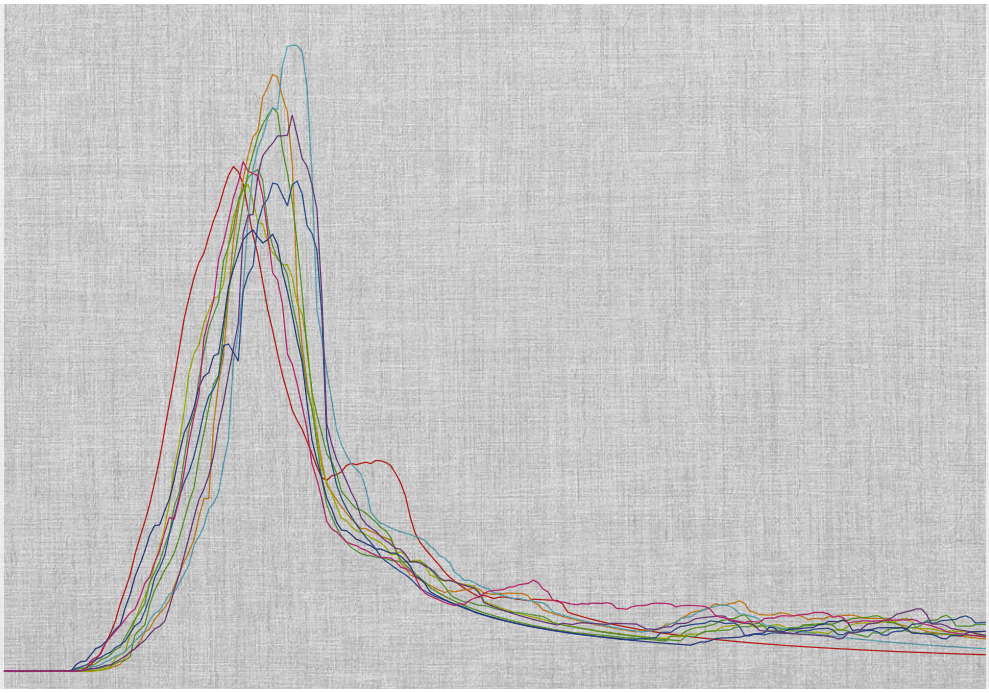


ESTIMATING REAL-TIME PREDICTIVE HYDROLOGICAL UNCERTAINTY

JAN VERKADE

ESTIMATING REAL-TIME PREDICTIVE HYDROLOGICAL UNCERTAINTY

JAN SIMON VERKADE

Estimating real-time predictive hydrological uncertainty

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. Ch.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op

woensdag 1 april 2015 om 10:00 uur

door Jan Simon VERKADE

Bachelor in Commerciële Economie
(Haagse Hogeschool, Den Haag)

Master of Arts in International Relations
(Dublin City University, Ierland)

civiel ingenieur
(Technische Universiteit Delft)

geboren te Maassluis
op 23 juli 1975

Dit proefschrift is goedgekeurd door de promotoren:
Prof. drs ir. J.K. Vrijling
Prof. dr. ir. P.H.A.J.M. van Gelder
Univ.–Prof. P. Reggiani, Ph.D.

Samenstelling van de promotiecommissie:
Rector Magnificus, voorzitter
Prof. drs. ir. J.K. Vrijling, Technische Universiteit Delft, promotor
Prof. dr. ir. P.H.A.J.M. van Gelder, Technische Universiteit Delft, promotor
Univ.–Prof. P. Reggiani, Ph.D., University of Siegen, promotor

Onafhankelijke leden:
Prof. dr. ir. H. Bijl, Technische Universiteit Delft
Prof. dr. M.–A. Boucher, Université du Québec à Chicoutimi
Prof. dr. H.L. Cloke, University of Reading
Dr. K.J. Franz, Iowa State University

Dr. ir. A.H. Weerts en dr. ir. H. van der Klis hebben als begeleider in belangrijke mate aan de totstandkoming van het proefschrift bijgedragen.

Keywords: hydrology, forecasting, predictive uncertainty

Published by Jan Verkade. Typeset using LaTeX and André Miede's `classicthesis` template. Printed by Gildeprint Drukkerijen, Enschede, the Netherlands. Cover design by Ilse van den Broek (WWW.ILSEVDBROEK.NL).

ISBN 978-94-6186-446-8
DOI 10.4233/uuid:a7e8ac36-4bdb-4231-a11e-d46778b2ae4a

SUMMARY

Flood early warning systems provide a potentially highly effective flood risk reduction measure. The effectiveness of early warning, however, is affected by forecasting uncertainty: the impossibility of knowing, in advance, the *exact* future state of hydrological systems. Early warning systems benefit from estimation of predictive uncertainties, i.e. by providing probabilistic forecasts. The present dissertation describes research in estimating the value of probabilistic forecasts as well as in skill improvement of estimates of predictive uncertainty.

A framework for estimating the value of flood forecasts, expressed in flood risk, is proposed in Chapter 2. The framework includes the benefits of damage reduction through early warning as well as the costs associated with forecasting uncertainty. The latter manifests itself through instances of missed floods and false alarms. Application of the framework to a case study to the White Cart basin — a small river in Scotland — shows that probabilistic forecasts have higher value than deterministic forecasts. It also allows for deciding on an optimal warning lead time, where the combined benefits of damage reduction (which increase with increasing lead time) and costs of forecasting uncertainty (that also increase with increasing lead time) are most beneficial.

Three post-processing approaches are investigated. The first approach (Chapter 3) comprises the statistical post-processing of meteorological forecasts and subsequent use thereof in hydrological forecasting. The analysis shows that while the quality of meteorological forecasts can be improved, the improvements do not proportionally propagate to the quality of the hydrological streamflow forecasts. It is believed that this is due to the inability of post-processing techniques to fully maintain the spatio–temporal correlations.

The second approach comprises an exploration of potential improvements to the application of Quantile Regression as described by Weerts et al. 2011. These include the application of an explicit requirement for non-crossing quantiles, the exploration of the benefit of deriving the statistical models in Gaussian space and the derivation of multiple statistical models on several sub-domains of the predictor. The results indicate that the non-crossing quantiles comprise an improvement and that the other two potential improvements do not actually result in observable increase in forecast skill, hence that the post-processor may be simplified for use in operation practice without losing skill.

The third approach explores the benefits – in terms of forecast skill – of a lumped post-processing approach versus separately addressing

meteorological and hydrological uncertainties. The latter approach was found to yield sharper forecasts, but at the expense of reliability. Combined, this resulted in very similar skill scores with the source-specific approach offering more scope for improvement.

The combined findings indicate that probabilistic forecasts have value and that there is scope for additional increase thereof. This is elaborated on in the Synthesis in Chapter 6. Also, recommendations for additional research are given. This includes research pertaining to value and skill of hydrological forecasts as well as to the *use* of forecasts in forecast, decision and response systems.

SAMENVATTING

Hoogwaterwaarschuwingssystemen vormen een potentieel bijzonder effectieve manier om hoogwaterrisico te reduceren. De effectiviteit van waarschuwingen wordt echter mede bepaald door de mate van *onzekerheid* in de hydrologische verwachtingen: het is onmogelijk om vooraf de *exacte* toekomstige staat van hydrologische systemen te kennen. Hoogwaterwaarschuwingssystemen zijn gebaat bij onzekerheidsschattingen, oftewel bij het maken van kansverwachtingen. Het voorliggende proefschrift beschrijft onderzoek naar het schatten van de waarde alsmede naar het verhogen van de kwaliteit van kansverwachtingen.

Hoofdstuk 2 beschrijft een raamwerk voor het schatten van de waarde van hoogwaterwaarschuwingen; die waarde is uitgedrukt in risicoreductie. Het raamwerk beschouwt schadereductie als gevolg van tijdige waarschuwingen alsmede de kosten die samenhangen met de onzekerheid in verwachtingen. Die onzekerheid toont zichzelf doordat sommige hoogwaters niet worden voorafgegaan door waarschuwingen of doordat sommige waarschuwingen niet worden gevolgd door hoogwaters. Het raamwerk wordt getoetst middels een toepassing op een casus: White Cart, een kleine rivier in Schotland. De casus laat zien dat kansverwachtingen een hogere waarde hebben dan deterministische verwachtingen, die geen expliciete schatting van onzekerheid bevatten. Kansverwachtingen maken het ook mogelijk dat er een optimale zichttijd bepaald wordt. Hier is de combinatie van baten (door schadereductie; deze nemen toe met toenemende zichttijd) en kosten (geassocieerd met onzekerheid; deze nemen ook toe met toenemende zichttijd) het gunstigst.

Het proefschrift beschrijft verder drie aanpakken voor het verhogen van de kwaliteit van verwachtingen middels statistisch nabewerken van meteorologische en hydrologische verwachtingen. De eerste aanpak (Hoofdstuk 3) behelst het statistisch nabewerken van neerslag- en temperatuurverwachtingen, en het successievelijke gebruik daarvan voor het maken van hydrologische verwachtingen. De analyse laat zien dat het inderdaad mogelijk is om de kwaliteit van de meteorologische verwachtingen te vergroten. Echter, deze toename in kwaliteit vertaalt zich niet in een evenredige toename in kwaliteit van de hydrologische afvoerverwachtingen. Een mogelijke reden daarvoor is dat de gebruikte statistische technieken geen rekening houden met de temporele en ruimtelijke correlaties in de oorspronkelijke meteorologische verwachtingen.

De tweede aanpak (Hoofdstuk 4) behelst een verkenning van de potentiële verbeteringen die gemaakt kunnen worden op de toepassing van Kwantielregressie, zoals beschreven door Weerts et al. 2011. Deze verbeteringen behelsen de toepassing van een expliciete eis dat kwantiellijnen niet mogen kruisen, het verkennen van de mogelijkheden van het toepassen van de statistische modellen in de Gaussische of Normale ruimte, en het afleiden van meerdere statistische modellen op sub-domeinen van de onafhankelijke variabele. De resultaten laten zien dat de niet-kruisende kwantiellijnen inderdaad een verbetering tot gevolg hebben, en dat de andere twee technieken niet leiden tot een daadwerkelijke verbetering van de kwaliteit van de gemaakte kansverwachtingen. Dit betekent dat de op dit moment in gebruik zijnde statistische techniek vereenvoudigd kan worden zonder dat dat leidt tot een vermindering in kwaliteit van de gemaakte verwachtingen.

De derde aanpak, beschreven in Hoofdstuk 5, verkent twee methoden om de kwaliteit van kansverwachtingen van toekomstige rivierafvoeren te vergroten. Bij de eerste methode worden meteorologische onzekerheden en hydrologische onzekerheden in gezamenlijkheid beschouwd middels het statistisch nabewerken van deterministische verwachtingen. Bij de tweede methode worden de twee bronnen van onzekerheid onafhankelijk van elkaar beschouwd door schattingen van hydrologische onzekerheid te combineren met meteorologische ensembleverwachtingen. De tweede methode resulteerde in scherpere verwachtingen (smallere betrouwbaarheidsintervallen) waarbij de kansen echter minder goed overeenkwamen met waargenomen relatieve frequenties dan bij de eerste methode. De kwaliteit van de gemaakte verwachtingen, uitgedrukt in een aantal veelgebruikte indicatoren, is bij beide methodes min of meer gelijk. De tweede methode echter biedt meer ruimte voor toekomstige verbeteringen.

Alle resultaten samen suggereren dat kansverwachtingen 'waarde' hebben en dat er potentie is om die waarde verder te vergroten. Hier wordt in de afsluitende Synthese (Hoofdstuk 6) op ingegaan. Hier worden ook aanbevelingen voor aanvullend onderzoek gegeven. Dit behelst onderzoek naar de waarde en de kwaliteit van verwachtingen, alsmede onderzoek naar het *gebruik* van kansverwachtingen in operationeel waterbeheer.

CONTENTS

"All those whose duty it is to issue regular daily forecasts know that there are times when they feel very confident and other times when they are doubtful as to the coming weather. It seems to me that the condition of confidence or otherwise forms a very important part of the prediction, and ought to find expression."

W. Ernest Cooke
Government Astronomer
Perth, Western Australia (1906)

1 INTRODUCTION

1.1 SETTING THE SCENE

Human settlements have always been sited on floodplains, for nearby rivers offer social, economic and environmental benefits. These benefits arose mainly from the transport opportunities afforded by a nearby river and from the fertile land that is often found in floodplains. At present, rivers also contribute to the 'tourism offer' of towns and cities (Fleming et al., 2001).

Siting settlements near rivers, however, exposes their communities to a periodic risk of flooding. Floods have the potential to adversely impact a community by causing casualties, by inflicting damage to physical property, by temporarily interrupting social and economic activities and by forcing a community to take emergency measures. Indeed, floods are natural disasters with a very high impact in terms of number of people affected, number of casualties and amount of damage (IFRC, 2013). A study of flood damage in the United States shows that this impact has increased over time as a result of both climate factors and societal factors: increased damage is associated with increased precipitation and with increasing population and wealth (Pielke Jr and Downton, 2000). Recorded history shows numerous floods including floods with a high impact in terms of economic and human losses (e.g. O'Connor and Costa, 2004). Recent high impact fluvial flood disasters include the 2007 summer floods in England and Wales, the 2009 Queensland floods in Australia, the 2010 and 2011 monsoon floods in Pakistan and Thailand, respectively, the 2013 Elbe floods in Germany and the 2014 Danube floods in Central Europe.

Because of the potential for major adverse consequences for society, humans have always tried to manage floods and their impacts. Traditionally, flood management plans were focused on the reduction of *flood hazards*: the magnitude, extent and probability of flooding. More recently, water managers shifted focus towards the management of *flood risk*, which is defined as the combination of the probability of flooding and its consequences, the latter consisting of a combination of exposure and vulnerability (Kron, 2002; Gouldby and Samuels, 2005; De Moel, 2012). Flood risk management thus addresses three factors: *hazard*, *exposure* and *vulnerability*. Exposure is a measure of the extent of communities and their assets that are potentially affected by a flood hazard and vulnerability refers to the potential of floods to afflict harm to those communities (Gouldby and Samuels, 2005).

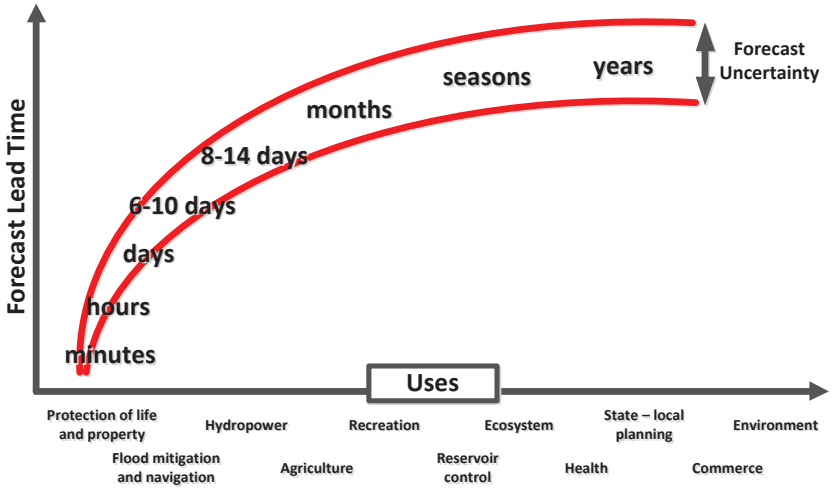


Figure 1: Various uses of hydrological forecasts, with typical time scales and a relative measure of forecasting uncertainty. Adapted from Seo and Demargne (2008).

Flood hazards are managed and mitigated by *structural* measures such as installing flood control reservoirs, raising levees and deepening and widening of the river bed. Exposure and vulnerability, however, are managed by *non-structural* measures. These include flood awareness raising, flood resistant construction, land use change and flood emergency management. The latter involves taking ad-hoc damage mitigating measures if and when a flood occurs to reduce exposure, vulnerability or both. For example, an at risk community can be temporarily evacuated or temporary barriers can be installed. Generally speaking, structural measures are considerably more expensive than non-structural measures (e.g. Jha et al., 2012).

Non-structural ad-hoc measures require advance notice of an upcoming flood. These advance notices are typically provided by hydrological forecasting systems. Such systems comprise the hardware, software and human forecasters required to produce an estimate of future streamflow and water level conditions on a river. The lead time provided depends on the hydrological properties of the basin considered but is typically in the range of hours to days, sometimes stretching to one or two weeks (Figure 1). Meteorological and hydrological observations usually originate from ground based measurement stations that are connected to the forecasting system by telemetry; some have their origin in remote sensing equipment. Meteorological forecasts are often provided by meteorological agencies in the form of Numerical Weather Predictions (NWP; see Inness and Dorling, 2013 for additional details on operational weather forecasting). Based on the available observations and forecasts, hydrological models produce estimates of future

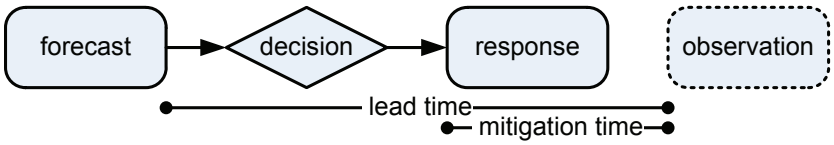


Figure 2: Forecast – decision – response system.

states of the hydrological system. These models consist of computer-based simplifications of streamflow generation and streamflow propagation processes. The models are 'tuned' by model parameters. The model output is assessed by a hydrological forecaster who, based on her expertise and experience, is likely to adjust the outputs to thus produce a hydrological forecast.

The combination of hydrological forecasting systems and non-structural ad-hoc measures is often referred to as *early warning systems* or forecast–decision–response systems (Figure 2). These systems are characterised by various timelines (Carsell et al., 2004). *Lead time* is the length of time between the production of a forecast and the onset of a flood event. The time that remains after decision-making and warning, i.e. between the onset of flood warning response and the arrival of the flood, is referred to as *mitigation time*. This is the time that can be used for actual damage mitigation. Assuming that the time required for decision making and notification remains unchanged, the potential for damage mitigation increases with increasing mitigation time. It is therefore beneficial to maximize lead time afforded by a forecasting system.

Early warning systems thus comprise a relatively inexpensive flood risk management measure. There is a catch, however: the future is *uncertain*. While this uncertainty is reduced by forecasting, it cannot be eliminated. Residual uncertainty remains and the forecast value is unlikely to be exactly equal to the observation that follows. This uncertainty originates in all of the elements of a forecasting system: in observations and measurements, in the model and its parameters, in initial conditions and in model inputs (also often referred to as model drivers or model forcings). In many cases, these model inputs – meteorological observations and forecasts – comprise a major source of uncertainty. Generally speaking, nonetheless, any of these sources contribute to overall uncertainty only insofar as the information they provide contributes to the forecasted variable at the location of interest at the required lead time. For example, uncertainty originating in a weather forecast is only relevant if the streamflow at the lead time of interest is affected by future precipitation and temperature. Conversely, if the lead time of interest is shorter than the time of concentration of a basin, future weather will have less of an impact on the uncertainty in

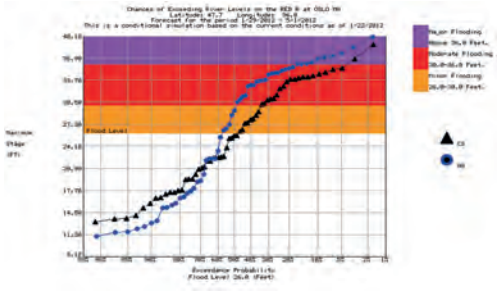
that specific forecast. The significance of individual sources of uncertainty thus varies with basin characteristics and required lead time.

Despite the presence of uncertainty, forecasting systems have traditionally produced deterministic forecasts which comprise a single estimate of future conditions only. While these forecasts reduce uncertainty, residual uncertainty remains. This uncertainty can be managed in a number of ways. First, decision makers and users can simply accept that the uncertainty is there and that this may result in a 'wrong' decision every so often. This may be acceptable if the forecasts are imperfect (they always are) yet skilful: the quality of the forecast is higher than that of an alternative forecast.

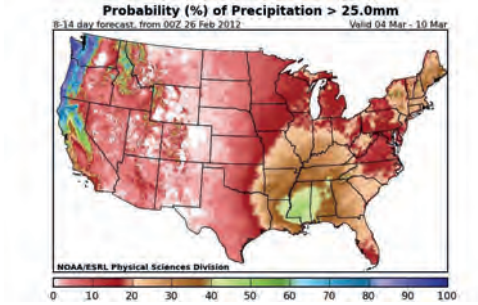
Secondly, attempts can be made to eliminate uncertainty as much as possible. Here, the distinction between epistemic and aleatory uncertainties is useful. Epistemic uncertainties can be reduced by improved understanding of physical meteorological and hydrological processes and increased ability to mathematically describe these. Increasing the number and quality of observations will contribute to this, too. There will be a random element to most of the processes that are modeled; these aleatory uncertainties are deemed irreducible, although the distinction between epistemic and aleatory uncertainties is, to some extent, arbitrary.

The third approach to managing uncertainty comprises estimation thereof. Short-term uncertainties that arise from reasonably well understood processes can be addressed by a probabilistic approach, thus yielding a probabilistic forecast: a probability distribution of the future value of a hydrological variable such as water level or streamflow rate. These estimates of predictive uncertainty can take many forms, including discretised and continuous probability density functions and cumulative probability distributions as well as probabilities of event occurrence. These events can be defined as the exceedence or non-exceedence of thresholds, or both, indicating the probability that a future value will be in a certain domain between a lower and an upper bound. Some examples of visualizations of hydrological probability forecasts are shown in Figure 3.

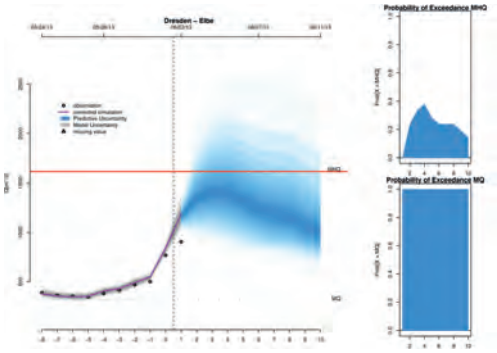
There is a strong theoretical rationale for probabilistic forecasting, comprising multiple arguments why estimates of predictive uncertainty are important for forecast sensitive decision making. Some of these arguments were put forward by Krzysztofowicz (2001), Montanari and Brath (2004) and Todini (2004). First, as there are always uncertainties about the future, any forecast that makes these explicit are more honest than forecasts that do not. If anything, this will cause the forecasts to better fit the beliefs of the expert forecaster who knows about the presence of uncertainty, and thus make for a better forecast (Murphy, 1993). This argument was put forward by the Australian "government astronomer" W. Ernest Cooke as early as 1906 (Cooke, 1906). Secondly,



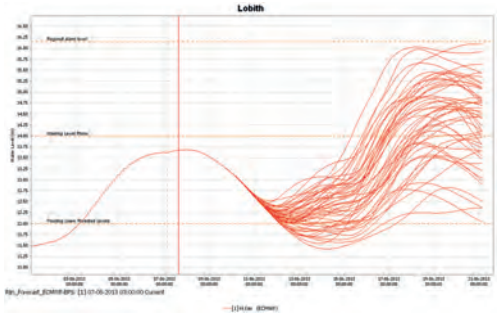
(a) Cumulative probability distribution of maximum river stage at Red River at Oslo in February through April 2012. Source: US National Weather Service, North Central River Forecast Centre



(b) Probability of precipitation over the contiguous United States exceeding 25mm. Source: NOAA Earth System Research Laboratory



(c) Discretized probability density forecast of streamflow at Elbe at Dresden (left) and exceedence probabilities of two thresholds MQ and MHQ. Source: European Flood Awareness System



(d) Ensemble forecast of streamflow, Rhine at Lobith. Source: Rijkswaterstaat's river forecast system RWsOS Rivers.

