introduces uncertainty in the normal operating state. Hence, the line power flows influenced by the stochastic generation introduce uncertainty in bus injections which might lead to transmission line overload. In addition to wind power, uncertainty in load growth contributes equally towards the problem.

The key contributions of this research can be listed as:

- A novel attempt to assess transmission line overloading risk while addressing spatio-temporal dependency using vine copula is made in this study. The sampling algorithm uses real-life data from nineteen spatially distributed load and two wind power zones spanning three years horizon from a U.S. utility to model the joint probability distribution.

- The sampled output is mapped onto a modified IEEE 39-bus system to achieve realistic results based on two case studies representing future scenarios of high wind power penetration with reduced fossil fuel and nuclear power generation. RBSA is performed on transmission line overloading by performing probabilistic AC optimal power flow and considering spatio-temporal dependence of load and wind power. Risk quantification of overloading is achieved as a product of probability and severity of overload.

5.3 DATABASE GENERATION AND PREPARATION

For this study, publicly available load and wind power data are taken from U.S. regional transmission operator [Web4, 2018]. Aggregated zonal load data (nineteen numbers) and wind power data (two numbers) spanning three years with hourly resolution is used in this study. To study the spatial correlation, geographical coordinates of zones are needed. Since the exact coordinates are treated confidential, an approximated weighted centroid has been defined to locate the 'center' of load zones and wind power generated zones. As in chapter 4, the three market zones are *MIDATL, WEST*, and *SOUTH*. To visualize the complexity, scatter plot with marginal histograms of four load and one wind power zone under the *WEST* zone is shown in Fig. 5.3. The marginal histograms (in the diagonal) reveal non-Gaussian nature while the scattered plots reveal the non-linear dependencies and suggest a weak correlation as well.

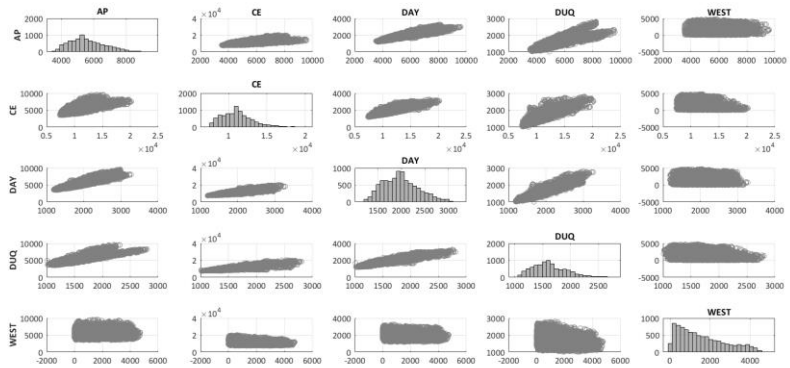


Fig. 5.3: Scatter plot with marginal histograms of original data of four load zones and one wind power zone

The proposed risk assessment is validated using the modified IEEE 39-bus system with a base value of 100 MVA. A modified version of the IEEE 39-bus system with WPPs and updated conventional generation capacities is considered in this study. The original test case consists of 39 buses, 10 conventional generators and two WPPs at bus 34 and 37, and 46 transmission lines with net demand in the network of $6254.23MW$ and net generation capacity of $7367MW$. The modified topology with wind farms and divided zones according to real data is shown in Fig. 5.4. Next step is to map the real-life data

onto the IEEE 39-bus system by scaling the real-life data to match test-case parameters. A scale ratio is defined as,

$$Scale\ Ratio = \frac{\text{maximum coincident peak demand of real} - \text{life data}}{\text{sum of active power demand across all buses in test} - \text{case}} \quad (5.3)$$

Taking into account large wind power penetration, generation capacity of conventional generation sources (such as nuclear and fossil fuels) is lowered and is met by wind power generation [Papaefthymiou & Dragoon, 2016, Web8, 2018]. Generation cost data is obtained from [Bukhsh et al., 2016]. In addition, there is no topological change in terms of addition of new transmission lines, which gives us the option to assess the overloading risk on existing grid network. The two cases studied in this research are:

Case I: The first case considers a 7.5% increase in system load to $6725.05MW$. The total generation is $8167.87MW$ including $2640MW$ of wind at bus 34 and 37 respectively. For the *MIDATL*, the net load is $2355.13MW$ and net generation is $4877.871MW$ including $1760MW$ of wind power. For the *WEST*, the net load is $3752.1MW$ and net generation is $3500MW$ including $880MW$ of wind power. For *DOM*, there is only an increase in demand to $617.82MW$.

Case II: The second case considers a 9% increase in system load to $6817.11MW$. The total generation is $7712.86MW$ including $2760MW$ of wind. For the *MIDATL*, the net load is $2447.19MW$ and net generation is $3792.861MW$ including $1760MW$ of wind power. For the *WEST*, the net load is same as *Case I* and net generation is $3360MW$ including $1000MW$ of wind power. For *DOM*, net load and generation are the same as in *Case I*.

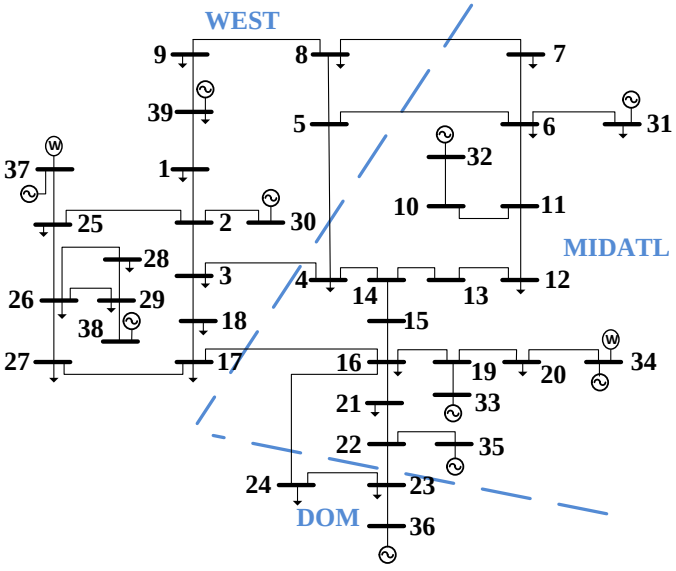


Fig. 5.4: Modified IEEE 39-bus system divided into three market zones (*MIDATL*, *WEST*, *DOM*)

For the two cases, a comparative analysis of uncorrelated and correlated load and wind power samples is performed. In the case of uncorrelated samples, synthetic wind power data using Weibull distribution and random samples of load taking the load at each bus as mean values is considered. This also serves as benchmark data to assess the correlated samples. The formula for the probability distribution function (PDF) of wind power is expressed mathematically as:

$$f(v;\gamma,\alpha) = \frac{\gamma}{\alpha}\left(\frac{v}{\alpha}\right)^{\gamma-1} exp\left\{-\left(\frac{v}{\alpha}\right)^{\gamma}\right\} \qquad (5.4)$$

$$v > 0, \gamma > 0, \alpha > 1$$

where, v is the wind speed, γ is the shape parameter and α is the scale parameter. Though α significantly depends on wind farm location, we consider a single value as the sole intention of this research work is to generate aggregated zonal wind power and thereby not considering different wind farm sizes. For the uncorrelated study, $\gamma = 2$ and $\alpha = 11$ are used to generate uncorrelated wind power for the probabilistic AC OPF [Carrillo et al., 2014]. Similarly, random load samples are generated from mean values of load at each bus [Bukhsh et al., 2016]. Technically, generated samples of load and wind power are uncorrelated and a correlation plot for *Case I* is shown in Fig. 5.5.

For generating correlated samples, real data is modeled using Algorithm 4.3 to generate a joint normal distribution with correlated samples. The correlation plot for *Case I* is shown in Fig. 5.6. Inferencing the plot suggests both positive and negative correlation. The correlation coefficients -0.2675, 0 and 1.0 represent slightly negative correlation, perfectly uncorrelated and perfectly correlated. It is important to understand the significance of correlation coefficients. If load and wind power generation were positively correlated, they would tend to increase and decrease at the same time, and adding wind would help the load following task of the power system. On contrary, if the correlation were negative, the wind would tend to decrease when load increases (and vice versa) and this would require more from the load following units in the system. The presence of a negative correlation between load and wind power is of physical significance. It explains the need to balance out the wind power fluctuations in different zones with corresponding load fluctuations to maintain a steady supply.

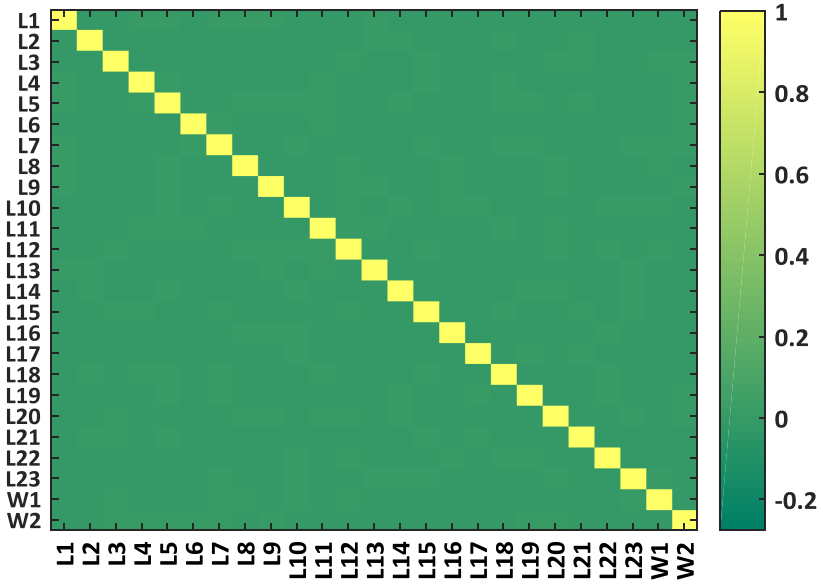


Fig. 5.5: Correlation plot for the uncorrelated load (L1…L23) and wind power (W1,W2) for *Case I*

