Estimating predictive hydrological uncertainty by dressing deterministic and ensemble forecasts; a comparison, with application to Meuse and Rhine

Verkade, J. S.; Brown, J. D.; Davids, F.; Reggiani, P.; Weerts, A. H.

DOI
10.1016/j.jhydrol.2017.10.024

Publication date
2017

Document Version
Final published version

Published in
Journal of Hydrology

Citation (APA)
Verkade, J. S., Brown, J. D., Davids, F., Reggiani, P., & Weerts, A. H. (2017). Estimating predictive hydrological uncertainty by dressing deterministic and ensemble forecasts; a comparison, with application to Meuse and Rhine. *Journal of Hydrology*, *555*, 257-277. https://doi.org/10.1016/j.jhydrol.2017.10.024

Journal of Hydrology 555 (2017) 257–277

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Estimating predictive hydrological uncertainty by dressing deterministic and ensemble forecasts; a comparison, with application to Meuse and Rhine

J.S. Verkade [a,b,c,*], J.D. Brown [d], F. Davids [a,b], P. Reggiani [e], A.H. Weerts [a,f]

[a] Deltares, PO Box 177, 2600 MH Delft, The Netherlands
[b] Ministry of Infrastructure and the Environment, Water Management Centre of The Netherlands, River Forecasting Service, Lelystad, The Netherlands
[c] Delft University of Technology, Delft, The Netherlands
[d] Hydrologic Solutions Limited, Southampton, United Kingdom
[e] University of Siegen, Research Institute for Water and Environment, Siegen, Germany
[f] Wageningen University and Research Centre, Hydrology and Quantitative Water Management Group, Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 17 January 2017
Received in revised form 29 July 2017
Accepted 8 October 2017
Available online 13 October 2017
This manuscript was handled by K. Georgakakos, Editor-in-Chief, with the assistance of Hamid Moradkhani, Associate Editor

Keywords:
Quantile Regression
Hydrological forecasting
Predictive uncertainty
Ensemble dressing
Statistical post-processing

ABSTRACT

Two statistical post-processing approaches for estimation of predictive hydrological uncertainty are compared: (i) 'dressing' of a deterministic forecast by adding a single, combined estimate of both hydrological and meteorological uncertainty and (ii) 'dressing' of an ensemble streamflow forecast by adding an estimate of hydrological uncertainty to each individual streamflow ensemble member. Both approaches aim to produce an estimate of the 'total uncertainty' that captures both the meteorological and hydrological uncertainties. They differ in the degree to which they make use of statistical post-processing techniques. In the 'lumped' approach, both sources of uncertainty are lumped by post-processing deterministic forecasts using their verifying observations. In the 'source-specific' approach, the meteorological uncertainties are estimated by an ensemble of weather forecasts. These ensemble members are routed through a hydrological model and a realization of the probability distribution of hydrological uncertainties (only) is then added to each ensemble member to arrive at an estimate of the total uncertainty.

The techniques are applied to one location in the Meuse basin and three locations in the Rhine basin. Resulting forecasts are assessed for their reliability and sharpness, as well as compared in terms of multiple verification scores including the relative mean error, Brier Skill Score, Mean Continuous Ranked Probability Skill Score, Relative Operating Characteristic Score and Relative Economic Value. The dressed deterministic forecasts are generally more reliable than the dressed ensemble forecasts, but the latter are sharper. On balance, however, they show similar quality across a range of verification metrics, with the dressed ensembles coming out slightly better. Some additional analyses are suggested. Notably, these include statistical post-processing of the meteorological forecasts in order to increase their reliability, thus increasing the reliability of the streamflow forecasts produced with ensemble meteorological forcings.

1. Introduction

The future value of hydrological variables is inherently uncertain. Forecasting may reduce, but cannot eliminate this uncertainty. Informed, forecast-sensitive decision making is aided by adequate estimation of the remaining uncertainties (see, for example, Verkade and Werner, 2011 and the references therein). Omission of relevant uncertainties would result in overconfident forecasts, hence all relevant uncertainties must be addressed in the estimation procedure. These include uncertainties related to the modeling of the streamflow generation and routing processes (jointly referred to as "hydrological uncertainties") and uncertainties related to future atmospheric forcing ("meteorological uncertainties"). Generally speaking, the total uncertainty can be estimated by separately modelling the meteorological and hydrological uncertainties or by lumping all uncertainties together (cf. Regonda et al., 2013).

* Corresponding author at: Deltares, PO Box 177, 2600 MH Delft, The Netherlands.
E-mail address: jan.verkade@deltares.nl (J.S. Verkade).

https://doi.org/10.1016/j.jhydrol.2017.10.024
0022-1694/

The source-specific approach identifies the relevant sources of uncertainty and models these individually before integrating them into an estimate of the total uncertainty. In this context, the hydrologic uncertainties may be treated separately (independently) from the meteorological uncertainties, because they depend only on the quality of the hydrologic modelling. This approach has been followed by, among others, Kelly and Krzysztofowicz (2000); Krzysztofowicz (2002); Krzysztofowicz and Kelly (2000); Seo et al. (2006) and Demargne et al. (2013). The approach has a number of attractive characteristics. The individual sources of uncertainty may each have a different structure, which can be specifically addressed by separate techniques. Also, some of the uncertainties vary in time, while others are time invariant. A disadvantage of source-based modelling is that developing uncertainty models for each source separately may be expensive, both in terms of the development itself as well as in terms of computational cost. Also, whether modelling the total uncertainty as a lumped contribution or separately accounting for the meteorological and hydrological uncertainties, hydrological forecasts will inevitably contain residual biases in the mean, spread and higher moments of the forecast probability distributions, for which statistical post-processing is important.

In the lumped approach, a statistical technique is used to estimate the future uncertainty of streamflow conditionally upon one or more predictors, which may include a deterministic forecast. Underlying this approach is an assumption that the errors associated with historical predictors and predictions are representative of those in future. This approach is widely used in atmospheric forecasting, where it is commonly known as Model Output Statistics (MOS) (Glahn and Lowry, 1972). Several reports of applications of the lumped approach in hydrology can be found in the literature. These include the Reggiani and Weerts (Reggiani and Weerts, 2008) implementation of the Hydrologic Uncertainty Processor (Kelly and Krzysztofowicz, 2000), the Model Conditional Processor (Todini, 2008; Coccia and Todini, 2011), Quantile Regression (Weerts et al., 2011; Verkade and Werner, 2011; López López et al., 2014), the "uncertainty estimation based on local errors and clustering" approach (UNEEC; Solomatine and Shrestha, 2009) and the Hydrologic Model Output Statistics approach (HMOS; Regonda et al., 2013). For a complete overview that is periodically updated, see Ramos et al. (2013). These techniques each estimate the total uncertainty in future streamflow conditionally upon one or more predictors, including the deterministic forecast. Of course, they vary in their precise formulation and choice of predictors. The lumped approach is attractive for its simplicity, both in terms of development and computational costs. The main disadvantage of the approach is that both meteorological and hydrological uncertainties are modeled together via the streamflow forecast, which assumes an aggregate structure for the modeled uncertainties (although the calibration may be different for particular ranges of streamflow). Also, in order to produce ensemble traces, these techniques must explicitly account for the temporal autocorrelations in future streamflow, which may not follow a simple (e.g. autoregressive) form.

In the text above, the source-specific and the lumped approach were presented as separate strategies. However, as the source-based approach may not fully account for all sources of uncertainty, statistical post-processing is frequently used to correct for residual biases in ensemble forecasts. In the present work, an intermediate approach is described, namely the 'dressing' of streamflow ensemble forecasts. Here, the meteorological uncertainties are estimated by an ensemble of weather forecasts. The remaining, hydrological uncertainties are lumped and described statistically. Subsequently, the streamflow ensemble members are, cf. Pagano et al. (2013), 'dressed' with the hydrological uncertainties. This approach has previously been taken by, among others, Reggiani et al. (2009); Bogner and Pappenberger (2011) and Pagano et al. (2013) and, in meteorological forecasting, by Fortin et al. (2006), Roulston and Smith (2003) and Unger et al. (2009). Most of these studies report skill of the dressed ensembles versus that of climatology; Pagano et al. (2013) explored the gain in skill when moving from raw to dressed ensembles and found this gain to be significant. In contrast, the present study compared dressed ensemble forecasts to post-processed single-valued streamflow forecasts.

The kernel dressing approach is akin to kernel density smoothing, whereby missing sources of uncertainty (i.e. dispersion) are introduced by dressing the individual ensemble members with probability distributions and averaging these distributions (cf. Bröcker and Smith, 2008). As ensemble dressing aims to account for additional sources of dispersion, not already represented in the ensemble forecasts, a "best member" interpretation is often invoked (Roulston and Smith, 2003). Here, the width of the dressing kernel is determined by the historical errors of the best ensemble member. The resulting distribution is then applied to each ensemble member of an operational forecast and the final predictive distribution given by the average of the individual distributions. In this context, ensemble dressing has some similarities to Bayesian Model Averaging (BMA; see Raftery et al., 2005 for a discussion).

In the ensemble dressing approach, one highly relevant source of uncertainty, namely the weather forecasts, is described using an ensemble Numerical Weather Prediction (NWP) model. This NWP model takes into account current initial conditions of the atmosphere and exploits the knowledge of physical processes of the atmosphere embedded in the NWP model, as well as any meteorological observations that are assimilated to improve estimates of the predicted states. The hydrologic uncertainties, which may originate from the hydrologic model parameters and structure (among other things) are then lumped, modelled statistically, and integrated with the meteorological contribution to the streamflow.

The objective of this work is to compare the quality and skill of the forecasts created through dressing of deterministic streamflow forecasts and through dressing of ensemble streamflow forecasts. *A priori*, the dressed ensemble forecasts are expected to have higher skill than the dressed deterministic forecasts. Both account for the effects of all relevant sources of uncertainty on the streamflow forecasts. However, in the ensemble case, the estimate of atmospheric uncertainties is based on knowledge of the physical system and its state at issue time of a forecast, whereas this knowledge is unused in the lumped approach. Nevertheless, the lumped approach accounts for any residual meteorological biases via the streamflow.

The context for this study is an operational river forecasting system used by the Dutch river forecasting service. This system models the total uncertainty in the future streamflow using a lumped approach, whereby a deterministic streamflow forecast is post-processed through quantile regression (following a procedure similar to that in Weerts et al., 2011). While this module performs reasonably well, there is a desire among operational forecasters to explore the benefits of (and account for) information in ensemble weather predictions, including information beyond the ensemble mean. This resulted in the operational implementation of the ensemble dressing approach using the same statistical technique (quantile regression). Thus, estimates of the meteorological uncertainties, which were previously modeled indirectly (i.e. lumped into the total uncertainty), are now disaggregated and included separately in the streamflow forecasts. This raises the question of whether the 'new' approach indeed increases forecast skill.

The novel aspects and new contributions of this work include (i) a direct comparison between the quality of the dressed deterministic forecasts and the dressed ensemble forecasts; (ii) the application of quantile regression to account for the hydrologic

uncertainties, and (iii) the application of the dressing technique to dynamic ensemble streamflow forecasts.

This paper is organised as follows. In the next section, the study approach is detailed, followed by a description of the study basins in Section 3. In Section 4 the results of the experiments are presented and analysed. In Section 5, conclusions are drawn and discussed.

2. Approach

2.1. Scenarios

The present study consists of an experiment in which verification results in two *scenarios* are inter-compared: dressed deterministic forecasts and dressed ensemble forecasts. These are tested in multiple *cases*, that is, combinations of forecasting locations and lead times.

2.2. Cross-validation

The results from the experiments that are described below are cross-validated by means of a leave-one-year-out analysis. The set of available data is divided into N available years. Models are trained using $N-1$ years and validated on the remaining one year. The experiment is carried out N times, every time with another year chosen as a validation year. Thus, the validation record that is verified comprises of N years of data that is obtained independently from the training data — and optimal use is made of the available length of record.

2.3. Hindcasting

The experiment comprises the process of hindcasting or *reforecasting*: forecasts are produced for periods of time that are currently in the past – while only using data that was available at the time of forecast (*issue time*, or *reference time* or *time zero*).

Hindcasts are produced using an offline version of the Delft-FEWS (Werner et al., 2013) based forecast production system "RWsOS Rivers" that is used by the Water Management Centre of The Netherlands for real-time, operational hydrological forecasting.

Hindcasting is a two-step process: first, the hydrological models are forced with observed temperature and precipitation for a period up to the forecast initialisation time. Thus, the internal model states reflect the basin's actual initial conditions as closely as possible. These initial states are used as the starting point for forecast runs, where the models are forced with the precipitation and temperature ensemble NWP forecasts. These are imported into the forecast production system as gridded forecasts and subsequently spatially aggregated to sub-basins. Basin-averaged precipitation or temperature is then equal to the average of the grid boxes within that basin, accounting for partial inclusion of grid boxes through a process of weighting.

In the case of St Pieter streamflow forecasts, an autoregressive-moving-average (ARMA) error-correction procedure was used to correct for biases in the raw streamflow forecasts and hence any residual biases contributed by the forcing (see below). This is in line with the procedures used in the operational system. The autoregressive (AR) correction scheme used here is based on estimation of the AR parameters using the Burg's method and ARMA-Sel (see Broersen and Weerts, 2005, and the references therein). The procedure is implemented in Delft-FEWS (Werner et al., 2013) and can be configured as self-calibrating (automated selection of order and/or AR parameter values) or with fixed AR parameters for post-processing of hydrological forecasts. Here, the latter option was chosen.

The effect of error correction is that forecast uncertainty will be reduced to zero at zero lead time; with increasing lead time, this uncertainty reduction will 'phase out'. To account for this in the streamflow hindcasts, error correction was used in simulation mode also. Effectively, the hindcasting procedure comprised the re-production of forecasts where the models were forced with precipitation *measurements* instead of precipitation *forecasts*. This introduces a lead time dependence in the quality of the streamflow simulations. This is described in more detail in Section 2.5.

2.4. 'Dressing' of streamflow forecasts

The dressing technique is similar across the lumped and the source-specific approaches in that the forecasts are dressed with predictive distributions of uncertainties that are not already explicitly addressed in the raw forecasts. Thus, deterministic hydrological forecasts are dressed with a predictive distribution that comprises both meteorological and hydrological uncertainties, and hydrological ensemble forecasts are dressed with a predictive distribution that comprises hydrological uncertainties only. In both approaches, the total uncertainty is computed by averaging over the number of ensemble members E (which $E=1$ in the case of deterministic forecasts),

$$\Phi_n(q_n|s_{n,1},s_{n,2},\ldots,s_{n,E}) = \frac{1}{E}\sum_{e=1}^{E}\phi_n(q_n|s_{n,e}), \tag{1}$$

where Φ is the aggregated distribution of observed streamflow q at lead time n, conditional on the raw streamflow forecast s that consists of ensemble members $e \in \{1,\ldots,E\}$, each of which are dressed with distribution ϕ.

In the ensemble dressing scenario, this means that each of the ensemble members is dressed with a predictive distribution of hydrological uncertainty, and that these multiple distributions are averaged to obtain a single distribution of predictive uncertainty. Note that here, we assume that the ensemble members are equiprobable (which generally applies to atmospheric ensembles generated from a single model, but not necessarily to multi-model ensembles, for example). If the members are not equiprobable then a weighting can easily be introduced.