Figure 3: Selected examples of probabilistic hydrologic forecasts.

deterministic forecasts may initially ignore the possibility of an extreme event that may then only be predicted close to its occurrence. This prevents early preparation for mitigation of casualties and damage. Probabilistic forecasts could have acknowledged the possibility of such an event much earlier, even though it may have initially been assigned a small probability of occurrence. Thirdly, expressing the uncertainty in terms of probability allows for risk-based decision making by weighting event consequences with their probability of occurrence. Finally, probabilistic forecasts allow for clear separation of responsibilities between forecasters and decision makers. To wit, in the absence of uncertainty estimates, a deterministic, single valued forecast would often implicitly or automatically result in a decision. A probabilistic

forecast, on the other hand, enables a decision maker to set her own level of certainty required to initiate a response.

Developments in science and technology increasingly allow for probabilistic forecasts to be produced. Uncertainty estimation is typically done using a combination of numerical meteorological or hydrological models and statistical techniques. These techniques comprise Monte Carlo analysis and statistical post-processing. Monte Carlo analysis is at the heart of ensemble techniques, where multiple plausible, equiprobable initial conditions or model parameters are used as inputs to multiple model runs. These initial conditions are sampled from a probability distribution; this is necessary as it is impossible to know the true state of the atmosphere. Statistical post-processing aims to characterise the relation between forecasts and observations and, assuming that this relation is valid in the future also, applies this relation to future forecasts.

These techniques aim to produce probabilistic forecasts that are *reliable* and as *sharp* as possible. Reliability pertains to the probabilistic nature of the forecasts: predicted probabilities have to be matched by observed relative frequencies. Sharpness pertains to the width — or rather, narrowness — of the predictive intervals. Ideally, these are as narrow as possible, with the ultimate but unattainable goal of having zero width.

Even though evidence suggests that humans are well able to intuitively manage uncertainty and risks, effectively using probabilistic forecasts in operational practice is not trivial. Probabilistic reasoning may be problematic for experts as well as non-experts (Murphy et al., 1980; Slovic, 1987; Gigerenzer et al., 2005; Spiegelhalter et al., 2011) and it is more difficult to assess the quality of probabilistic forecasts and to communicate and understand this forecast quality (Werner et al., 2015). The 'extra' dimension (probability or likelihood) to an already highly dimensional forecast (space, time, event) complicates visualisation. It also poses additional requirements to the language used in communicating (about) forecasts. Decision criteria have to specifically take into account probabilities of event occurrence rather than certain event occurrence. This will need to be laid down in process descriptions and procedures and all stakeholders will need to be trained on the use of probabilistic forecasts. Addressing these issues requires expertise on forecasting, cognitive processing, decision science and communication.

This additional complexity may seem cumbersome, but may prove very worthwhile. One of the characteristics of a 'good' forecast is that it has *value*: an incremental economic and/or other benefit realized by a decision maker through the use of the forecast (Murphy, 1993). Studies into the value to society of forecasts confirm that probabilistic forecasts have higher value than deterministic forecasts (see, for example, US-ACE 1994; Katz and Murphy 1997; Zhu et al. 2002; Carsell et al. 2004; Roulin 2007; Buizza 2008; Boucher et al. 2012). The reason for this is

that the former allow for a decision maker to choose her own optimal decision threshold against which to initiate mitigation. For example, mitigation measures with relatively low costs may be initiated at low probability of event occurrence; should the event not occur then the (in hindsight!) unnecessary investment was only low. On the contrary, if a mitigation measure is very costly — compared to its benefit — then she would require a high degree of certainty of event occurrence as a false alarm would be a relatively costly affair.

1.2 DEFINITIONS AND FOCUS

The main theme of the present dissertation is the estimation of *predictive hydrological uncertainty*: a probability distribution of the future value of hydrologic variables such as water level and streamflow rate. Predictive hydrological uncertainty is synonymous with stochastic predictions, probabilistic forecasts or probability forecasts. It is also often referred to as simply predictive uncertainty. These terms are used interchangeably throughout the text of this dissertation.

Probabilistic forecasts are also closely related to statistics and hence sometimes referred to as statistical forecasts. Statistics and probability are closely related mathematical disciplines. Statistics studies the cause and frequency of events, based on which a probabilistic estimate of future frequency can be made. In layman's terms: statistics answers questions about what *did* happen and probability answers questions about what *will* happen (e.g. StackExchange, 2010).

Deterministic forecasts are single estimates of future conditions, hence sometimes also referred to as single valued forecasts or, somewhat cynically, 'best guesses'. Implicitly, they hold the assumption or promise that this is the only possible future condition — thus obscuring the presence of uncertainty about the future.

The meaning of the words *forecast* and *prediction* is more or less the same (Oxford University Press, 2014) and they are used as synonyms in the present dissertation.

The approaches, results and conclusions are limited to fluvial forecasting applications on the short to medium range. On these timescales, forecasts are affected by both initial conditions at issue time of a forecast and by future weather rather than by future climate. Specifically, the research does not address uncertainties related to system behaviour in far-away futures (often referred to as foresight studies, Van Asselt 2000) that have to take into account uncertainties in the future climate and the socio-economic system (Haasnoot, 2013).

The research is geared towards application in flood forecasting, even though some of it may apply to other flow regimes such as low flow forecasts. Some of the research may even be applicable to disciplines other than hydrology; this will be revisited in the closing chapter.

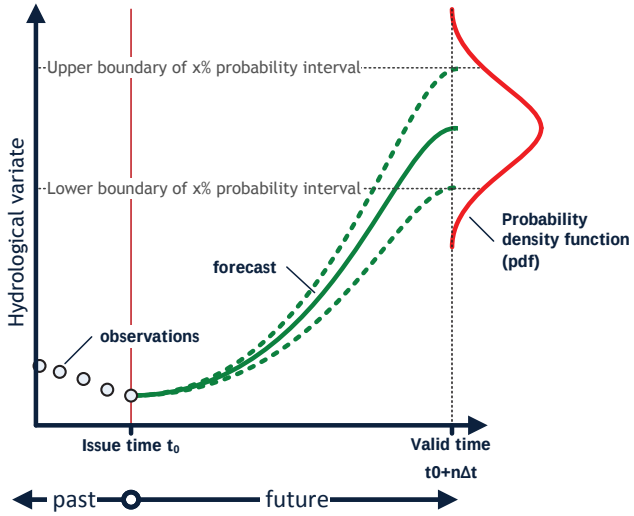


Figure 4: Schematised example of a probabilistic forecasts showing issue time and valid time. In this example, the predictive distribution is symmetric. In reality, this is seldom the case.

Real-time hydrological forecasting is the production of estimates of future states of hydrological variables, conditional on current states at issue time of the forecast. The latter is often referred to as 't-zero' (t0), with the time in the future being referred to as 'valid time' (Figure 4). This type of forecasting is sometimes referred to as *operational forecasting* (Pagano et al., 2014), referring to its application in operational forecasting centres. *Hindcasting* or *reforecasting* (Hamill et al., 2006) is the process of forecasting for a time that is now in the past — but without the benefit of hindsight. Thus, issue times are still prior to valid times, no observations dating after the issue time are used and the uncertainties are identical to those in real-time forecasting. *Backcasting* is the process of forecasting in reverse time (Hyndman, 2014), where valid times are prior to issue times. To the best of the author's knowledge, backcasting is not used in hydrological applications. *Nowcasting* comprises the detailed description of the current weather along with forecasts obtained by extrapolation for a period of 0 to 6 hours ahead (World Meteorological Organization, 2014).

A *forecast production system* comprises the hardware and software that allows for the data storage and data flow that are required to feed and run forecasting models and disseminate the results. A notable example is the Delft-FEWS system (Werner et al., 2013) which has been used for most of the research described in the present dissertation. *Forecasting techniques* are the theories underlying the forecast runs. These comprise data assimilation techniques, numerical modelling techniques and statistical techniques.

Ensemble forecasting and *post-processing* are statistical techniques that are used to estimate predictive uncertainty and/or to correct biases in probabilistic forecasts. Ensemble forecasting is a form of Monte Carlo simulation, where multiple plausible initial states, boundary conditions and/or model parameterizations are used to create an ensemble of multiple model outcomes (see Figure 3 for an example). As such, ensemble techniques constitute statistics that are applied prior to the forecast runs, contrary to statistical post-processing which is applied posterior to a forecast run. Statistical post-processing attempts to characterise the joint forecast, observation distribution with the aim to bias-correct future forecasts or to estimate predictive uncertainty. Based on an extensive record of past forecasts and observations, their relation is statistically described and subsequently applied to future forecasts — on the assumption that the past relation applies in the future, too.

There are multiple synonymous terms for statistical post-processing. Within the context of hydrological forecasting, statistical post-processing of meteorological forecasts is sometimes referred to as statistical *pre*-processing, to indicate that the post-processing takes place prior to a run of the hydrological forecast models. Within the meteorological sciences, post-processing is often referred to as *calibration*, which has a different meaning in the hydrological sciences, namely that of finding optimal values of model parameters. Sometimes the term *bias-correction* is used instead of post-processing. The latter comprises the estimation of uncertainties as well as correcting for biases in existing forecasts and is hence wider in scope.

Verification is the quantitative assessment of the relation between forecasts and their verifying observations (Stanski et al., 1989). Many aspects of this relationship can be described; these are referred to as forecast quality aspects (Murphy, 1993). When these quality metrics are expressed on a scale relative to another forecast (the reference, or baseline), they are referred to as forecast skill. Verification also comprises the assessment of forecast value: the economic benefit accomplished by using the forecast (Murphy, 1993). In the commercial sector, this is sometimes referred to as *forecast value added* (FVA; Gilleland 2013).

1.3 RESEARCH OBJECTIVE AND RESEARCH QUESTIONS

The objective of this research project is to contribute — in two distinct ways — to the use of probabilistic hydrologic forecasts in flood early warning systems: (i) by providing a valuing technique for estimating the value of probabilistic flood forecasts in terms of flood risk so that the value of flood early warning systems can be compared to the value of other risk reduction measures; and (ii) by the development of various post-processing approaches for improving the skill of probabilistic

hydrological forecasts. In order to reach this objective, the following research questions are addressed:

1. How can the value of probabilistic forecasts be expressed in terms of flood risk?

Flood risk management requires investments in risk mitigation measures. In order to efficiently allocate scarce resources, the costs and benefits of various available measures need to be estimated. This would allow, for example, to decide between raising a levee and implementing a flood forecasting, warning and response system. Estimates of the value of forecasting systems require that the (adverse) effect of forecasting uncertainty — that manifests itself through missed events and false alarms — are included in the analysis, in addition to the reduction in flood damage that can be effected by appropriate warnings (i.e., hits or true positives).

2. Can statistical post-processing further improve the skill of estimates of probabilistic forecasts?

Most probabilistic forecasts are skilful yet not perfect. Forecasts may be biased in mean, spread or both. Some of the biases may be removed through statistical post-processing, where past forecast performance is used to make a probabilistic estimate of future performance. Many different post-processing approaches are possible; here, three approaches are taken. These address the following research questions:

2a. Can the skill of ensemble streamflow forecasts be improved by post-processing ensemble NWP temperature and precipitation forecasts?

Hydrologic models are often forced by output from ensemble NWP models. The latter are often biased in mean, spread or higher moments. These biases propagate to ensemble streamflow predictions. Relatively little is known about the effects of post-processing NWP for hydrologic applications. Hence the inputs to the hydrological model will be post-processed in order to improve quality of resulting streamflow forecasts.

2b. Can estimates of predictive hydrological uncertainty be improved by changing the configuration of a post-processor?

Post-processing is a popular technique for estimating predictive hydrological uncertainty based on one or more predictors. Earlier work reported on using Quantile Regression to estimate predictive uncertainty based on single valued forecast as predictor. This was a relatively straightforward implementation, using a Gaussian transform to manage nonlinearities in the joint forecast — observation distribution.

However, alternative approaches were not reported and these will have to be explored to ensure that the best possible results are obtained.

2c. Can the skill of raw ensemble streamflow forecasts be improved by 'dressing' the ensemble members with distributions that describe the hydrologic uncertainties?

Hydrologic Ensemble Prediction Systems often route ensemble NWP products through a hydrologic model to arrive at an ensemble streamflow forecast. However, the spread of this ensemble is indicative of uncertainty in meteorological forcings only, and not of all relevant uncertainties. Recently, ensemble dressing techniques have been proposed, where members are dressed with distributions that describe hydrological uncertainties. This raises question of how well the technique performs against post-processing of deterministic forecasts — that is currently often used to estimate the 'total uncertainty' hence in many ways the technique to beat.

1.4 RESEARCH CONTEXT

The research described in this dissertation was carried out as part of the Deltares R&D programmes on real-time forecasting for flood risk management and water resources management (for details, see Deltares, 2013). While their scope is wider, these programmes also address mission critical research needs for Rijkswaterstaat, the national water management authority in the Netherlands. The research was done in part–time, in addition to the author's work as a hydrologist at Deltares, hence some of the cases were necessarily chosen for pragmatic reasons also, i.e. coincided with other studies carried out by Deltares. The research is also linked to the author's work as a forecaster in Rijkswaterstaat's River Forecasting Service. Finally, the author, the co-authors of the journal papers based on this dissertation and some members of the supervisory committee maintain strong links with the HEPEX community for researchers and practitioners in hydro-meteorology (Schaake et al., 2007; HEPEX community, 2013). The community has provided inspiration, valuable suggestions, co-authors and peer reviewers.

1.5 APPROACH AND OUTLINE

The research questions are addressed by two parallel, linked approaches (Figure 5). The one approach focuses on forecast value whereas the other focuses on forecast skill. Ultimately, an increase in forecast skill will result in an increase in forecast value — or at least in theoretical value.

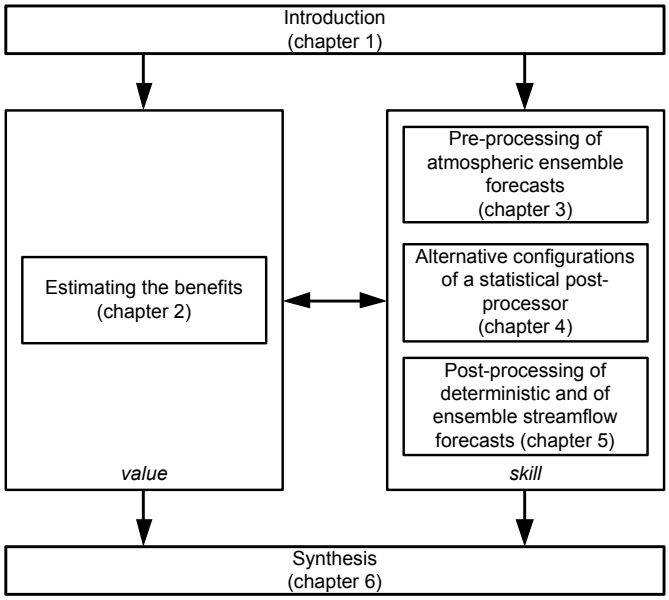


Figure 5: Outline of the research presented in this dissertation.

The main part of this thesis consists of five chapters that have been published as papers or have been submitted to a peer reviewed, scientific journal. As a result there is some overlap in the content between the chapters (papers). This mainly concerns sections that are part of the appendices of the papers; these have been moved to the appendix in the present dissertation as to remove some of this overlap.

Each chapter (paper) addresses one of the research questions or subquestions (see Figure 5). Chapter 2 reflects on the question of how to estimate the value — expressed in flood risk — of both deterministic and probabilistic flood forecasts, taking into account forecasting uncertainty and its adverse consequences. Chapters 3 through 5 address the question of how to produce skilful probabilistic forecasts. Each of these three chapters explores a different technique for doing so: post-processing of atmospheric forecasts (Chapter 3), varying the configuration of a statistical post-processor for streamflow forecasts (Chapter 4) and the 'dressing' (post-processing) of deterministic and ensemble streamflow forecasts (Chapter 5). Finally, Chapter 6 revisits the research questions and reflects on the research by discussing the contributions to the value and skill of real-time probabilistic forecasting in hydrology.

Case studies are an important element in each of the chapters 2 through 5. The collection of study basins is relatively diverse. The White Cart basin was used for the 'estimating the value' chapter: it is relatively flood prone and consequences of flooding are reasonably well known, hence it was possible to estimate potential flood damage

and reduction thereof through flood warning response. For the 'pre-processing' study, there was a need for both a relatively long record of meteorological forecasts and observations at spatial and temporal resolutions befitting the choice of study basin - this made the Rhine a suitable case study. The 'alternative configurations' study was essentially a continuation from an earlier study (Weerts et al., 2011) hence it was decided to reuse the Severn as a study basin. The 'dressing' study, similar to the 'pre-processing' study, required a basin that befitted meteorological observations and forecasts. For that reason, the Rhine basin was used again, and to make the results more robust to the choice of NWP product, the Meuse basin - which is an order of magnitude smaller than the Rhine basin - was included in the study also.

2

ESTIMATING THE BENEFITS OF SINGLE VALUE AND PROBABILITY FORECASTING FOR FLOOD WARNING

ABSTRACT

Flood risk can be reduced by means of flood forecasting, warning and response systems (FFWRS). These systems include a forecasting subsystem which is imperfect, meaning that inherent uncertainties in hydrological forecasts may result in false alarms and missed events. This forecasting uncertainty decreases the potential reduction of flood risk, but is seldom accounted for in estimates of the benefits of FFWRSs. In the present chapter, a method to estimate the benefits of (imperfect) FFWRSs in reducing flood risk is presented. The method is based on a hydro-economic model of expected annual damage (EAD) due to flooding, combined with the concept of Relative Economic Value (REV). The estimated benefits include not only the reduction of flood losses due to a warning response, but also consider the costs of the warning response itself, as well as the costs associated with forecasting uncertainty. The method allows for estimation of the benefits of FFWRSs that use either deterministic or probabilistic forecasts. Through application to a case study, it is shown that FFWRSs using a probabilistic forecast have the potential to realise higher benefits at all lead-times. However, it is also shown that provision of warning at increasing lead-time does not necessarily lead to an increasing reduction of flood risk, but rather that an optimal lead-time at which warnings are provided can be established as a function of forecast uncertainty and the cost-loss ratio of the user receiving and responding to the warning.

This chapter has been published as Verkade, J. S. and Werner, M. G. F., 2011. Estimating the benefits of single value and probability forecasting for flood warning, Hydrology and Earth System Sciences, 15(12), 3751–3765, DOI: 10.5194/HESS-15-3751-2011

2.1 INTRODUCTION

Floods are an act of God but flood damage is an act of Man (White, 1942). For long though, flood management has primarily focused on managing flood hazards, e.g. on reducing the frequency of flooding, flood extent, depth and duration and flow velocities. Recent years have seen an increased emphasis on the management of flood *risk*, where risk is defined as the combination of the probability of occurrence of a flood event, and its consequences in terms of casualties and economic damage (Merz et al., 2010). This shift from flood hazard management to flood risk management has led to an increased emphasis on non-structural measures including, for example, spatial planning, raising flood awareness, flood proofing and the use of flood forecasting, warning and response systems (FFWRSs).

Of these flood risk management measures, flood warning is regarded as being one of the most effective (UNISDR, 2004). Considerable attention has been given to the effectiveness of these systems. These studies generally focus on estimating flood losses, the potential reduction of these losses through warning response and the relationship between flood warning lead-time and loss reduction (e.g. Parker, 1991; Carsell et al., 2004; Parker et al., 2008; Molinari and Handmer, 2011).

Flood forecasts, which form an essential element in the flood forecasting, warning and response process are, unfortunately, affected by inherent uncertainties. These pertain to the forecasting model structure, parameter values and initial conditions, to meteorological forcing (especially when this forcing is forecast rather than observed), and to measurements and interpolations of these measurements as for example in deriving catchment average rainfall. This forecasting uncertainty can be explicitly accounted for if the forecasting sub-system of a FFWRS produces an estimate of predictive uncertainty as in the case of probabilistic forecasting.

Irrespective of the nature of the forecasting system, this forecasting uncertainty can lead to "wrong" decisions: floods that occur may not have been predicted in time, or floods that are predicted may not occur. The costs associated with this forecasting uncertainty can be considerable. An analysis of the role of benefits of FFWRSs should therefore also include these costs, consisting of an opportunity cost in the case of a flood that was not predicted, and the cost of unnecessary warning response in the case of a false alarm.

Flood risk can be defined as the expected value of flood related damage and costs. Floods are random events and therefore flood damage is a random event. Although the exact amount of damage in any given year cannot be predicted, the expected annual value of flood damage can be determined if the probability distribution of flood damage, or damage-frequency curve is known. This expected annual damage is a

measure of flood risk. Flood risk may be estimated using a *hydro-economic Expected Annual Damage (EAD) model* (USACE, 1994; Dingman, 2002; Loucks et al., 2005), which uses three basic relationships to establish the probability distribution of flood damage: the flood frequency curve, the rating curve and the stage-damage curve.

To evaluate the benefit of measures taken to reduce flood risk, the cost of these measures should be taken into account. In the case of flood warning systems, such an analysis should include the expected reduction of flood losses due to the provision of warning and subsequent response, as well as the costs of operating such systems and the costs associated with uncertainty. Whilst the first two of these can be readily incorporated in analysing the benefit of flood warning, the latter is less straightforward.

In meteorological applications, *Relative Economic Value* (e.g. Murphy, 1985; Zhu et al., 2002) is often used to establish the value of forecasting systems relative to two benchmark situations. These are the situations in which no warning system is present, and the situation in which a perfect warning system is present. In the latter, forecasting uncertainty is absent and hence no "wrong" decisions are ever made.

To the best of our knowledge, no flood risk analyses have been published that include the damage mitigating effects of flood warning, the costs of the warning system, *and* the costs associated with forecasting uncertainty. In the present chapter, a method is proposed that can be used to estimate flood risk in the presence of an imperfect FFWRS. The method consists of combining the hydro-economic EAD model with the theory of Relative Economic Value. This combines expected annual damage, loss reduction, cost of warning response *and* the costs associated with forecasting uncertainty into an estimate of the benefit of flood forecasting and warning in reducing flood risk.

This method allows for the comparison of the effect of flood risk management measures of different nature. For example, the flood risk reduction attained by the implementation of a flood warning system can be compared with that attained by the raising of levees, installation of flood retention areas or increasing flow conveyance. Additionally, the method allows for an intercomparison of FFWRSs. For example, the benefit of systems based on deterministic forecasting can be compared with those that are based on probabilistic forecasting. This allows explicitly estimating the benefit of probabilistic forecasting in terms of flood risk reduction, which so far has only been described in terms of their potential for improved decision making in flood event management (e.g., Krzysztofowicz, 2001; Todini, 2004).