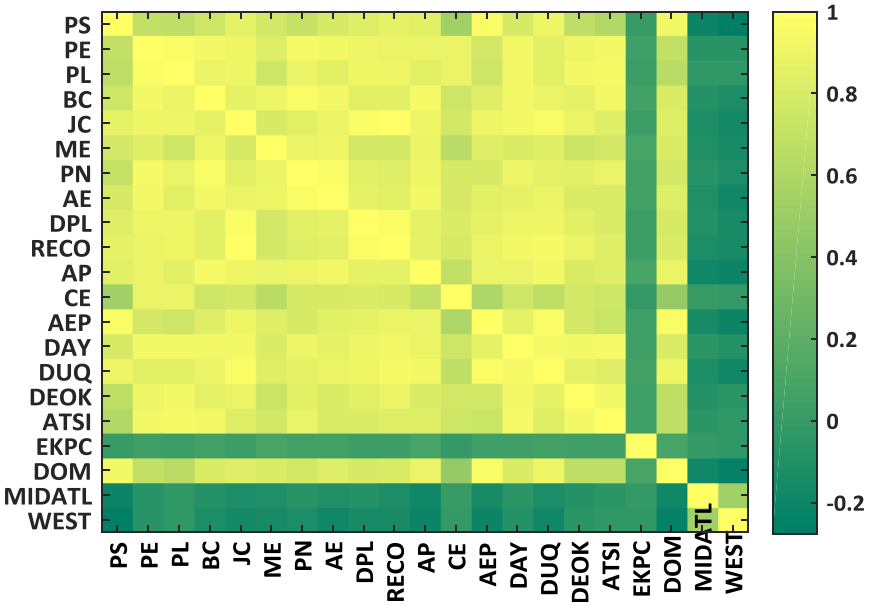


Fig. 5.6: Correlation plot for the correlated load (L1…L23) and wind power (W1,W2) for *Case I*

5.4 RESULTS AND DISCUSSION

For the inter-spatial dependencies and temporal correlations between load and wind power, two extremes of *completely uncorrelated* and *completely correlated* samples are considered. Such extreme scenario evaluations will enable us to gain insight into the importance of considering correlation for assessing transmission line overloading risk. Thus, for each *Case I* and *II*, there are two cases of uncorrelated and correlated samples. All modeling work is performed in MATLAB (version 2017b) environment using the Matpower package [Zimmerman et al., 2011] on an Intel Core i7 with 8 cores and 8GB RAM. Matpower's deterministic power flow is used to implement the probabilistic OPF. Probabilistic OPF is built based on a deterministic one because the OPF that Matpower uses is a deterministic OPF. In general, the function of deterministic power flow study can be stated as,

$$z = g(x) \tag{5.5}$$

where x is the vector of input variables which includes active power injection P_i at each bus, reactive power injection Q_i at each PQ bus and voltage magnitude V_i at each PV bus and slack bus; z is the vector of output variables which include bus voltage V_i at each PQ bus, bus angle θ_i (except slack bus), branch active power flow P_{ij}, reactive power flow Q_{ij}, and apparent power flow S_{ij}. For the probabilistic power flow problem, the input random variables $x_1, \dots, x_K$ are probabilistic distributions of P_i and Q_i. When wind power is included in probabilistic power flow, an additional random input variable is introduced as the wind power of the WPPs. The output information is probabilistic distributions of V_i , θ_i , P_{ij}, Q_{ij}, and S_{ij}.

The problem of OPF is the allocation of given load amongst the generating units in operation so that the overall cost of generation is minimum. In OPF, the entire set of equality and inequality constraints, all the necessary and sufficient conditions of control parameters etc. must be satisfied thoroughly. The objective function can take various forms such as fuel cost, transmission losses, and reactive sources allocation. The objective function of interest in this research is the minimization of the total production cost of scheduled generating units. Various techniques have been proposed to solve the OPF problem, for example, non-linear programming, quadratic programming, linear programming, and interior point methods. In this research, the interior point method is used to solve the OPF problem by using Matpower Interior Point Solver (MIPS). For a thorough understanding of MIPS, readers are referred to [Zimmerman & Wang, 2016]. A flowchart showing data collection, dataset preparation, generating correlated samples using Algorithm 4.3 described in chapter 4 and running power flow with the calculation of risk indices is shown in Fig. 5.7.

All lines are monitored for overload risk though special attention is given to line connecting WPPs and the next nearest bus to rest of the grid, i.e., line 20-34 and line 25-37 for the modified IEEE 39-bus system. Both the lines have a rated capacity of 900MVA. The impacts of load and generation correlation on line overload risk are studied. It can be understood that fluctuation of high wind power can be easily compensated by the grid, provided it is distributed among the strong lines connecting to immediate load sites or the presence of a conventional generator bus that can regulate its generation depending on the needs of high or low wind power generation.

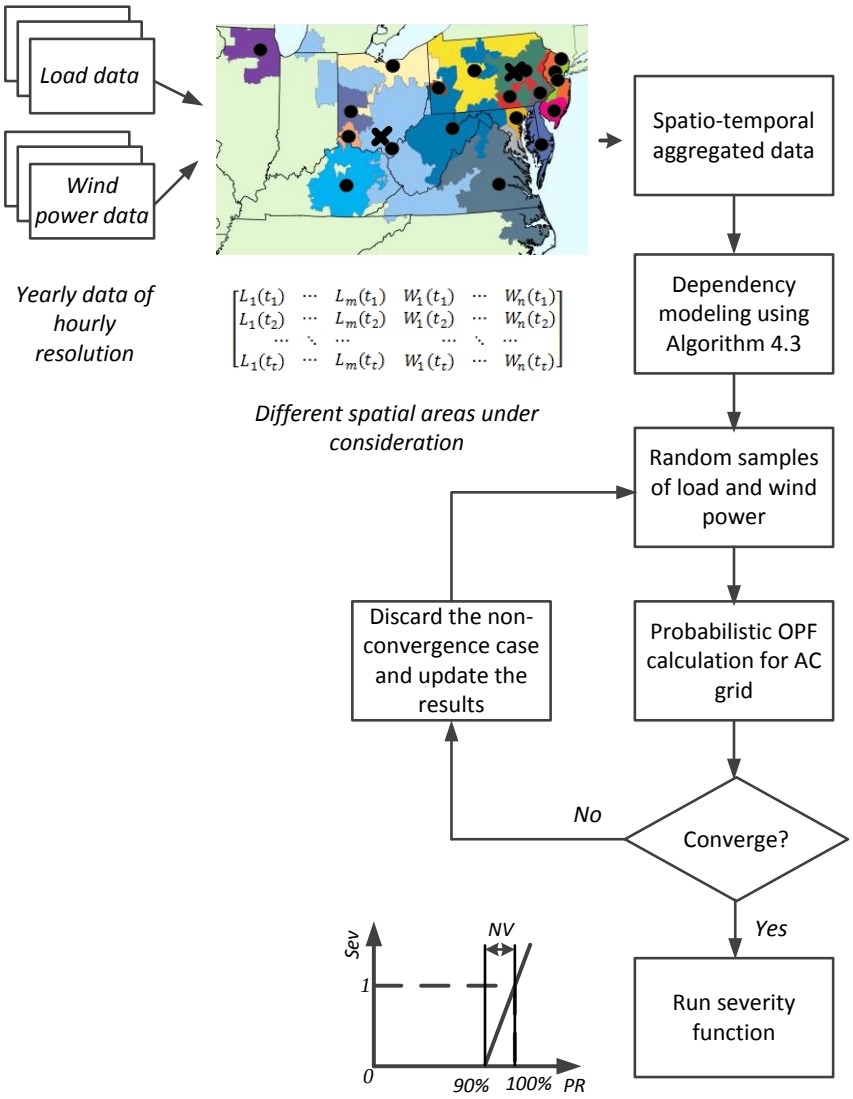
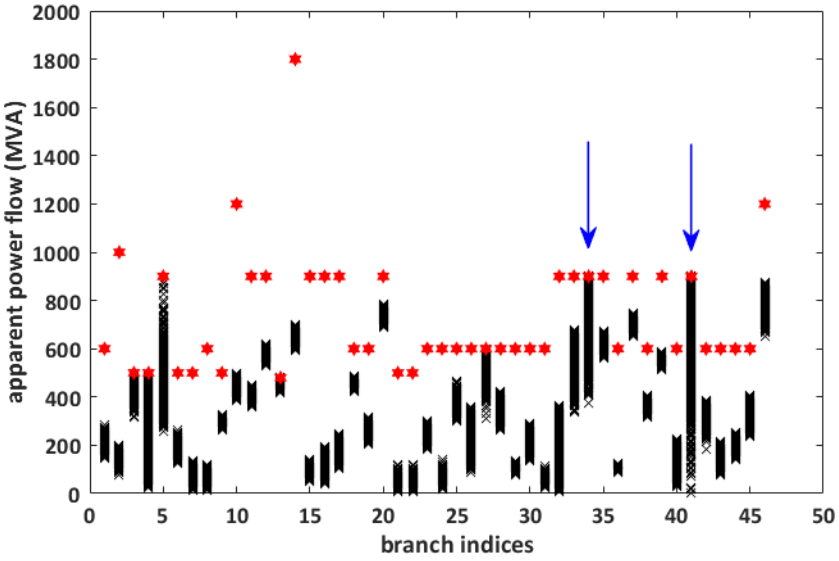


Fig. 5.7: Flowchart showing the computation of severity function starting from data collection to running power flow. The map shows the control areas of PJM with load (L_1…L_m) and wind power (W_1…W_n) zones' for *t* time frames and approximated load centroid (●) and wind power centroid (**X**)