Here, the distribution, Φ, aims to capture the historical residuals between the observed and simulated streamflows (i.e. streamflows produced with observed forcing). A "best member" interpretation does not apply here, because the dressing kernel is aiming to capture a new source of uncertainty (the hydrologic uncertainty) and not to account for under-dispersion in (the hydrologic effects of) the meteorological uncertainties. In short, we assume that the meteorological ensembles are unbiased and correctly dispersed. This is a necessary assumption: uncertainty originating in future weather is only estimated by the spread of the streamflow ensemble and not in any way through statistical post-processing, nor does post-processing in any way bias-correct the streamflow ensemble to better account for meteorological uncertainty. This constitutes a disclaimer to the ensemble dressing approach: if the assumption of a correctly dispersed meteorological ensemble is invalid then this will have an adverse effect on the quality of the dressed ensemble!

By construction, the raw deterministic forecasts are dressed with a single distribution only, which aims to account for the total uncertainty of the future streamflow, not the residual uncertainty of a best ensemble member,

$$\Phi_n(q_n|s_{n,1}) = \phi_n(q_n|s_{n,1}). \tag{2}$$

The dressing procedures are schematically visualised in Fig. 1.

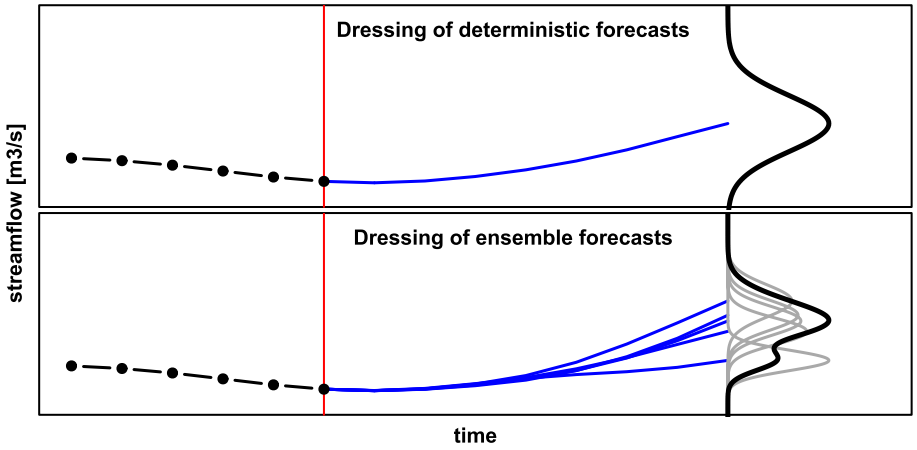

Fig. 1. Schematic representation of the dressing procedures. The vertical red line denotes the issue time of the forecast, with the observations (black dots) in the past and the raw forecasts (blue lines) in the future. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.5. Uncertainty models

As mentioned above, the source-specific and lumped approaches differ in the predictive distributions that the raw forecasts are dressed with. In the lumped approach, the deterministic forecast is dressed by a distribution of both hydrological and atmospheric uncertainties, conditional on the value of the deterministic forecast itself. The "errors" in the deterministic forecast are thus a measure of uncertainties originating in both the meteorological forcing as well as the hydrological modeling.

Ensemble streamflow forecasts are dressed using a predictive distribution of hydrological uncertainty only. This is achieved by fitting a probability distribution to the historical residuals between the hydrological model simulations and observations (see Section 2.6 for details on how this was done). The simulations are derived by forcing a hydrological model with observed precipitation and temperature. As such, the simulations are independent of lead time. This approach is quite widely reported in the literature, for example by Montanari and Brath (2004), Seo et al. (2006), Chen and Yu (2007), Hantush and Kalin (2008), Montanari and Grossi (2008), Todini (2008), Bogner and Pappenberger (2011), Zhao et al. (2011) and Brown and Seo (2013).

Time invariant estimates of hydrological uncertainty using these hydrological simulations, however, do not take into account any error correction or data assimilation procedures that, in a real-time setting, reduce predictive uncertainty. In the Meuse at St Pieter case, such an error correction method is an integral component of the forecasting system. Hence, in the Meuse case, hydrological uncertainty is not based on observations and simulations, but on observations and "perfect forcing hindcasts". Similar to the forecasts produced by the operational system, these hindcasts benefit from the ARMA error correction procedure that is applied to streamflow forecasts at St Pieter at the onset of every hindcast. However, the hindcasts are forced with observed precipitation and streamflow. Hence the hindcasting record is similar to a simulation record, but with the added benefit of an ARMA error correction. This introduces a lead time dependency in the skill of the hindcast. At zero lead time, the hindcast has perfect skill; with increasing lead time, forecast errors increase in magnitude and skill deteriorates. These resulting forecast errors are largely due to hydrological uncertainties. When the effect of the ARMA procedure has worn out, forecast skill reduces to that of the hydrological simulations. The procedure is similar to that followed by Bogner and Pappenberger, 2011; an alternative approach to this would be to use the latest available observation as a predictor in the uncertainty model; this approach was taken by, among others, Seo et al. (2006).

As the experimental setup is quite complex, the Table 1 should be helpful in distinguishing the various types of model runs from each other.

2.6. Quantile Regression

In the present paper, in both scenarios, uncertainty is estimated using Quantile Regression (QR; Koenker and Bassett, 1978; Koenker and Hallock, 2001; Koenker, 2005). QR is a regression technique for estimating the quantiles of a conditional distribution. As the sought relations are conditional quantiles rather than conditional means, quantile regression is robust with regards to outliers. Quantile Regression does not make any prior assumptions regarding the shape of the distribution; in that sense, it is a non-parametric technique. It is, however, highly parametric in the sense that, for every quantile of interest, parameters need to be estimated. QR was developed within the economic sciences, and is increasingly used in the environmental sciences (see, for example, Bremnes, 2004; Nielsen et al., 2006). Specific applications in the hydrological sciences include Weerts et al. (2011), Roscoe et al. (2012) and López López et al. (2014).

In the present work, Quantile Regression is used to estimate lead time n specific conditional distributions of streamflow,

Table 1
Summary of characteristics of the various types of model runs used in the present manuscript

Type of model run	Forcings	Characterised by lead-time?	Use
Simulations	Observed meteorological parameters	No	Estimation of hydrological uncertainty in cases where no data assimilation is applied
Perfect forcing (hydrological) forecasts	Observed meteorological parameters	Yes	Estimation of hydrological uncertainty in cases where data assimilation is applied
Deterministic and ensemble (hydrological) forecasts	Meteorological forecasts	Yes	These comprise the raw hydrological forecasts

$$\phi_n = \{Q_{n,\tau_1}, Q_{n,\tau_2}, \ldots, Q_{n,\tau_T}\} \quad (3)$$

where T is the number of quantiles τ ($\tau \in [0,1]$) considered. If T is sufficiently large and the quantiles τ cover the domain $[0,1]$ sufficiently well, we consider ϕ_n to be a continuous distribution. Here, $T = 50$ and $\tau \in \{0.01, 0.03, \ldots, 0.99\}$,

$$\phi_n = \{Q_{n,\tau=.01}, Q_{n,\tau=.03}, \ldots, Q_{n,\tau=.99}\} \quad (4)$$

We assume that, cf. Weerts et al., 2011, separately for every lead time n considered and for every quantile τ, there is a linear relationship between the streamflow forecast S and its verifying observation Q,

$$Q_{n,\tau} = a_{n,\tau} S_n + b_{n,\tau} \quad (5)$$

where $a_{n,\tau}$ and $b_{n,\tau}$, are the slope and intercept from the linear regression. Quantile Regression allows for finding the parameters $a_{n,\tau}$ and $b_{n,\tau}$ of this linear regression by minimising the sum of residuals,

$$\min \sum_{j=1}^{J} \rho_{n,\tau}(q_{n,j} - (a_{n,\tau} s_{n,j} + b_{n,\tau})) \quad (6)$$

where $\rho_{n,\tau}$ is the quantile regression weight for the τ^{th} quantile, $q_{n,j}$ and $s_{n,j}$ are the jth paired samples from a total of J samples, and $a_{n,\tau}$ and $b_{n,\tau}$ the regression parameters from the linear regression between streamflow forecast and observation. By varying the value of τ, the technique allows for describing the entire conditional distribution.

In the present work, solving Eq. 6 was done using the `quantreg` package (Koenker, 2013) in the R software environment (R Core Team, 2013). Figs. 5 and 6 show the joint distributions of forecast-observation pairs as well as a selection of estimated quantiles; these plots are discussed in the Results and Analysis section.

2.7. Verification strategy

Forecast quality in the two scenarios is assessed using visual exploration of forecast hydrographs, examination of graphical measures of reliability and sharpness as well as a selection of metrics (presented as skill scores) for both probabilistic forecasts and single valued derivatives thereof. The metrics and skill scores are described in detail in Appendix A.

Reliability is the degree to which predicted probabilities coincide with the observed relative frequencies, given those forecast probabilities. The degree to which the forecasts are reliable is indicated by reliability plots that plot the conditional relative frequency of observations against forecast probabilities. Proximity to the 1:1 diagonal, where observed frequency equals predicted probability, indicates higher reliability.

The refinement of a set of forecasts refers to the dispersion of the marginal distribution of predicted probabilities. A refinement distribution with a large spread implies refined forecasts, in that different forecasts are issued relatively frequently and so have the potential to discern a broad range of conditions. This attribute of forecast refinement often is referred to as sharpness in the sense that refined forecasts are called sharp (Wilks, 2011). Often, sharpness is summarized as the ability to forecast extreme probabilities such as 0% and 100%. This is closely related to the width, or spread, of an ensemble forecast. Large spreads are unlikely to yield 0% or 100% event probabilities. Hence, here we have measured the width of the ensemble forecasts. These "sharpness plots" show the empirical cumulative distribution of the width of the 10th–90th quantiles of the probability forecasts. In this context, sharpness measures the degree of confidence (narrowness of spread) afforded by the forecasts.

Metrics that describe the quality of single valued forecasts include the correlation coefficient (COR), relative mean error (RME), mean absolute error (MAE) and the root mean squared error (RMSE). The correlation coefficient describes the degree of linear dependence between the observation and the forecast. The RME or fractional bias measures the average difference between the forecast and the observation, relative to the mean of the observations. MAE measures the mean absolute difference between a set of forecasts and corresponding observations. RMSE provides the square root of the average mean square error of the forecasts. It has the same unit as the forecasts and the observations. In each of these four metrics the forecast mean is used as a single valued forecast.

In terms of the probabilistic characteristics of forecasts, the overall accuracy is measured with the Brier Score (BS) and the mean Continuous Ranked Probability Score (CRPS). The BS comprises the mean square error of a probability forecast for a discrete event, where the observation is an indicator variable. The CRPS measures the integral square error of a probability forecast across all possible event thresholds, again assuming that the observation is deterministic.

In the present paper, both BS and CRPS of the forecasts under consideration are presented as a skill relative to the BS and CRPS of the climatological forecast, that is, the climatology of streamflow observations as derived from the sample of verification pairs.

For discrimination, the trade-off between correctly predicted events (true positives, or hits) and false alarms (false positives) is considered. Hit rate is plotted versus false alarm rate in the ROC curves. The area under the curve (AUC) is a measure of discrimination; this is expressed as the Relative Operating Characteristic score (ROCS), which factors out the climatological AUC of 0.5, i.e. $2AUC - 1$. Forecast *value* is measured using the Relative Economic Value (REV; Murphy, 1985; Zhu et al., 2002). REV ($V[-]$) is calculated by matching the occurrences of hits, misses, false alarms and correct negatives ('quiets') with their consequences (Table 2). It is expressed on a scale from negative infinity to 1, where $V = 0$ is the situation in which there is no forecasting and warning system present and $V = 1$ is the situation in which there is a perfect forecasting and warning system present. Negative values imply that the warning system introduces more costs than benefits. The REV is expressed as a function of the users cost-loss rate r.

Verification was performed at the same timestep as the post-processing; results are shown for a selection of lead times only. For verification, the open source Ensemble Verification System (Brown et al., 2010) is used. EVS takes ensemble forecasts as input. Here, predictive uncertainty is expressed by quantiles rather than ensemble members, but the 50 quantiles are equally spaced, and ensemble members may be interpreted as quantiles of the underlying probability distribution from which they are sampled (e.g. Bröcker and Smith, 2008).

Conditional quality and skill is determined by calculating verification metrics for increasing levels of the non-exceedence climatological probability, P, ranging from 0 to 1. Essentially, $P = 0$ constitutes an unconditional verification for continuous measures, such as the CRPSS, as all available data pairs are considered

Table 2
Contingency table. The consequences of the items listed are in brackets. Taken from Verkade and Werner, 2011.

	Event observed	Event NOT observed	$\sum$
Warning issued	Hits $h(C + L_u)$	False alarms $f(C)$	w
Warning NOT issued	Missed events $m(L_a + L_u)$	Quiets/correct negatives $q(-)$	w'
$\sum$	o	o'	N

(Bradley and Schwartz, 2011). Conversely, at $P = 0.95$, only the data pairs with observations falling in the top 5% of sample climatology are considered; this amounts to approx. 60 pairs here for the Meuse case and approx. 150 pairs for the Rhine case (Fig. 2).

The BSS, ROCS and REV measure forecast skill for discrete events. The BSS, ROCS and REV are, therefore, unknown for thresholds corresponding to the extremes of the observed data sample, nominally denoted by $P = 0$ and $P = 1$).

Sampling uncertainties were quantified with the stationary block bootstrap (Politis and Romano, 1994). Here, blocks of adjacent pairs are sampled randomly, with replacement, from the J available pairs in each basin. Overlapping blocks are allowed, and the average length of each block is determined by the autocorrelation of the sample data. In both cases, an average block length of 10 days was found to capture most of the autocorrelation (some interseasonal dependence remained). The resampling was repeated 1,000 times, and the verification metrics were computed from each sample. Confidence intervals were derived from the bootstrap sample with a nominal coverage probability of 0.9, i.e. $[0.05, 0.95]$. The intervals should be treated as indicative and do not necessarily provide unbiased estimates of coverage probabilities, particularly for rare events (Lahiri, 2003). Also, observational uncertainties were not considered.

These sampling uncertainty intervals provide information as to the 'true value' of the metric or skill considered. Unfortunately, the intervals cannot be used for a formal statistical analysis as the verification samples are not strictly independent. Hence in the present paper, the comparison between scenarios is (necessarily) based on a qualitative description of the uncertainty intervals.

3. Study basins, models and data used

To enhance the robustness of the findings presented in this paper, the experiment was carried out on two separate study basins. These comprise forecasting locations in two basins with different characteristics, where hydrological models are forced with different atmospheric ensemble forcing products.

3.1. Meuse

3.1.1. Basin description

The river Meuse (Fig. 3) runs from the Northeast of France through Belgium and enters the Netherlands just south of Maastricht. It continues to flow North and then West towards Dordrecht, where it meets the Rhine before discharging into the North Sea near Rotterdam. Geology and topography vary considerably across the basin. The French Meuse basin is relatively flat and has thick soils. The mountainous Ardennes are relatively high and steep and the area's impermeable bedrock is covered by thin soils. Average annual basin precipitation varies around 900 mm. The Meuse is a typically rain-fed river; long lasting, extensive snowpacks do not occur. Fig. 4 shows the distribution of streamflow at the forecasting locations considered in this study. Near Maastricht, average runoff equals approx. 200^3/s. Temporal variability can be large as, during summer, streamflow can be less than 10^3/s, while the design flood, associated with an average return period of 1250 years, has been established at approx. 3000^3/s.