In the next section, the proposed method is explained in detail. In Sect. 2.3, results of a case study are presented where the method is demonstrated by application to a small basin. The results are discussed

in Sect. 2.4. Finally, a summary and brief conclusions are presented in Sect. 2.5.

2.2 MATERIALS AND METHODS

2.2.1 *Flood forecasting, warning and response systems*

A properly working flood forecasting, warning and response system (FFWRS) gives property owners, floodplain residents and responsible authorities time to respond to a flood threat before flooding occurs. FFWRSs usually consist of a number of sub-systems (Fig. 6). The forecasting sub-system produces forecasts of hydrological variables such as water levels or flow rates, either as a deterministic single value forecast or as a probability distribution. Based on these forecasts, a decision is taken whether or not to initiate warning response. The warning-response sub-system then consists of warning procedures and subsequent mitigation action that can be taken to reduce flood losses.

Although in actual operational forecasting the decision to warn will be taken by the forecaster using guidance from the forecasting subsystem, in the present chapter it is assumed that decisions are based on forecasts only. Depending on the nature of the forecasting sub-system, the decision sub-system is deterministic or probabilistic. In the case of deterministic forecasts, it is assumed that forecast water levels that are higher than the flooding threshold will automatically initiate a warning response. Essentially, this decision is then taken implicitly by the forecaster. If the forecasting system provides explicit estimates of predictive uncertainty, the decision will have to be based on a probabilistic decision rule. If the probability of forecast water levels exceeding the flooding threshold is higher than a probability threshold, a warning response will be initiated. This allows users to choose an optimal threshold (in terms of probability threshold) at which mitigating action is initated (Krzysztofowicz, 2001), but it is again assumed here that forecast probabilities higher than the selected probability threshold will automatically initiate a response.

The warning-response sub-system pertains to the damage-mitigating actions that can be taken after a flood warning has been issued. During the time between a flood warning and the arrival of flood waters – the mitigation time – floodplain residents can move themselves and/or their property out of reach of the pending flood. Increasing the available mitigation time intuitively allows for increased loss reduction, and therefore this mitigation time should be maximised (but note that with increasing mitigation time, response costs may increase as well). Forecasting lead-time and mitigation time are different due to the time needed to produce and disseminate a forecast and to take a decision whether or not to initiate a warning response (Fig. 6) (Carsell et al.,

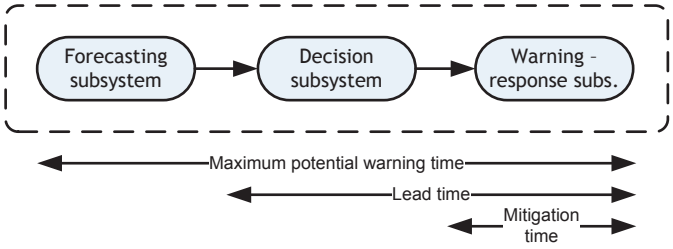


Figure 6: Flood forecasting, warning and response system (FFWRS) subsystems. Adapted from Parker and Fordham (1996) and Carsell et al. (2004).

2004). However, in the context of the present chapter the time taken in the decision sub-system is negligible and lead-time and mitigation time are used synonymously.

Maximum potential reduction of flood damage by flood warning response is rarely attained as it is unlikely that all floodplain residents will be notified in time, nor that all residents will heed the warnings and act effectively. To account for this, Parker (1991) and Green and Herschy (1994) defined the actual flood damage avoided L_a [GBP] as a product of the maximum potential flood damage avoided with a fully effective system (L_p [GBP]), the probability that a forecast is made in time (R [−]), the fraction of residents available to respond to a warning (P_a [−]), the fraction of residents who will respond to a warning (P_r [−]) and the fraction of households who respond effectively (P_e [−]). Together, these probabilities and dimensionless factors, each ranging from 0 to 1, represent the effectiveness of the response: $L_a = L_p \times R \times P_a \times P_r \times P_e$. In the UK, the Department for Environment, Food and Rural Affairs (DEFRA) indicated the values for the factors and probabilities (R, P_a, P_r, P_e) the Environment Agency seeks to achieve (DEFRA, 2004). These would result in $L_a = 0.5 \times L_p$, which is the value used in the present chapter.

2.2.2 *Expected annual flood damage*

Flooding is a random process and therefore flood damage is a random process. The expected value of annual direct, tangible flood damage can be estimated from the probability distribution of flood damage:

$$EAD = \int_0^1 D(P)\, dP \tag{1}$$

where P is the annual probability of exceedence of a certain flood level and $D(P)$ is the direct, tangible flood damage caused by that flood event (e.g. Van Dantzig and Kriens, 1960; USACE, 1994; Carsell et al.,

2004; De Bruijn, 2005; Loucks et al., 2005). To determine the probability distribution of flood damage, the hydro-economic EAD model (USACE, 1994; Davis et al., 2008; Dingman, 2002; Loucks et al., 2005) links the flood frequency distribution through flood stages to flood damage. The model can best be explained graphically (Fig. 7). The starting point of the analysis is the probability distribution of flow rates (or flood frequency curve, bottom left). A rating curve (top left) links flow rates to flood stages. Stages higher than the flooding threshold will cause damage, described by the stage – damage curve shown in the top right quadrant. By linking the probability of each flood discharge to the stage in the river to the damage occurring, the probability distribution of flood damage $D(P)$ can be established (bottom right). The expected annual flood damage can then be easily established as the area enveloped by the probability-damage curve (Eq. 1).

The effect of flood risk management measures can easily be shown in the graphical model. Measures that reduce flood frequencies push the flood frequency curve (bottom left) towards the origin. Measures aimed at a reduction of flood stage, e.g. by river bed deepening or widening, change the rating curve (top left). The reduction of flood damage, either by structural or by non-structural measures, reduce damage associated with flood stages (top right). Ultimately, measures that are effective in reducing flood risk will move the probability – damage curve towards the origin (Dingman, 2002), thus reducing the expected annual damage.

Figure 7 shows an example of the effect of a flood risk management measure. Here, a measure was implemented that reduces flood damage. Such a measure could be, for example, flood-proofing private properties. The measure does not affect either the probability of flooding or the rating curve, but does change the stage – damage relationship, with a reduced damage expected at the same stage. This results in a probability – damage relationship that lies closer to the origin, with the expected annual damage being reduced.

2.2.3 *Cost of flood warning response and cost-loss ratio*

Flood forecasting, warning and response systems come at a cost, consisting of initial costs for setting up the system, fixed costs for operation and maintenance, and variable event costs for flood warning response; the latter are incurred every time a warning is issued. The fixed costs can be included in the EAD analysis by adding these to flood damage, and shifting the stage-damage curve to the right. Strictly speaking, the term "damage" is then incorrect as it also includes the cost of measures. In this chapter, it is assumed for simplicity that the fixed costs are included in the event costs. Additionally, the event costs are considered independent of the height of the flood stage (contrary to

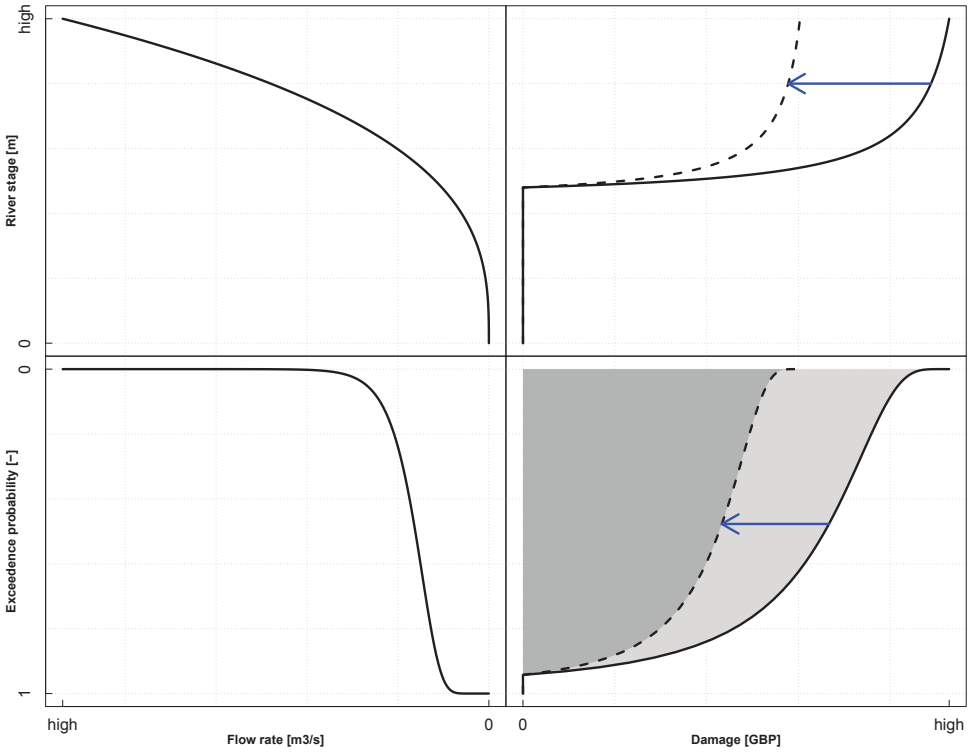


Figure 7: Schematic representation of the hydro-economic EAD-model. The bottom left quadrant shows the probability distribution of flow rates. The stage-discharge relationship is shown in the top left quadrant, and flood damage curve in the top-right quadrant. These three relationships yield the damage-probability curve (bottom right). The figure shows how a flood risk management measure affects flood risk, with the ex-ante situation as a solid, and the ex-post as a dotted line.

event *damage* which is explicitly correlated with stage). This is considered reasonable as the cost of response is incurred based on a forecast (probability) threshold being exceeded, and therefore independent of the actual height with which the threshold is exceeded. The cost-loss ratio r in Eq. (2) can be used to express the costs of warning response C as a fraction of the avoidable losses L_a. It is clear that where $r > 1$ there is no benefit in flood warning response, whilst for a very low C the ratio r approaches 0,

$$r = \frac{C}{L_a}. \tag{2}$$

Table 1: Contingency table. The consequences of the items listed are in brackets.

	Event observed	Event NOT observed	$\sum$
Warning issued	hits $h\,(C + L_u)$	false alarms $f\,(C)$	w
Warning NOT issued	misses $m\,(L_a + L_u)$	quiets $q\,(-)$	w'
$\sum$	e	e'	N

2.2.4 *Costs associated with forecasting uncertainty*

2.2.4.1 *Relative economic value*

If a decision to initiate warning and response procedures is solely based on an imperfect forecast, forecasting uncertainty may lead to false alarms and missed events. Both false alarms and missed events are instances of imperfect system performance and adversely impact the potential reduction in flood risk. Combining the hydro-economic EAD model with the theory of Relative Economic Value (e.g. Murphy, 1985; Zhu et al., 2002) offers a convenient way of incorporating the costs associated with forecasting uncertainty in estimates of expected annual damage.

Using the hydro-economic EAD model, flood risk can be estimated for the *No Warning* and for the *Perfect Warning* cases. Zhu et al. (2002) define the Relative Economic Value (REV) as a dimensionless factor to scale between these estimates. The maximum value of 1 is assigned to the *Perfect Warning* case, while a warning system that has the same skill as the climatology (here meaning the long-term average frequency of flooding) is assigned 0. Given the low climatological frequency of flood threshold exceedance, this can be considered equivalent to the case with *No Warning* being present. The REV can be calculated based on the skill of the FFWRS.

The performance of a FFWRS can be captured in a two-by-two contingency table that shows forecast/observation pairs for dichotomous events (Wilks, 2011). In this case, the table shows in how many cases a flood warning was followed by a flood event (Table 1). A contingency table is based on a record of forecasts and events and should be made for every decision rule that is used.

In the absence of a FFWRS, a user's flood losses will be determined by the climatological frequency of flooding and consist of unmitigated losses, which is the sum of the losses avoided through warning response L_a, and the losses that cannot be avoided L_u for every flood event e:

$$EAD_{nowarn} = e\ (\ L_a\ +\ L_u\ ). \tag{3}$$

If a FFWRS is based on perfect forecasts, a flood event is always preceded by a warning and flood damage can always be reduced by mitigating action. False alarms and missed events do not occur. The expected damage then consists of the sum of cost for warning response and unavoidable losses for every flood event:

$$\mathrm{EAD}_{\mathrm{perfect}} = e\,(C + L_u). \tag{4}$$

The performance of a FFWRS based on imperfect forecasts can be assessed using a contingency table. Missed events result in unmitigated flood losses, which equal the sum of avoidable and unavoidable losses $L_a + L_u$. Loss mitigation through warning response can only be achieved at a cost C. In case of false warnings, these are the only costs incurred by a user. A user's expected costs and losses consist of those associated with hits, misses and false alarms:

$$\begin{aligned}\mathrm{EAD}_{\mathrm{FFWRS}} &= h\,(C + L_u) + f\,C + m\,(L_a + L_u)\\ &= e\,L_u + (h + f)\,C + m\,L_a.\end{aligned} \tag{5}$$

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 ($V = 1$):

$$V = \frac{\mathrm{EAD}_{\mathrm{nowarn}} - \mathrm{EAD}_{\mathrm{FFWRS}}}{\mathrm{EAD}_{\mathrm{nowarn}} - \mathrm{EAD}_{\mathrm{perfect}}}. \tag{6}$$

Note that REV can be less than 0 if the cost of false alarms is higher than the benefits attained by the warning system.

Substituting Eqs. (3), (4) and (5) in (6), subsequent division by L_a and substitution of C/L_a by r (Eq. 2) yields:

$$\begin{aligned}V &= \frac{e\,L_a - (h + f)\,C - m\,L_a}{e\,L_a - e\,C}\\ &= \frac{e - (h + f)\,r - m}{e - e\,r}\\ &= \frac{e - (h + f)\,r - m}{e\,(1 - r)}.\end{aligned} \tag{7}$$

This derivation of relative economic value slightly differs from that of, for example, Zhu et al. (2002). The difference is in the expected expense in the absence of a warning system. Zhu et al. include an additional decision where, based on a minimisation of cost, a user may decide either to *never*, or to *always* take action. In the latter case, a single

warning-response action is assumed to have an impact that is unlimited in time, leading to an expected expense of $C + eL_u$. Including this $EAD_{nowarn} = \min\left[e\left(L_a + L_u\right),\ C + eL_u\right]$ in the analysis would yield relative economic value as a function of $\min\left(e,\ r\right)$ which is discontinuous at $r = e$. In the present application, the climatological frequency of flooding e approaches 0 and most if not all users' cost-loss ratio r is greater than e. For that reason, the present derivation may be simplified. It may be noted that flood risk in the "always take action" option may be estimated by using the hydro-economic EAD-model.

2.2.4.2 *Optimal warning rule*

It is assumed that a decision to issue a warning will only be taken if the expected value of the warning response is less than the expected value of *not* issuing a warning. This yields the optimal warning rule:

$$\begin{aligned} C + P \times L_u &< P \times (L_a + L_u) \\ P &> \frac{C}{L_a} \\ P &> r, \end{aligned} \tag{8}$$

with P the predicted probability of flooding. Only if a user applies the optimal warning rule to flood event decision making, will the benefits of probability forecasting be fully realised.

2.2.4.3 *Combining expected annual damage with relative economic value*

Flood risk in the No Warning and Perfect Forecasts cases can be calculated using the hydro-economic EAD-model. This equally yields EAD_{nowarn} and $EAD_{perfect}$ respectively. To calculate EAD_{FFWRS}, REV is subsequently used to scale between the flood risk of benchmark cases using Eq. (6):

$$EAD_{FFWRS} = EAD_{nowarn} - V\left(EAD_{nowarn} - EAD_{perfect}\right). \tag{9}$$

In words: the flood risk in case of a warning system being present equals the flood risk in the absence of such a system minus the avoidable risk, which is scaled by the warning system performance. A perfect system (where $V = 1$) brings the full benefits of a warning system ($EAD_{FFWRS} = EAD_{perfect}$). A system that performs as well as acting on climatological information ($V = 0$) does not bring any additional benefits, and is equivalent to no warning system being present: $EAD_{FFWRS} = EAD_{nowarn}$. A system that brings benefits compared to the absence of a warning system ($0 < V < 1$) will result in an expected annual damage between that of the benchmark cases: $EAD_{nowarn} >$

$EAD_{FFWRS} > EAD_{perfect}$. If the warning system performance is worse than that in the No Warning case ($V < 0$), flood risk will increase to levels higher than that in the No Warning case: $EAD_{FFWRS} > EAD_{nowarn}$. In that case, there is no economic rationale for flood warning.

As the potential for loss mitigation increases with increasing lead-time provided by the warning system, flood risk in the presence of a FFWRS is different for different lead-times: $EAD_{FFWRS} = f(n)$ (where n is lead-time). Additionally, Eq. (7) shows that relative economic value is expressed as a function of the users' cost-loss ratios: $V = f(r)$. Explicitly including these dependencies in Eq. (9) gives:

$$EAD_{FFWRS}(n, r) = EAD_{nowarn} - V(r)\left(EAD_{nowarn} - EAD_{perfect}(n)\right). \tag{10}$$

The assumption that was made here is that flood forecasting performance, as expressed by V, does not depend on the height of the flood wave. This is considered a reasonable assumption because the warning system performance is based on the exceedence of a flooding threshold only, and not on the prediction of the height of the flood wave.

2.2.5 *Case study: White Cart Water*

The combination of hydro-economic EAD model with relative economic value is used to estimate flood risk in a small basin in Scotland. The White Cart Water is a river located in the greater Glasgow area and a tributary of the river Clyde. This case study focuses on Overlee gauging station, which is where the White Cart Water enters the city of Glasgow, and the nearby flood warning locations at which flood damage to residential properties has been known to occur. The White Cart Water at Overlee has an upstream area of $106\,km^2$, with an average flow in the order of $3.5\,m^3\,s^{-1}$. The upper parts of the catchment are mainly rural catchment, while the lower catchment is predominantly urban. The White Cart is a very fast responding catchment, with a time of concentration of approximately 3 h. Flooding frequently occurs in the reaches downstream of Overlee, where the river flows through dense residential areas of southern Glasgow. The data record used in this study contains a dozen or so events, even though this number is obscured somewhat in relation to the number of forecasts in the same period (e.g., Table 3).

To mitigate the adverse consequences of flooding, a flood warning scheme is in place. The forecasting and warning system (Cranston et al., 2007; Werner et al., 2009) is operated by the Scottish Environmental Protection Agency (SEPA). It is a statutory requirement to SEPA to issue flood warnings no less than three hours in advance (Werner

and Cranston, 2009). The operational forecasting system includes one source of forecast precipitation only (radar now/forecasts) which has a maximum lead time of six hours. While this does not allow the at risk community to take extensive mitigating action, some actions can (and indeed are) taken. Empirical evidence suggests that the initial four hour warning period is associated with the greatest savings (Parker, 1991; NHRC, 2002; Carsell et al., 2004).

Flood risk is estimated for four cases. The two benchmark cases – No Warning and Perfect Forecasts – are investigated first. Subsequently, two imperfect FFWRSs are investigated: one in which deterministic forecasts are used and one in which probabilistic forecasts are used.

Re-forecasting analyses were carried out using an off-line version of an existing forecast production system: FEWS Scotland, which is based on the Delft-FEWS shell (Werner et al., 2013). Deterministic hydrological forecasts for White Cart at Overlee are produced using a sequence of a PDM rainfall runoff model (Moore, 1985), a kinematic wave routing model and an ARMA error correction model (Moore et al., 1990).

Predictive hydrological uncertainty was estimated using Quantile Regression (QR) (Koenker and Bassett Jr, 1978; Koenker and Hallock, 2001; Koenker, 2005; Weerts et al., 2011). QR is a post-processing method that can be used to characterise the relationship between water level forecasts and water level observations in terms of quantiles, or probabilities of exceedence or non-exceedence. See Appendix A for details. The use of a post-processor in near real-time forecasting systems is attractive as the computation time required is limited; in this case, the post-processor takes less than ten seconds to estimate the predictive distribution.

For the White Cart case study, QR was calibrated using a five year period (1 April 1991 through 31 March 1996), and subsequently validated on a period covering nearly eleven years (1 April 1996 through 20 February 2007). For both calibration and validation periods, records of deterministic water level re-forecasts were constructed using FEWS Scotland. The hydrological model was forced using observed precipitation. While using so-called perfect forcing significantly reduces uncertainty compared to a situation in which precipitation forecasts are used (Werner and Cranston, 2009), this equally affects both probability forecasts and deterministic forecasts. It does therefore not affect the demonstration of the method presented in this chapter.

Deterministic water level forecasts from the calibration period were paired with observations and from these two time series, the quantile regression relationship $h_\tau = f(s)$ was determined for all quantiles $\tau \in (.01, .02, ..., .99)$. For the validation period, a probabilistic reforecast was established through application of the quantile regression relationship to each deterministic forecast to derive water levels corresponding to the 99 quantiles $\tau \in (.01, .02, ..., .99)$. From this discre-

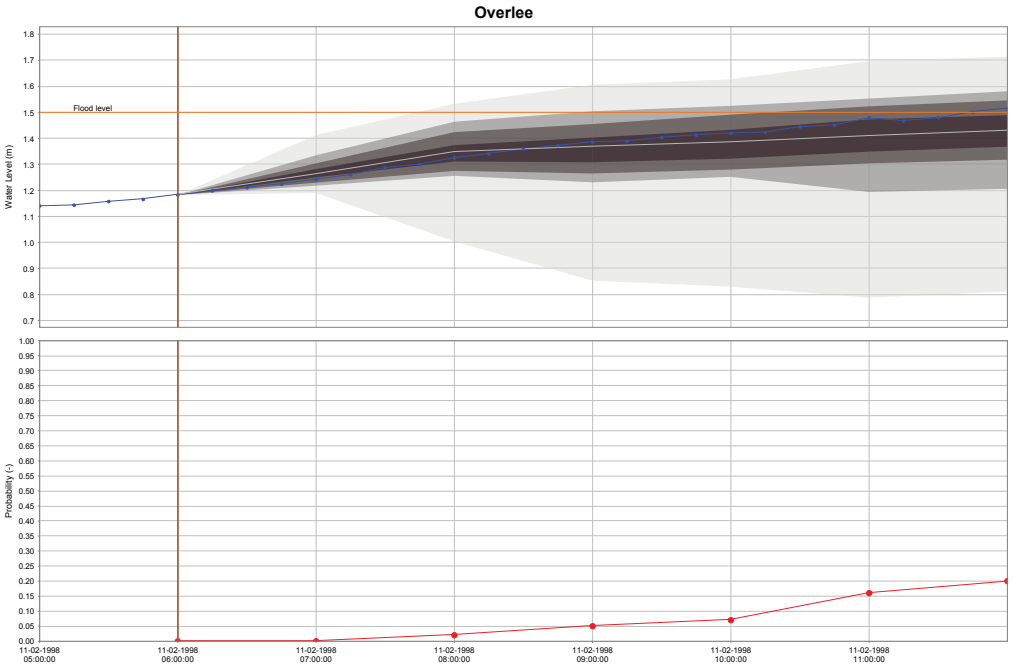


Figure 8: Sample probability forecast as produced by the research version of the forecast production system FEWS Scotland. Two graphs are shown: a discretised predictive probability distribution of water levels at quantiles $\tau \in (.01, .05, .10, .25, .50, .75, .90, .95, .99)$ (top) and the probability of exceedence of the flooding threshold (bottom). In both graphs, the vertical red line indicates the forecast issue time (t_0). In this case, it was forecast that there was a 20 % probability of threshold exceedence at $t_0 + 6$ h. The posterior water level observation (dotted blue line) showed that the threshold was exceeded at this time.

tised predictive probability distribution, the probability of exceedence of the flooding threshold (local datum + 1.5 m) was determined (Fig. 8). This threshold coincides with the water level at which flood damage starts to occur.