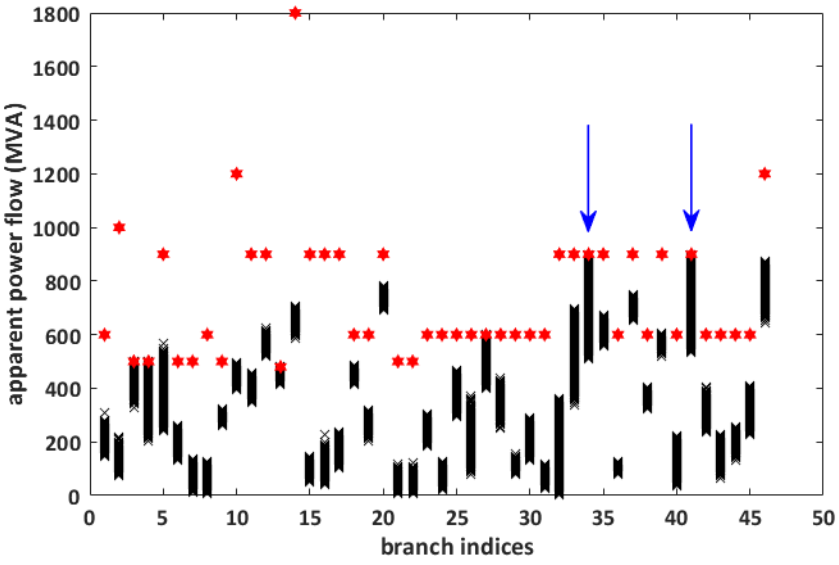
5.4.1 CASE I

In *Case I*, we consider WPPs at bus 34 (1760MW) and bus 37 (880MW) to compensate the reduced conventional generation as well as load growth. Running probabilistic AC load flow, Fig. 5.8(a) and Fig. 5.8(b) show the loading of all the lines for uncorrelated and correlated load and wind power samples. Branch indices correspond to Matpower branch indices. The figures give an overview of most loaded lines and we focus on lines 20-34 and 25-37 (indices 34 and 41) as shown in Fig. 5.9 and Fig. 5.10. Because the number of convergence changes for each power flow case, the probability of occurrence vs. apparent power flow of line was mapped. A closer look at the figures reveal the advantage of considering spatio-temporal dependence in terms of line overloading occurences.

Fig. 5.9 shows a considerable decrease in line overloading risk where the probability of overload during both near violations as well as overload is less than halved. On the other hand, Fig. 5.10 reveals not much success in considering spatio-temporal dependency. Loading of line 25-37 at 100% is marginal. This could be understood as the net load being higher than generation capacity. To understand the risk index of the overall system, Table 5.1 shows the risk indices for *Case I*. Recalling from equation 5.1, risk indices correspond to values when the lines are overloaded by 90% or more of their rated capacity. The overload risk index is measured by the probability of line overload and corresponding severity. Line 20-34 which connects 1600MW of wind power to rest of the grid shows a remarkable decrease of line overload by more than 50% when the correlation is considered. Similarly, overload risk of line 25-37 decreases by 12.3%. Since the two lines are considered vital when connecting the massive WPPs into rest of the grid, such a decrease in overload risk can be considered beneficial in comparison to the construction of a new line in the same corridor. For the entire system, there is a decrease in the overload risk index from 4.6010 to 3.7840.



(a)



(b)

Fig. 5.8: Line apparent power flow (MVA) vs. branch indices after running OPF for (a) uncorrelated load and wind power in *Case I*, (b) correlated load and wind power in *Case I*. Red marks represent the maximum line capacity. Blue arrows represent the line index under consideration.

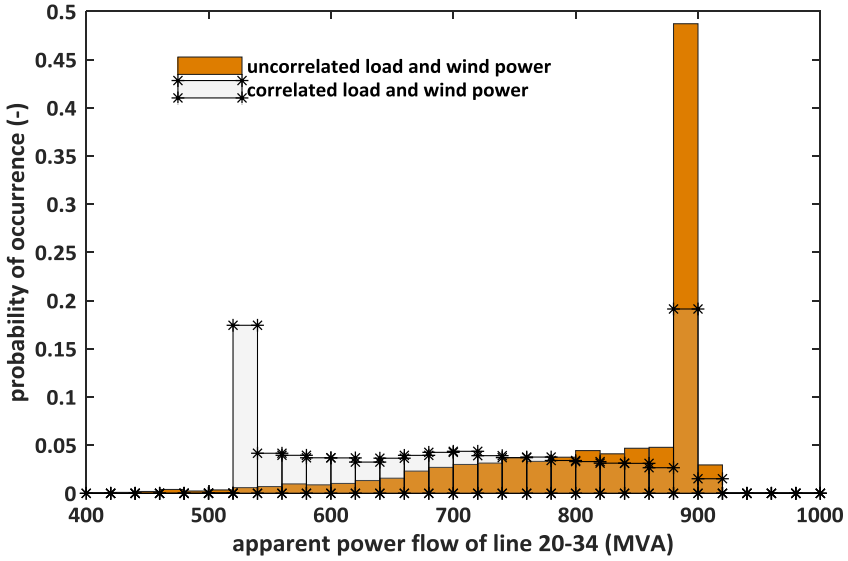


Fig. 5.9: Empirical probability distribution of power flow of line 20-34 (index 34) for uncorrelated and correlated load wind power for *Case I*. Line rating of line 20-34 is 900MVA.

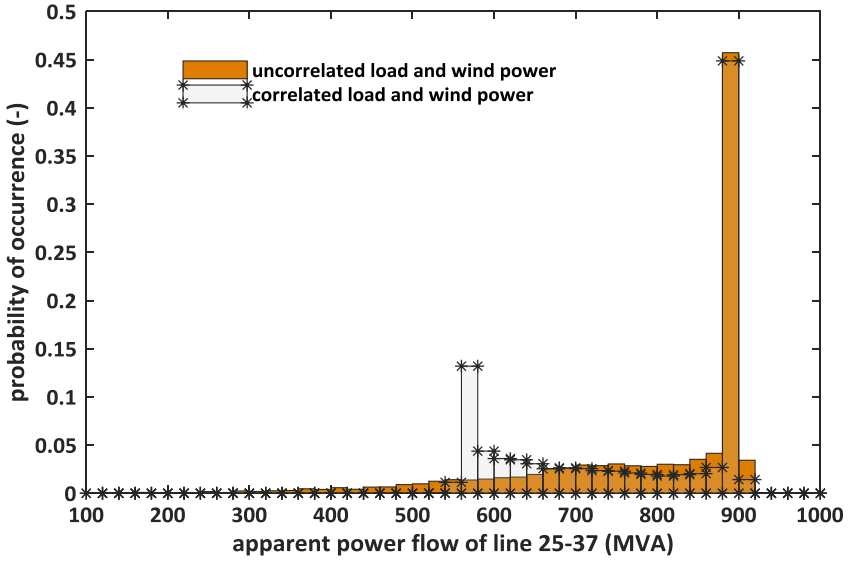


Fig. 5.10: Empirical probability distribution of power flow of line 25-37 (index 41) for uncorrelated and correlated load wind power for *Case I*. Line rating of line 25-37 is 900MVA.

If we consider the generation facility in *WEST*, influx of 880MW of wind at bus 37 is accompanied with other generation sources at bus 37, 30 and 39 which is responsible for the heavy loading of lines 2-3, 2-25 and 25-37. The risk index of line 25-37 fairly decreases with spatio-temporal modeling. Moreover, variation in wind production is compensated by already available generation at bus 37 that accounts for a slightly low-risk index. Net risk indices of the three lines decrease by 10.38% with consideration of correlation. The other two lines that are heavily loaded are 6-11 and 16-19. The risk index of 6-11 remains more or less constant with or without considering correlation. It is to be noted that a new load site at bus 6 and an increase of load at buses 7 and 31 without any changes in the generation can be considered as responsible for such overloading. One proposed solution for such case is building an interconnection between bus 34 and 11 or other nearby buses. Overloading of line 16-19 is explained by the addition of WPPs at bus 34 accompanied by generation at bus 33. Consideration of correlation leverages overloading which decreases by 25% and it can be considered advantageous for TSOs.

Table 5.1: Comparison of line overload risk indices for *Case I*

Index	Line	Uncorrelated	Correlated
3	$\mathbf{2-3}$	0.7688	0.7075
4	$\mathbf{2-25}$	0.6024	0.5331
13	$\mathbf{6-11}$	0.9841	0.9815
27	$\mathbf{16-19}$	0.9571	0.7127
34	$\mathbf{20-34}$	0.6746	0.3118
41	$\mathbf{25-37}$	0.6128	0.5374
	Total	$\mathbf{4.6010}$	$\mathbf{3.7840}$

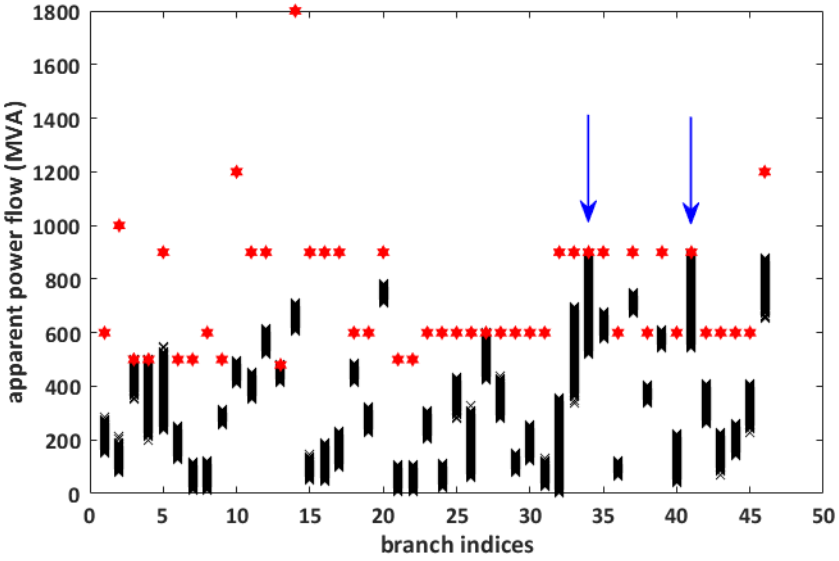
An extensive suite of 30000 Monte Carlo simulation (MCS) was performed. The choice of sample size is kept the same as in vine copula sampling from chapter 4. MCS is one way to solve probabilistic OPF which often serves as accuracy reference. MCS firstly samples the random variables and then for each sample a load flow case is solved to obtain all states. Based on the load flow results of all samples, scenarios are generated randomly from PDF. Table 5.2 compares the number of samples converging in the probabilistic OPF problem. The number of correlated cases converging is on a higher side as compared to uncorrelated samples. This is certainly helpful as the empirical probability distribution of power flow of lines contains a higher number of feasible states.