This study explicitly looks at St Pieter, which is near where the river enters The Netherlands. Water levels in the Belgian stretch of

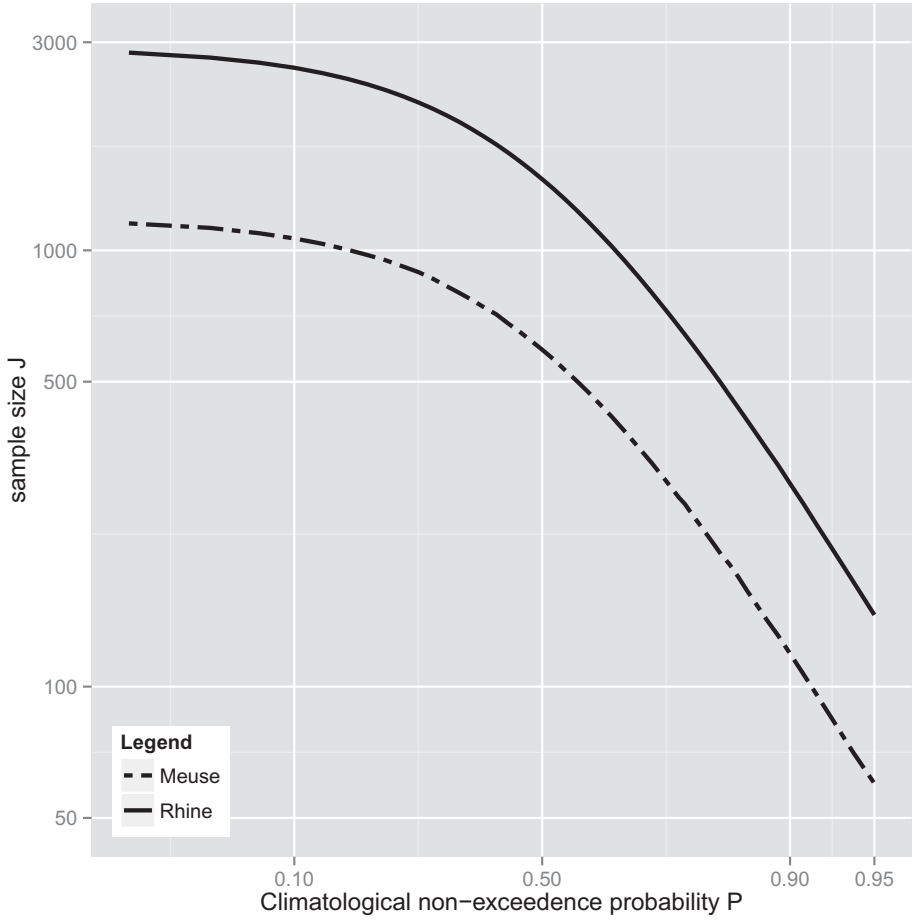

Fig. 2. Sub-sample size as a function of the probability P of non-exceedence of the observation in the climatological record.

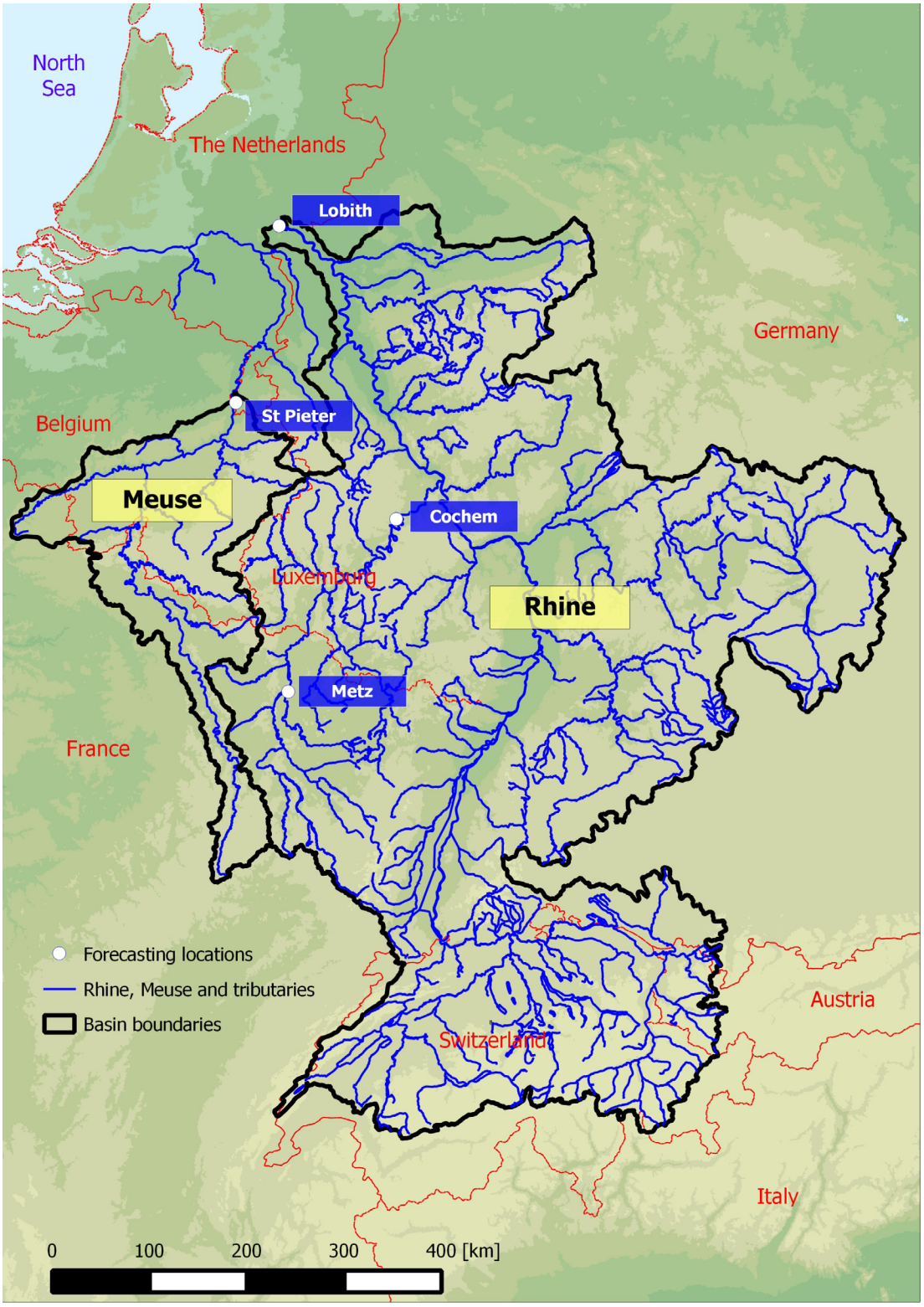

Fig. 3. Map of the Meuse and Rhine basins and the forecasting locations that are considered in this manuscript.

the Meuse, just upstream of the Dutch-Belgian border, are heavily regulated by large weirs. These, together with the locks that have been constructed to allow ships to navigate the large water level differences, cause relatively high fluctuations in discharge. The manual operations that lead to these fluctuations are not communicated with the forecasting agency across the border in The Netherlands, which introduces additional uncertainties with respect to future streamflow conditions.

3.1.2. Models used

The forecasting system contains an implementation of the HBV rainfall-runoff model (Bergström et al., 1995). This is a semi-

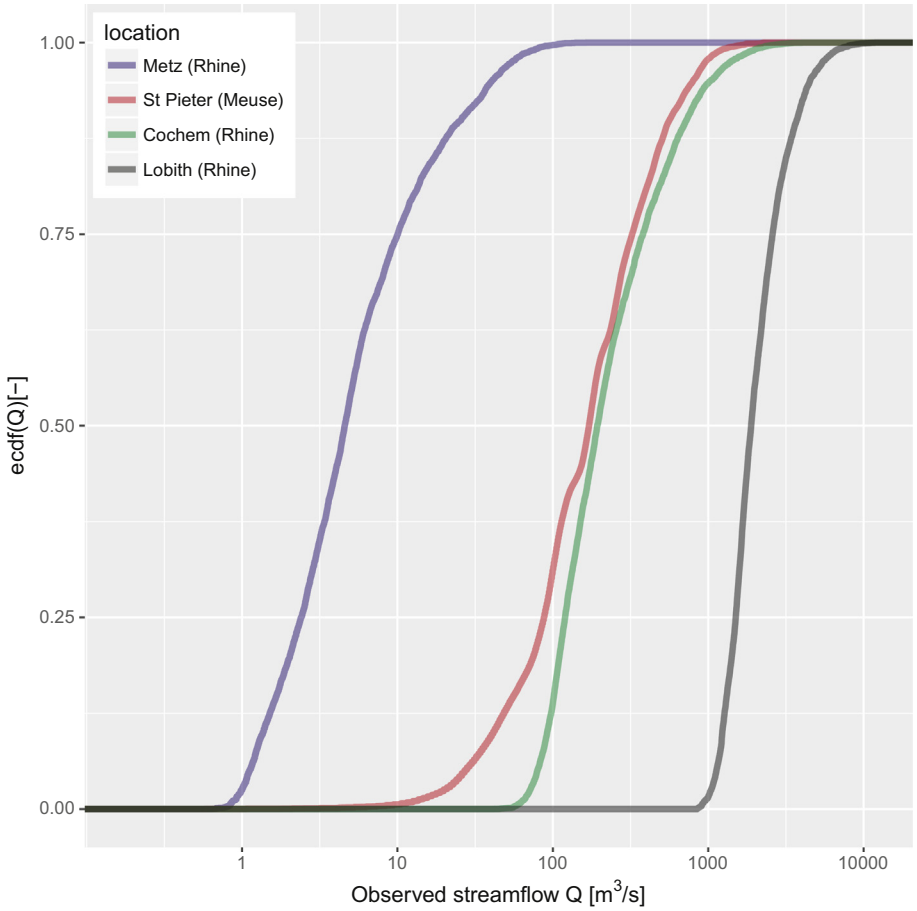

Fig. 4. Distribution of streamflow observations at the forecasting locations considered in this study.

lumped, conceptual hydrological model, which includes a routing procedure of the Muskingum type. The model schematisation consists of 15 sub-basins jointly covering the Meuse basin upstream of the Belgian-Dutch border, which is very near the St Pieter forecasting location. The model runs at a one-hour time step. Inputs to the model consist of temperature and precipitation forcings; actual evaporation is estimated from a fixed annual profile that is corrected using temperature forecasts. The model simulates both streamflow generation and streamflow routing in natural flow conditions only. Thus, it does not include models of human interference that occurs at weirs and sluices. This interference occurs mainly at low flows; at high flows, weirs are drawn. Hence, at low flows, considerable uncertainty is associated with model outcomes.

3.1.3. Forecasts and observations used

COSMO-LEPS is the ensemble implementation of the COSMO model, a non-hydrostatic, limited-area atmospheric prediction model. Its 16 members are nested on selected members of the ECMWF-EPS forecasts. COSMO-LEPS runs twice daily on a 10 km grid spacing and 40 vertical layers. It covers large parts of continental Europe including the Meuse basin. For the present experiment, approx. 1400 historical COSMO-LEPS forecasts were available (Fig. 2): one every day between mid 2007 and early 2011. The forecasts have a 1-h time step and have a maximum lead time of 132-h, i.e. 5.5 days. Within the operational forecasting system, the lead time is artificially extended to 7 days through assuming zero precipitation and 8° C temperature for the lead times ranging from 132-h through 168-h. The 36-h lead time gain more or less coincides with the time required for a flood wave to cover the distance from Chooz (near the French–Belgian border) to St Pieter. As a general rule, about half of the streamflow volume originates from the basin upstream from Chooz.

From the 16 members, a single member was isolated to serve as the deterministic forecast. Note that while the results for a single deterministic forecast are presented here, the dressing was in fact done 16 times, using each of the available 16 members as a single deterministic forecasts. Each of these 16 dressed deterministic forecasts behaves similarly with respect to the 'competing' scenario — hence only one of these is presented in this paper.

Hourly streamflow observations St Pieter as well as temperature and precipitation observations within the Meuse basin were obtained from the Water Management Centre of The Netherlands.

3.2. Rhine

3.2.1. Basin description

The river Rhine runs from the Swiss Alps along the French-German border, through Germany and enters The Netherlands near Lobith, which is situated upstream of the Rhine-Meuse delta, and is often considered the outflow of the Rhine. At Lobith, the basin area equals approx. 160,000 km^2. During spring and early summer, a considerable fraction of flow at the outlet originates from snowmelt in the Swiss Alps. Fig. 3 shows the basin location, elevations and the forecasting locations that were used in this work. These are Metz, Cochem and Lobith. Metz is located in the headwaters of the river Moselle, of which Cochem is the outlet.

3.2.2. Models used

The forecast production system that was used to create simulations and hindcasts for the Rhine is a derivative of the operational system that was mentioned above. The system contains an implementation of the HBV rainfall runoff model (Bergström et al., 1995). The Rhine model schematisation consists of 134 sub-basins jointly covering the entire basin. The models run at a daily time step. Inputs to the model consist of temperature and precipitation forcings; actual evaporation is estimated from a fixed annual profile that is corrected using temperature forecasts.

3.2.3. Forecasts and observations used

For observations of precipitation, the CHR08 dataset (Photiadou et al., 2011) was used. This dataset was prepared specifically for the HBV model used here and covers the period 1961 through 2007. The spatial scale of the observations coincides with the 134 HBV sub-basins. Temperature observations originate from version 5.0 of the E-OBS data set (Haylock et al., 2008), and are available from 1951 through mid 2011. These forcings were available at a time step of one day. The observations are used to force the hydrological model in historical mode to estimate the initial conditions at the onset of a hydrological forecast, as well as in simulation mode.

Predicted forcings consisted of the ECMWF reforecast dataset, comprising medium-range EPS forecasts with 5 ensemble members (Hagedorn, 2008). These reforecasts were produced using the current operational model (Cy38r1 with a 0.25 degrees horizontal resolution). The forecasts were temporally aggregated to a one day time step, which coincided with that of the hydrological model used, and go out to a maximum lead time of 240-h, i.e. 10 days. The gridded forecasts were spatially averaged to the HBV sub-basin scale. For this work, approx. 2,900 reforecasts were available (Fig. 2), covering the period 1990–2008.

Similar to the Meuse case, the deterministic forecasts used in this study consist of a randomly chosen ensemble member from each of the available ensemble forecasts. Each of the members was used to create a deterministic forecast which was subsequently dressed and analysed. However, results for one of these forecasts is presented only.

Hourly streamflow observations for the hydrological stations within the Rhine basin were obtained from the Water Management Centre of The Netherlands.

4. Results and analysis

4.1. Post-processing of single valued forecasts

Scatter plots of single-valued forecasts and observations are shown in Figs. 5 and 6 for St Pieter and Rhine locations, respectively. Two datasets are shown: (i) the forecasts with perfect forcings (*simulations* in the Rhine case) and (ii) the deterministic forecasts. In all plot panels, the horizontal axes are identical to the vertical axes and the 1:1 diagonal is emphasised. The forecast-observation pairs are plotted in a transparent colour; areas that show up darker comprise multiple pairs plotted on top of one another. A selection of estimated quantiles is superimposed on the scatter plots, with the median in red and the 5th, 10th, 25th, 75th, 90th and 95th percentiles in blue.