2.3 CASE STUDY RESULTS

2.3.1 *Case 1: no warning*

For Overlee gauging station, an 18-year record of 15-minute water level observations and a rating curve were available. Observed water levels were rated and from this record the flood duration curve was established. A stage-damage relation was not available and was established. First, the number of properties affected as a function of flood stage at

Overlee was estimated. For simplicity it was assumed that inundation depth is linearly correlated with river stage at Overlee, i.e. that an increase in river stage at Overlee leads to a similar increase in river stage at these properties. The damage to individual properties as a function of inundation depth was determined from Penning-Rowsell et al. (2005). Combining the number of properties affected as a result of a level at the Overlee gauging station and the flood damage per individual property yields the flood damage as a function of stage at Overlee. Using the hydro-economic EAD-model, the depth-damage probability distribution was established (black line in Fig 9). From this distribution, the expected annual flood damage can be calculated. In this case, this expected damage (EAD_{nowarn}) amounts to 394 695 GBP a^{-1}.

2.3.2 *Case 2: perfect forecasts*

One of the primary aims of a FFWRS is to reduce flood losses. Flood damage for individual properties can be considered as the sum of damage to building fabric and damage to household inventory. It is assumed that in the White Cart basin, given the relatively short time available for mitigating action, damage to building fabric cannot be avoided and constitutes an unavoidable loss. Carsell et al. (2004) investigated which categories of household items may be saved given a certain length of mitigation time. This information was combined with the stage-damage relationships from Penning-Rowsell et al. (2005), which is conveniently broken down into similar categories. This allows for estimating new stage-damage curves for single residential properties, conditional on the length of mitigation time available. These can be used to determine new stage-damage curves for the White Cart basin, which are subsequently used to plot the probability – damage curves for the Perfect Forecasts case (Fig. 9). Calculating the area below these curves yields the expected annual flood damage, conditional on the presence of a perfect warning system and given a certain mitigation time. These amounts are listed in Table 2. The ΔEAD column shows the flood risk reduction (losses avoided) achieved by the (perfect) warning system. The table shows that losses avoided increase with lead-time as expected, although the relationship is not smooth due to increments in the categories of items being potentially saved at increasing lead-times.

Loss reduction comes at a cost, namely that of flood warning response. In the hydro-economic EAD model, costs can be added to flood damage in the stage – damage relationship (top right quadrant of Fig 7). This leads to a changed probability – damage relationship, thus yielding new estimates of flood risk which now includes the cost of warning response. Assuming that the response cost may be expressed as a fraction of avoidable losses (Eq. 2), resulting flood risk (original flood risk minus loss reduction plus response costs) may be plotted

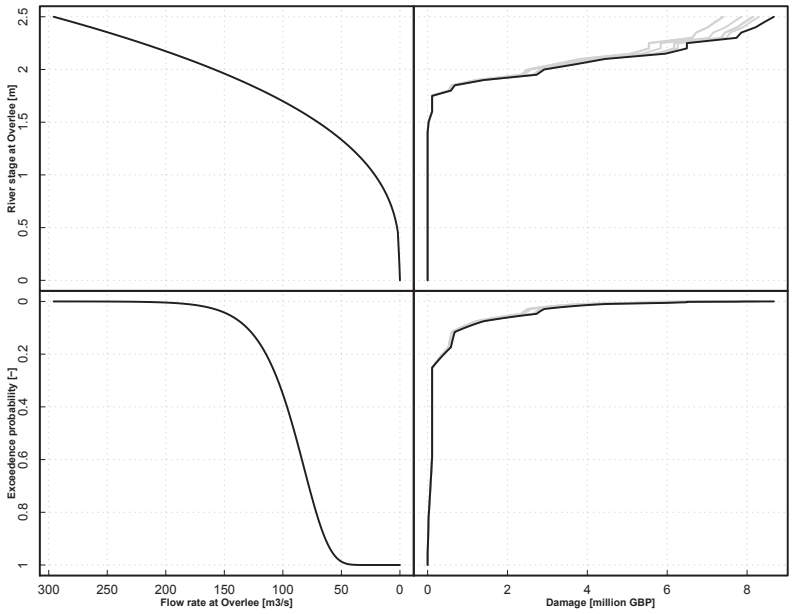


Figure 9: Hydro-economic EAD model for the No Warning (black) and Perfect Forecasts (grey) cases. Reduced damage as a result of flood warning response results in reduced expected damage; from right to left, grey lines show damage curves as a function of increasing mitigation times from 1 to 6 hours. Resulting flood risk is listed in Table 2.

as a function of r (Fig. 10). This shows that for users whose costs are negligible (which means that $r \approx 0$), the maximum loss reduction is attained (i.e. that of the Perfect Forecasts case), with an increase of losses avoided as lead-time increases (Eq. 2). Flood risk increases with cost-loss ratio; if the cost of warning response approaches the amount of potential loss reduction ($r \rightarrow 1$), flood risk approaches original, No Warning levels. For values of $r > 1$, where the cost of response is larger than the losses avoided, the total flood risk would increase when compared to the case of No Warning. This is not considered here as it would then clearly not be rational to employ a flood warning service. Note that lines for 5- and 6-h lead-times coincide as the potential losses avoided are equal.

2.3.3 *Case 3: deterministic forecasts*

In reality, the forecasting component of a FFWRS is unlikely to be perfect and predictive uncertainty will result in both missed events and false alarms occurring. For White Cart, the frequency of these was determined using the re-forecasting analysis. The available record of

Table 2: Loss reduction in terms of expected annual damage (response costs not included).

Case	mitigation time [h]	EAD [GBP]	ΔEAD [GBP]	ΔEAD [%]
No Warning		394 695		
Perfect forecasts	1	386 871	− 7 824	−2%
Perfect forecasts	2	384 640	−10 055	−3%
Perfect forecasts	3	384 129	−10 566	−3%
Perfect forecasts	4	359 473	−35 221	−9%
Perfect forecasts	5	349 913	−44 782	−11%
Perfect forecasts	6	349 913	−44 782	−11%

precipitation observations (April 1996–January 2007) was used to force the hydrological forecasting model for White Cart. Forecasts were produced four times daily with a maximum forecast horizon of 6 h and paired with their corresponding observations.

This information was subsequently used to create contingency tables (one for every lead-time, Table 3). This table shows the number of occurrences of hits h, missed events m, false alarms f and quiets q respectively, adding up to the total number of decisions made N. This is a high number as the re-forecasting analysis covered almost 11 years with a re-forecast being produced four times daily. While this re-forecasting frequency seems high, it still causes some sampling issues, as shown by the performance of the 3-h lead-time re-forecasts versus that of the 5-h lead-time re-forecasts: the latter has a better ratio of hits to false alarms than the former.

The information from the contingency tables was used to determine REV as a function of cost-loss ratio and lead-time (Fig. 11). The figure shows that REV for the 1-h forecasts is unaffected by false alarms as none were observed in the re-forecasting period at this short lead-time. This results in the REV being independent of the cost-loss ratio. As there were misses at these short lead-times, the REV is lower than that of the perfect forecast. Longer lead-times all show declining REV with increasing cost-loss ratios. This is due to false alarms which become increasingly expensive with increasing values of r. It can now be seen that the REV for forecasts at 5- and 6-h lead-time no longer coincide – as the uncertainty increases with lead-time, resulting in an increasing number of false alarms and misses.

Flood risk in the present case can be calculated by scaling the flood risk estimates from benchmark cases No Warning and Perfect Forecasts with REV, using Eq. (10). This gives EAD_{FFWRS} as a function of lead-time and of cost-loss ratio. Flood risks for the Deterministic Fore-

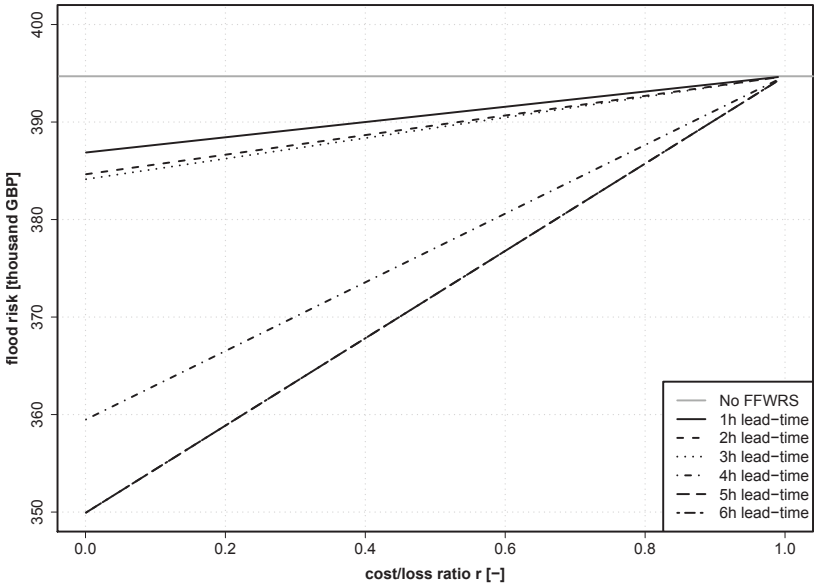


Figure 10: Flood risk in the Perfect Forecasts case, as a function of cost-loss ratio and lead-times. This flood risk includes unavoidable flood damage and the cost of flood warning response.

casting case for all lead-times and all users are shown in Fig. 12. The figure also shows the original flood risk (i.e. from the No Warning case). It can be seen that for users with a cost-loss ratio $0 \leqslant r \leqslant .8$, issuing warnings with 5-h lead-time leads to the lowest flood risk. Users with higher cost-loss ratios benefit most from warnings based on a 1-h lead-time. For these users, false alarms are costly and minimising forecasting uncertainty yields more benefits than a longer mitigation time.

For all lead-times larger than 1 h, the resulting flood risk increases beyond that of the case using No Warning for the higher values of r. This is again attributed to the increasing expense of false warning-response. At 6-h lead-times a much higher residual flood risk is found than at the 5-h lead-times, meaning that considering 6-h lead-time forecasts in making a decision to initiate a warning response is detrimental for values of $r \succeq 0.75$. Clearly this is a result of the lack of additional potential of avoiding losses at this increased lead-time (Table 2), combined with the occurrence of fewer hits and more misses and false alarms.

Table 3: Performance of the FFWRS based on deterministic forecasts, expressed in the elements of a contingency table.

lead-time [h]	h [–]	m [–]	f [–]	q [–]	N [–]
1	10	2	0	15 860	15 872
2	10	4	2	15 856	15 872
3	6	9	2	15 855	15 872
4	7	7	2	15 856	15 872
5	8	6	2	15 856	15 872
6	6	9	3	15 854	15 872

2.3.4 *Case 4: probabilistic forecasts*

For the forecasting sub-system based on probabilistic forecasts, users may choose their own decision rule. This means that they may either raise or lower the probability threshold at which a decision whether or not to initiate a warning response is taken. While probabilistic forecasting and associated decision rules do not affect flood losses that can be avoided at different lead-times, it does affect probabilities of detection and false alarm rates and therefore allows the user to optimise residual flood risk by tuning the costs associated with forecasting uncertainty.

In this case, a hindcast was made using the hydrological model and the QR post-processor. The same hindcasting period and forecasting frequencies as in the Deterministic Forecasts case (Sect. 2.3.3) were used. For every forecast, the probability of exceeding the flooding threshold was determined, and these were paired with the observed threshold exceedences. From these pairs of forecasts and observations, for every decision rule, the number of resulting hits, misses, false alarms and quiets was determined. Table 4 shows these numbers for forecasts with a 3-h leadtime. For the decision rule 'warn if forecasted event probability is equal to or higher than 0 per cent' (i.e. always issue a warning) the number of hits is equal to the number of observed events ($h = e = 15$) with the number of false alarms being equal to the number of forecasts made, minus the number of hits ($f = N - h = 15\,857$). At the other extreme, the decision rule 'warn if forecasted event probability equals 100 per cent' results in zero hits, zero false alarms and all events missed.

These in turn were used to determine REV as a function of cost-loss ratio and lead-time. Figure 13 shows the REV for forecast with a leadtime of 3 h. Note that the figure shows multiple REV-curves; one for every decision rule, where probability of flooding exceeds 0, .1, .2, ..., .9, and 1. The upper enveloping curve is printed in black, showing the optimal decision rule as a function of the cost-loss ratio r. It is assumed

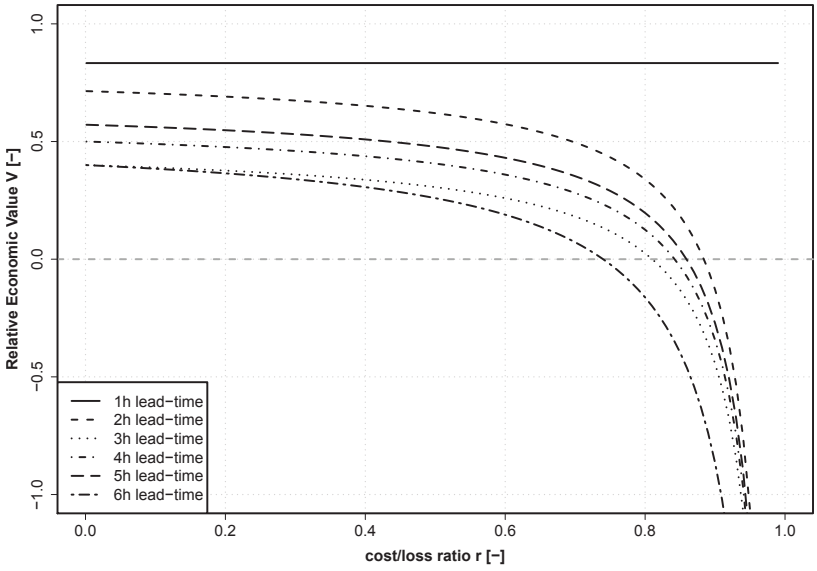


Figure 11: Relative economic value as a function of cost-loss ratio and lead-time in the Deterministic Forecasts case.

that every user will optimises the REV by choosing the decision rule coinciding with their own cost-loss ratio. The procedure to calculate flood risk is identical to that used in the previous case. Figure 14 shows the resulting flood risk. For higher values of r, the increasing cost of response to false alarms reduce the benefit of flood warning, ultimately resulting in a higher residual flood risk than in the No Warning case. The increasing number of false alarms for decision rules with decreasing thresholds compounds this effect. The decision rule with a probability threshold of 1 converges to the same residual risk as for the No Warning case for all r, with the added cost of operating the (useless) FFWRS.

2.3.5 *Summary of results*

The flood risk estimates for different scenarios are summarised in Fig. 15. The figure contains six plots, one for each lead-time considered. All plots show results from the four cases investigated. The No Warning case results in flood risk values that are independent of either lead-time or forecasting uncertainty and therefore constant for all users. In case of a perfect FFWRS, increased lead-time results in increased loss mitigation and decreasing flood risk. Maximum loss mitigation, i.e. minimum flood risk, is attained for those users whose

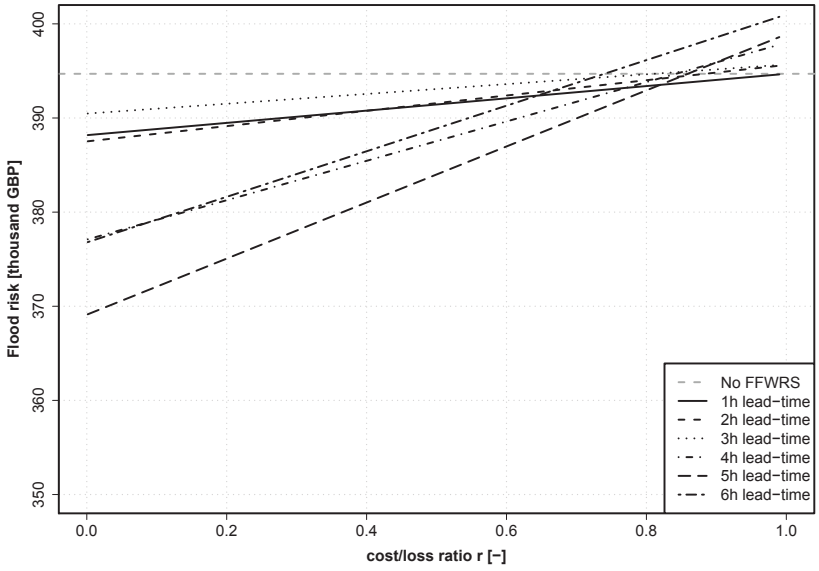


Figure 12: Flood risk as a function of cost-loss ratio and lead-time in the Deterministic Forecasts case.

actions come at little or no cost ($r \approx 0$). For all other users, the costs of mitigating action increases flood risk. If the cost of flood response equals the mitigated losses ($r \to 1$), flood risk is equal to that in the No Warning situation.

Results of the imperfect FFWRS cases show that there is a trade-off between the benefits of loss mitigation and the costs associated with forecasting uncertainty. Both increase with lead-time, while that benefit decreases with increasing cost-loss ratio. In all cases, the envelope curve of the probabilistic forecasts results in a lower residual risk than for the deterministic forecast, irrespective of the lead-time and cost-loss ratio of the user. As the cost-loss ratio approaches zero, the probabilistic forecast converges to the perfect forecast system. This is in a sense meaningless, as the low cost of response results in probability thresholds being set to zero so that the response decision is positive for every forecast made. This artefact disappears with increasing response costs. For users with high cost-loss ratios, the costs associated with forecasting uncertainty can be so high that the resulting flood risk is higher than it would be if no system were in place. It is interesting to note that the cost-loss ratio at which this occurs is very similar for both the probabilistic and deterministic forecasting sub-system.

Table 4: Performance of a warning system based on probabilistic forecasts, expressed in the elements of a contingency table. This table pertains to decisions based on forecasts with a 3-h lead-time.

threshold [–]	h [–]	m [–]	f [–]	q [–]	N [–]
0	15	0	15 857	0	15 872
0.1	15	0	17	15 840	15 872
0.2	13	2	7	15 850	15 872
0.3	13	2	6	15 851	15 872
0.4	11	4	5	15 852	15 872
0.5	11	4	5	15 852	15 872
0.6	8	7	2	15855	15 872
0.7	6	9	2	15 855	15 872
0.8	3	12	1	15 856	15 872
0.9	3	12	1	15 856	15 872
1	0	15	0	15 857	15 872

2.4 DISCUSSION

2.4.1 *Probabilistic versus deterministic forecasting*

The method presented allows for estimating the costs of forecasting uncertainty given different decision rules. Thus, deterministic forecasts and associated decision rules can be compared with probabilistic forecasts and decisions. The analysis shows that when optimising on long-term flood risk, probabilistic forecasts yield higher flood risk reductions than deterministic forecasts. This is due to the fact that a user can choose a probabilistic decision rule that is befitting of the user's cost-loss ratio, thus optimising on expected costs and benefits. In the case of deterministic forecast, this is not possible due to the absence of uncertainty information and therefore a lack of information for risk-based decision-making.

In the application of the method to the White Cart, *observed* precipitation was used in the forecast re-analysis period. Forecast uncertainty was estimated through Quantile Regression, with the regressions derived based on the deterministic model performance using these data. There is interdependency between the two types of forecasts: the uncertainties in the deterministic forecasts are made explicit in the probabilistic forecast. The main difference, of course, is that these uncertainties remain "hidden" in the case of single value forecasting, thus preventing risk-based decision making.

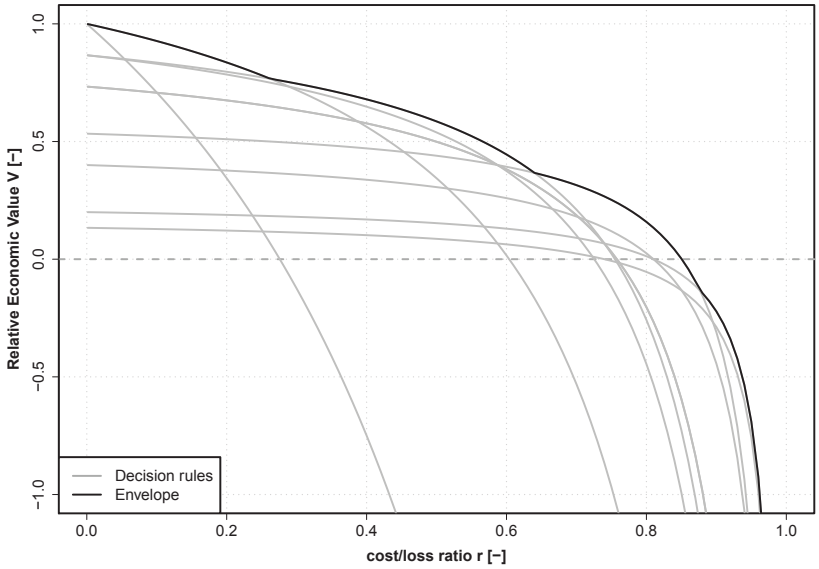


Figure 13: Relative economic value as a function of cost-loss ratio for decisions based on probabilistic forecasts with a 3-h lead-time. The grey lines correspond to the values of V for each of the decision rules $P \geqslant 0,\ 0.1,\ 0.2,\ \ldots\ ,\ 1$ (from top to bottom). The black line is the envelope of these curves.

If the technique for producing probability forecasts would depend on the use of different forcing data than that used for producing deterministic forecasts, this interdependency could well be different. It is not uncommon, for example, for flood forecasting agencies to use a high-resolution deterministic meteorological forecast for a deterministic forecast, and a meteorological ensemble product of lesser resolution for a probability forecast. In that case, the uncertainties could be different and the relative performance of the two cases could be different also.

2.4.2 *Limitations and assumptions*

The hydro-economic EAD model and the theory of Relative Economic Value are tools that value systems in terms of direct, tangible damage only. Indirect and/or intangible flood damage is not included in the flood risk estimates, nor in the estimates of cost-loss ratios. Notably, there may be a wish to estimate the number of flood casualties and the reduction thereof by flood risk management measures (e.g. Molinari, 2011). Possibly, the model can be adapted to include casualties and

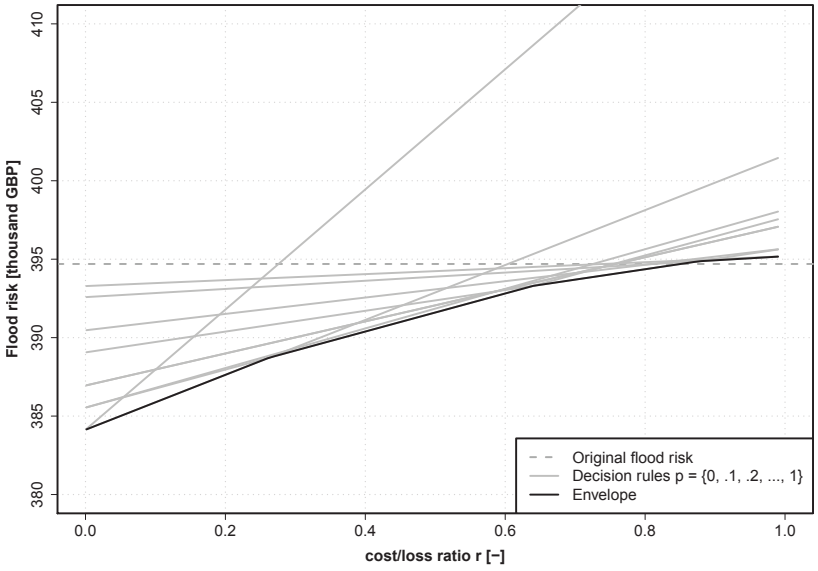


Figure 14: Flood risk as a function of cost-loss ratio for decisions based on probabilistic forecasts with a 3-h lead-time. The grey lines correspond to the values of flood risk for each of the decision rules $P \geqslant 0,\ 0.1,\ 0.2,\ \ldots\ ,\ 1$ (from bottom upwards). The black line is the envelope of these curves.

other types of flood damage but in the present chapter, no attempt to do that has been made as it was deemed to be outside of its scope.