Table 5.2: Comparison of uncorrelated and correlated samples converging in *Case I*

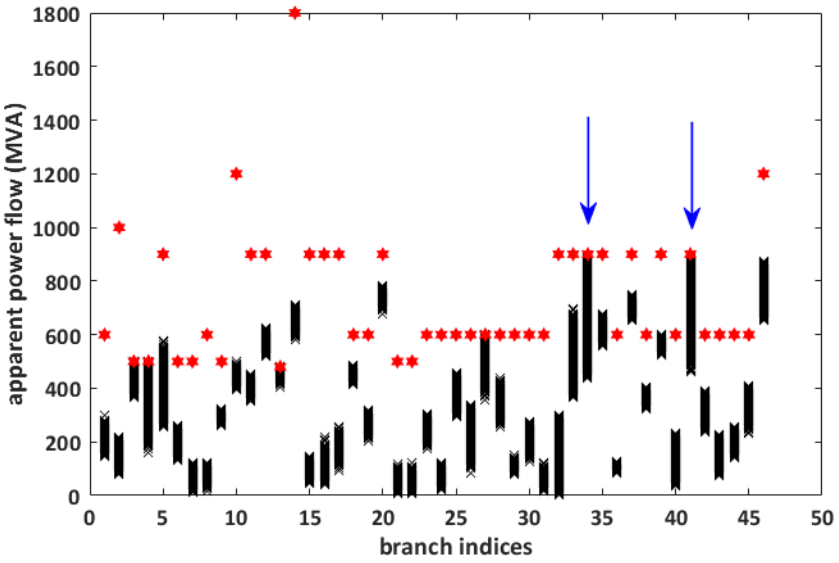
Total samples	Uncorrelated	Correlated
30000	15914	28493

5.4.2 CASE II

Compared to *Case I*, *Case II* considers a slight increase in net wind power penetration while a slight decrease in net system load. WPPs of 1760MW at bus 34 and 1000MW at bus 37 are used to compensate for lowering of conventional (nuclear and fossil fuel) generation units and load growth. Numerically, a 1.5% equivalent increase of wind power penetration and 5.5% decrease in overall generation compared to *Case I* is considered for probabilistic AC load flow. Fig. 5.11(a) and Fig. 5.11(b) shows the loading of all the lines after running the power flow. A comparative figure showing the loading of lines 20-34 and 25-37 is shown in Fig. 5.12 and Fig. 5.13. From Fig. 5.12, it is evident that the loading of line 20-34 at 100% significantly decreases when spatio-temporal dependency is considered. A remarkable decrease in loading proves the advantage of addressing correlation. However, Fig. 5.13 shows the heavy loading condition of line 25-37 at slightly higher than 100%. Though there is a decrease in line loading at nearly 100%, an increase in wind power generation in *WEST* still does not compensate for the high zonal net load.



(a)



(b)

Fig. 5.11: Line apparent power flow (MVA) vs. branch indices after running OPF for (a) uncorrelated load and wind power in *Case II,* (b) correlated load and wind power in *Case II*. Red marks represent the maximum line capacity. Blue arrows represent the line index under consideration.

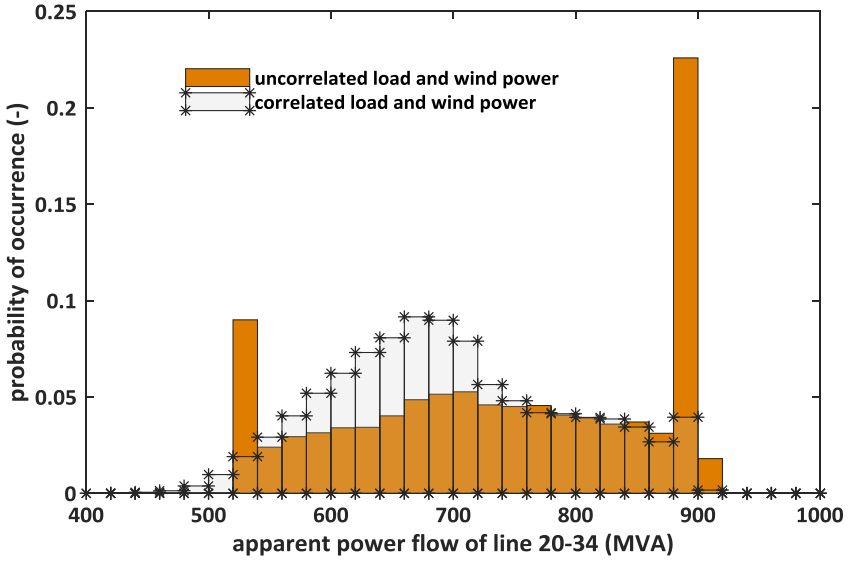


Fig. 5.12: Empirical probability distribution of power flow of line 20-34 (index 34) for uncorrelated and correlated load wind power for *Case II*. Line rating of line 20-34 is 900MVA.

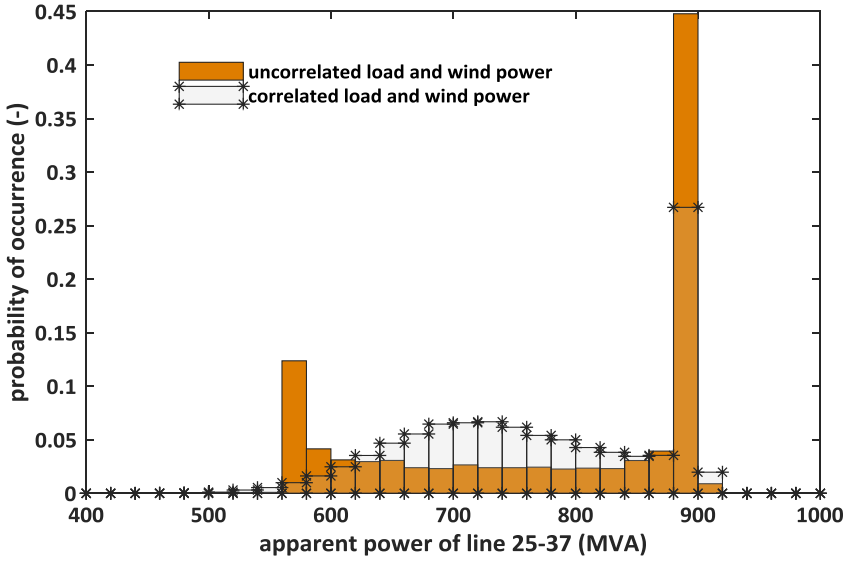


Fig. 5.13: Empirical probability distribution of power flow of line 25-37 (index 41) for uncorrelated and correlated load wind power for *Case II*. Line rating of line 25-37 is 900MVA.

Table 5.3 shows the overload risk indices for the total system and some heavily loaded lines. The net system risk index is lowered by 15.23% when the correlation is considered. In both *Case I* as well as *Case II*, it is noticed that the same set of lines are

often overloaded. In the *WEST*, risk indices for lines 2-3, 2-25 and 25-37 decreases considerably after accounting for correlation although there is an increase of wind power generation at bus 37. In the *MIDATL*, lines 6-11 and 16-19 show an increase in risk index when correlation is considered. It can be understood that risk indices are seriously affected with consideration of spatio-temporal dependency. The proposed solution of adding an interconnection between bus 34 and 11 or other nearby buses is still considered to be a proposed solution to combat the increase in risk index. An increase of load at bus 16 accounts for a marginal increase in risk index for line 16-19. The positive aspect is seen in the form of low overall risk index of the total system with consideration of spatio-temporal dependency.

Table 5.3: Comparison of line overload risk indices for *Case II*

Index	Line	Uncorrelated	Correlated
3	$\mathbf{2-3}$	0.8553	0.7009
4	$\mathbf{2-25}$	0.5582	0.4025
13	$\mathbf{6-11}$	0.9660	0.9846
27	$\mathbf{16-19}$	0.8467	0.8564
34	$\mathbf{20-34}$	0.3670	0.1616
41	$\mathbf{25-37}$	0.5617	0.4158
	Total	$\mathbf{4.1548}$	$\mathbf{3.5217}$

Similar to *Case I*, the number of correlated samples converging is higher than the uncorrelated samples. The performance of correlated samples is nearly the same with the similar number of converging states ($\sim 28xxx$ states). This indicates the advantage of considering correlated samples as a higher number of feasible states correspond to better insight on system operating condition.

Table 5.4: Comparison of uncorrelated and correlated samples converging in *Case II*

Total samples	Uncorrelated	Correlated
30000	11943	28562

5.5 CONCLUSIONS

To answer the urgent need of relevant tool and risk quantification measures for transmission line overloading, a novel attempt of using vine copula for spatio-temporal

modeling to perform a risk-based security assessment of transmission line overloading is presented. Real load and wind power data are mapped onto a modified IEEE 39-bus system representing different market zones of U.S. utility. Use of real data with probabilistic AC OPF analysis gives a more realistic behavior of grid performance in terms of realistic risk indices. The main outcomes of this chapter can be listed as:

- Use of vine copula to model joint normal distribution addresses spatio-temporal dependency. In order to quantify the risk, this is seen important in future because of the massive integration of wind power into existing grid infrastructure. Joint normal distribution is important as it points out to two key properties: the non-Gaussianity of marginal distributions and the complex dependence structures. The reproducible sampling algorithm generates correlated samples that are then mapped onto the test case for risk assessment. It was also observed that the number of converging states for correlated samples is nearly the same for both the cases. High number of converging states corresponds to exploring high number of operating states.
- Probabilistic AC OPF allows measuring the probability of line overload. Overload probabilities contribute significantly to the risk index of both lines and also the whole system. Risk quantification is achieved by combining the probability with the severity of line overload. For studying the overload risk indices for 90% loading or more, the probability of overload is taken into account and the corresponding probabilistic risk indices are calculated. The proposed risk quantification technique is able to qualitatively interpret the numerical values corresponding to risk indices.
- The two cases studied in this research can be regarded as future scenarios aiming at low-carbon electricity generation in the form of massive integration of WPPs. Risk indices for the overall system vary significantly for the two cases and a high number is seen for *Case I*. The reason can be understood as an increase in both load and wind power as compared to *Case II*. For both the cases, overloading of lines 6-11 and 16-19 indicate the need for future expansion planning to address the issue.