The pairs and the estimated quantiles in the St Pieter figure (Fig. 5) show that the perfect forcing pairs (bottom row) are closer to the diagonal than the deterministic forecast pairs (top row). This is because the residuals between the perfect forcings forecast and the observations comprise the hydrological uncertainties only. The plots also show that the median quantile of the pairs comprising the deterministic forecasts has a shallower slope than the diagonal. This indicates an overforecasting bias: the majority of pairs is located below, or to the right of the diagonal. The median of the pairs comprising the perfect forcing forecasts shows a slope almost equal to, or higher than that of the 1:1 diagonal. The latter indicates underforecasting: most of the pairs are located above, or to the left of the 1:1 diagonal. Both sets of pairs show that the spread increases with increasing forecast lead time and that higher values of flow have higher spread in real units.

In the Rhine case (Fig. 6), the simulations are independent of forecast lead time. The difference between the spread of pairs based on the simulations and that of the deterministic forecasts is less obvious, especially when the shorter lead times are considered. Without exception, the median forecast is located below the diagonal which indicates an overforecasting bias.

4.2. Forecast hydrographs

Sample forecasts or predictive distributions for both scenarios are shown in Fig. 7. The rows show the cases that use deterministic (top) and ensemble (bottom) forecasts, with the raw forecasts indicated by thick blue lines and the dressed forecasts by thin grey lines. Note that the raw cases show one or multiple *traces*, whereas for the dressed cases, *quantiles* are shown (which should not be confused with ensemble traces).

By construction, ensemble dressing corrects for under-dispersion and, therefore, increases the ensemble spread. In this example, the spread of the dressed single-valued forecasts is larger than the spread of the dressed ensemble forecasts. It is also noticeable that the raw ensemble forecast fails to capture many observations, whereas the dressed forecasts capture all.

Please note that this is an example on a particular date and not necessarily the general behaviour of all forecasts that are post-processed.

The example forecasts also show an artefact associated with statistical post-processing, namely that the most extreme quantiles are relatively noisy This originates from the increased sampling uncertainty associated with estimating extreme quantiles.

4.3. Single-valued forecast verification

Generally speaking, COR, RME and RMSE follow similar patterns. Each worsens with increasing lead time and with increasing value of the verifying observation as indicated by P. In this manuscript, only the RME is shown (Fig. 8).

The correlations (plot not shown) are highest for Lobith, followed by Cochem, Metz and St Pieter. While the correlations are generally positive, they approach zero at St Pieter for higher P at longer forecast lead times. Both the patterns and values of the correlation coefficient (as function of lead time and P) are similar across the two scenarios. Only at the longer forecast lead times and at St Pieter do they differ. In those cases, the dressed ensembles outperform the dressed deterministic forecasts.

The RME plots (Fig. 8) show that, at $P = 0$, the dressed deterministic forecasts have near-perfect RME, that is, RME ≈ 0, at all forecast lead times. The dressed ensemble forecasts show a larger fractional bias, with St Pieter and Metz showing positive values and Cochem and Lobith showing negative values. For higher values of P and at longer forecast lead times, RME becomes increasingly negative. Consequently, at higher values of P and at longer forecast lead times, the dressed ensembles at St Pieter and Metz show smaller fractional bias than the dressed deterministic forecasts. The converse is true for Cochem and Lobith, where the dressed deterministic forecasts have smaller RME. The difference in RME between scenarios increases with increasing forecast lead time.

The RMSE (not shown in plots) worsens (i.e., increases) with increasing forecast lead time, increasing threshold amount, and

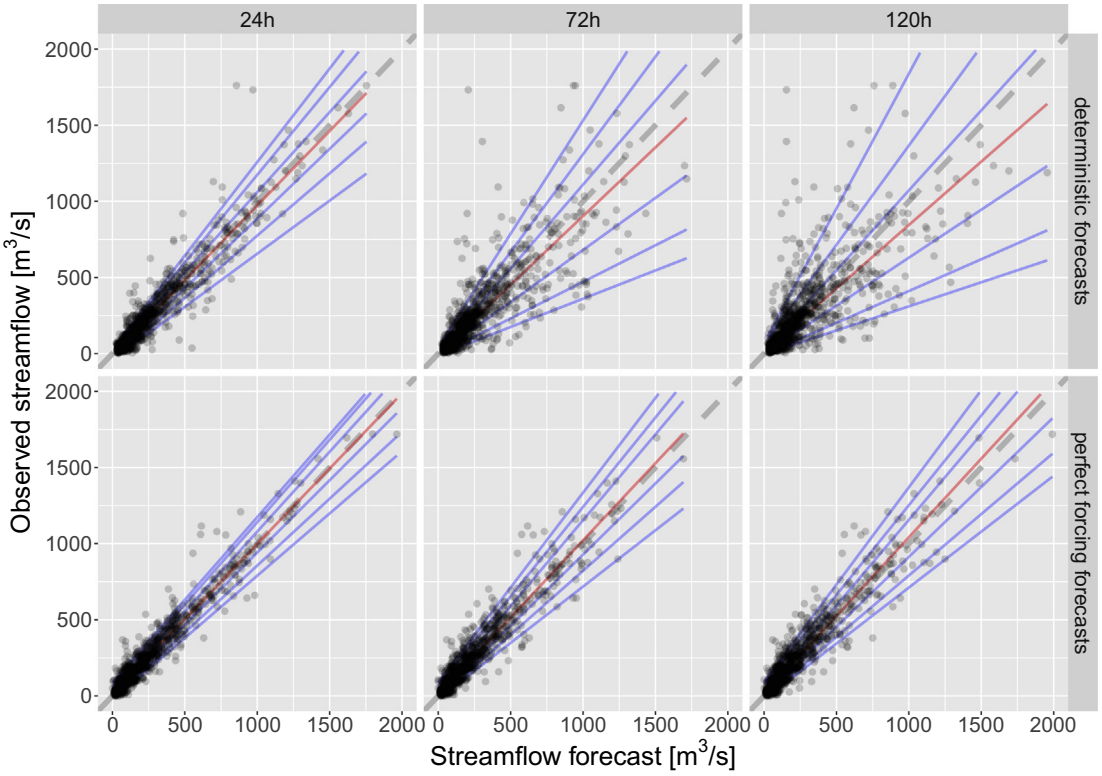

Fig. 5. Quantile Regression plots for St Pieter for deterministic (top row) and perfect forcing (bottom row) forecasts with 24-h, 72-h and 168-h forecasts (columns). The blue lines indicate the estimated 5%, 10%, 25%, 75%, 90% and 95% quantiles; the red line indicates the median quantile. As a reminder, the reader is referred to Section 2.5 for a description of how simulations differ from forecasts and perfect forcings forecasts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

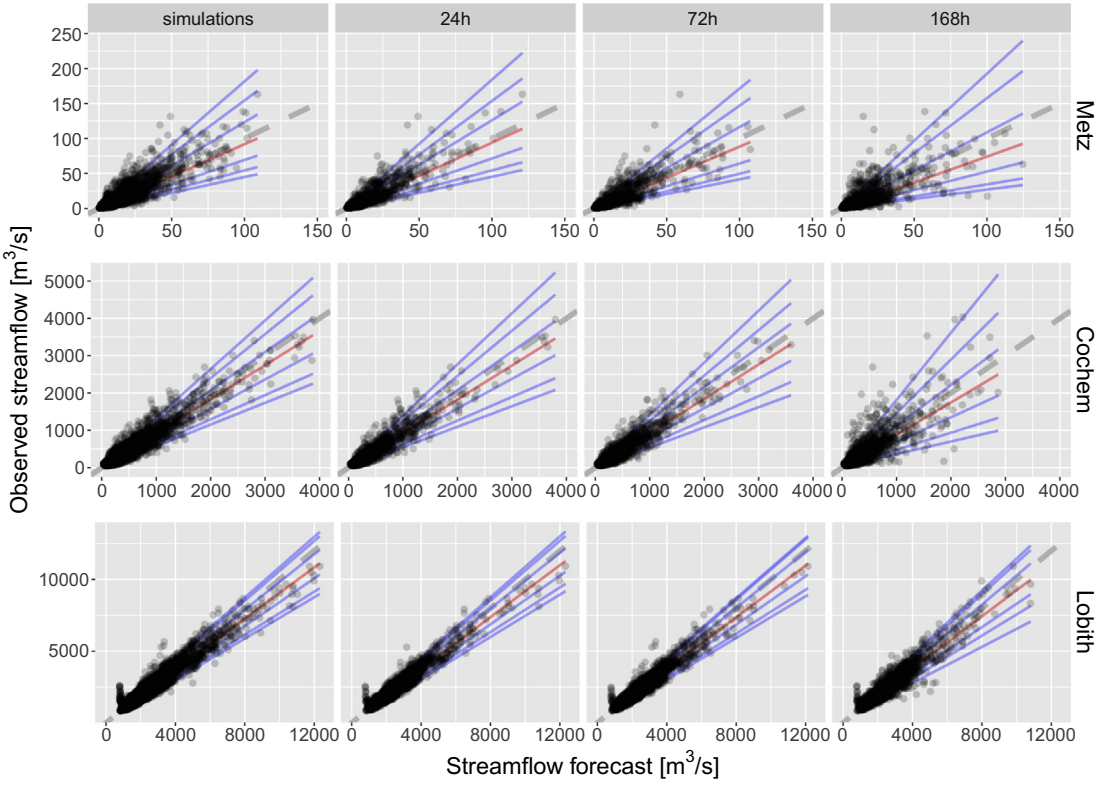

Fig. 6. Quantile Regression plots for Metz (top), Cochem (middle) and Lobith (bottom) for simulations (leftmost column) and deterministic forecasts with 24-h, 72-h and 168-h forecasts (rightmost three columns). The blue lines indicate the estimated 5%, 10%, 25%, 75%, 90% and 95% quantiles; the red line indicates the median quantile. As a reminder, the reader is referred to Section 2.5 for a description of how simulations differ from forecasts and perfect forcings forecasts.

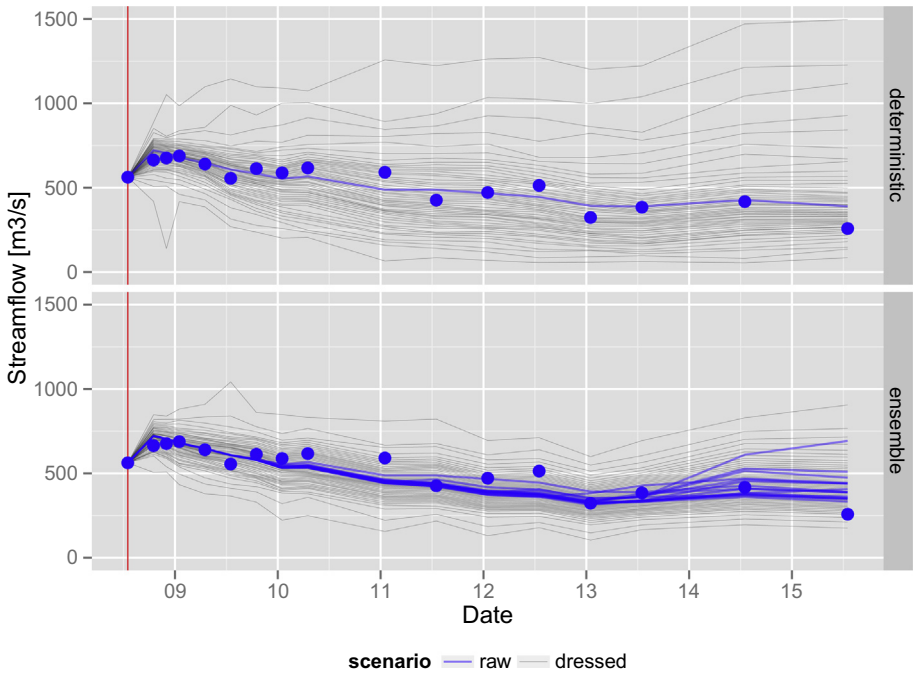

Fig. 7. Sample forecasts of the scenarios: deterministic and ensemble based forecasts in top and bottom row, respectively. The vertical red lines on the horizontal time axis indicate forecast issue time. Observed values are indicated by blue dots. Note that the lines for the raw forecasts represent one or multiple traces, whereas the lines for the dressed forecasts represent quantiles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

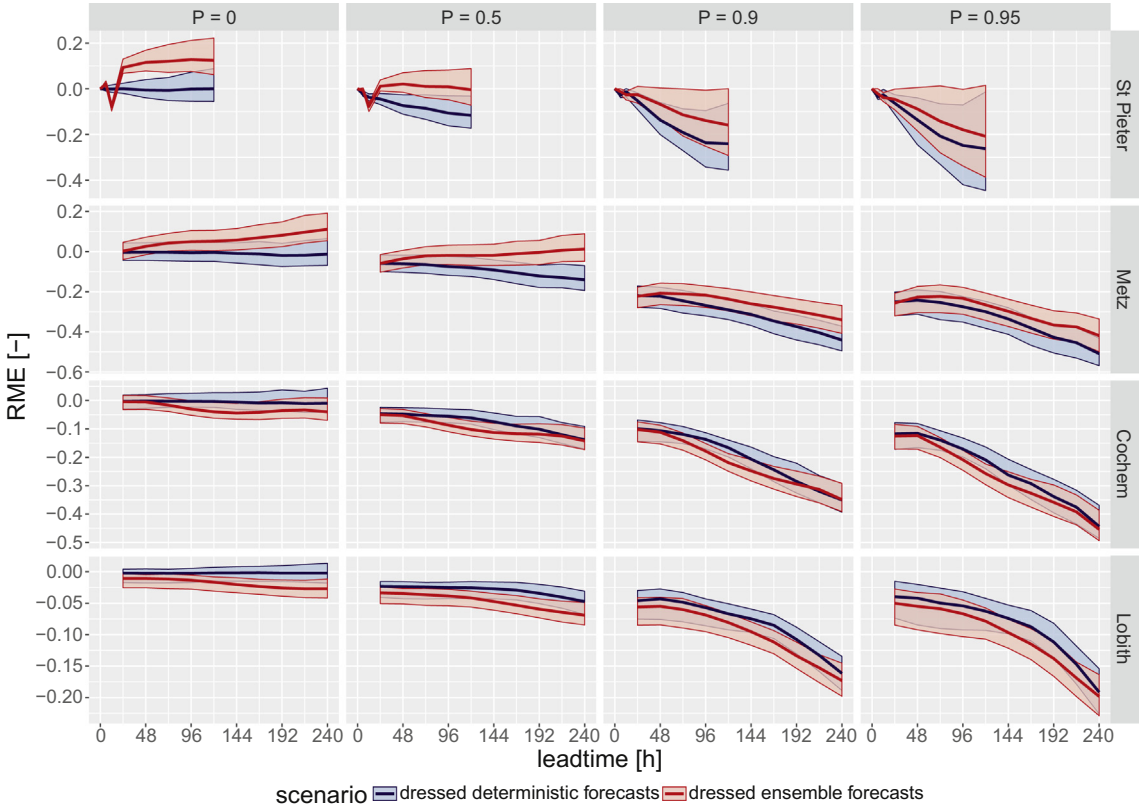

Fig. 8. Relative Mean Error (RME) as a function of lead time for several subsamples of the verification pairs (columns) and several locations (rows). Note that the vertical scales used in the various columns differ. The shading shows the width of the intervals obtained by bootstrapping the verification metric (see Section 2.7 for details).

with declining basin size. The RMSE is lower for the dressed ensemble forecasts than the dressed single-valued forecasts at most values of P. Only at Cochem and Lobith and for some values of P is the RMSE higher for the dressed ensemble forecasts.