Another limitation to the hydro-economic EAD model is the assumption that direct, tangible flood damage can be estimated as a function of flood depth only. This omits other important determinants such as flow velocity, flood duration and flood water quality. Merz et al. (2010) suggest that flood depth is the most important indicator of flood damage, as is considered here. Penning-Rowsell et al. (cited in Messner et al., 2007) propose a simple method to include additional parameters such as duration of flooding by increasing the damage at a given depth. Other factors can equally be incorporated to create a "compound" depth-damage curve.

In this chapter, it was assumed that decisions are based on forecasts only. In reality, forecasters will add an important element to the forecast model output: expert judgement. Very likely, this expert judgement will introduce a probabilistic element to deterministic model outputs. Forecasters will only issue a warning if they think there is a high probability of flooding. In that sense, the deterministic system that is assessed in this chapter is a stereotype that may not be easily found in reality.

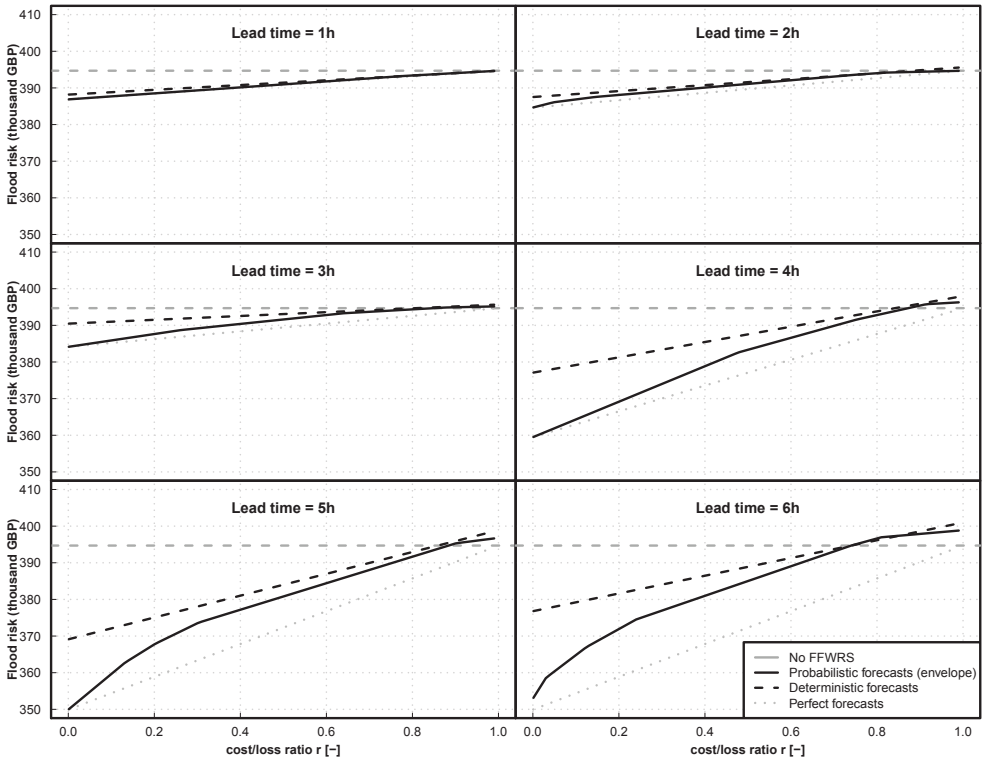


Figure 15: Flood risk as a function of cost-loss ratio, for all lead-times and all cases.

Flood warning systems introduce costs, including initial costs for designing and implementing a system, recurring costs for operation and maintenance, and variable per-event costs. The approach that is presented in this chapter assumes that these costs can all be included in the per-event costs. Alternative attributions of costs may exist though. Possibly, these alternative methods can be included in the method. For example, recurring costs may be included in flood risk estimates by shifting the stage – damage curve to the right. Initial costs can then be included in annually recurring costs. This is in line with best practices on depreciation of assets, where the investment in the warning system is allocated to its expected useful life.

In calculating the benefits of the provision of warning, it would seem that the reduction of losses in this case are modest. These have been derived using only a rough estimate of damage to inundation in the flood warning area downstream of Overlee, and a more complete flood risk assessment would be required to provide more reliable figures. When considering the possible benefits of flood warning, it is important to consider the economy of scale. Operational costs for forecasting are in-

curred in FEWS Scotland at the national level (Werner et al., 2009), which provides warnings across Scotland. Whilst the costs of modelling increase with every warning scheme considered, it is clear that many costs are shared – thus increasing the relative benefit of flood warning.

2.4.3 *Possible implications for policymakers*

The present study shows that FFWRS that are based on probabilistic forecasting bring higher benefits than FFWRS that are based on deterministic forecasting. These benefits can only be realised, of course, if forecasting authorities include probability forecasting in their standard operating procedures. In England and Wales, such a move was recently suggested by Pitt (2008). However, the Pitt Review also suggested that "*...the Met Office and the Environment Agency should produce an assessment of the options for issuing warnings against a lower threshold of probability*". The present study, however, shows that this may not be a good option for *all* forecast users.

Probabilistic forecasting allows for a decision maker to choose a decision rule in terms of the required minimum probability of threshold exceedence. This assumes that the user is capable of optimal decision making in the presence of uncertainty, but also that the cost-loss ratio is known. Especially the latter is not trivial and may be subject to considerable uncertainties. Also, a user's cost loss ratio may change over time and may depend on flood stage and lead-time.

The benefits of FFWRSs depend to a high degree on system efficiency, which consists of a number of factors pertaining to other elements of a FFWRS than its forecasting component. Here, it was assumed that damage mitigation is half of the potential damage mitigation ($L_a = .5\, L_p$). Note that this affects all 'with warning' cases equally. Increasing system efficiency is outside of the scope of the present chapter but currently the topic of scientific research (e.g. Parker et al., 2008; Molinari and Handmer, 2011).

The benefits of probabilistic forecasting can only be attained if forecast users apply optimal decision rules, i.e. if they are able to *manage* predictive hydrological uncertainty. This may pose substantial requirements to decision makers. Possibly, they will have to be trained in decision making. Also, it is likely that a shift to probabilistic forecasting will require forecasting procedures to be adjusted.

The approach that was presented may help a decision maker in prioritising available flood risk management measures. The present chapter shows that these may include measures aimed at reducing either the cost of warning response, at increasing the potential loss reduction, or both. For example, increasing the potential loss reduction may be achieved by increasing the efficiency of flood warning (Sect. 2.2.1)

through awareness raising or flood response exercises. The flood risk analysis now allows for these non-structural measures to be compared with structural engineering measures.

2.4.4 *Open questions and future research*

Probabilistic forecasts used in the present study have not been evaluated in terms of reliability or sharpness. Whereas here, the envelope of multiple probabilistic forecasting risk curves was used, it was not checked whether these coincide with optimal decision rules. Should the probability forecasts show poor calibration, this may not be the case. Additionally, while it is known that the value of a FFWRS does not always increase with forecasting accuracy (Murphy and Ehrendorfer, 1987), it is assumed that the value will increase with increasing sharpness. It would be worthwhile to have a clearer idea of what qualities of a forecast need to be improved for maximisation of value.

The benefits of probability forecasting stem from the possibility of tuning a decision rule so that an optimal balance between forecasting lead-time and forecasting uncertainty is attained. This is assumed not be the case in deterministic forecasting as only a single decision rule is deemed possible. Theoretically however, this assumption may be relaxed and warnings may be issued against a single value threshold different from the flood level. This calibration of deterministic warnings may bring identical benefits.

In reality, FFWRS rarely use a single threshold only. Often, a phased warning and response approach is used. These phases may range from an increase in forecasting frequency to evacuation of floodplain residents. In principle, a phased approach will also benefit from a move to probabilistic forecasting.

2.5 SUMMARY AND CONCLUSIONS

A method for estimating the benefits of flood forecasting and warning, and comparing this against the benefit of other flood risk reduction measures is presented. The method is based on the established hydro-economic expected annual damage (EAD) model. This model is extended with the concept of Relative Economic Value (REV), which is a metric for verifying probability forecasts in terms of economic benefits relative to scenarios where a forecasting system is either absent, or perfect. This allows the cost of predictive uncertainty in estimating the benefit of an uncertain (or imperfect) flood warning to be considered. The method allows for comparing the benefits of warning systems relying on deterministic, single value forecasts with those using probability forecasts. In addition, the method may be used to estimate flood risk

reduction through *improving* flood forecasts, e.g. by using more reliable forcings, better models, improved model parametrisation and/or data assimilation.

In the probabilistic case, the probability threshold at which a response is initiated can additionally be optimally chosen as a function of the cost-loss ratio of the forecast user. As uncertainty can be expected to increase with lead time, the method allows an optimal forecast lead-time to be determined, based on the minimisation of long-term flood risk

The method is applied in a case study to the White Cart Water, a small catchment on the outskirts of Glasgow, Scotland. In this case study it is shown that:

- Using probability forecasts (in combination with the optimal warning rule) results in lower values of residual flood risk when compared to using deterministic, single value forecasts. This is noted throughout different lead times and cost-loss ratios of the user of the forecast.
- The optimal lead-time for warning is not necessarily equal to the longest lead-time that can be provided by the forecasting system, but that it is a function of the cost-loss ratio of the user of the forecast, as well as the uncertainty of the forecast.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Michael Cranston at the Scottish Environment Protection Agency, who allowed water level data and hydrological models of the White Cart Water to be used for this study. Also, the authors would like to acknowledge the Flood Control 2015 programme for part funding the research presented in this chapter. Finally, we are especially grateful to Dr Nathalie Voisin at Pacific Northwest National Laboratory and to one anonymous referee for their referee comments, to Dr Daniela Molinari at Politecnico di Milano for her suggestions given during the Open Discussion and to Dr Yuqiong Liu at the National Aeronautics and Space Administration for editing the journal paper based on this chapter.

3

POST-PROCESSING ECMWF PRECIPITATION AND TEMPERATURE ENSEMBLE REFORECASTS FOR OPERATIONAL HYDROLOGIC FORECASTING AT VARIOUS SPATIAL SCALES

ABSTRACT

The ECMWF temperature and precipitation ensemble reforecasts are evaluated for biases in the mean, spread and forecast probabilities, and how these biases propagate to streamflow ensemble forecasts. The forcing ensembles are subsequently post-processed to reduce bias and increase skill, and to investigate whether this leads to improved streamflow ensemble forecasts. Multiple post-processing techniques are used: quantile-to-quantile transform, linear regression with an assumption of bivariate normality and logistic regression. Both the raw and post-processed ensembles are run through a hydrologic model of the river Rhine to create streamflow ensembles. The results are compared using multiple verification metrics and skill scores: relative mean error, Brier skill score and its decompositions, mean continuous ranked probability skill score and its decomposition, and the ROC score. Verification of the streamflow ensembles is performed at multiple spatial scales: relatively small headwater basins, large tributaries and the Rhine outlet at Lobith. The streamflow ensembles are verified against simulated streamflow, in order to isolate the effects of biases in the forcing ensembles and any improvements therein. The results indicate that the forcing ensembles contain significant biases, and that these cascade to the streamflow ensembles. Some of the bias in the forcing ensembles is unconditional in nature; this was resolved by a simple quantile-to-quantile transform. Improvements in conditional bias and skill of the forcing ensembles vary with forecast lead time, amount, and spatial scale, but are generally moderate. The translation to streamflow forecast skill is further muted, and several explanations are considered, including limitations in the modelling of the space-time covariability of the forcing ensembles and the presence of storages.

This chapter has been published as Verkade, J.S., J.D. Brown, P. Reggiani, and A.H. Weerts. Post-processing ECMWF Precipitation and Temperature Ensemble Reforecasts for Operational Hydrologic Forecasting at Various Spatial Scales. Journal of Hydrology 501 (September 2013): 73–91. DOI: 10.1016/j.jhydrol.2013.07.039. Supplemental information related to this chapter is available at an online data repository via http://dx.doi.org/10.4121/uuid:56637037-8197-472b-b143-2f87adf49abc

3.1 INTRODUCTION

Hydrologic forecasts are inherently uncertain. Uncertainties originate from the forcing data and from the initial conditions, the model structure and its parameters. Estimating the uncertainties in hydrologic forecasts yields probabilistic forecasts that form one input to risk-based decision making. While "best practice" for using these probabilistic forecasts attracts ongoing debate, there is good evidence to suggest that probabilistic forecasts could improve decision-making if used appropriately (e.g. Krzysztofowicz, 2001; Raiffa and Schlaifer, 1961; Ramos et al., 2012; Todini, 2004; Verkade and Werner, 2011).

Hydrologic models are often forced with the output from numerical weather prediction (NWP) models. As hydrologic models are sensitive to the forcing inputs, and meteorological forecasts often contain significant biases and uncertainties, the forcing data is typically an important source of bias and uncertainty in streamflow forecasting. Meteorological ensemble prediction systems (EPS) are increasingly used in hydrologic prediction (see, for example, Cloke and Pappenberger 2009 for an overview of ensemble use in flood forecasting). Examples of meteorological EPS include the National Centers for Environmental Prediction's Global Ensemble Forecast System (GEFS; Hamill and Whitaker 2006), the UK Met Office's Global and Regional Ensemble Prediction System (MOGREPS; Bowler et al. 2008; Schellekens et al. 2011) and the European Centre for Medium-Range Weather Forecasts' Ensemble Prediction System (ECMWF-EPS; Buizza et al. 2007).

Due to limitations of the models and associated data, forecasts from meteorological EPS generally contain biases in the mean, spread and higher moments of their forecast distributions. These biases are manifest at temporal and spatial scales that are relevant to hydrologic prediction. The information content in the raw forcing may contain valuable information for post-processing. A variety of techniques may be used for this, including techniques that use single-valued predictors, such as the ensemble mean of the forcing forecast (e.g. Kelly and Krzysztofowicz, 2000; Reggiani and Weerts, 2008b; Zhao et al., 2011), and techniques that use additional moments or all ensemble members, as well as auxiliary variables.

Biases in forcing ensembles propagate through the hydro-meteorological system and may, therefore, introduce biases into the streamflow predictions. Biases in streamflow forecasts are often removed through statistical post-processing[1] where, based on the historical per-

[1] In this chapter, the term post-processing is used to indicate reduction of biases and/or estimation of uncertainties using statistical techniques that are applied subsequently to a model run. As such, post-processing is synonymous with bias-correction, forecast calibration, statistically correcting, and preprocessing. In hydrology, the term preprocessing is sometimes used to indicate the post-processing of meteorological forcings prior to being used in a hydrologic model.

formance of the forecasting system, operational streamflow forecasts are statistically corrected in real-time (e.g. Bogner and Pappenberger, 2011; Brown and Seo, 2013; Krzysztofowicz, 1999; Reggiani and Weerts, 2008a; Todini, 2008; Weerts et al., 2011). This correction may lump together the hydrologic and meteorological uncertainties or factor them separately (Brown and Seo, 2013). The two sources of uncertainty are lumped together by calibrating the streamflow post-processor on observed streamflow. The hydrologic uncertainties are factored out by calibrating the streamflow post-processor on simulated streamflow, i.e. on streamflow predictions with observed forcing (Seo et al., 2006; Zhao et al., 2011). In both cases, the streamflow forecasts may benefit from post-processing of the forcing forecasts. However, in separately accounting for the hydrologic uncertainties (the first case), it is assumed that the meteorological uncertainties and biases have been adequately addressed. In contrast, corrections to the streamflow should indirectly account for the meteorological biases and uncertainties if the forcing and hydrologic uncertainties are lumped together into a streamflow postprocessor.

Important questions remain about the combined benefits of forcing and streamflow post-processing in this context. For example, lumping together the forcing and streamflow uncertainties may lead to strongly heterogeneous behaviours that are difficult to model statistically. However, post-processing of forcing forecasts is generally complex and resource intensive, requiring statistical models of temporal, spatial and cross-variable relationships to which streamflow is often sensitive and for which sample sizes may be limited; in short, forcing bias correction may leave substantial residual biases and invoke imperfect models of space-time covariability.

Indeed, initial attempts to address this issue have been reported in the scientific literature. Kang et al. (2010) focused on the reduction of uncertainties by applying post-processing to predicted forcings, to predicted streamflow and both. In their study, post-processed ensemble members were re-ordered using the Schaake Shuffle prior to being used in the hydrologic and hydrodynamic models. The Schaake Shuffle aims to capture spatio-temporal patterns in the observed meteorological forcings that are lost following post-processing of the marginal distributions. The authors found that the forecasts were most skillful when combining post-processing of the forcings with post-processing of the streamflow forecasts. However, they also note that post-processing of the streamflow forecasts more effectively reduced the total uncertainty than post-processing the forcings alone. Clearly, this will depend on the relative importance of the forcing and hydrologic uncertainties in any given basin.

Zalachori et al. (2012) compared the skill of, and biases, in ensemble streamflow forecasts that were produced using different combinations

of forcing and streamflow post-processing. Post-processing of meteorological forcings was performed by dressing the ensemble members with 50 analog scenarios that naturally included appropriate space-time relationships. They found that, while post-processing the forcings increased the skill of the forcing ensembles, there was little improvement in the skill of the streamflow ensembles. Also, those improvements were obscured by the effect of streamflow post-processing.

Similarly,Yuan and Wood (2012) explored the benefits of post-processing of forcing ensembles versus post-processing of streamflow ensembles, but in a different context, namely that of seasonal forecasting. They found that both post-processing of forcings and post-processing of streamflow adds skill, and when techniques are combined, skill is highest.

Several techniques have been proposed for reducing bias in forcing forecasts (Hamill, 2012). These techniques use past forecasts and observations (and possibly auxiliary variables) to estimate the parameters of a statistical model that is subsequently applied in real-time to estimate the "true" (unbiased) probability distribution of the forecast variable, conditionally upon the raw forecast (and any other predictors). Techniques include linear regression with an assumption of joint normality (e.g. Gneiting et al., 2005; Hagedorn et al., 2008; Wilks, 2011), logistic regression (Hamill et al., 2008; Wilks, 2011), quantile regression (Bremnes, 2004) and indicator co-Kriging (Brown and Seo, 2010, 2013), among others. Unsurprisingly, Wilks and Hamill (2007) conclude that no single post-processing technique is optimal for all applications.

Statistical correction of numerical weather forecasts requires a long historical record of forecasts and observations, from which the joint distribution can be estimated with reasonably small sampling uncertainty and bias. Unless explicitly accounting for non-stationarity with additional model parameters, the joint distribution should be relatively homogeneous in time. Forecasting systems, however, generally improve over time, rendering archived operational forecasts inhomogeneous. In contrast, weather forecasts that are retrospectively generated with a fixed numerical model ("reforecasts" or "hindcasts"), provide a reasonable platform for statistically correcting weather forecasts (Hamill et al., 2006). Available reforecast datasets include the ECMWF-EPS (Hagedorn, 2008), GFS (Hagedorn et al., 2008; Hamill and Whitaker, 2006; Hamill et al., 2008), and the more recent GEFS, for which hindcasts were recently completed (Hamill et al., 2013) and TIGGE (Hamill, 2012).

The extent to which the skill of, and biases in, streamflow forecasts can be improved through post-processing of the forcing ensembles, separately or together with streamflow post-processing, is an ongoing question and the focus of this chapter. For example, these issues must be explored in basins with different hydrologic characteristics

and for which the total uncertainties comprise different contributions from the meteorologic and hydrologic uncertainties, including a mixture of headwater and downstream basins. First, we evaluate the biases in the forcing ensembles at the scales used to force the hydrologic models, and how these biases translate into the streamflow ensemble forecasts. Secondly, a number of bias-correction techniques are applied to the temperature and precipitation ensembles. The post-processed forcing ensembles are used to drive the hydrologic models, which are then evaluated for any reduction in bias and increase in skill associated with the forcing post-processing. These post-processing techniques include the unconditional quantile-to-quantile transform (a correction to the forecast climatology) as well as conditional techniques such as linear regression in the bivariate normal framework and logistic regression. The streamflow ensembles are evaluated at multiple spatial scales and, crucially, by verifying against simulated streamflows (predictions made with observed forcings), in order to isolate the contribution of the forcing biases and uncertainties to the streamflow forecasts.

The structure of this chapter is as follows. The Materials and Methods section describes (i) the techniques that have been used for post-processing of forcing ensembles, (ii) the study basin, (iii) the models and data that are used and (iv) a detailed setup of the different experiments. The results are presented in Section 3.3 and subsequently discussed in Section 3.4. Finally, some conclusions are drawn together with suggestions for future studies (Section 3.5).

3.2 MATERIALS AND METHODS

3.2.1 *Post-processing techniques*

Several techniques were used to post-process the temperature and precipitation ensembles. Temperature ensemble forecasts were post-processed using the quantile-to-quantile transform and, separately, using linear regression. For precipitation, the quantile-to-quantile transform was used, as well as logistic regression. A brief description of each technique is provided below; more details can be found in Appendix A.

The quantile-to-quantile transform (QQT, sometimes also called Quantile Mapping or cdf-matching, e.g. Brown and Seo 2013; Hashino et al. 2007; Madadgar et al. 2012; Wood et al. 2002) is an unconditional technique insofar as the unconditional climatology of the forecasts is re-mapped to the unconditional climatology of the observations. QQT is not expected to provide post-processed ensembles that are equally skilful as those resulting from a conditional correction. However, the skill of a conditional correction may largely stem from an improvement in forecast climatology and an unconditional correction provides a valuable baseline for a more complex, conditional correction.

The conditional post-processing techniques are often applied in similar ways. For each of the forcing variables, the post-processor is configured for each lead time and each location (basin-averaged quantity) separately. A distribution of the predictand Y (observed temperature or precipitation) is sought, conditional upon a vector of predictors $\mathbf{X} = X_1, \ldots, X_m$. In this case, the predictors comprise the five (possibly biased) ensemble members of the raw forecast.

$$F(y|x_1, \ldots, x_m) = \Pr[Y \leqslant y | X_1 = x_1, \ldots, X_m = x_m] \quad \forall y \qquad (11)$$

For post-processing temperature ensemble predictions, the observed and forecast temperatures are frequently assumed joint normally distributed. Linear regression is then used to estimate the mean and spread (and hence full probability distribution) of the observed variable conditionally upon the predictors (Gneiting et al., 2005; Hagedorn et al., 2008; Wilks, 2011).