REFERENCES

[Abdullah et al., 2013] Abdullah, M. A., Agalgaonkar, A. P., & Muttaqi, K. M. (2013). Probabilistic load flow incorporating correlation between time-varying electricity demand and renewable power generation. *Renewable energy*, *55*, 532-543.

[Arya et al., 2006] Arya, L. D., Titare, L. S., & Kothari, D. P. (2006). Determination of probabilistic risk of voltage collapse using radial basis function (RBF) network. *Electric power systems research*, *76*(6-7), 426-434.

[Bukhsh et Bukhsh, W. A., Zhang, C., & Pinson, P. (2016). An integrated multiperiod OPF

al., 2016] model with demand response and renewable generation uncertainty. *IEEE Transactions on Smart Grid*, *7*(3), 1495-1503.

[Carrillo et al., 2014] Carrillo, C., Cidrás, J., Díaz-Dorado, E., & Obando-Montaño, A. F. (2014). An approach to determine the Weibull parameters for wind energy analysis: The case of Galicia (Spain). *Energies*, *7*(4), 2676-2700.

[Dai et al., 2001] Dai, Y., McCalley, J. D., Abi-Samra, N., & Vittal, V. (2001). Annual risk assessment for overload security. *IEEE Transactions on Power Systems*, *16*(4), 616-623.

[Jong et al., 2018] de Jong, M., Papaefthymiou, G., & Palensky, P. (2018). A framework for incorporation of infeed uncertainty in power system risk-based security assessment. *IEEE Transactions on Power Systems*, *33*(1), 613-621.

[Khuntia et al., 2018c] Khuntia, S. R., Rueda, J. L., & van der Meijden, M. A. (2018). Risk-based security assessment of transmission line overloading considering spatio-temporal dependency of load and wind power using vine copula. *IET Generation, Transmission & Distribution.* Under review.

[Li et al., 2015] Li, X., Zhang, X., Wu, L., Lu, P., & Zhang, S. (2015). Transmission line overload risk assessment for power systems with wind and load-power generation correlation. *IEEE Transactions on Smart Grid*, *6*(3), 1233-1242.

[McCalley et al., 1999] McCalley, J. D., Vittal, V., & Abi-Samra, N. (1999, July). An overview of risk based security assessment. In *Power Engineering Society Summer Meeting, 1999. IEEE* (Vol. 1, pp. 173-178). IEEE.

[McCalley, 2005] McCalley, J. D. (2005). Security assessment: Decision support tools for power system operators. *Ames, IA: Iowa State University*.

[Milligan et al., 2015] Milligan, M., Kirby, B., Acker, T., Ahlstrom, M., Frew, B., Goggin, M., ... & Osborn, D. (2015). Review and status of wind integration and transmission in the united states: Key issues and lessons learned. *National Renewable Energy Laboratory (NREL), Golden, CO, USA, Tech. Rep. NREL/TP-5D00-61911*.

[Morales et al., 2010a] Morales, J. M., Minguez, R., & Conejo, A. J. (2010). A methodology to generate statistically dependent wind speed scenarios. *Applied Energy*, *87*(3), 843-855.

[Morales et al., 2010b] Morales, J. M., Baringo, L., Conejo, A. J., & Mínguez, R. (2010). Probabilistic power flow with correlated wind sources. *IET generation, transmission & distribution*, *4*(5), 641-651.

[Ni et al., 2003a] Ni, M., McCalley, J. D., Vittal, V., Greene, S., Ten, C. W., Ganugula, V. S., & Tayyib, T. (2003). Software implementation of online risk-based security assessment. *IEEE transactions on power systems*, *18*(3), 1165-1172.

[Ni et al., 2003b] Ni, M., McCalley, J. D., Vittal, V., & Tayyib, T. (2003). Online risk-based security assessment. *IEEE Transactions on Power Systems*, *18*(1), 258-265.

[Papaefthymiou & Dragoon, 2016] Papaefthymiou, G., & Dragoon, K. (2016). Towards 100% renewable energy systems: Uncapping power system flexibility. *Energy Policy*, *92*, 69-82.

[TechRep, 2016] Danish TSO experiences with large scale integration of DERs, *IEA*, Paris February 29, 2016.

[TechRep, Eirgrid grid code, version 6.0. *EirGrid*, Tech. Rep., 2015

2015]

[TechRep, 2010] *Energinet.dk* (2010). Technical regulation 3.2. 5 for wind power plants with a power output greater than 11 kW. Technical Report.

[Usaola, 2010] Usaola, J. (2010). Probabilistic load flow with correlated wind power injections. *Electric Power Systems Research*, *80*(5), 528-536.

[Xie et al., 2011] Xie, L., Carvalho, P. M., Ferreira, L. A., Liu, J., Krogh, B. H., Popli, N., & Ilic, M. D. (2011). Wind integration in power systems: Operational challenges and possible solutions. *Proceedings of the IEEE*, *99*(1), 214-232.

[Xie et al., 2014] Xie, L., Gu, Y., Zhu, X., & Genton, M. G. (2014). Short-term spatio-temporal wind power forecast in robust look-ahead power system dispatch. *IEEE Transactions on Smart Grid*, *5*(1), 511-520.

[Zimmerman et al., 2011] Zimmerman, R. D., Murillo-Sánchez, C. E., & Thomas, R. J. (2011). MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on power systems*, *26*(1), 12-19.

[Zimmerman & Wang, 2016] Zimmerman, R.D. & Wang, H. (2016). Matpower Interior Point Solver MIPS 1.2.2 User's Manual, PSERC. Available online: http://www.pserc.cornell.edu/matpower/docs/MIPS-manual-1.2.2.pdf

CHAPTER 6

CONCLUSION AND FUTURE RESEARCH

6.1 CONCLUSION

The primary objective of this research was to develop a statistical model to address the spatio-temporal dependence of load and wind power and to use the model to perform a risk-based security assessment of transmission line overloading in a future power system with high wind power penetration and lower fossil fuel or nuclear power generation. The conceptual framework developed in this research has been described in this way:

"The developed spatio-temporal model addresses the dependence between electricity load and wind power. Consideration of this dependency in system security assessment (short-term horizon) gives a better picture of the risk value associated with transmission line overloading. The quantified risk can support the alleviation of long-term investments to a certain extent."

The results derived from this research can be summarized in three parts:

FIRST PART (RE-VISITING TSO ACTIONS IN OPERATION AND PLANNING):

The first part of this research focused on preparing a solid background on state-of-the-art TSO actions in three time-horizons. Time-horizons and corresponding activities are long-term (grid development), mid-term (asset management), and short-term (system operation) horizons. With respect to time-horizons in the planning and operation of transmission networks, a critical review has been performed to identify the gaps in current practices and requirements for future planning. The review is accompanied by an extended analysis of the asset management decision-making process. The extended analysis is performed for two important reasons: (i) the ageing of existing assets, and (ii) the advent of new technologies that are incompatible with the existing infrastructure. Uncertainty related to load growth and wind power penetration exerts stress on the transmission grid infrastructure which causes them to work at operating limits. In addition, the presence of uncertainty induces a risk of overloading which is described by a severity function in this research. A survey on "assets considered under

asset management" with pan-European TSOs revealed that overhead transmission lines, busbars, transformers, circuit breakers and protection systems are considered equally important. For this research, risk associated with the overloading of transmission lines due to load growth and wind power penetration is considered with a special focus on spatio-temporal dependency because of the spatial distribution of load sites and WPPs that results in stress on transmission lines for energy transmission. In addition, building a new transmission corridor is more challenging in terms of its economic and social impact when compared to investments in other transmission assets. This leads to the second part of the result analysis where statistical models developed in this research are described.

SECOND PART (DEVELOPMENT OF STATISTICAL MODELS FOR FUTURE POWER SYSTEMS):

Statistical models for modeling and forecasting electricity load in the short-term and long-term horizons were developed in this research. For the short-term horizon, a neural network-based load forecasting technique was developed, aiming at better representation of forecast error distribution. With this research, truncated normal distribution was proposed to represent forecast error distribution. An important advantage of truncated normal distribution is that it plays the role of a normal distribution after removing the outliers, which also helps describe the statistical properties of forecast error. The long-term forecast considered aggregated load and volatility in long-term forecast as driving factors. A multiplicative error model was developed with a forecast horizon of 4 years conforming to the fact that off-shore wind farm construction (long-term grid development activity) takes 4 years to complete. With the developed load models, it was possible to account for uncertainty in load growth in terms of temporal scale only. Upon accounting for temporal as well as spatial correlation and considering wind power as another variable, a multivariate modeling approach was deemed necessary. Thus, a vine copula modeling technique was developed to account for the spatio-temporal dependency of load and wind power. A temporal horizon spanning three years and spread across ten different locations consisting of load and wind power is used in this research. In addition, hourly load and wind power data obtained from a U.S. utility is used to further develop the statistical model. The joint probability distribution incorporates the one-dimensional marginal distributions of load and wind power. In reality, one-dimensional marginal distributions are insufficient to calculate the joint probability distribution. They are sufficient if load and wind power are independent random variables. However, studies reveal that load and wind power demonstrate dependency and cannot be treated as independent random variables. However, obtaining the joint probability distribution, given the marginal distributions, is a non-trivial problem since an infinite number of joint probability distributions exist with the same marginal distributions, corresponding to an

infinite number of stochastic dependence structures between random variables such as load and wind power. The use of vine copula facilitates building multi-dimensional copulas out of bivariate copulas as they are easy to estimate and interpret. Copulas provide a more detailed description of the dependence structure between several random variables since they represent multidimensional functions that link marginal probability distributions functions of the random variables to their joint probability distributions functions. The final step in this research is performance evaluation of the vine copula model by performing a risk-based security assessment of the transmission line overloading, described in the third and final part.