Overall, in terms of the single valued verification measures, neither the dressed ensemble forecasts nor the dressed deterministic forecasts perform consistently better. At St Pieter and Metz, the mean of the dressed ensembles has higher quality in terms of

COR, RME and RMSE, whereas at Cochem and Lobith, the reverse is true in terms of RME and, at small ranges of P, for RMSE.

4.4. Reliability and sharpness

Reliability plots for the $P = 0.50$ and $P = 0.90$ subsamples (Figs. 9 and 10) show that both approaches are reasonably and more or less equally reliable at all leadtimes and at all locations. The one exception here is St Pieter, where the dressed deterministic forecasts appear to be more reliable than the dressed ensemble forecasts.

Reliability varies slightly with the level of extremeity of the event: the $P = 0.50$ reliability plot (Fig. 9) follows the diagonal more closely than the $P = 0.90$ reliability plot (Fig. 10). In the $P = 0.50$ reliability plot (Fig. 9), the samples are distributed much more evenly across the horizontal (as indicated by the size of the plotting positions) compared to the $P = 0.90$ reliability plot (Fig. 10). In the latter, the distribution is much less even, with a relatively large number of forecasts to be found at the 0% and 100% forecasts.

The width of the forecast distributions (Figs. 11 and 12) increases with increasing lead time and with increasing value of P. Sharpness, compared to reliability, varies more sharply (we couldn't resist the pun) with the value of P. At lead times longer than 24-h, the differences in forecast width between scenarios becomes noticeable. In all cases, the dressed ensembles result in more narrow predictive distributions than the dressed deterministic forecasts. These differences are more pronounced at higher values of P. Note that while 'narrowness' is a desirable property, it is only valuable in a decision making context if the forecasts are also reliable.

4.5. Probabilistic skill scores

For the skill scores (Fig. 13), the patterns are more important than the absolute values, as the baseline is unconditional climatology. The patterns of the Brier Skill Score are similar to those observed for other metrics: skill is highest for the largest basins and reduces with increasing forecast lead time. The BSS is generally very similar for both scenarios. Only in the case of St Pieter is there a consistent difference between the scenarios, with the dressed ensemble forecasts outperforming the dressed deterministic forecasts, but not beyond the range of sampling uncertainty.

The patterns of the mean CRPSS (Fig. 14) are similar to those of the BSS. The difference is that the CRPSS improves with increasing P. This is understandable because sample climatology is much less skilful at higher thresholds.

Often, the CRPSS is similar across the two scenarios. Again, any differences are more pronounced at longer forecast lead times and higher values of P, where the dressed ensemble forecasts are somewhat more skilful, particularly at St Pieter and Metz (but again, not beyond the range of sampling uncertainty).

4.6. Forecast value

Relative Operating Characteristic plots for the event defined by the exceedence of the 90th percentile of the observational record are shown in Fig. 15. The plots show that, in all cases, the ROC curves for both scenarios are well above the diagonal, indicating that these forecasts improve upon climatology.

At the shortest lead time shown, the curves for the two scenarios are very similar. Differences, if any, increase with increasing forecast lead time. At longer forecast lead times, the dressed

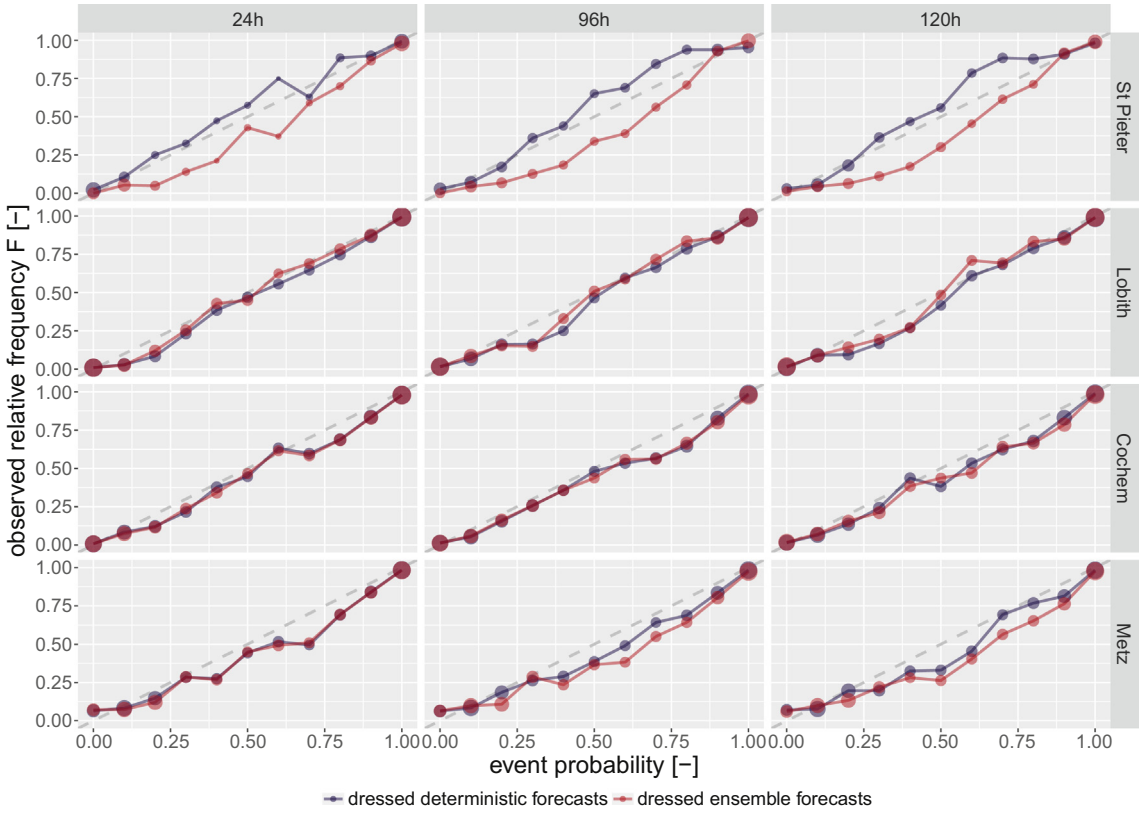

Fig. 9. Reliability plots for various lead times (columns) for several locations (rows). The plot is conditional on the observation exceeding the 50th percentile of the climatological exceedence probability (i.e., $P = 0.50$).

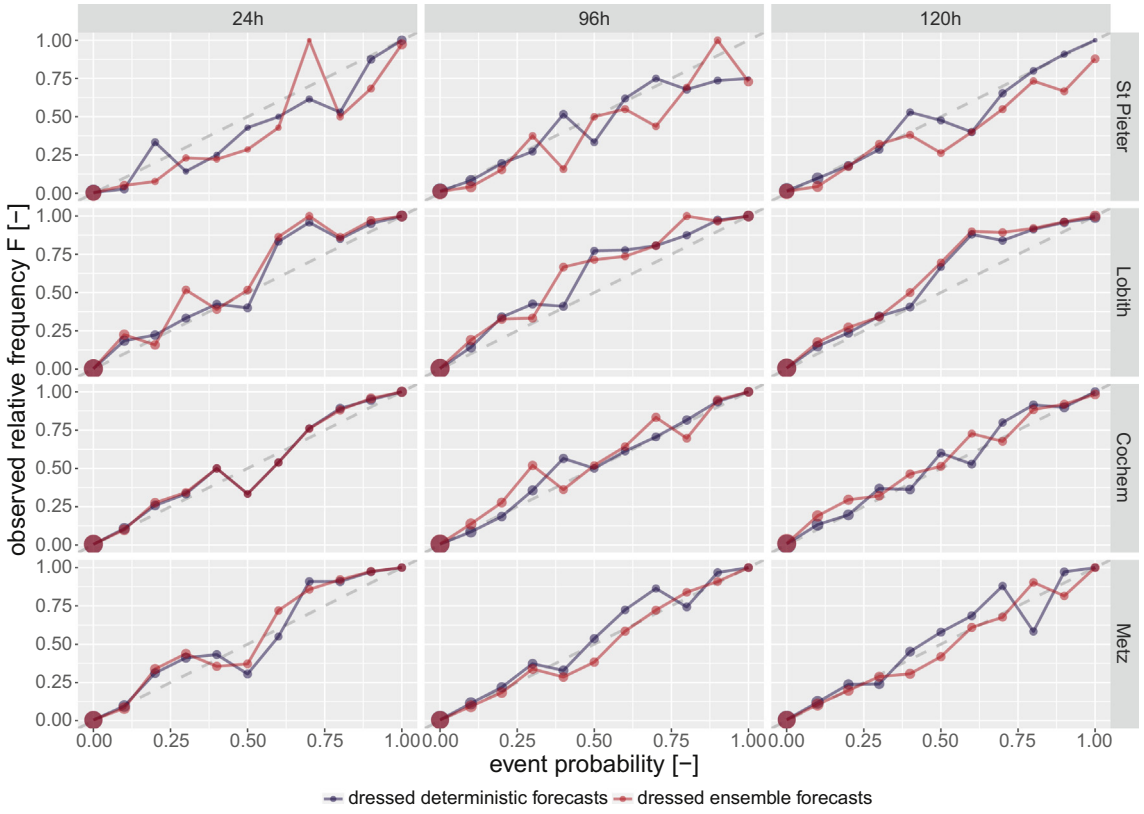

Fig. 10. Reliability plots for various lead times (columns) for several locations (rows). The plot is conditional on the observations exceeding the 90th percentile of the climatological exceedence probability (i.e., $P = 0.90$).

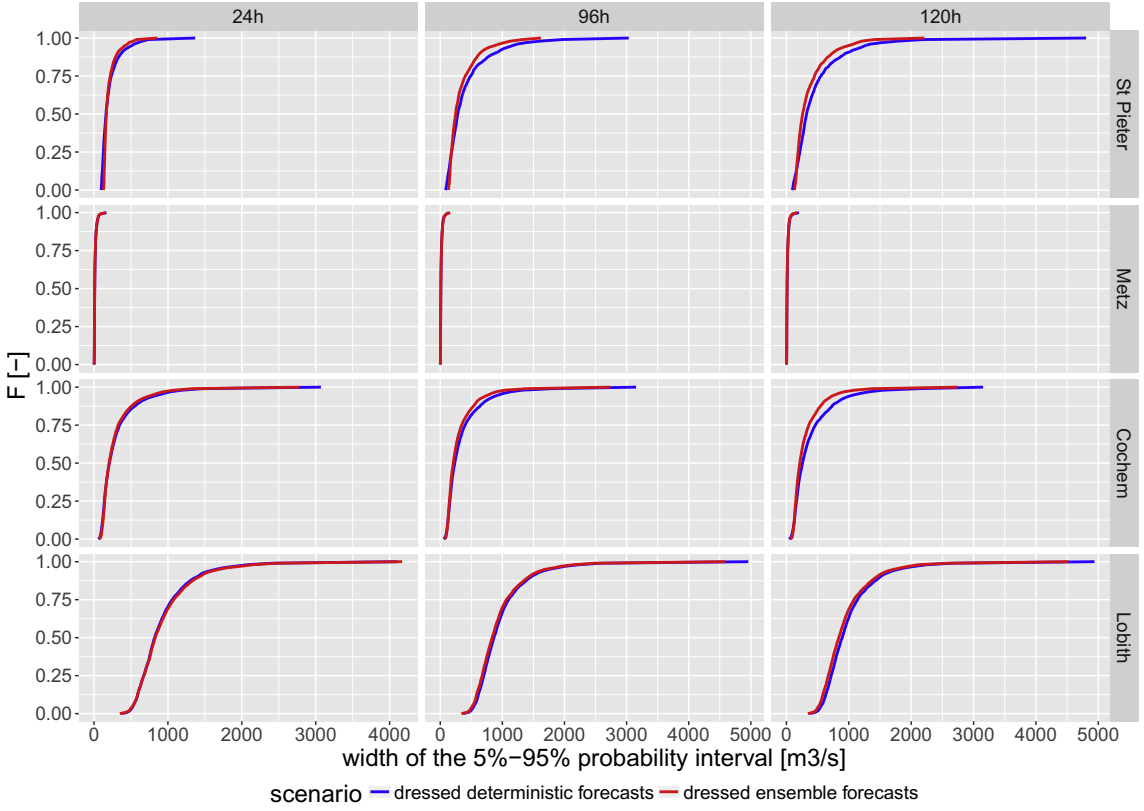

Fig. 11. Sharpness plots for various lead times (columns) for several locations (rows). The plot is unconditional, i.e. for the full data sample (i.e., $P = 0$).

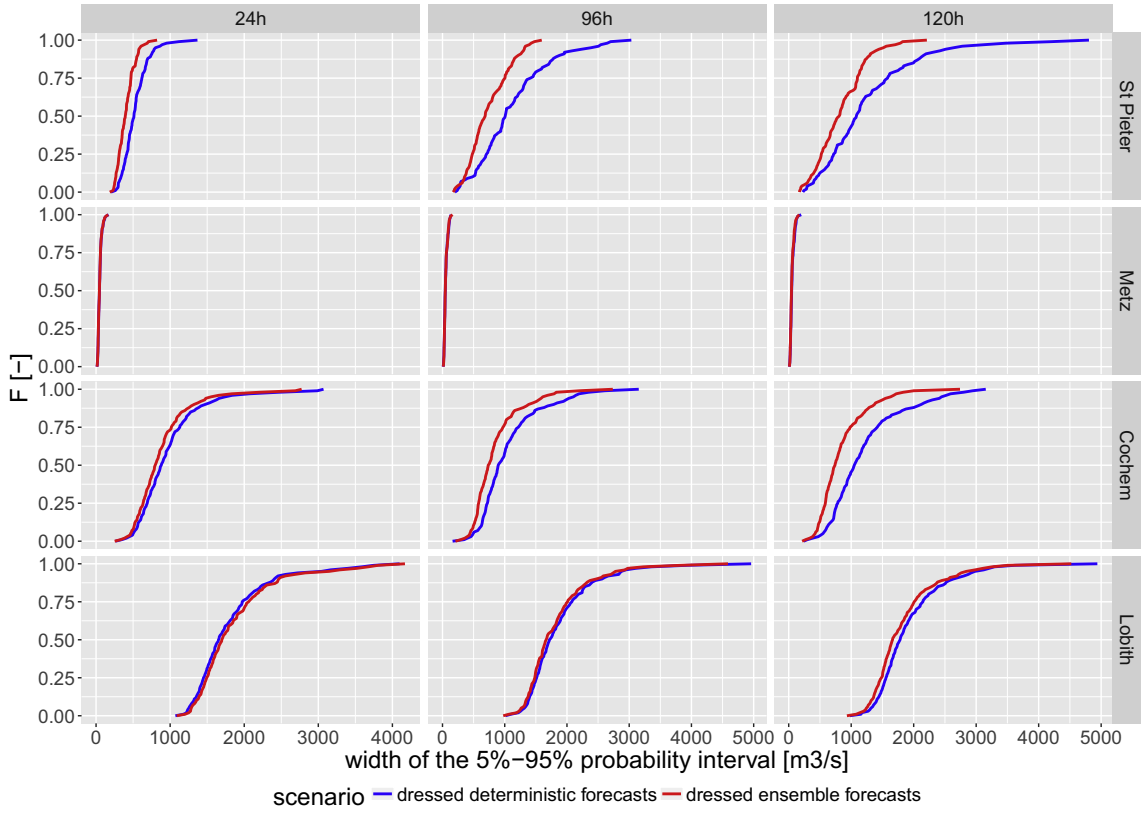

Fig. 12. Sharpness plots for various lead times (columns) for several locations (rows). The plot is conditional on the observations exceeding the 90th percentile of the climatological exceedence probability (i.e., $P = 0.90$).