Predictive distributions of precipitation are non-Gaussian (e.g. Hamill et al., 2008), and threshold-based or "non-parametric" techniques are often applied, although a meta-Gaussian approach is also possible (Wu et al., 2011). Precipitation forecasts are often biased conditionally upon observed precipitation amount (a so-called Type-II conditional bias), with overestimation of smaller observed precipitation and underestimation of larger observed precipitation. These amounts are typically important for practical applications of hydrologic forecasts (e.g. for drought and flood forecasting; see Brown and Seo 2013). Logistic regression is a common approach for post-processing of precipitation forecasts and is known to perform reasonably well in a variety of contexts (e.g. Hamill et al., 2008; Schmeits and Kok, 2010; Wilks, 2011). The technique involves estimating the probability of not-exceeding several discrete thresholds, for which the parameters of the logistic regression may be estimated separately at each threshold (standard logistic regression) or fixed across all thresholds (Wilks, 2009). In estimating the parameters separately at each threshold, the cumulative probabilities are not guaranteed to be valid in combination, and some post-correction smoothing is typically required.

A potential problem with statistically post-processing temperature and precipitation forecasts separately at each of multiple forecast lead times and locations is that space-time covariability is not adequately captured. For hydrologic applications, the space-time covariability of the forcing is important as the hydrologic model integrates the forcing both in time and in space (Clark et al., 2004).

In order to introduce appropriate space-time covariability into the post-processed forcing ensembles, the so-called "Schaake shuffle" was used here (Clark et al., 2004). For each ensemble trace, a corresponding observed time-series was obtained from the same start date in a ran-

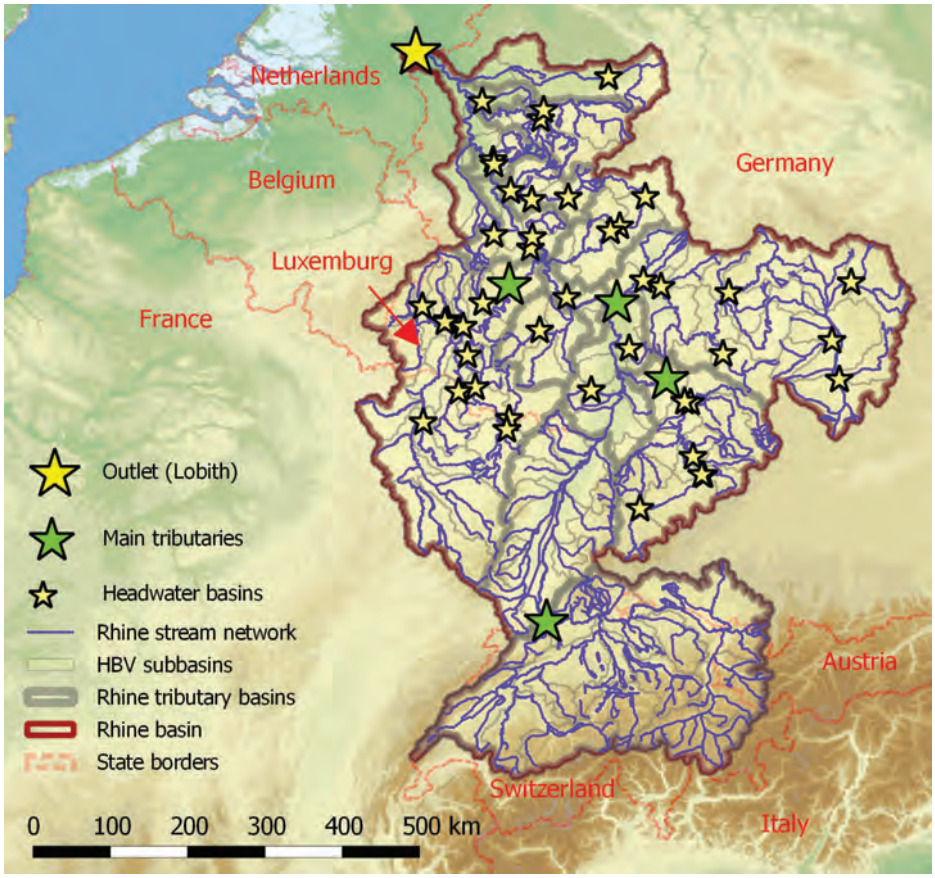


Figure 16: Location of the Rhine basin in continental Europe.

domly chosen historical year. The ensemble members at each forecast lead time were then assigned the same rank positions as the observations from the corresponding (relative) times in their associated historical years. The Schaake shuffle introduces (observed) rank correlations to the forecast ensemble members on the basis that spatial and temporal covariability will lead to ensemble members at nearby locations and proximate times having similar ranks within their own probability distributions. The Schaake shuffle does not, however, capture this space-time covariability conditionally upon the state of the atmosphere at the forecast issue time. Rather, it introduces space-time covariability conditionally upon forecast issue date alone (as formulated in Clark et al. 2004). Clearly, other implementations are possible, such as preservation of the rank order-relations in the raw forecasts.

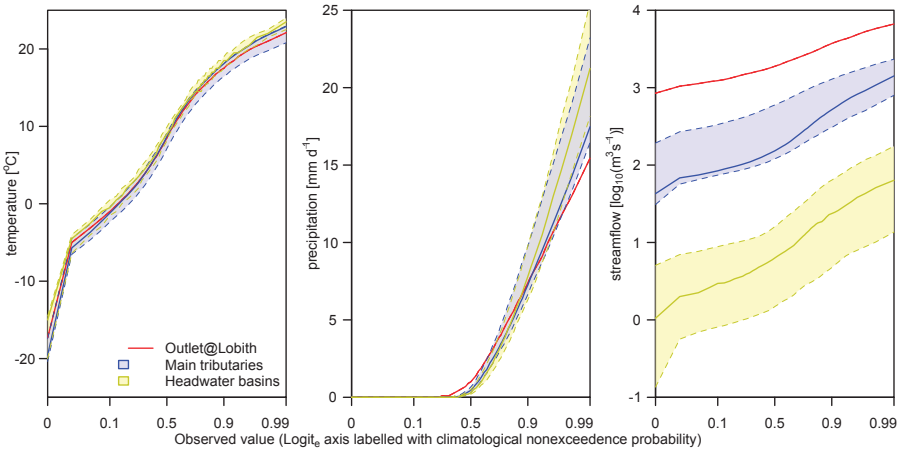


Figure 17: Distribution of daily averaged temperature, daily total precipitation and daily averaged streamflow in Rhine basins and stations. Three spatial scales are shown: 43 headwater basins, four large tributaries and the Rhine outlet at Lobith. For scales containing multiple locations, the median location is shown as a thick line and the 10^{th} and 90^{th} percentiles bound a shaded area.

3.2.2 *Study basin: Rhine*

The river Rhine runs from the Swiss Alps along the French-German border, through Germany and enters the Netherlands near Lobith, which is often considered the outflow. At Lobith, the basin area equals approx. 160,000 km^2. Snow and snowmelt have a large effect on the river Rhine's temporal streamflow patterns. During spring and early summer, more than half of the river's flow at the outlet originates from snowmelt in the Swiss Alps. Figure 16 shows the basin location, elevations and the gauged outlets of tributaries that were used in this study; the three different symbols used for the gauging stations coincide with the three spatial scales used in the analysis.

Clearly, the quality of the streamflow predictions at downstream locations is affected by the quality of the streamflow predictions at upstream locations. Ensemble streamflow predictions are therefore analysed at three spatial scales: (i) 43 outlets of basins that each have a contributing area of less than 2500 km^2; in the remainder of this chapter, these are referred to as headwater basins (ii) four outlets of relatively large Rhine tributaries: the Main, Moselle, Neckar and Swiss Rhine, and (iii) the outlet of the river Rhine, at Lobith. Some summary statistics of the magnitudes of the contributing areas of these outlets are shown in Table 5.

Table 5: Contributing areas of the spatial scales that are analysed.

	Contributing area $[km^2]$				
Spatial Scale	10^{th} perc	median	90^{th} perc	mean	sum
43 headwaters	370	929	2,008	1,142	49,125
4 tributaries	17,972	27,767	34,035	26,507	106,029
Rhine basin					159,559

Figure 17 shows the non-exceedence climatological probabilities of observed daily mean temperature, daily total precipitation and daily averaged streamflow for the three spatial scales used in the analysis. Both the "tributaries" and the "headwater" scales comprise of multiple outlets (four and 43 respectively). For these scales, the thick line designates the median location, and the thin lines designate the 10^{th} and 90^{th} percentiles. In the case of the four main tributaries, determining the quantiles required linear interpolation between four available data points.

Determination of temperature and precipitation at larger spatial scales has a modulating effect on extreme values of temperature and precipitation. The relatively fat tail of precipitation over the four tributaries originates from relatively high precipitation levels over the Swiss Rhine. As none of the headwater basins considered are located in that tributary basin, this fat tail is not observed in the curve for the smaller, headwater basins.

3.2.3 *Models and data*

For the temporal and areal aggregation of ensemble forcing forecasts and corresponding observations, and for retrospective generation of streamflow predictions, a Delft-FEWS forecast production system (Werner et al., 2013) was used. The system is an adapted version of the forecast production system FEWS Rivers, which is used by the Water Management Centre of the Netherlands for real-time forecasting of streamflow and water levels in the Rhine and Meuse rivers.

The system contains an implementation of the HBV rainfall-runoff model (Bergström and Singh, 1995). This is a semi-lumped, conceptual hydrologic model, which includes a routing procedure of the Muskingum type. The model schematisation consists of 134 sub-basins jointly covering the entire Rhine basin. The model runs 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.

The forecasting system runs in two operating modes: historical and forecast mode. In historical mode, the hydrologic models are forced with meteorological observations for a period leading up to the forecast issue time. This ensures that the internal model states reflect the actual initial conditions of the basin as closely as possible. In forecast mode, these model states are the starting point for the model run, where the models are now forced by numerical weather predictions.

For observations of precipitation, the CHR08 dataset was used. This dataset covers the period 1961 through 2007. The CHR08 dataset was prepared specifically for the HBV model used here (Photiadou et al., 2011). The spatial scale of these CHR08 observations coincides with the 134 sub-basins used in the HBV model schematisation for the Rhine basin. Temperature observations originate from version 5.0 of the E-OBS data set; these were available from 1951 through mid 2011 (Haylock et al., 2008). Both precipitation and temperature data were available at a daily time step.

The ECMWF reforecast dataset, comprising medium-range EPS forecasts with 5 ensemble members (Hagedorn, 2008), was used for retrospective predictions of temperature and precipitation. At ECMWF, a retrospective forecast is produced every week for the same date in the 18 years preceding the current year, using the current operational model. To illustrate, on March 13, 2009, reforecasts were produced with initial conditions of March 13, 1991, March 13, 1992, and so forth until March 13, 2008. The reforecasts are produced using the operational model (currently Cy38r1 with a T639 horizontal resolution, i.e. 0.25 degrees in either direction). The set of reforecasts was thus produced using an operational model which, since the inception of the reforecasting scheme, has changed only slightly. This has little or no effect on the hydrologic model outcomes though, as was shown by Pappenberger et al. (2011). By July 2011, over 3,100 retrospective forecasts were available for use in the present study. While the forecast horizon extends to 30 days at a six hour time step, for the present study only the first 10 days were available. Forecasts were temporally aggregated to a daily time step to match the time step used by the hydrologic model. The gridded forecasts were spatially averaged to the HBV sub basin scale.

Hourly streamflow observations for hydrologic stations within the Rhine basin were obtained from the Water Management Centre of the Netherlands. These observations were temporally aggregated to daily averages.

3.2.4 *Experiment*

Streamflow forecasts were produced with raw and post-processed forcings and verified against simulated streamflow, in order to establish the

Table 6: Overview of cases.

	Temperature correction	Precipitation correction
Baseline case	none (RAW)	none (RAW)
Case 1	quantile-to-quantile transform (QQT)	quantile-to-quantile transform (QQT)
Case 2	linear regression (LIN)	logistic regression (LOG)

contribution of the forcing post-processing to the streamflow forecasts independently of any biases in the hydrologic model.

The baseline scenario comprised no post-processing of the forcing ensembles. Raw ensemble predictions of precipitation and temperature were used to generate streamflow ensemble predictions. In subsequent cases, temperature and precipitation ensemble predictions were statistically corrected using the techniques described in Section 3.2.1. These post-processed forcing ensemble predictions were then used to generate streamflow ensemble predictions. Thus, three cases were considered (Table 6): a baseline case, a case where an unconditional quantile-to-quantile transform (QQT) was applied to each variable (Case 1), and a case in which the forcing ensemble predictions were corrected using conditional techniques (Case 2). In terms of the latter, temperature ensemble predictions were statistically corrected using linear regression in the bivariate normal framework (LIN) and precipitation ensemble predictions were corrected using logistic regression (LOG). Variants of these two techniques were also considered, but not adopted. Specifically, for temperature, the assumption of homogeneous spread of the post-processed ensembles was relaxed to allow for a linear dependence on the raw ensemble spread (Gneiting et al., 2005), but without discernible benefits. For precipitation, a variant of LOG involving homogeneous parameters across all thresholds (Wilks, 2009) was evaluated, but this incurred an appreciable loss of skill.

3.2.5 *Post-processing strategy*

The parameters of any post-processor must be estimated with sample data. Both ensemble predictions and verifying observations were available for the period 1991–2007. This amounted to roughly 2,920 pairs of forecasts and observations at each forecast lead time. These pairs were not evenly distributed over the period of record due to the reforecasting procedure adopted by ECMWF.

The forcing ensembles were post-processed using the approaches described in Section 3.2.1 and Appendix A. Post-processing was performed separately for each of the 10 forecast lead times and 134

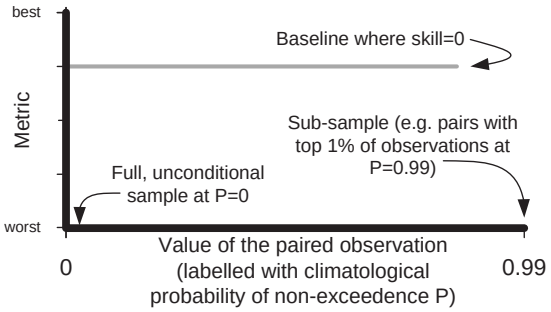


Figure 18: Sample plot showing how the skill score plots are defined. For event skills, the value of P constitutes the threshold for discrete events.

subbasins. Spatio-temporal covariability was then introduced via the Schaake Shuffle (Section 3.2.1). The post-processing was conducted within a cross-validation framework whereby separate periods of record were used to estimate the model parameters and independently verify the post-processed forecasts. Specifically, a leave-one-year-out cross-validation approach was adopted. This led to 17 separate calibrations of each post-processor, each comprising 16 years of calibration data and one year of independent prediction. The 17 years of independent predictions were then collated, verified, and used to force the streamflow models.

3.2.6 *Verification strategy*

The verification strategy focused on identifying the skill and biases in the forcing ensembles, as well as in the streamflow ensembles generated using these forcings. Skill and bias were identified with five well-known verification metrics. The correlation coefficient and the Relative Mean Error (RME) are measures of, respectively, the linear association of the forecast ensemble mean and observations and the relative bias of the ensemble mean. The (half) Brier Score (BS), the mean Continuous Ranked Probability Score (CRPS) and the area under the Relative Operating Characteristic (ROC) curve measure different attributes of the probabilistic quality of the forecasts. A short description of the latter scores is provided below, with accompanying equations given in Appendix B. Verification was performed with the Ensemble Verification System (Brown et al., 2010). The data that constituted input for verification, is posted to an online data repository (Verkade et al., 2013a).

The Brier Score (Brier, 1950; Murphy, 1973; Wilks, 2001) measures the average square error of a probabilistic forecast of a discrete event. The mean CRPS (Hersbach, 2000; Stanski et al., 1989) is an integral measure of (square) probabilistic error in the forecasts across all possible discrete events. Both the BS and CRPS may be decomposed into further

attributes of forecast quality by conditioning on the forecast variable (the calibration-refinement factorization). In addition, the BS may be decomposed by conditioning on the verifying observation (the likelihood-base-rate factorization). The area under the ROC curve (AUC) is a measure of event discrimination; that is, the ability of the forecasts to adequately discriminate between the exceedence and non-exceedence of a discrete threshold, such as the flood threshold.

Skill scores provide a convenient method for summarizing an improvement (or reduction) in forecast quality over a wide range of basins and conditions, as they are normalized measures. Here, both the BS and CRPS are formulated as skill scores with sample climatology as the baseline. These scores are denoted by the Brier Skill Score (BSS) and the Continuous Ranked Probability Skill Score (CRPSS), respectively. Rather than using the raw ensembles as the reference forecast, the scores are shown for the raw and post-processed ensembles with a consistent baseline, namely sample climatology. Likewise, the ROC Score (ROCS) comprises the AUC of the main forecasting system normalized by the AUC of the climatological forecast, i.e. 0.5 (Appendix B). This allows for the relative improvement of the forcing and streamflow forecasts to be identified in the context of background skill. However, some care is needed with interpretation, as sample climatology is unconditional and, therefore, increasingly (conditionally) biased towards the tails.

For continuous measures such as the CRPSS, conditional quality and skill was 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, as all available data pairs are considered (Bradley and Schwartz, 2011). Conversely, at $P = 0.99$, only the data pairs with observations falling in the top 1% of sample climatology are considered; this amounts to approx. 30 pairs here. For BSS and ROCS – which are event skills – the verification metrics are calculated for increasingly high events which are defined as the exceedence of the value of P. For example, at $P = 0.9$ the event is the occurrence of an observation falling in the top 10% of the climatological distribution. For these discrete measures, $0 \leqslant P \leqslant 1$ as event skills are unknown for thresholds corresponding to the extremes of the observed data sample, nominally denoted by $P = 0$ and $P = 1$. See Figure 18 for a graphical description of this procedure.

While the sampling uncertainties of the verification metrics were not explicitly evaluated here (see Brown and Seo, 2013), the results were not interpreted for thresholds larger than the 0.99 climatological probability or $\backsim$ 30 pairs.

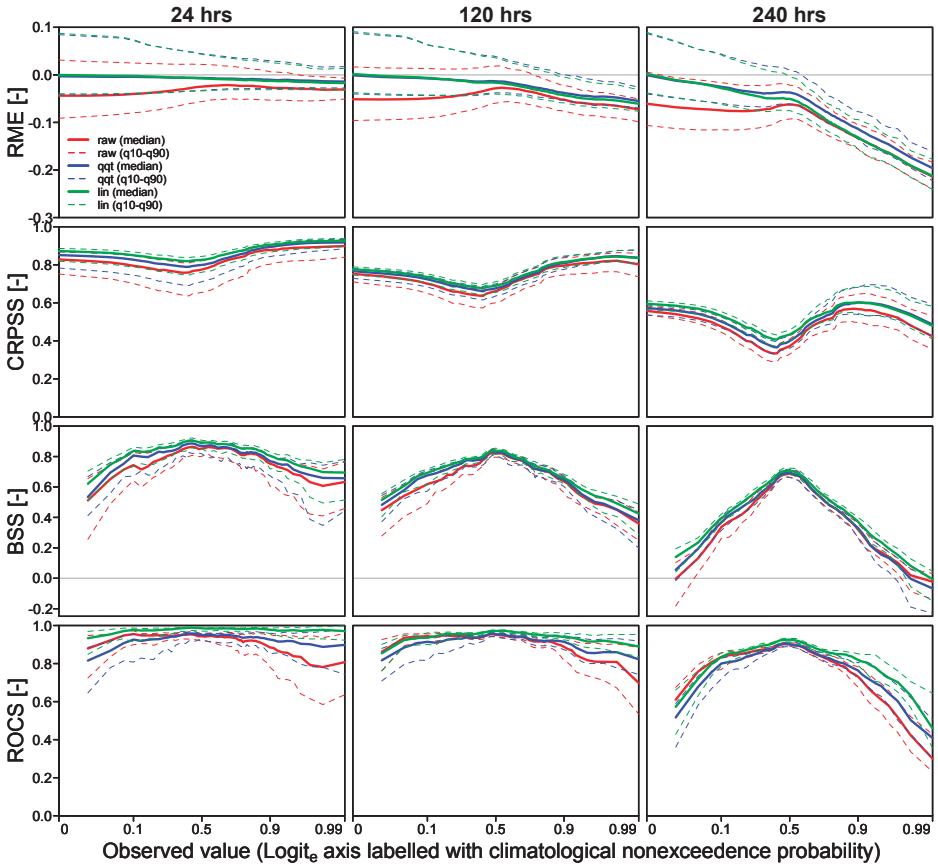


Figure 19: RME, CRPSS, BSS and ROCS for ensemble temperature forecasts at 24-hour, 120-hour and 240-hour ahead forecasts. For CRPSS, BSS and ROCS, the baseline is formed by sample climatology. The results pertain to 134 basins: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

3.3 RESULTS

The results are presented in three subsections, each coinciding with one of the variables considered: temperature, precipitation and streamflow. Within those subsections, a discussion of the baseline case is followed by a discussion of the post-processed cases 1 (QQT) and 2 (conditional corrections LIN and LOG).

Correlation coefficients are very similar across cases and are mentioned in the text but not shown in tables or plots. Verification results are plotted in a series of multi-panel figures, showing RME, BSS, CRPSS and ROCS for the forecasts with lead times of 24-hours, 120-hours and 240-hours. The metrics are plotted as a function of the value

of the verifying observation, expressed as a climatological probability of non-exceedence P, to allow for comparison across different locations. Most figures show results for multiple locations: thick lines indicate median values and thin lines denote the 10% and 90% quantiles of metrics over those multiple locations. Metrics pertaining to streamflow ensemble forecasts are shown across several plots, each corresponding to a spatial scale defined in Section 3.2.2. Note that, for ease of interpretation, all skill scores and associated decompositions are oriented to show the "best" scores at the top of the range axis and the "worst" at the bottom (Figure 18).

3.3.1 *Ensemble temperature forecasts*

Verification metrics for the ensemble temperature forecasts are shown in Figure 19. The metrics indicate that forecast quality decreases with increasing lead time, that it is conditional on the magnitude of the verifying observation and that this conditionality is more pronounced at longer lead times. This is true for both raw and post-processed temperature ensembles.

3.3.1.1 *Raw temperature ensembles*

Raw temperature ensembles show reasonably good correlation with observations. Values for the unconditional sample (at $P = 0$) range from 0.99 at the 24-hour lead time to 0.90 at the 240-hour lead time. At $P = 0.95$, these values are 0.87 and 0.18 respectively. Relative Mean Error plots indicate that for most basins, the ensemble mean underestimates the observation; this under-forecasting increases with higher values of the verifying observation. The CRPSS is largely constant, with a small dip in CRPSS values near to the median observed value. The BSS and ROCS show similar patterns, different from the CRPSS; both scores are consistently lowest at the extreme ends of the distribution.