THIRD PART (PERFORMANCE EVALUATION):

One of the challenges identified in wind power generation is that places with high potential for bulk or large WPPs (say, several hundred of MW) are often located far from demand sites. Bundled with available capacity in transmission lines and its actual loading level, it imposes significant stress on transmission assets for transporting energy from generating to load sites as they work more frequently at operating limits. The choice of transmission line for this research was explained in the first part.

Therefore, why is there a need to consider the spatio-temporal dependence of both load and wind power in operational planning studies? The third and last part of this research aimed to answer this question by evaluating the vine copula model for security assessment in case of transmission line overloading. Two case studies representing future scenarios with massive wind power penetration and reduced fossil fuel or nuclear power generation were described and the developed model was evaluated. A severity function of line overload was used to calculate the overloading risk and the overload risk index was treated as a security indicator in this research. Considering both the temporal and spatial correlation of load and wind power proved to be beneficial in alleviating the overloading risk in terms of quantifying the overload risk index.

6.2 ANSWERS TO RESEARCH QUESTIONS

Q.1. *What is the implication of different TSO actions taken in different time-horizons on power system reliability? Will the traditional concept of time-horizons be valid in the future when there is uncertain load growth and high penetration of renewable energy into the existing transmission grid infrastructure?* (chapter 2)

A.1. The first research question forms the cornerstone of this research. In the planning and operation of the electric power system, different independent actions are taken by TSOs in different time-horizons to secure a high reliability level. The nature of actions varies from one TSO to another, as

stated in the literature review conducted in this research. It was revealed that the old concept of time-horizons, which refers to the sequential approach adopted by TSOs, is challenged by uncertainty in load growth, massive penetration of wind power and other DERs and development and integration of new technologies. A critical review of current practices identified the challenges that TSOs encounter, with only a few of these answered in this research. The system is expected to perform in ways and in a context for which it was not designed, resulting in an increasingly stressed performance. The system not only becomes heavily loaded and vulnerable to disturbances, but also places security of supply at risk. Surveys involving European TSOs revealed transmission lines to be of critical importance, and hence were subsequently chosen to be analyzed in this research on overloading risk. To study the impact of transmission line overloading when there is uncertainty in demand growth and massive integration of wind power, a spatio-temporal model to learn the dependencies among load and wind power has been developed. An accurate model will help TSOs reconsider their decisions to construct new lines or invest in other transmission infrastructure to combat the stress level of transmission lines.

Q.2. *How should uncertainty in load growth be addressed and what are the associated modeling challenges in the short-term and long-term horizons? How can forecast error be accounted for in terms of error distribution in the short-term horizon? What is the role of volatility in long-term forecasting and how does it impact the modeling framework?* (chapter 3)

A.2. To address the first part of the research question, a neural network based load forecasting model was designed, implemented and trained with real data and results were obtained with a high degree of accuracy. With this model, *truncated normal distribution* to model forecast error is proposed and used in the *GARPUR* project studies. Compared to short-term load forecasting, it was realized that long-term forecasting *requires a different approach* which is based on (i) *identifying and extrapolating mega-trends going as far back in time as necessary, thus resulting in a rich historical database* (ii) *addressing volatility*, and (iii) *constructing scenarios to consider future possibilities*. Moreover, it was realized in this thesis that volatility is forecastable because of a number of persistent properties: (i) *it appears in clusters*, (ii) *it changes over time and has unusual jumps*, (iii) *it does not grow to infinity and is persistent in a specific time-span*, and (iv) *it reacts differently to an increase or decrease of the considered entity*. In this thesis, the novel attempt of using a multiplicative error model to forecast load in long-term horizon is presented with the aim of achieving an accurate long-term forecast methodology. The reason behind choosing a multiplicative error model to

forecast in long-term horizon is supported by the fact of dealing with volatility that cannot be addressed with 'conventional' time series methods. The multiplicative error model inherits many properties from the theory of ARMA models like volatility clustering, fat tails and mean reversion, which assists in learning the statistics of volatility in time series. The term conditional variance in the multiplicative error model denotes its dependency on a past sequence of events in contrast to the unconditional, which implies long-term behavior assuming null knowledge of past events. In this study, consideration of conditional variance has resulted in improved forecast (both low forecast error and directional forecast as well) for long-term horizon. Though both conditions are vital in volatility forecasting, usage of conditional variance in multiplicative error model improves forecast accuracy. A relevant result is the inclusion of heteroskedastic errors that improves forecast performance and also shows that it is possible to predict the direction of change of residuals in the presence of conditional heteroskedasticity, even if the residuals themselves cannot be predicted. The two performance indicators for this long-term forecasting study are: point forecast with a low error percentage as proved by mean error metrics for both in-sample and out-of-sample forecasts, and directional accuracy during the Great Recession of 2008.

Q.3. *How should load variability and wind power generation for spatially distributed locations in a large-scale system be modeled? How can both spatial as well as temporal correlations be effectively addressed? How can high dimensional data be accounted for when the future will be data-centric?* (chapter 4)

A.3. To tackle the multiple variables and the non-Gaussian distribution as well as non-linear relationship of load and wind power, it was concluded in A.2. that dependencies over correlation must be learned and it also addresses the need to model spatio-temporal correlations. With the increased penetration of wind power into the existing grid infrastructure, it is vital to model the complex interdependencies introduced by the stochastic generation sources along with the system load. A combination of dependency modeling and machine learning techniques forms the basis of this research to model high-dimensional spatio-temporal dependency. The modeling of stochastic dependence has been a cornerstone in the research on multivariate uncertainty analysis and the most promising approach is the use of the vine copula model to address spatio-temporal dependency. Among the varieties of copula types and vine copula models, a canonical vine-based high-dimensional spatio-temporal modeling approach is presented in this study. Use of copula type is validated by performing *Goodness of Fit* tests, namely,

Kolmogorov-Smirnov and *Cramer-von Mises* tests. Followed by copula type selection, construction of the canonical vine is made on cluster weight for each cluster. Prior to copula type and vine copula model selection, machine learning techniques are employed for data mining and feature extraction to reduce the high-dimensional data to low-dimensional. Clustering helps in partitioning the data into groups of similar statistical characteristics and choosing the right number of clusters is important to validate the clustering algorithm. In this study, *k-means*, *Gaussian mixture model* (GMM) and *hierarchical linkage* (HL) clustering techniques are assessed. Determining the optimal number of clusters and appropriate clustering method is based on the performance of clustering validation indicators, namely, the *Davies-Bouldin index* and *gap statistics index*. Followed by clustering, upon predicting that future power systems will be data-centric, the computational burden can be predicted for high dimensional modeling and this thesis addresses the problem in terms of efficient modeling by using singular value decomposition and principal component analysis to extract the features for each cluster and thereby reduce the problem to one that is low-dimensional. The proposed sampling methodology is reproducible and able to capture the spatio-temporal dependency among twenty one different sites of load and wind power over a time span of three years by employing a conditional density function calculated using the multi-dimensional canonical vine copula function. In addition, tail dependencies are addressed by copulas with lower values of p-value.

Q.4. *Does the consideration of spatio-temporal dependence of load and wind power prove beneficial to quantify the risk of overloading transmission lines? How does the correlation impact the risk values of line overload? How do the risk values of individual lines or the entire system enable the system operator to assess system operation conditions?* (chapter 5)

A.4. A new attempt to use vine copula for spatio-temporal modeling to perform a risk-based security assessment of transmission line overloading is presented. Real load and wind power data is mapped onto modified IEEE 39-bus system, representing different market zones of U.S. utility. Use of real data with probabilistic AC OPF analysis gives a more realistic indication of grid performance. The main outcomes of this research can be listed as such:

i. Use of vine copula to model joint normal distribution addresses spatio-temporal dependency. Developing risk quantifying strategies is important to the future because of the massive integration of wind power into the existing grid infrastructure. Joint normal distribution is also significant as it points out two key properties: the non-Gaussianity of marginal distributions and the complex dependence structures. The

reproducible sampling algorithm generates correlated samples which are then mapped onto the test case for risk assessment.

ii. Probabilistic AC OPF allows the probability of line overload to be measured. Risk quantification is achieved by combining the probability with the severity of line overload. When studying the overload risk indices for 90% loading or more, the probability of overload is taken into account and the corresponding probabilistic risk indices are calculated. The proposed risk quantification technique is able to qualitatively interpret the numerical values corresponding to risk indices. Using real data on the modified IEEE 39-bus system shows more realistic risk indices, and simulation results show the advantage of considering spatio-temporal dependency.

The two cases considered in this study can be regarded as future scenarios aiming for low-carbon generation in the form of massive integration of WPPs. Risk indices for the overall system vary significantly for the two cases though a high number is seen in *Case I* due to an increase in both load and wind power as compared to *Case II*. In both cases, overloading the lines 6-11 and 16-19 indicates that the issue needs to be addressed in future plans.

6.3 Recommendations and Future Work

This thesis addressed the research questions in terms of statistical modeling and risk assessment but there were additional challenges encountered during the course of research with respect to data, models and acceptance of methodology.

6.3.1 In Terms of Forecasting and Error Modeling

The neural network model developed in this thesis has the possibility of further improvements in terms of optimizing the neural network weights by using heuristic optimization of neural network weights which will result in improving forecast accuracy. In the future, inclusion of random disturbances, consumer class, and demand side management as input parameters for forecast is also possible. Paving the way towards future research by including distributed generation while forecasting load in long-term horizon can be beneficial. In this thesis, the authors implicate that load forecast can correspond to total load in presence of distributed generation like wind and solar, which is treated as negative load.