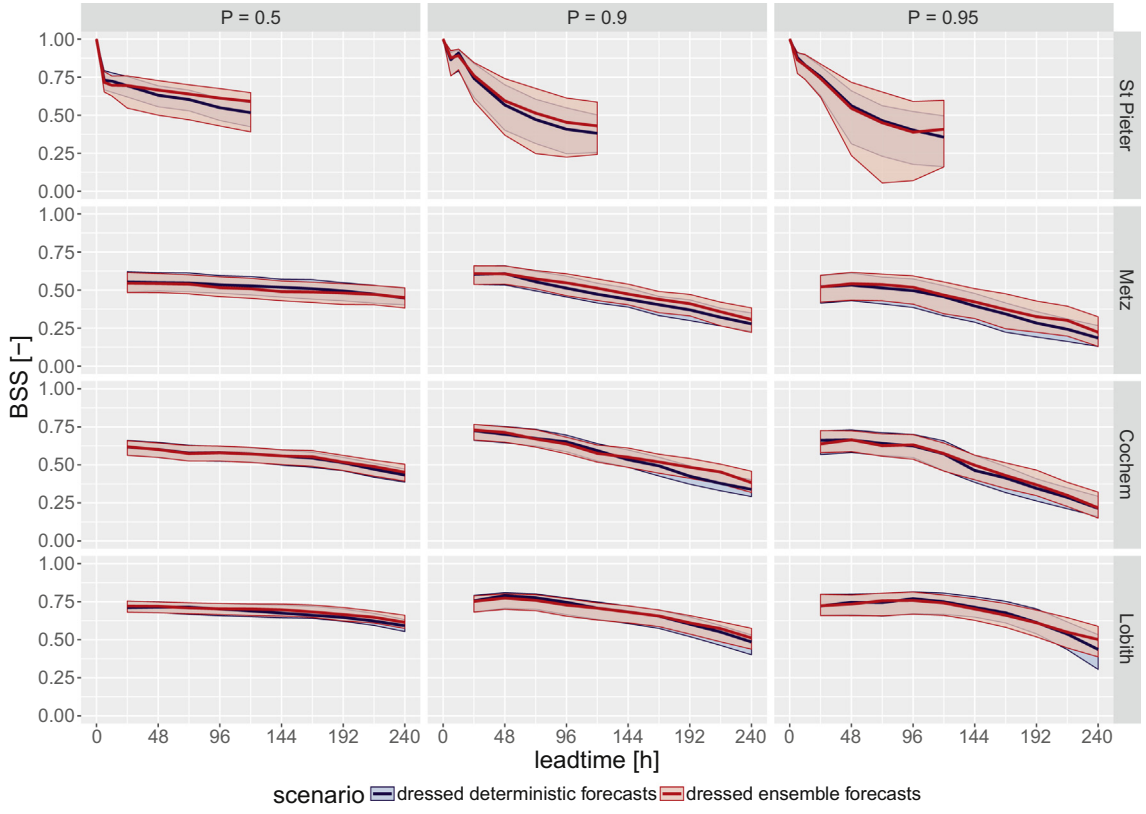

Fig. 13. Brier Skill Score (BSS) as a function of lead time for several events (columns) and several locations (rows). The shading shows the width of the intervals obtained by bootstrapping the verification metric (see Section 2.7 for details).

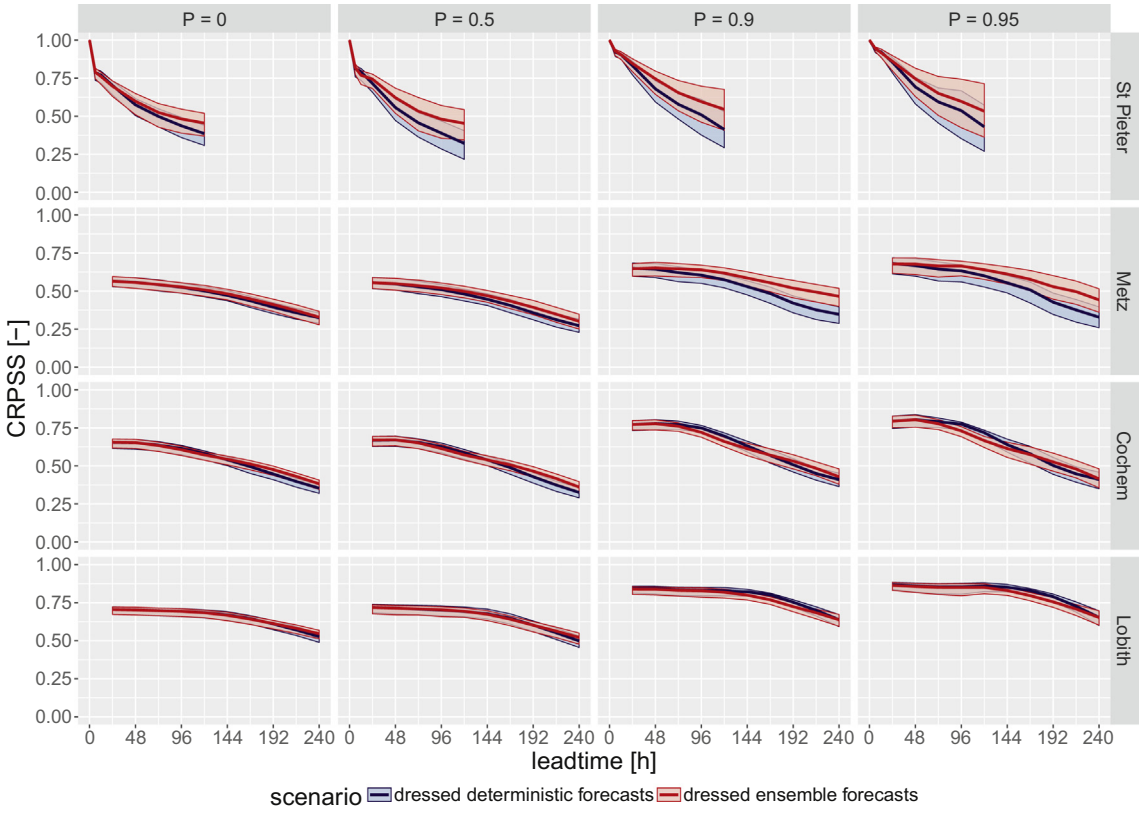

Fig. 14. Mean Continuous Ranked Probability Skill Score (CRPSS) as a function of lead time for several subsamples of the verification pairs (columns) and several locations (rows). The shading shows the width of the intervals obtained by bootstrapping the verification metric (see Section 2.7 for details).

ensemble forecasts are slightly more discriminatory than the dressed deterministic forecasts.

The associated ROC scores (Fig. 16) are very similar at most locations and forecast lead times and generally decline with increasing forecast lead time and increase with threshold amount.

The Relative Economic Value also relies on the ability of a forecasting system to discriminate between events and non-events, but assigns a cost-loss model to weight the consequences of particular actions (or inaction). In most cases, the REV of the dressed ensemble forecasts is similar to, or slightly higher than, the dressed deterministic forecasts for different values of the cost-loss ratio. Again, these differences are more pronounced at longer forecast lead times, higher thresholds, and larger values of the cost-loss ratio Figs. 17 and 18.

4.7. Analysis

The results show that, at $P = 0$, the dressed deterministic forecasts improve (albeit only marginally) on the dressed ensemble forecasts in terms of reliability and RME. However, the dressed ensemble forecasts are somewhat sharper. On balance, the dressed ensemble forecasts have slightly better RMSE and CRPSS scores.

The dressed ensemble forecasts are only slightly less reliable than the dressed deterministic forecasts at the three Rhine locations Metz, Cochem and Lobith. The differences are larger at St Pieter, where the dressed ensemble forecasts show a substantial wet bias. In this context, the dressed deterministic forecasts account for both the atmospheric and hydrologic uncertainties and correct for biases via the quantile regression, whereas this dressed ensemble forecasts do not account for under-dispersion of the meteorological forecasts. In other words: the dressing assumes that the meteorological uncertainties are well described by the ensembles whereas in reality this isn't necessarily so.

At $P = 0$, the fractional bias of the dressed deterministic forecasts is small at all forecast lead times. This is understandable, because post-processing techniques, such as quantile regression, are generally good at correcting for unconditional biases and biases conditional upon forecast value/probability (i.e. lack of reliability).

The dressed ensemble forecasts are sharper than the dressed deterministic forecasts. However, sharpness is only meaningful when the forecasts are also reliable. In the case of St Pieter at 0-h lead time, forecasts are infinitely sharp and of near perfect quality — this is due to the error correction just upstream of St Pieter.

This error correction introduces some erratic effects on the fractional bias at the shortest lead time (Fig. 8 and a relatively steep jump to less-than-perfect skills at those lead times when the effect of error correction has worn off. This is but one of the differences between the St Pieter catchment on the one hand and the Rhine catchments on the other. In general, the St Pieter results are slightly more erratic and dramatic in nature: more pronounced dependency on lead time, bigger difference in sharpness between the two scenarios. We believe this to be due to the nature of the basin which is more complex (in terms of geology and topography) than the other basins, in combination with the fact that less information is available for the Meuse: the hydrometeorological monitoring network is a lot less dense.

At higher values of P, both sets of forecasts are consistently less reliable. The dressed deterministic forecasts show a 'dry bias' where the observed relative frequency (or quantile exceedence) is higher than the predicted probability. However, at St Pieter, this conditional dry bias is offset by an unconditional wet bias, leading to reasonably reliable forecasts at higher thresholds and early forecast lead times.

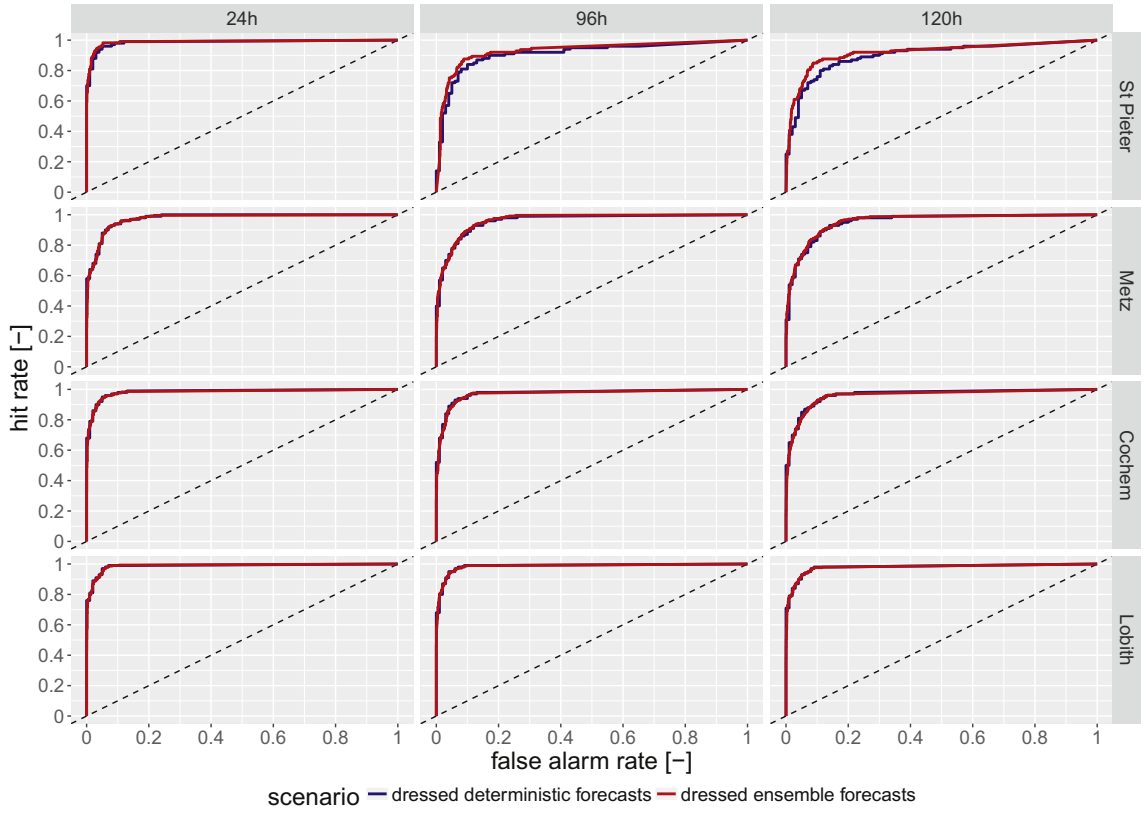

Fig. 15. ROC plots for various lead times (columns) for several locations (rows). The plot is for the event that the posterior water level exceeds the 90th percentile of the climatological exceedence probability (i.e., $P = 0.90$). The shading shows the width of the intervals obtained by bootstrapping the verification metric (see Section 2.7 for details).

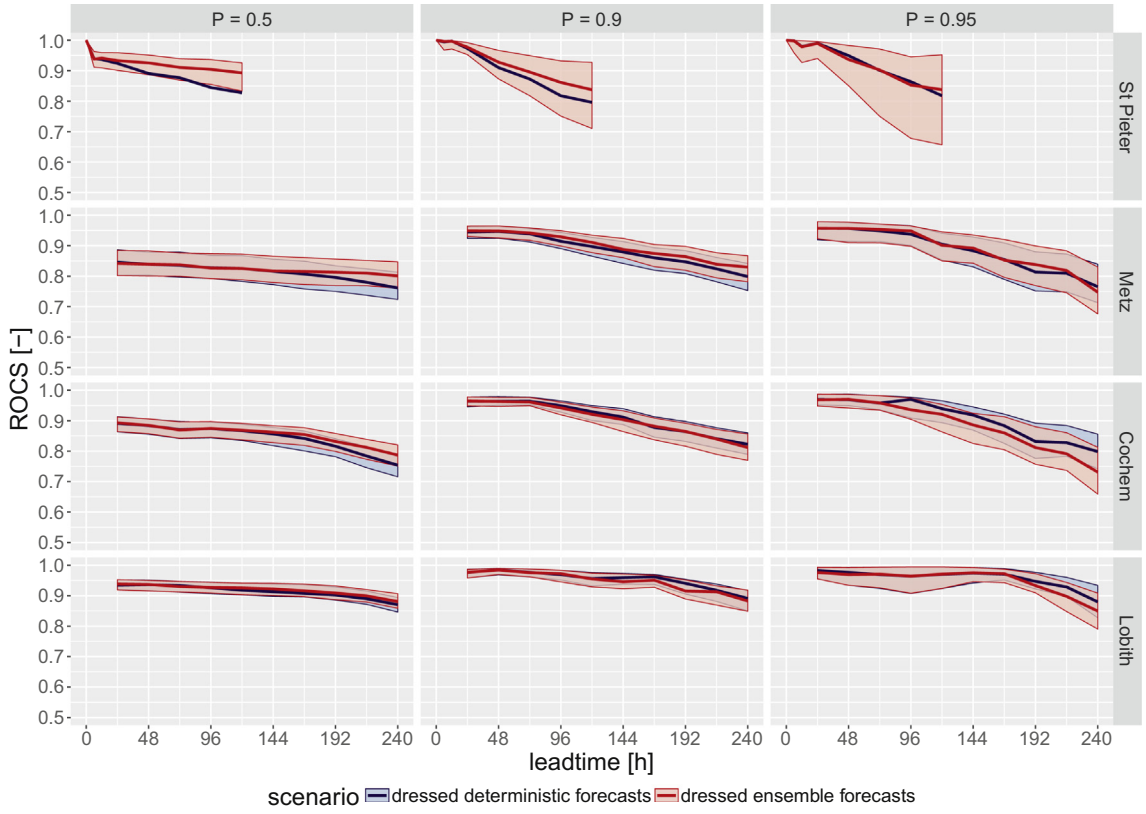

Fig. 16. Relative Operating Characteristic Score (ROCS) as a function of lead time for several subsamples of the verification pairs (columns) and several locations (rows).