These patterns reflect the different formulations of the verification scores and the choice of reference forecast. The BSS and ROCS measure the quality of discrete predictions, with contributions to the score being dominated by the corollary (i.e. non-occurrence) at extreme (low and high) thresholds. At longer lead times, the residual skill of the temperature forecasts is concentrated towards the median temperature, where the forecasts have least conditional bias and greatest correlation (and the occurrences and non-occurrences, by definition, contribute equally). In contrast, the CRPS is a smooth, continuous measure that factors skill across all possible thresholds for each paired sample. Since the sample climatology is unconditional by construction, the baseline forecasts will be least reliable in the tails of the climatological distribution, with large conditional biases contributing to poorer quality of the reference

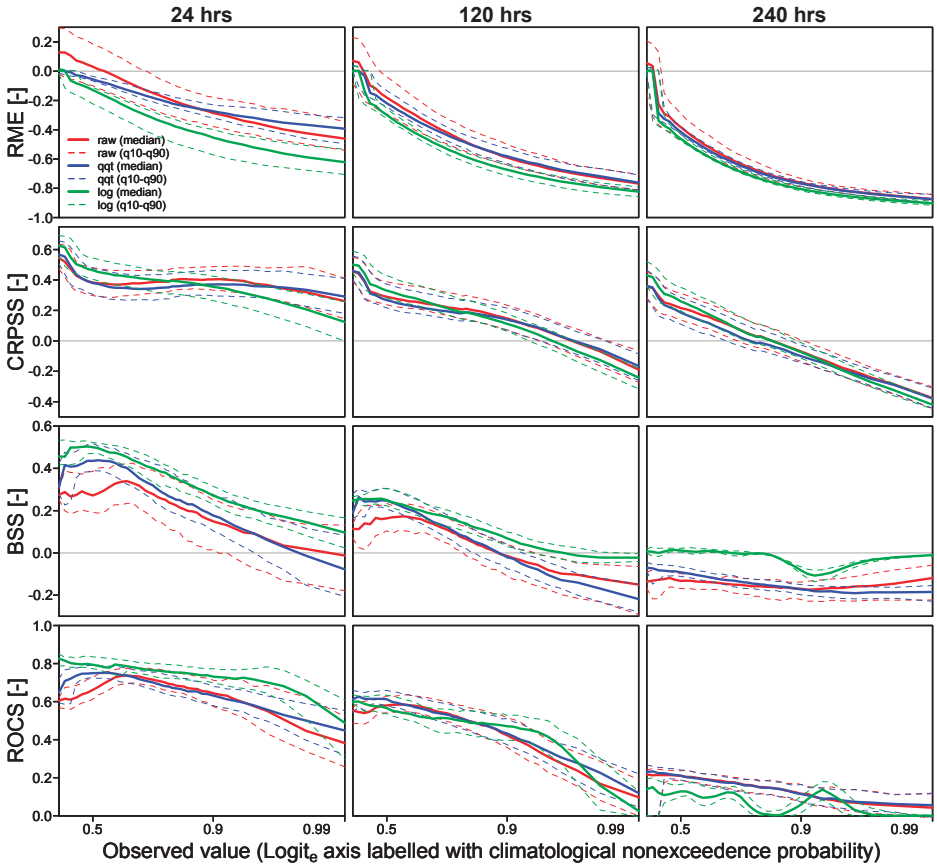


Figure 20: RME, CRPSS, BSS and ROCS for ensemble precipitation forecasts at 24-hour, 120-hour and 240-hour ahead forecasts. For CRPSS, BSS and ROCS, the baseline is formed by sample climatology. The results pertain to 134 basins: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

forecast in the tails (and hence greater relative quality of the ECMWF forecasts, whether post-processed or not).

3.3.1.2 *Post-processed temperature ensembles*

After post-processing, the correlation of the temperature ensembles with the verifying observations was virtually unchanged from the raw case. In terms of RME, BSS, CRPSS and ROCS, LIN almost always outperformed QQT (noting that QQT is a non-linear transform and may not preserve correlation), which in turn outperformed the raw ensembles. For the latter three metrics, the differences in quality are most pronounced at large values of the verifying observation.

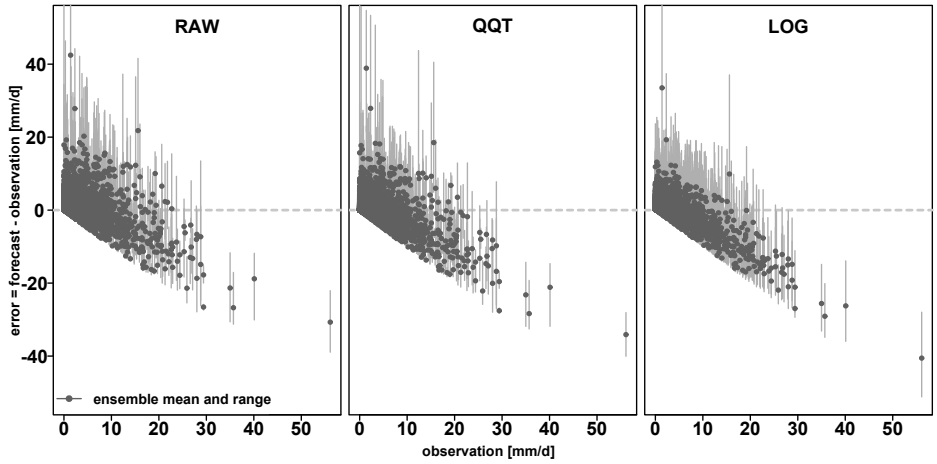


Figure 21: Forecast error versus observations for uncorrected (left), QQT-corrected (middle) and LOG-corrected precipitation ensembles for a single location (basin I-RN-0001, which is located in the Neckar sub-basin) at 120-hour lead time.

3.3.2 *Ensemble precipitation forecasts*

Verification metrics for the ensemble precipitation forecasts are shown in Figure 20. Subsequent figures show the calibration-refinement decomposition of the CRPSS (Figure 22) and the BSS (Figure 23) as well as the likelihood-base rate decomposition of the BSS (Figure 24). Similar to the temperature figures, verification metrics are plotted as a function of observed amount, expressed as a climatological probability of non-exceedence, P. In the case of precipitation, however, the domain axis range is $[0.4, 1.0]$. As the probability of precipitation (PoP) is approx. 60% for all basins, smaller probabilities all correspond to the PoP threshold of zero preciptiation and produce identical scores. As in the case of temperature ensembles, forecast quality is seen to decrease with increasing lead time, and to be strongly conditional on the amount of precipitation.

3.3.2.1 *Raw ensemble precipitation forecasts*

Correlation between the mean of the raw precipitation ensembles and observations is largely positive, but distinctly lower than that of the temperature ensembles. Correlation deteriorates with forecast lead time and with increasing value of the observation. At $P = 0$, correlation ranges from 0.71 to 0.13 for lead times of 24-hour and 240-hours respectively. At $P = 0.95$, these values are 0.36 and 0.04 respectively.

The RME shows that the ensemble mean overestimates zero and small precipitation amounts. For increasing values of the observation,

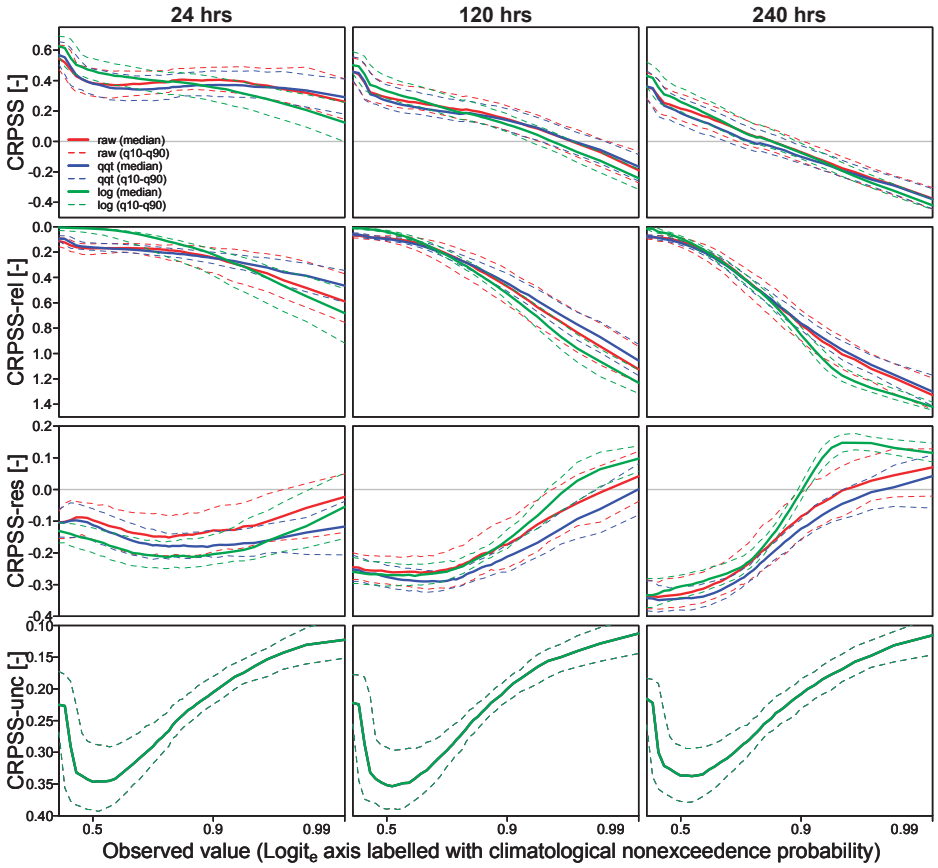


Figure 22: CRPSS calibration-refinement decomposition for ensemble precipitation forecasts at 24-hour, 120-hour and 240-hour ahead forecasts. The baseline is formed by sample climatology. The results pertain to 134 basins: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

the ensemble mean increasingly underestimates precipitation. For example, at a lead time of 120 hours, the RME equals 0.07, −0.18 and −0.59 at $P = 0$, $P = 0.5$ and $P = 0.9$ respectively.

These conditional biases stem from the inability of the raw predictors used in the post-processor to correctly predict when large events occur (large relative to other events in the climatological distribution). This leads to a real-time adjustment that reflects the assumed, but wrong, conditions. Also, statistical post-processors are calibrated for good performance under a range of conditions (i.e. for unconditional skill and unbiasedness), which inevitably leads to some conditional biases. In short, some conditional bias is a "natural" consequence of post-processing with imperfect predictors and with a focus on global optimal-

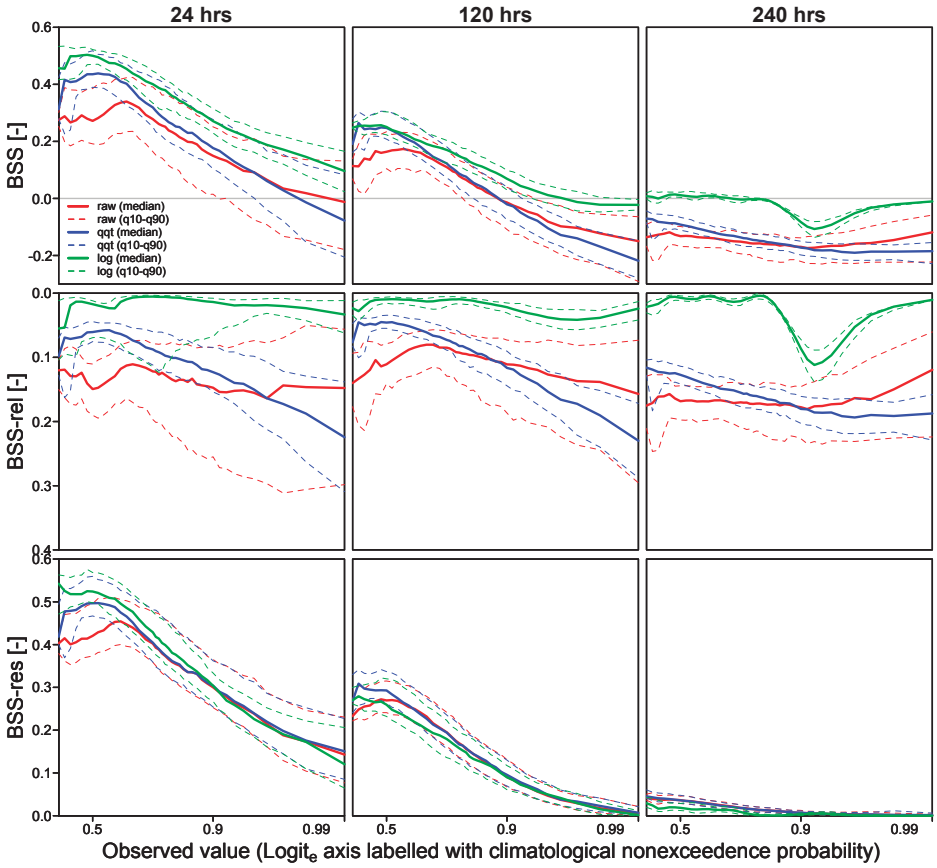


Figure 23: BSS calibration-refinement (Type I) decomposition for ensemble precipitation forecasts at 24-hour, 120-hour and 240-hour ahead forecasts. The baseline is formed by sample climatology. The results pertain to 134 basins: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

ity. However, it is also a practically significant feature of these and other post-processed ensemble forecasts. While the precise description of these conditional biases will depend on the choice of measure (e.g. the RME is sensitive to skewness), the conditional biases are present, regardless of the choice of measure. Figure 21 shows the 120-hour lead time forecast error as a function of the verifying observation for a single basin. Clearly, at higher values of the observation, the ensembles consistently, and increasingly, underestimate the observed value, with insufficient spread to offset this conditional bias.

The CRPSS declines with both lead time and increasing amount of observed precipitation. The BSS and ROCS plots show similar patterns; both metrics are lowest at the tails, indicating that it is relatively dif-

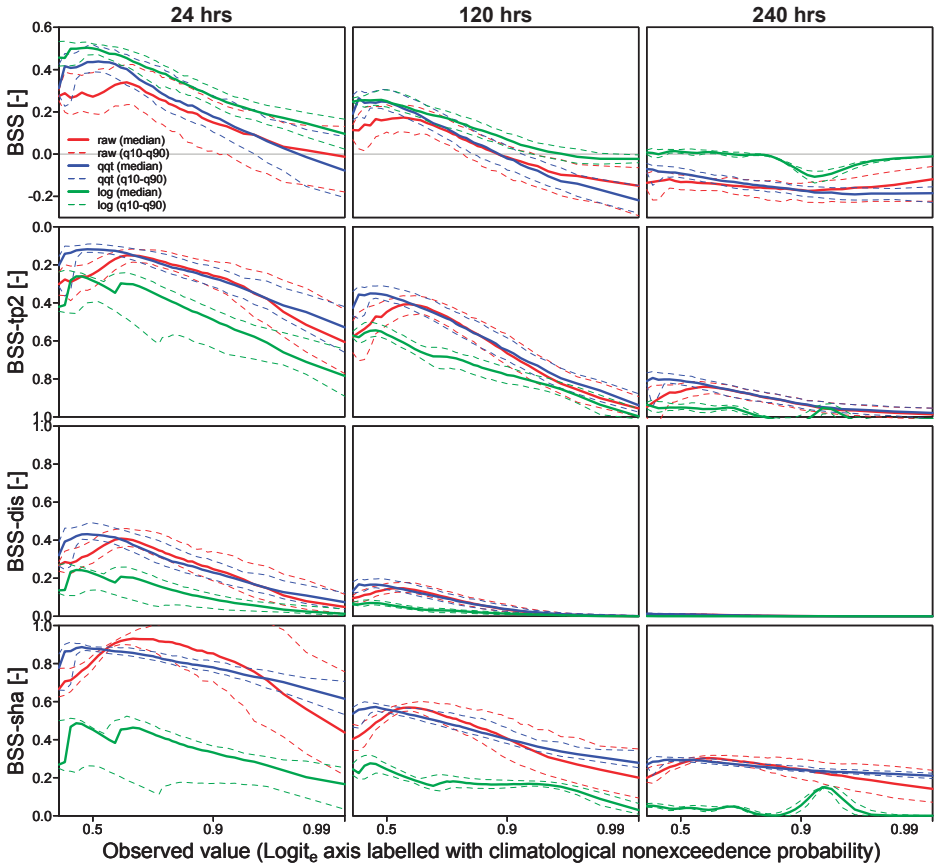


Figure 24: BSS likelihood-base rate (Type II) decomposition for ensemble precipitation forecasts at 24-hour, 120-hour and 240-hour ahead forecasts. The baseline is formed by sample climatology. The results pertain to 134 basins: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

ficult to distinguish between zero and non-zero precipitation and to correctly predict the occurrence of large precipitation amounts.

3.3.2.2 *Post-processed precipitation ensembles*

When moving from raw to post-processed precipitation ensembles, the correlation between the ensemble forecast and the observation is largely conserved. Only in the case of LOG does correlation drop slightly, and only at higher precipitation amounts.

Both the QQT and LOG techniques produce ensemble forecasts that are unconditionally unbiased. However, in all cases, there is an increasingly large conditional negative bias at higher precipitation amounts. At longer lead times, the RME across all cases is very similar. The raw

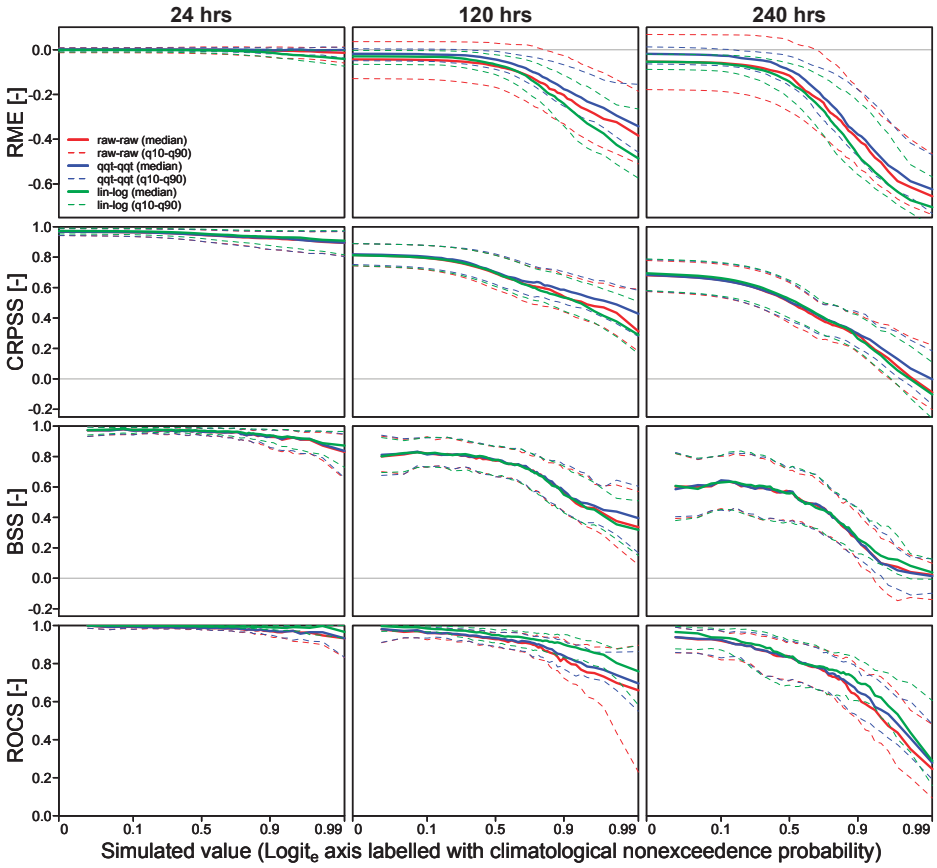


Figure 25: RME, CRPSS, BSS and ROCS for ensemble streamflow forecasts for the headwater basins at 24-hour, 120-hour and 240-hour lead times. For CRPSS, BSS and ROCS, the baseline is formed by sample climatology. The results pertain to 43 locations: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

ensembles initially show a small positive RME, which at some value of P becomes negative and then continues to drop. For non-zero precipitation, LOG shows the highest negative RME at all lead times. From Figure 21, it is clear that the post-processing methods were unable to correct for the Type-II conditional biases at high observed precipitation amounts.

For both techniques, the gain in CRPSS following post-processing is only modest or marginal at all lead times and precipitation amounts. In terms of unconditional CRPSS ($P = 0$), LOG shows the highest increase in skill at all lead times. At higher observed precipitation amounts, LOG does markedly worse than the raw and QQT ensembles due to a large, negative, conditional bias in the ensemble mean. The CRPSS

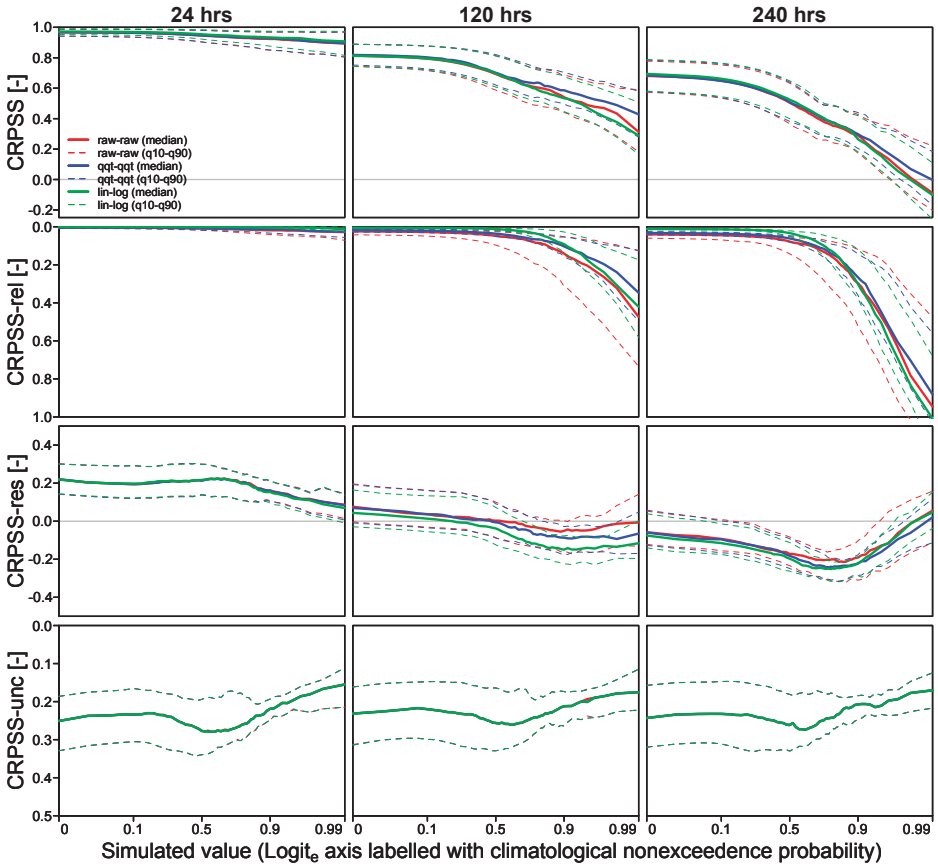


Figure 26: CRPSS calibration-refinement decomposition for ensemble streamflow forecasts for the headwater basins at 24-hour, 120-hour and 240-hour lead times. The baseline is formed by sample climatology. The results pertain to 43 locations: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

of the QQT corrected ensembles are largely similar to that of the raw ensembles. The CRPSS decomposition (Figure 22) and the BSS decomposition (Figure 23) show that none of the post-processing techniques was able to consistently improve both the reliability and resolution of the precipitation ensembles. Rather, there is a trade-off whereby the post-processing generally results in improved reliability at the expense of some loss in resolution. This is different from the post-processed temperature ensembles, which showed improved reliability while consistently maintaining or improving resolution (results not shown). For precipitation, the combination of lower quality of the raw forecasts and a larger number of parameters to estimate for LOG leads to greater sampling uncertainty and weaker performance overall.