6.3.2 In Terms of Practical Realization and TSO-DSO Interaction

The forecast methodology for both the short-term as well as long-term horizon and modeling of high-dimensional spatio-temporal dependency presented in this thesis can

be foreseen to be widely adapted with the advent of smart grids and increased participation of stochastic generation sources. It paves way for the system operators to move beyond decision making under past observations. And, yes the biggest challenge of any modeling technique is the feasibility in terms of practical realization. The selection of a suitable method is restricted by the data availability and the daily grid operation processes. A number of factors influence the uncertainty in load and RES forecasting.

Since the modeling technique is presented from TSOs' point of view, a possible extension of this study is the inclusion of solar power that will require an active interaction of DSOs. Thus, it can be concluded that for a more pragmatic methodology, it is recommended to devise a multivariate problem taking load, wind and solar power together. Temporal analysis of multiple data while considering the spatial complexity will provide useful insights and better coordination of TSO-DSO activities in terms of load patterns in various locations across different time-frames (calendar seasons). Output from such a spatio-temporal analysis will facilitate both the operators to plan their grid development and/or maintenance activities in a more optimized way. The penetration of power electronic devices introduces stability issues and with the presented modeling technique, it can be extended to transient stability studies.

6.3.3 IN TERMS OF BIG DATA ANALYTICS

With respect to high-dimensional spatio-temporal modeling, it will be beneficial to include asset failure data for future research. The impact of weather on asset failure rate is already a key research topic and inclusion of possible impact of load and wind power on assets will be advantageous. The huge amount of data needed for such modeling will be handled as big data. In conclusion, big data will have a large impact on the management of utilities in case of fast deployment of ICT and intelligent sensing within the transmission network. There are many challenges which would affect the success of big data applications in future for utilities. Reduced cost of storage with advancement of cloud based data analytic technology will enable the data analysts and scientists to easily extract the information from large volumes of data. Presently, experience in integrating big data with current framework is limited. In particular, analytics must be supported by true optimization models to automate the maintenance planning and outage scheduling. It is intended to discover correlations or patterns to make holistic decisions and with the help of analytics utilities can consider all aspects of a decision – the financial side, the maintenance side, as well as the operations side. Also, real application of big data is the ability to understand what data sample is required, ways to analyze and interpret, and then use it. Without completed fields, or validated data, analysis is not possible. So good amount of effort is needed in the future to be spent to develop more advanced and efficient algorithms for data analysis that can be easily accepted by utilities. In the end, effective maintenance will be a result of

quality, timeliness, accuracy and completeness of information related to machine degradation state, based on which decisions are made. In terms of tractability and scalability, reducing the computational burden is possible by exploring machine learning techniques in terms of advanced clustering algorithm and feature extraction.

APPENDIX A1

This appendix supplements data for Chapters 4 and 5. The data supplied here is primarily taken from PJM website, namely:

1. Load data: http://www.pjm.com/markets-and-operations/ops-analysis/historical-load-data.aspx
2. Wind power data: http://www.pjm.com/markets-and-operations/ops-analysis.aspx
3. Zone map: http://www.pjm.com/-/media/about-pjm/pjm-zones.ashx?la=en
4. https://www.pjm.com/-/media/committees-groups/subcommittees/irs/postings/pjm-pris-task-3a-part-a-modeling-and-scenarios.ashx?la=en
5. http://www.pjm.com/committees-and-groups/subcommittees/irs/pris.aspx

All links were last accessed on 3rd April, 2018.

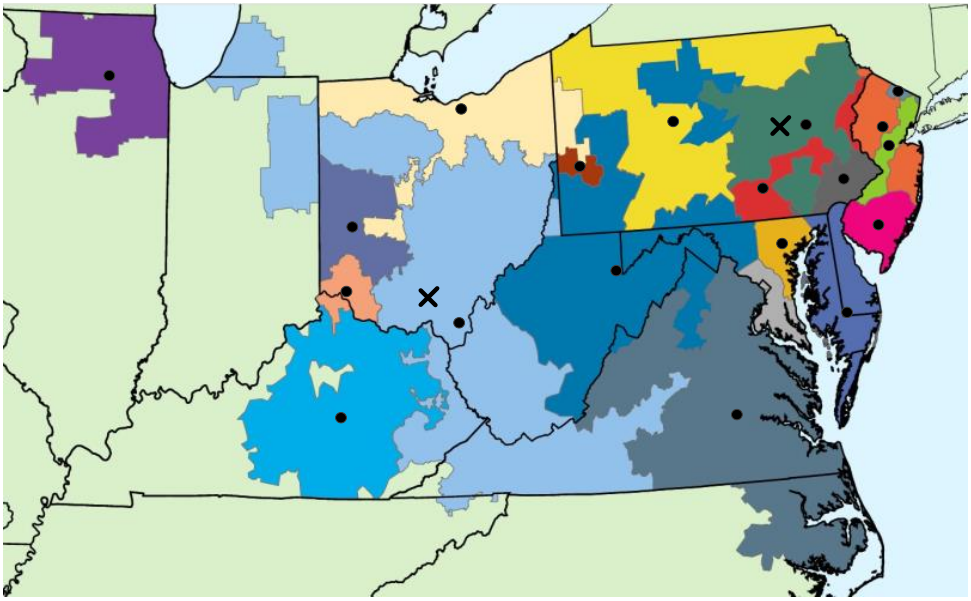


Fig. A.1: PJM zones showing the different load zones with spatial centroid (●) and wind zones with spatial centroid (X)

Fig. A.1 shows the nineteen load zones and two wind power zones considered in the study. The exact latitude and longitude of different zones are treated confidential by PJM and hence an approximated centroid (latitude and longitude) is calculated detailed in Table A.1. Based on the approximated centroid, distances between two locations are calculated and shown in Table A.2.

Table A.1: Zones with corresponding latitude and longitude

Zones	Latitude	Longitude
AP	39.200221	-79.460021
AEP	38.670387	-82.316467
ATSI	41.181128	-81.422454
AE	39.638103	-74.862243
BC	39.454384	-76.463511
CE	41.220909	-88.926116
DAY	39.892241	-84.300872
DPL	38.429811	-75.658752
DOM	37.277908	-77.713205
DEOK	39.205276	-84.508239
DUQ	40.631050	-80.088999
EKPC	37.564059	-84.096252
JC	40.806355	-74.502451
ME	40.078942	-76.507456
PL	40.997343	-75.875742
PE	40.179742	-75.524180
PN	40.940380	-78.394347
PS	40.498296	-74.474985
RECO	41.085461	-74.150889
MIDATL	40.664836	-76.232764
WEST	38.943206	-82.868506

Table A.2: *Distance (in kms.) between different zones*

	AP	AEP	ATSI	AE	BC	CE	DAY	DPL	DOM	DEOK	DUQ	EKPC	JC	ME	PL	PE	PN	PS	RECO	MI	WE
AP	0	254	276	398	259	834	422	340	262	435	168	443	458	271	364	354	214	449	497	320	295
AEP	254	0	289	651	512	630	218	579	432	199	289	198	708	523	607	606	419	701	746	565	57
ATSI	276	289	0	581	462	627	282	579	538	342	128	463	582	432	465	509	255	589	608	439	277
AE	398	651	581	0	139	1201	806	151	361	829	457	834	133	149	174	82	332	101	172	163	693
BC	259	512	462	139	0	1073	672	133	265	692	335	696	224	70	179	114	233	205	267	136	554
CE	834	630	627	1201	1073	0	417	1173	1059	437	745	580	1209	1054	1092	1134	882	1216	1235	1066	574
DAY	422	218	282	806	672	417	0	762	641	78	366	259	836	664	723	747	513	836	867	689	162
DPL	340	579	579	151	133	1173	762	0	221	771	452	745	282	197	286	195	364	251	322	253	628
DOM	262	432	538	361	265	1059	641	221	0	630	425	564	480	328	443	374	411	454	523	397	487
DEOK	435	199	342	829	692	437	78	771	630	0	409	186	870	691	760	775	554	867	904	723	144
DUQ	168	289	128	457	335	745	366	452	425	409	0	485	471	309	357	389	147	474	502	325	302
EKPC	443	198	463	834	696	580	259	745	564	186	485	0	901	714	803	796	617	892	939	760	187
JC	458	708	582	133	224	1209	836	282	480	870	471	901	0	188	117	111	327	34	43	147	743
ME	271	523	432	149	70	1054	664	197	328	691	309	714	188	0	115	84	186	178	228	69	560
PL	364	607	465	174	179	1092	723	286	443	760	357	803	117	115	0	96	211	130	145	48	638
PE	354	606	509	82	114	1134	747	195	374	775	389	796	111	84	96	0	257	96	153	81	644
PN	214	419	255	332	233	882	513	364	411	554	147	617	327	186	211	257	0	334	356	184	441
PS	449	701	589	101	205	1216	836	251	454	867	474	892	34	178	130	96	334	0	71	149	738
RECO	497	746	608	172	267	1235	867	322	523	904	502	939	43	228	145	153	356	71	0	181	779
MI	320	565	439	163	136	1066	689	253	397	723	325	760	147	69	48	81	184	149	181	0	598
WE	295	57	277	693	554	574	162	628	487	144	302	187	743	560	638	644	441	738	779	598	0

Other useful links used in extraction of spatial data and calculation of spatial lags:

1. https://www.latlong.net/
2. http://www.pjm.com/markets-and-operations/etools/data-miner-2/data-availability.aspx
3. https://www.nhc.noaa.gov/gccalc.shtml

APPENDIX A2

A2.1 TWO SAMPLE KOLMOGOROV-SMIRNOV TEST

The two sample Kolmogorov-Smirnov (K-S) test is a nonparametric test that compares the cumulative distributions of two data sets. The test statistic reports the maximum difference between the two cumulative distributions, and calculates a p-value from that and the sample sizes. Some key properties of KS-test are:

- K-S test is non-parametric in nature.
- It does not assume that data are sampled from Gaussian distributions (or any other defined distributions).
- The results of K-S test do not change if any data transformation is applied (such as logarithmic). Because the maximum distance between any two frequency distribution remains the same irrespective of transformation.