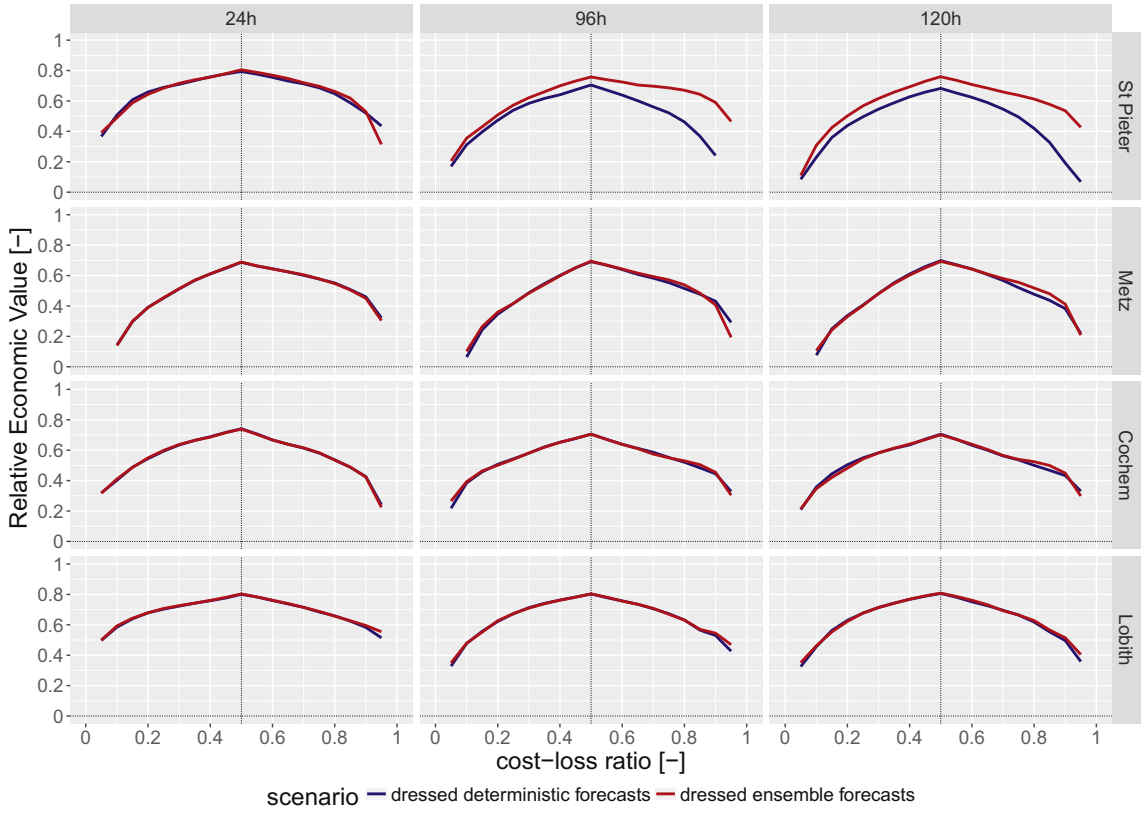

Fig. 17. Value plots for various lead times (columns) for several locations (rows). The plot is conditional on the observations exceeding the 50th percentile of the climatological exceedence probability (i.e., $P = 0.50$).

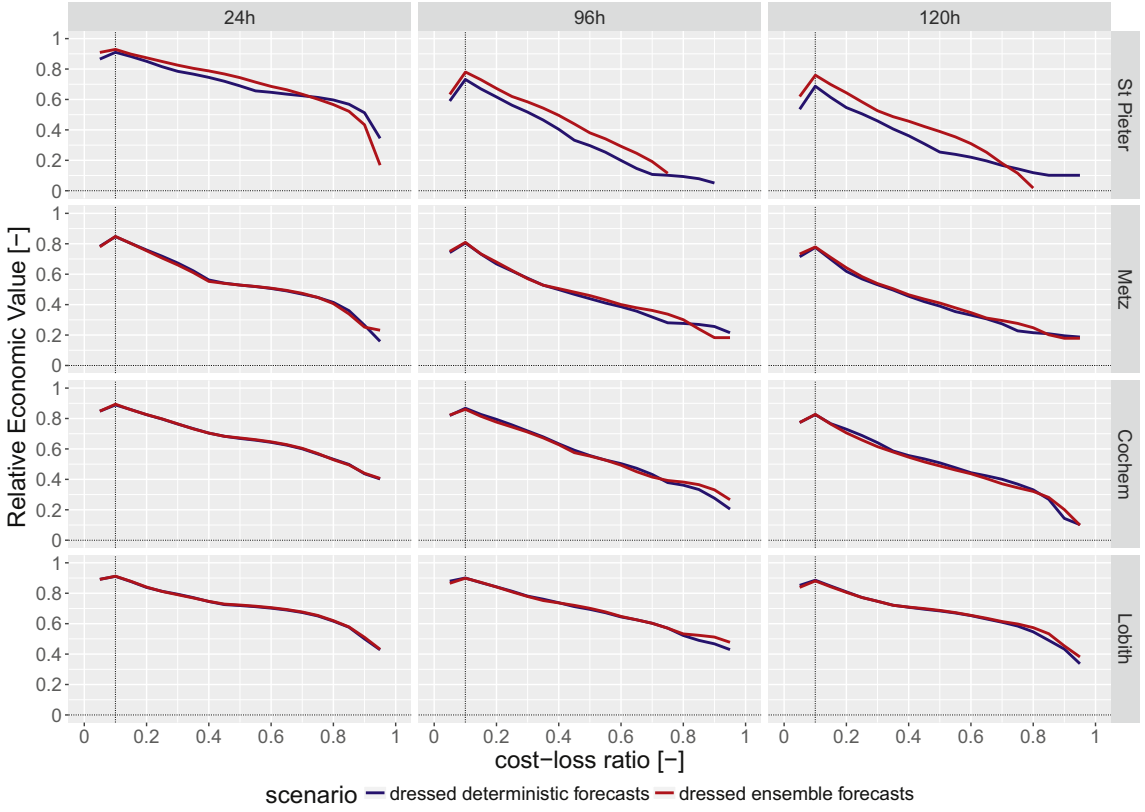

Fig. 18. Value plots for various lead times (columns) for several locations (rows). The plot is conditional on the observations exceeding the 90th percentile of the climatological exceedence probability (i.e., $P = 0.90$).

In general, the fractional negative bias (RME) increases with increasing threshold. This is consistent with the RME of the precipitation forecasts, which systematically underestimate the largest observed precipitation amounts (Verkade et al., 2013). At higher thresholds, the differences in RME between the two scenarios are similar in pattern to those in the unconditional sample. In other words, at St Pieter and Metz, the fractional bias of the dressed ensemble forecasts is smaller, while at Cochem and Lobith, the dressed deterministic forecasts show smaller fractional bias. Again, this is due to the RME in the precipitation forecasts.

The BSS, ROCS and REV, which assess the quality of the forecasts at predicting discrete events, are very similar between the two scenarios at all forecast lead times and all values of *P*. One exception is St Pieter, where the dressed ensemble forecasts improve somewhat on the dressed deterministic forecasts in terms of ROCS and REV, but not at the highest thresholds. At St Pieter and at these higher values of *P*, the dressed ensembles were both sharper and more reliable than the dressed deterministic forecasts.

5. Conclusions

We compared a source-based approach and a lumped approach for estimating predictive uncertainty in terms of various aspects of forecast quality, skill and value. The analysis shows that the dressed ensemble forecasts are sharper, but slightly less reliable than the dressed deterministic forecasts. On balance, this results in skill and value that is very similar across the two scenarios, with the dressed ensemble forecasts improving slightly on the dressed deterministic forecasts at St Pieter and Metz and the reverse being true at Cochem and Lobith.

6. Discussion

While the analysis revealed quite similar results between the scenarios, further studies or different approaches to quantifying the various sources of uncertainty could reveal larger differences. For example, a larger hindcast dataset would help to reduce the sampling uncertainties and identify any marginal differences in forecast quality, as well as supporting an analysis of higher thresholds. The quantile regression technique could be configured in alternative ways (some configurations were tested by López López et al. (2014)), or could be replaced by an alternative technique altogether. Alternative basins can be used for the experiment, and/or alternative ensemble NWP products.

As noted earlier, the quality of the dressed streamflow ensembles is dependent on the quality of the raw streamflow ensembles - and therefore on the quality of the underlying NWP ensembles. Any lack of reliability will transfer to the dressed streamflow ensembles. Such meteorological biases could be addressed through meteorological post-processing or by accounting for the effects of under-dispersion on the streamflow forecasts. Conceivably, meteorological post-processing could address error structures that are specific to the forecast variable (i.e., precipitation, temperature or other) while hydrologic post-processing then addresses error structures specific to the hydrological forecasts. Presumably, the latter errors will have been reduced by the presence of meteorological post-processing techniques. Note that in practice, a meteorological post-processing technique that is effective within the context of hydrological forecasting is yet to be found. For example, the approach taken by Verkade et al. (2013) did not result in improved hydrological forecasts, which was in line with the findings of Kang et al. (2010). Approaches such as those followed by Bennett et al. (2016), Schefzik (2016), Schefzik (2016) and Hu et al. (2016) appear to be promising.

In terms of choosing an approach, the results presented here are quite similar for both techniques. However, there are additional considerations, including the relative simplicity of dressing a deterministic forecast, data availability, and the expected additional information contributed by an ensemble of weather forecasts (versus a single forecast or the ensemble mean) in different contexts. As indicated by Pagano et al. (2014), combining ensemble and other forecasting technologies with the subjective experience of operational forecasters is an ongoing challenge in river forecasting.

Essentially, statistical modelling relies on the stationarity of the model errors or the ability to account for any non-stationarity with a reasonably simple model. In practice, however, basin hydrology changes over time with changes in climate and land-use, among other things. The lumped approach cannot easily account for this, while the source-based approach may, using information about the individual sources of uncertainty, better isolate (and model) the causes of non-stationarity.

Also, by definition, extreme events are not well represented in the observational record and frequently change basin hydrology. Thus, for extreme events in particular, basing estimates of predictive uncertainty on the observational record is fraught with difficulty. In this context, ensemble approaches to propagating the forcing and other uncertainties through hydrological models should (assuming the models are physically reasonable) capture elements of the basin hydrology that are difficult to capture through purely statistical approaches.

Acknowledgements

The authors would like to thank Marc van Dijk at Deltares for running the historical simulations for the river Meuse. Florian Pappenberger at ECMWF provided the ECMWF-EPS reforecast data. The Water Management Centre of the Netherlands is thanked for allowing us to use an off-line version of the operational system RWsOS Rivers to do this research. The "Team Expertise Meuse" at Rijkswaterstaat Zuid-Nederland provided valuable comments on the implementation of the ensemble dressing technique for St Pieter. The HEPEX community (http://www.hepex.org) is thanked for providing inspiration, and for bringing additional fun to doing research into predictive hydrological uncertainty. The digital elevation model used as a background map in Fig. 3 was made available by the European Environment Agency on a Creative Commons Attribution License (www.eea.europa.eu). We are also grateful to no less than four anonymous referees for the time taken to review earlier versions of the manuscript, and to Konstantine Georgakakos and Hamid Moradkhani for acting as editors.

Appendix A. Verification metrics

For ease of reference, the probabilistic verification metrics used in this study are briefly explained; this description is partly based on that in Verkade and Werner (2011); Brown and Seo, 2013 and Verkade et al., 2013. Additional details can be found in the documentation of the Ensemble Verification System (Brown et al., 2010) as well as in reference works on forecast verification by Jolliffe and Stephenson (2012) and Wilks (2011).

A.1. Reliability plots

One desired property of probabilistic forecasts is that forecasted event probabilities *f* coincide with observed relative frequencies *F*. This is visualized in *reliability plots* that plot the latter variable versus the former,

$$F_i = \frac{\sum o | f = i}{J_i} \tag{A.1}$$

where F is the observed relative frequency, i is one of eleven allowable values of the forecast $f : i \in \{0, 0.1, 0.2, \ldots, 1\}$, o is an indicator variable that is assigned a value of 1 if the event was observed and a value of 0 if the event was not observed, f is the forecasted probability of event occurrence and J_i is the total number of forecasts made within bin i (Wilks, 2011).

A.2. Sharpness

Here, sharpness is indicated by the width of the centred 80% interval of the predictive distribution,

$$w_j = S_{\tau=0.90,j} - S_{\tau=0.10,j}, \tag{A.2}$$

for all J forecasts. Again, sharpness is determined for each lead time n separately and the lead time indicators have been omitted from above equation. The combined record $w_{j=1,2,\ldots,J}$ is shown as an empirical cumulative distribution function.

A.3. Correlation coefficient

The Correlation Coefficient (COR) is a measure for the statistical relationship between two variables. Here, the Pearson product-moment correlation coefficient is a measure of the strength and direction of the linear relationship between two variables that is defined as the (sample) covariance of the variables divided by the product of their (sample) standard deviations,

$$\mathrm{COR}_{X,Y} = \frac{\mathrm{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \tag{A.3}$$

where E indicates an expected value, μ is the mean and σ is the standard deviation.

A.4. Relative mean error

The Relative Mean Error (RME, sometimes called *relative bias*) measures the average difference between a set of forecasts and corresponding observations, relative to the mean of the latter,

$$\mathrm{RME} = \frac{\sum_{j=1}^{J} \left(\overline{S}_j - q_j\right)}{\sum_{j=1}^{J} q_j}, \tag{A.4}$$

where q is the observation and $\overline{S}$ is the mean of the ensemble forecast. The RME thus provides a measure of relative, first-order bias in the forecasts. RME may be positive, zero, or negative. Insofar as the mean of the ensemble forecast should match the observed value, a positive RME denotes overforecasting and a negative RME denotes underforecasting. Zero RME denotes the absence of relative bias in the mean of the ensemble forecast.