Given two samples, it tests if their distributions are the same. It starts with computing the observed CDFs of the two samples and computing their maximum difference. For two samples,

$$X: 1.2, 1.4, 1.9, 3.7, 4.4, 4.8, 9.7, 17.3, 21.1, 28.4$$
$$Y: 5.6, 6.5, 6.6, 6.9, 9.2, 10.4, 10.6, 19.3$$

Now, the combined sample is sorted in order to compute the ECDFs as shown in table below:

Table A2.1: Combined sample of X and Y and corresponding ECDFs

Combined X and Y in ascending order	F_X	F_Y
1.2	0.1	0.0
1.4	0.2	0.0
1.9	0.3	0.0
3.7	0.4	0.0
4.4	0.5	0.0
4.8	0.6	0.0
5.6	0.6	0.1
6.5	0.6	0.2
6.6	0.6	0.4

6.9	0.6	0.5
9.2	0.6	0.6
9.7	0.7	0.6
10.4	0.7	0.8
10.6	0.7	0.9
17.3	0.8	0.9
19.3	0.8	1.0
21.1	0.9	1.0
28.4	1.0	1.0

In this research, MATLAB function kstest2 reports the test statistics. For this example, p-value with a significance level of 95% is used to reject the null hypothesis.

MATLAB script:

```
clear all
clc

x=[1.2, 1.4, 1.9, 3.7, 4.4, 4.8, 9.7, 17.3, 21.1, 28.4]';
y=[5.6, 6.5, 6.6, 6.9, 9.2, 10.4, 10.6, 19.3]';
[h1,p1,k1] = kstest2(x,y)
cdfplot(x)
hold on
cdfplot(y)
```

Output:

```
p1 = 0.0473
k1 = 0.6000
```

Fig. A2.1 illustrates the overlay plots of two ECDFs. The difference between their distributions is significant at the 5% level (p = 4%). The K-S test statistic is the maximum difference between these functions. And in this case, it is 0.6.

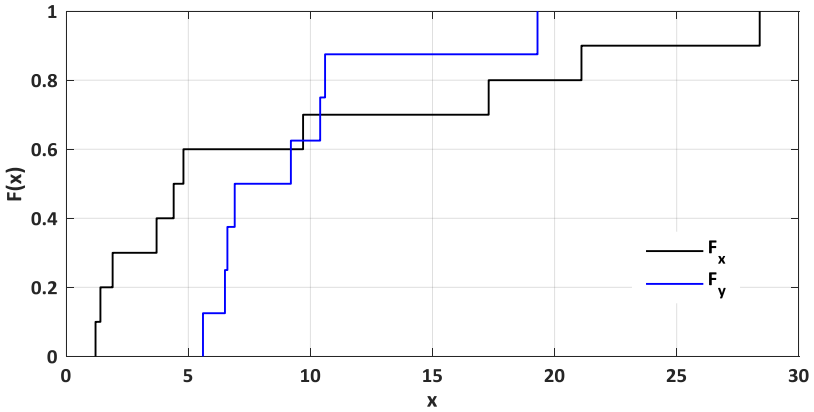


Fig. A2.1: Plots of the two ECDFs (F_X and F_y)

A2.2 GOODNESS OF CLUSTERING (GOC) TEST

A2.2.1 DAVIES-BOULDIN INDEX (DBI)

The Davies-Bouldin Index identifies clusters which are far from each other and compact. For n_c clusters, the DBI is defined as:

$$DBI = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i \tag{1}$$

where, $R_i = \max_{j=1\ldots n_c, i\neq j}(R_{ij})\,, i = 1 \ldots n_c$

$R_{ij} = \frac{s_i+s_j}{d_{ij}}$, $d_{ij} = d(v_i, v_j)$, $s_i = \frac{1}{\|c_i\|}\sum_{x\in c_i} d(x, v_i)$

where, $d(x, y)$ is the Euclidean distance between x and y, c_i is the cluster i, v_i is the centroid of cluster i, and $\|c_i\|$ refers to the norm of c_i. Since the objective is to obtain clusters with minimum intra-cluster distances, small values for DBI are interesting. Therefore, this index is minimized when looking for the best number of clusters.

A2.2.2 GAP STATISTICS INDEX (GSI)

The gap statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data, i.e., a distribution with no obvious clustering. First, assume that there are a set of samples $\{x_i\}$, then by the use of the clustering method, the resultant clusters $C_1, C_2, \ldots, C_k$ can be obtained. For any cluster C_r, the sum of the pair wise distance $d^2(x_i, x_j)$, for all points in cluster r is calculated. And the sum of within-cluster dispersion W_k is defined the following equation

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} \sum_{ij \in C_r} d^2(x_i, x_j) \quad (2)$$

As to the central concept of the gap statistic, it is to compare the $\log(W_k)$ with its expectation under a reference distribution. It is defined

$$Gap_n(k) = E_n^*\{\log(W_k)\} - \log(W_k) \quad (3)$$

where E_n^* denotes expectation with a sample of size n of the reference distribution. The optimized number of clusters k is decided by $Gap_n(k)$. Note that, the logarithm of the W_k values is used, as they can be quite large. The gap statistic measures the deviation of the observed W_k value from its expected value under the null hypothesis. The estimate of the optimal clusters k will be value that maximize $Gap_n(k)$ (i.e, that yields the largest gap statistic). This means that the clustering structure is far away from the uniform distribution of points.

APPENDIX A3

A3.1 ARMA MODEL

Definition 1: *An autoregressive process of order* $p \geq 1$ *is defined as*

$$X_t = b_1 X_{t-1} + \cdots + b_p X_{t-p} + \epsilon_t$$

where $\{\epsilon_t\} \sim N(0, \sigma^2)$*. The time series* $\{X_t\}$ *generated from this model is called* $AR(p)$ *process.*

Autoregressive (AR) model represents the current value of the process X_t as a linear combination of past values of the process together with some white noise. That is, current state is completely determined by the past values $X_{t-1}, \dots, X_{t-p}$ and some random error ϵ_t. In the above definition, a stochastic process $\{y_t\}$ is called a white noise, denoted as $\{y_t\} \sim N(0, \sigma^2)$, if $Ey_t = 0$, $Var(y_t) = \sigma^2$ and $Cov(y_t, y_s) = 0$ for all $s \neq t$.

The model can be viewed as a classical linear regression model without intercept. In practice, modeling the time series without intercept is no restriction, since it is common in time series analysis to subtract the mean from the data before proceeding with further analyses.

Definition 2: *A moving average process with order* $q \geq 1$ *is defined as*

$$X_t = \epsilon_t + a_1 \epsilon_{t-1} + \cdots + a_q \epsilon_{t-q}$$

where $\{\epsilon_t\} \sim N(0, \sigma^2)$*. The time series* $\{X_t\}$ *generated from this model is called* $MA(q)$ *process.*

Moving Average (MA) models represents the current value as a linear combination of a white noise process realizations, so the value of X_t can be considered completely random.

By combining AR and MA processes, we arrive at an autoregressive moving average process.

Definition 3: *The autoregressive moving average (ARMA) process of orders p and q is defined as:*

$$X_t = b_1 X_{t-1} + \cdots + b_p X_{t-p} + \epsilon_t + a_1 \epsilon_{t-1} + \cdots + a_q \epsilon_{t-q}$$

where $\{\epsilon_t\} \sim N(0, \sigma^2)$, $p, q \geq 0$ are integers. We write $\{X_t\} \sim ARMA(p, q)$. The time series $\{X_t\}$ generated from this model is called $ARMA(p, q)$ process.

It is often useful to represent linear ARMA processes using backshift operators. Namely, denote

$$b(z) = 1 - b_1 z - \cdots - b_p z^p \text{ and } a(z) = 1 + a_1 z + \cdots + a_q z^q ,$$

for $z \in \mathbb{C}$ and further define the backshift operator B as

$$BX_t = X_{t-1}, B^k X_t = (B^{k-1}) BX_t = X_{t-k}, k \in \mathbb{N}$$

We can then rewrite the ARMA process in a simple form as

$$b(B) X_T = a(B) \epsilon_t$$

One advantage of using the polynomial representation of ARMA model is that lot of properties of ARMA models can be determined by exploring the polynomials $b(z)$ and a(z).

BIOGRAPHY

Swasti Ranjan Khuntia was born in Rourkela, India. After finishing his high school at Cuttack, India, he moved to Berhampur, India for his undergraduate studies in Electrical and Electronics Engineering at National Institute of Science and Technology (08/2006-07/2010). His undergraduate project was awarded the prestigious Young Scientist Award from Department of Science and Technology, Government of India when he presented his work at 2010 Power Control and Optimization conference in Malaysia. Later, he moved to Chicago, USA to pursue his Master of Science in Electrical Engineering at Illinois Institute of Technology (08/2011-12/2013). During his Master study, he was awarded the *Deutscher Akademischer Austauschdienst* (DAAD) RISE Professional scholarship to intern at Corporate Research center of Robert Bosch GmbH, Germany. From April 2014 till September 2018, he was a Doctoral researcher in the Intelligent Electrical Power Grids (IEPG) research group of Dept. Electrical Sustainable Energy at Delft University of Technology. He participated and lead tasks in EU-funded FP7 project named "GARPUR" which developed a new probabilistic reliability criteria at pan-European level.

Swasti has been actively involved in IEEE and CIGRE during his studies, and has been IEEE Student Member since 2010. He was the President of IEEE Student Chapter at Delft (06/2014-03/2018) and an active board member of Young CIGRE Netherlands (02/2017-03/2018). While serving Young CIGRE Netherlands, he was also the student Ambassador at TU Delft on behalf of Young CIGRE board. He served as a board member and Vice-President of PromooD (10/2014-03/2018), PhD social group at TU Delft as well.

Currently, Swasti is working as a Scientific Researcher at Centrum Wiskunde & Informatica (CWI), the national research institute for mathematics and computer science in the Netherlands.