A.5. Brier skill score

The (half) Brier score (BS, Brier, 1950) measures the mean square error of J predicted probabilities that Q exceeds q,

$$\mathrm{BS} = \frac{1}{J} \sum_{j=1}^{J} \left\{ F_{Q_j}(q) - F_{S_j}(q) \right\}^2, \tag{A.5}$$

where

$$F_{S_j}(q) = \Pr[Q_j > q]$$

and

$$F_{Q_j}(q) = \begin{cases} 1 & \text{if } Q_j > q \\ 0 & \text{otherwise} \end{cases}.$$

The Brier Skill Score (BSS) is a scaled representation of forecast quality that relates the quality of a particular system BS to that of a perfect system $\mathrm{BS}_{\mathrm{perfect}}$ (which is equal to 0) and to a reference system $\mathrm{BS}_{\mathrm{ref}}$,

$$\mathrm{BSS} = \frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{ref}}}{\mathrm{BS}_{\mathrm{perfect}} - \mathrm{BS}_{\mathrm{ref}}} \tag{A.6}$$

BSS ranges from $-\infty$ to 1. The highest possible value is 1. If $\mathrm{BSS} = 0$, the BS is as good as that of the reference system. If $\mathrm{BSS} < 0$ then the system's Brier score is less than that of the reference system.

A.6. Mean continuous ranked probability skill score

The Continuous Ranked Probability Score (CRPS) measures the integral square difference between the cumulative distribution function (cdf) of the forecast $\mathrm{F}_S(q)$, and the corresponding cdf of the observed variable $\mathrm{F}_Q(q)$,

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} \{\mathrm{F}_S(q) - \mathrm{F}_Q(q)\} \mathrm{d}q. \tag{A.7}$$

The mean CRPS comprises the CRPS averaged across J pairs of forecasts and observations,

$$\overline{\mathrm{CRPS}} = \frac{1}{J} \sum_{j=1}^{J} \mathrm{CRPS}_j. \tag{A.8}$$

The Continuous Ranked Probability Skill Score (CRPSS) is a function of the ratio of the mean CRPS of the main prediction system, $\overline{\mathrm{CRPS}}$, and a reference system, $\overline{\mathrm{CRPS}}_{\mathrm{ref}}$,

$$\mathrm{CRPSS} = \frac{\overline{\mathrm{CRPS}} - \overline{\mathrm{CRPS}}_{\mathrm{ref}}}{\overline{\mathrm{CRPS}}_{\mathrm{perfect}} - \overline{\mathrm{CRPS}}_{\mathrm{ref}}} \tag{A.9}$$

A.7. Relative Operating Characteristic score

The relative operating characteristic (ROC; Green and Swets, 1966) plots the hit rate versus the false alarm rate. These are calculated using the elements of a contingency table, which is valid for a single probabilistic decision rule, and are defined as follows

$$\text{hit rate} = \frac{\#\text{hits}}{\#\text{ observed events}} = \frac{h}{o} \tag{A.10}$$

$$\text{false alarm rate} = \frac{\#\text{false alarms}}{\#\text{events not observed}} = \frac{f}{o^p rime},$$

The ROC score measures the area under the ROC curve (AUC) after adjusting for randomness, i.e.

$$\mathrm{ROCS} = 2 \times (\mathrm{AUC} - 0.5). \tag{A.11}$$

A.8. Relative economic value

The Relative Economic Value ($V[-]$) of an imperfect warning system is defined as the value relative to the benchmark cases of No Warning (V=0) and Perfect Forecasts (V=1). REV can be less than 0 if the cost of false alarms is higher than the benefits attained by the warning system:

$$V = \frac{V_{\mathrm{noFWS}} - V_{\mathrm{FWS}}}{V_{\mathrm{noFWS}} - V_{\mathrm{perfect}}}. \tag{A.12}$$

Using the consequences of contingency table elements (Table 2) and the cost-to-loss ratio r (Verkade and Werner, 2011), REV can be expressed as a function of r,

$$V = \frac{o - (h+f)r - m}{o(1-r)}. \quad \text{(A.13)}$$

References

Bennett, J.C., Wang, Q.J., Li, M., Robertson, D.E., Schepen, A., 2016. Reliable long-range ensemble streamflow forecasts: combining calibrated climate forecasts with a conceptual runoff model and a staged error model. Water Resour. Res. 52 (10), 8238–8259. URLhttp://doi.wiley.com/10.1002/2016WR019193.

Bergström, S., Singh, V.P., 1995. The HBV model. In: Singh, V. (Ed.), Computer models of watershed hydrology. Water Resources Publications, Highlands Ranch, Colorado, United States, pp. 443–476.

Bogner, K., Pappenberger, F., 2011. Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. Water Resour. Res. 47 (7), W07524. URLhttp://www.agu.org/pubs/crossref/2011/2010WR009137.shtml.

Bradley, A.A., Schwartz, S.S., 2011. Summary verification measures and their interpretation for ensemble forecasts. Mon. Weather Rev. 139 (9), 3075–3089. URLhttp://journals.ametsoc.org/doi/abs/10.1175/2010MWR3305.1.

Bremnes, J.B., 2004. Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. Mon. Weather Rev. 132, 338–347.

Brier, G., 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78, 1–3.

Bröcker, J., Smith, L.A., 2008. From ensemble forecasts to predictive distribution functions. Tellus A 60 (4), 663–678.

Broersen, P., Weerts, A., 2005. Automatic error correction of rainfall-runoff models in flood forecasting systems. Instrumentation and Measurement Technology Conference, 2005. IMTC 2005. Proceedings of the IEEE, vol. 2, pp. 963–968.

Brown, J.D., Demargne, J., Seo, D.-J., Liu, Y., 2010. The ensemble verification system (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Environ. Model. Softw. 25 (7), 854–872.

Brown, J.D., Seo, D.-J., 2013. Evaluation of a nonparametric post-processor for bias correction and uncertainty estimation of hydrologic predictions. Hydrol. Process. 27 (1), 83–105. URLhttp://doi.wiley.com/10.1002/hyp.9263.

Chen, S., Yu, P., 2007. Real-time probabilistic forecasting of flood stages. J. Hydrol. 340 (1–2), 63–77.

Coccia, G., Todini, E., 2011. Recent developments in predictive uncertainty assessment based on the model conditional processor approach. Hydrol. Earth Syst. Sci. 15 (10), 3253–3274.

Demargne, J., Wu, L., Regonda, S., Brown, J., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H.D., Fresch, M., Schaake, J., Zhu, Y.,. 2013. The science of NOAA's operational hydrologic ensemble forecast service. Bull. Am. Meteorolog. Soci., 130611111953000. http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-12-00081.1.

Fortin, V., Favre, A.-C., Said, M., 2006. Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. Q. J. R. Meteorolog. Soc. 132 (617), 1349–1369.

Glahn, H.R., Lowry, D.A., 1972. The use of model output statistics (MOS) in objective weather forecasting. J. Appl. Meteorol. 11 (8), 1203–1211.

Green, D.M., Swets, J.A., 1966. Signal Detection Theory and Psychophysics. John Wiley & Sons Inc, New York.

Hagedorn, R., 2008. Using the ECMWF reforecast dataset to calibrate EPS forecasts. ECMWF Newsletter 117, 8–13.

Hantush, M.M., Kalin, L., 2008. Stochastic residual-error analysis for estimating hydrologic model predictive uncertainty. J. Hydrol. Eng. 13 (7), 585–596.

Haylock, M., Hofstra, N., Klein Tank, A., Klok, E., Jones, P., New, M., 2008. A european daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. J. Geophys. Res 113, D20119.

Hu, Y., Schmeits, M.J., Jan van Andel, S., Verkade, J.S., Xu, M., Solomatine, D.P., Liang, Z., 2016. A stratified sampling approach for improved sampling from a calibrated ensemble forecast distribution. J. Hydrometeorol. 17 (9), 2405–2417. URLhttp://journals.ametsoc.org/doi/10.1175/JHM-D-15-0205.1.

. Jolliffe, I.T., Stephenson, D.B. (Eds.), 2012. orecast Verification: A Practitioner's Guide in Atmospheric Science, Second ed. John Wiley & Sons. URLhttp://onlinelibrary.wiley.com/book/10.1002/9781119960003.

Kang, T.-H., Kim, Y.-O., Hong, I.-P., 2010. Comparison of pre- and post-processors for ensemble streamflow prediction. At. Sci. Lett. 11 (2), 153–159. ULhttp://doi.wiley.com/10.1002/asl.276.

Kelly, K., Krzysztofowicz, R., 2000. Precipitation uncertainty processor for probabilistic river stage forecasting. Water Resour. Res. 36 (9).

Koenker, R., 2005. Quantile Regression. Cambridge University Press.

Koenker, R., 2013. Quantreg: quantile regression. R Package Version 5, 05. URLhttp://CRAN.R-project.org/package=quantreg.

Koenker, R., Bassett Jr, G., 1978. Regression quantiles. Econometrica: J. Econ. Soc., 33–50

Koenker, R., Hallock, K., 2001. Quantile regression. J. Econ. Perspect. 15 (4), 143–156.

Krzysztofowicz, R., 2002. Bayesian system for probabilistic river stage forecasting. J. Hydrol. 268 (1–4), 16–40.

Krzysztofowicz, R., Kelly, K., 2000. Hydrologic uncertainty processor for probabilistic river stage forecasting. Water Resour. Res. 36 (11).

Lahiri, P., 2003. On the impact of bootstrap in survey sampling and small-area estimation. Stati. Sci. 18 (2), 199–210.

López López, P., Verkade, J., Weerts, A., Solomatine, D., 2014. Alternative configurations of quantile regression for estimating predictive uncertainty in water level forecasts for the upper severn river: a comparison. Hydrol. Earth Syst. Sci. Discuss. 11, 3811–3855.

Montanari, A., Brath, A., 2004. A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. Water Resour. Res. 40 (1), W01106.

Montanari, A., Grossi, G., 2008. Estimating the uncertainty of hydrological forecasts: a statistical approach. Water Resour. Res. 44 (12).

Murphy, A., 1985. Decision making and the value of forecasts in a generalized model of the cost-loss ratio situation. Mon. Weather Rev. 113 (3), 362–369.

Nielsen, H.A., Madsen, H., Nielsen, T.S., 2006. Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts. Wind Energy 9 (1–2), 95–108.

Pagano, T.C., Shrestha, D.L., Wang, Q.J., Robertson, D., Hapuarachchi, P., 2013. Ensemble dressing for hydrological applications. Hydrolog. Processes.

Pagano, T.C., Wood, A.W., Ramos, M.-H., Cloke, H.L., Pappenberger, F., Clark, M.P., Cranston, M., Kavetski, D., Mathevet, T., Sorooshian, S., Verkade, J.S., 2014. Challenges of operational river forecasting. J. Hydrometeorol. URLhttp://journals.ametsoc.org/doi/abs/10.1175/JHM-D-13-0188.1.

Photiadou, C.S., Weerts, A.H., van den Hurk, B.J.J.M., 2011. Evaluation of two precipitation data sets for the rhine river using streamflow simulations. Hydrol. Earth Syst. Sci. 15 (11), 3355–3366. URLhttp://www.hydrol-earth-syst-sci.net/15/3355/2011/.

Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. J. Am. Stat. Assoc. 89 (428), 1303–1313.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URLhttp://www.R-project.org/.

Raftery, A., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133 (5), 1155–1174.

Ramos, M., Voisin, N., Verkade, J., 2013. HEPEX Science and Implementation plan: hydro-meteorological post-processing. URLhttp://hepex.irstea.fr/science-and-implementation-plan/.

Reggiani, P., Renner, M., Weerts, A., van Gelder, P., 2009. Uncertainty assessment via bayesian revision of ensemble streamflow predictions in the operational river rhine forecasting system. Water Resour. Res. 45 (2).

Reggiani, P., Weerts, A., 2008. A bayesian approach to decision-making under uncertainty: an application to real-time forecasting in the river rhine. J. Hydrol. 356 (1–2), 56–69.

Regonda, S.K., Seo, D.-J., Lawrence, B., Brown, J.D., Demargne, J., 2013. Short-term ensemble streamflow forecasting using operationally-produced single-valued streamflow forecasts – a hydrologic model output statistics (HMOS) approach. J. Hydrol. 497, 80–96. URLhttp://linkinghub.elsevier.com/retrieve/pii/S0022169413003958.

Roscoe, K.L., Weerts, A.H., Schroevers, M., 2012. Estimation of the uncertainty in water level forecasts at ungauged river locations using quantile regression. Int. J. River Basin Manage. 10 (4), 383–394.

Roulston, M.S., Smith, L.A., 2003. Combining dynamical and statistical ensembles. Tellus A 55 (1), 16–30. URLhttp://onlinelibrary.wiley.com/doi/10.1034/j.1600-0870.2003.201378.x/abstract.

Schefzik, R., 2016. Combining parametric low-dimensional ensemble postprocessing with reordering methods: Parametric low-dimensional ensemble postprocessing and reordering. Q. J. R. Meteorolog. Soc. URL http://doi.wiley.com/10.1002/qj.2839.

Schefzik, R., 2016. A similarity-based implementation of the Schaake shuffle. Mon. Weather Rev. 144 (5), 1909–1921. URLhttp://journals.ametsoc.org/doi/10.1175/MWR-D-15-0227.1.

Seo, D., Herr, H., Schaake, J., 2006. A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. Hydrol. Earth Syst. Sci. Discuss. 3 (4), 1987–2035.

Solomatine, D.P., Shrestha, D.L., 2009. A novel method to estimate model uncertainty using machine learning techniques. Water Resour. Res. 45 (12).

Todini, E., 2008. A model conditional processor to assess predictive uncertainty in flood forecasting. Int. J. River Basin Manage. 6 (2), 123–137.

Unger, D., van den Dool, H., O'Lenic, E., Collins, D., 2009. Ensemble regression. Mon. Weather Rev. 137 (7), 2365–2379.

Verkade, J., Brown, J., Reggiani, P., Weerts, A., 2013. Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. J. Hydrol. 501, 73–91. URLhttp://linkinghub.elsevier.com/retrieve/pii/S0022169413005660.

Verkade, J.S., Werner, M.G.F., 2011. Estimating the benefits of single value and probability forecasting for flood warning. Hydrol. Earth Syst. Sci. 15 (12), 3751–3765. URLhttp://www.hydrol-earth-syst-sci.net/15/3751/2011/.

Weerts, A., Winsemius, H., Verkade, J., 2011. Estimation of predictive hydrological uncertainty using quantile regression: examples from the national flood forecasting system (england and wales). Hydrol. Earth Syst. Sci. 15 (1), 255–265. URLhttp://www.hydrol-earth-syst-sci.net/15/255/2011/.

Werner, M., Schellekens, J., Gijsbers, P., van Dijk, M., van den Akker, O., Heynert, K., 2013. The Delft-FEWS flow forecasting system. Environ. Model. Softw. 40, 65–77. URLhttp://linkinghub.elsevier.com/retrieve/pii/S1364815212002083.

Wilks, D., 2011. Statistical Methods in the Atmospheric Sciences. Academic Press, Oxford, United Kingdom.

Zhao, L., Duan, Q., Schaake, J., Ye, A., Xia, J., 2011. A hydrologic post-processor for ensemble streamflow predictions. Adv. Geosci. 29, 51–59. URLhttp://www.adv-geosci.net/29/51/2011/.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., Mylne, K., 2002. The economic value of ensemble-based weather forecasts. Bull. Am. Meteorol. Soc. 83 (1), 73–83.