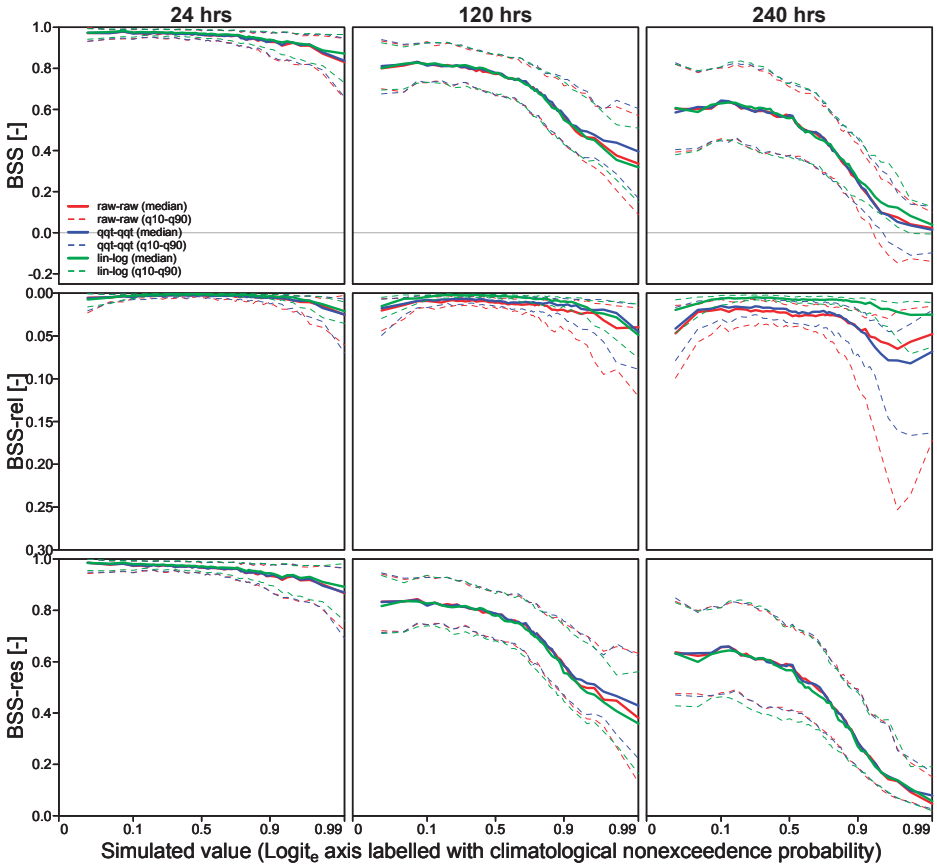


Figure 27: BSS calibration-refinement (Type I) decomposition for ensemble streamflow forecasts for the headwater basins at 24-hour, 120-hour and 240-hour lead times. The baseline is formed by sample climatology. The results pertain to 43 locations: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

In terms of BSS, LOG consistently outperforms the raw and QQT-post-processed precipitation ensembles. As indicated in Figure 24, this is largely explained by an increase in the reliability (or reduction in Type-I conditional bias) of the precipitation ensembles following LOG. However, the RME and the likelihood-base-rate decomposition of the BSS (Figure 25) show a greater tendency of the LOG ensembles to under-forecast high observed precipitation amounts, i.e. they display a larger Type-II conditional bias.

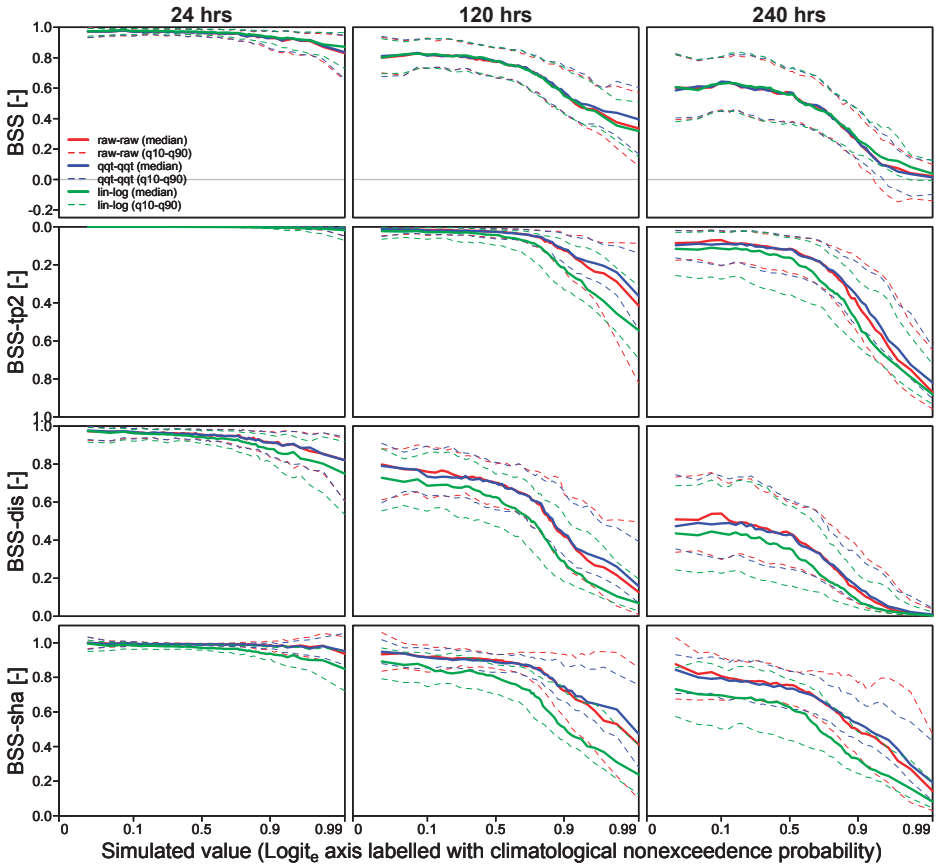


Figure 28: BSS likelihood-base rate (Type II) decomposition for ensemble streamflow forecasts for the headwater basins at 24-hour, 120-hour and 240-hour lead times. The baseline is formed by sample climatology. The results pertain to 43 locations: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

3.3.3 *Streamflow ensemble forecasts*

Verification results for the streamflow ensembles are presented for multiple spatial scales. For the 43 headwater basins, Figure 25 shows RME, CRPSS, BSS and ROCS values. Figures 26 and 27 show calibration-refinement decompositions of the CRPSS and BSS respectively; Figure 28 shows the likelihood-base-rate decomposition of the BSS. The RME, CRPSS, BSS and ROCS values for the Main, Neckar, Moselle and Swiss Rhine tributaries and for the Rhine outlet at Lobith are shown in Figures 29 and 30 respectively.

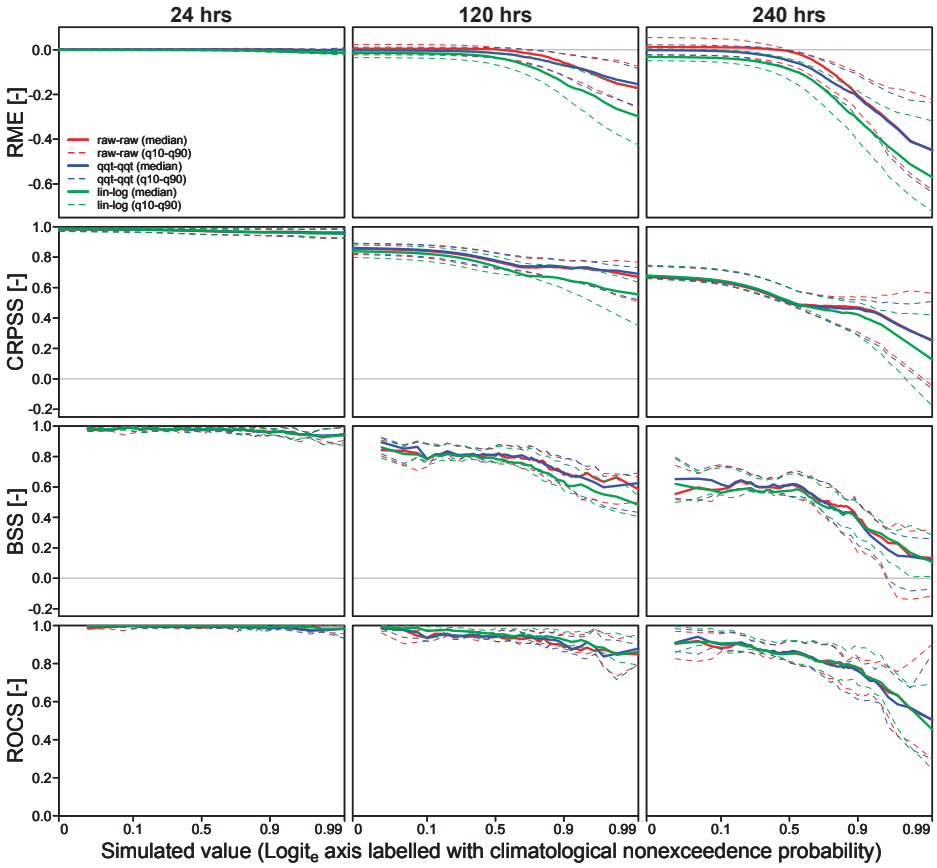


Figure 29: RME, CRPSS, BSS and ROCS for ensemble streamflow forecasts for the four main tributaries at 24-hour, 120-hour and 240-hour lead times. For CRPSS, BSS and ROCS, the baseline is formed by sample climatology. The results pertain to 4 locations: solid lines show the (interpolated) median value; dashed lines show the (interpolated) 0.10 and 0.90 quantiles.

3.3.3.1 *Streamflow ensemble forecasts based on raw forcings*

The ensemble mean of the streamflow forecasts is highly correlated with the simulated streamflow at short lead times. For example, the correlation exceeds 0.98 at $P = 0$ and, at $P = 0.95$, ranges from 0.90 to 0.98 to 0.99 for the smallest to largest spatial scales, respectively. Generally, correlation reduces with decreasing spatial scale: it is lowest for the collection of headwater basins and highest at the outlet, where the aggregate response has a modulating effect on the errors from individual basins. Correlation declines with increasing lead time and with increasing value of the streamflow simulation.

At all spatial scales, the unconditional RME is negligible at the earliest lead times, but increases with increasing lead time. For stream-

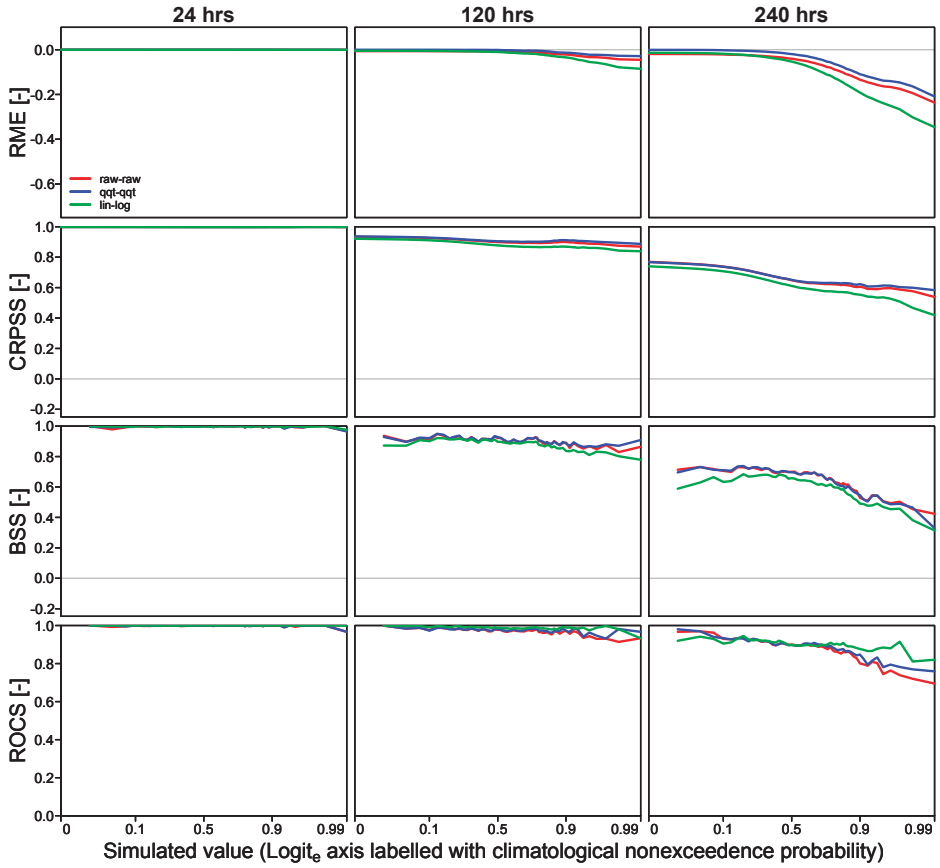


Figure 30: RME, CRPSS, BSS and ROCS for ensemble streamflow forecasts for the outlet at Lobith at 24-hour, 120-hour and 240-hour lead times. For CRPSS, BSS and ROCS, the baseline is formed by sample climatology.

flows larger than the median climatological flow, the forecast ensemble mean increasingly underestimates the simulated streamflow. For example, the RME for the headwater basins (Figure 25) at a lead time of 120 hours shows a median RME of -0.07, -0.20 and -0.27 at $P = 0.5$, $P = 0.9$ and $P = 0.95$ respectively. At the outlet at Lobith (Figure 30) the corresponding values are -0.001, -0.02 and -0.03 respectively.

The patterns in BSS, CRPSS and ROCS are similar to one another and across all spatial scales and lead times. The skill is greatest for the unconditional flows at the shortest lead times and declines with increasing value of the verifying simulation, particularly above the median climatological streamflow where the conditional bias increases. The skill also increases with increasing spatial scale. For example, the

median CRPSS values at $P = 0.90$ at a lead time of 120-hours are 0.54, 0.73 and 0.90 for headwaters, tributaries and outlet respectively.

3.3.3.2 *Streamflow ensembles based on post-processed forcings*

Correlations between the simulated streamflow and the forecast ensemble means are hardly affected by post-processing of the forcings. A slight reduction is observed at longer lead times and at higher quantiles of the simulation for the LIN-LOG case. At $P = 0.90$ and a 240-hour lead time, correlation drops from 0.34 for the raw case to 0.31 for the LIN-LOG case.

Unconditionally, the combinations of QQT-QQT and LIN-LOG result in RME values that are closer to 0 than those of the RAW-RAW case. However, there is a tendency for all techniques to under-forecast the higher simulated streamflows, with the greatest conditional bias for the LIN-LOG case. For example, in the LIN-LOG case at a lead time of 120 hours, the median RME for the four main tributaries increases (negatively) from -0.04 at $P = 0.50$ to -0.15 at $P = 0.90$ and -0.19 at $P = 0.95$ (Figure 29).

In terms of the BSS, CRPSS and ROCS, the streamflow ensembles derived from quantile-to-quantile transformed forcings generally show higher skill than those derived from raw forcings. However, the differences are small. The QQT-QQT ensembles show similar skill at low and moderate values of the verifying simulation only, since QQT is unable to correct for conditional biases, whether Type-I or Type-II in nature.

Generally, post-processing of the forcing variables using conditional techniques (LIN-LOG) does not result in increased skill in terms of CRPSS and BSS. Below the median climatological streamflow, skills are largely similar to those of streamflow ensembles derived from raw forcings. At higher quantiles, there is actually a small decrease of skill. An increase of skill is observed in terms of the reliability component of the BSS and in terms of the ROCS; that is, in the ability of the forecasts to discriminate between the occurrence and non-occurrence of discrete events.

3.4 DISCUSSION

Several questions were posed in this case study: are the raw ECMWF-EPS temperature and precipitation ensembles biased and if so, how? Do these biases translate into streamflow biases and reduced skill? Does post-processing of the temperature and precipitation ensembles improve the quality of the forcing ensembles, and is this improvement noticeable in the streamflow ensembles?

The raw temperature and precipitation ensemble forecasts are biased in both the mean and spread. However, they are more skilful than sample climatology for shorter lead times and moderate thresholds, with reduced skill at longer lead times and for larger amounts (and for zero precipitation). The temperature ensemble forecasts are less biased and more skilful than the precipitation ensemble forecasts. The effects of these biases on the streamflow ensemble forecasts depend on the concentration time of the basins considered with a more rapid deterioration with leadtime in skill for headwaters than for downstream basins. This is largely due to the absence of hydrologic biases and uncertainties in the verification results, of which those in the initial conditions are an important part (i.e. verification was conducted against simulated streamflow). Thus, the skill of streamflow predictions is strongly affected by the initial conditions; this effect lasts longer in larger basins.

Overall, the improvements in the ensemble forcing predictions were modest; this was especially the case for the precipitation ensembles. However, this does not imply that the forcing ensembles are nearly perfect. Rather, it suggests that no additional signal can be found in the forecast-observed residuals to improve forecast quality with the statistical techniques considered. In some cases, post-processing reduces skill; it attempts to use a signal that, in hindsight, turns out to be noise with no predictive information for future forecasts.

Post-processing of the temperature ensembles resulted in greater improvements than post-processing of the precipitation ensembles. This is not surprising, because temperatures are relatively more predictable than precipitation and the gain in skill from post-processing (with a conditional technique) partly depends on the strength of association between the forecasts and observations. Much of the improvement in the precipitation ensemble forecasts is unconditional in nature. Possibly, the improvements from conditional post-processing would be greater when calibrating on a larger dataset, as there were only $\backsim$ 2,900 pairs available in this study, when using a more parsimonious statistical technique (e.g. Wu et al., 2011) or when supplementing the training data set at a particular location with data from other locations with similar climatologies (Hamill et al., 2008).

Application of the forcing post-processors generally results in a reduction in bias and an improvement in skill of the forcing ensembles, although the precise effects depend on forecast lead time, threshold, spatial scale and the types of bias considered. For example, while LOG generally improves the reliability of the precipitation ensembles, the ensemble mean is negatively biased with increasing observed precipitation amount, i.e. a Type-II conditional bias. Post-processing does not improve on all qualities at all lead times and at all levels of the verifying observation. Generally, but not always, post-processing improves

the reliability of the forecasts, but this is sometimes accompanied by a loss of resolution or an increase in the Type-II conditional biases.

Changes in the biases and skill of the forcing ensembles cascade to the ensemble streamflow forecasts. No combination of techniques improves all forecast qualities considered at all lead times and all levels of the verifying simulation. A reduction in the unconditional bias and in the reliability of the ensemble precipitation forecasts is followed by improvements in the reliability of the streamflow ensemble forecasts. However, the trade-off between reliability and resolution is also observed in the streamflow predictions.

The improvements in precipitation and temperature do not translate proportionally into the streamflow forecasts. This may be partly explained by the strong non-linearity of the Rhine basin (due to substantial storage of water in the subsurface, in extensive snowpacks and, to a lesser degree, in reservoirs) and, accordingly, the hydrologic model. Possibly, the effects of post-processing would be stronger in basins where streamflow has a more linear response to forcing variables, e.g. in basins with less storage ,or when leadtimes are sufficiently long to allow for the stored water to reach the streamflow network. This may explain why Yuan and Wood (2012) found that in their seasonal forecasting case, post-processing of forcings leads to a more noticeable improvement of streamflow forecast skill than was found in the case described in the present chapter.

Another potential cause of muted signal resulting from the forcing bias-correction may also be explained by inadequate modelling of the space-time covariability of the forcing forecasts. Forcing verification (as presented here, but more generally) is sensitive to the joint distribution of the forecasts and observations at specific times, locations and for specific variables. In contrast, hydrologic models are sensitive to the space-time covariability of the forcing forecasts. In this context, the use of the Schaake shuffle to recover some of this space-time covariability may be limiting. The Schaake shuffle introduces rank association only, and it introduces this only insofar as these patterns appear historically on the same or nearby dates. For example, it cannot account for more complex statistical dependencies, novel structures, or structures that are conditional upon the state of the atmosphere at the forecast issue time. These weaknesses are likely exaggerated when the forecasts have greater spread because the Schaake shuffle has greater scope to affect the space-time patterns of the ensemble traces. In order to account for more complex structures, post-processors with explicit models of space-time covariability are needed, such as geostatistical models (Kleiber et al., 2011), together with parsimonious verification techniques that are sensitive to these space-time and cross-variable relationships.

Verification against simulated streamflows allows for the hydrologic biases to be factored out of the streamflow skill associated with forcing post-processing. However, it also magnifies the resulting streamflow ensemble skill. When verifying against observations, the overall biases and uncertainty will be larger due to inclusion of the hydrologic biases and uncertainties, including those in the streamflow observations. Relatively speaking, the change in skill due to the post-processed forcings will be more difficult to detect.

The research questions posed in the introduction were addressed by looking at a selection of verification metrics. While reasonably broad, the results may be sensitive to the choice of metric. In addition, the parameters of each post-processing techniques are estimated with a particular objective function. If these objective functions are similar to the verification metrics used, it should not be surprising that a particular technique scores well in terms of that metric.

The available reforecast dataset allowed for testing our hypothesis using a reasonable number of retrospective forecasts (just over 3,100). Conditional verification however, especially of extreme events, quickly reduces the size of the subsample. In this study, the cut-off of the climatological nonexceedence probability was chosen at $P = 0.99$, which is where 1% of the available data is used for verification. This coincides with approx. 30 data pairs. If the present study would be repeated and extended by stratifications, for example on a two season basis, then the $P = 0.99$ quantile would equate to approx. 15 verification pairs, which is deemed too small for verification purposes. Conversely, if, in case of stratification, the minimum number of pairs would be kept fixed at 30, this would mean that less extreme events can be analysed only. Ideally, longer sets of reforecasts (hindcasts) would be available. Note that by the time the present chapter was submitted for publication, the ECMWF reforecast set had been extended considerably. Even so, the authors support the call for reforecast datasets, eloquently voiced by Hamill et al. (2006).

In the current study, the improvements to streamflow accrued by post-processing of the forcing predictions were modest. Moreover, these effects may be negligible when verifying against streamflow observations. Since forcing post-processing is both labour intensive and inherently difficult for precipitation, particularly in accounting for appropriate space-time covariability, it is worth considering other methods to improve the skill of the forcing and streamflow ensembles, such as multi-model combinations, data assimilation (to improve the hydrologic initial conditions), and streamflow post-processing. For example, under conditions where forcing post-processing contributes significant skill to streamflow, it needs to be established whether that skill remains after streamflow post-processing or whether statistical post-processing

can adequately remove the forcing biases via the streamflow, despite the aggregation of multiple sources of bias and uncertainty.

3.5 SUMMARY AND CONCLUSIONS

Ensemble forecasts of temperature and precipitation were tested for biases and an attempt was made to reduce these biases through statistical post-processing. This resulted in modest improvements in the quality of the forcing ensembles. The effects on streamflow were explored by factoring out the effects of bias in the hydrologic model; that is, by verifying against simulated streamflow. In general, the improvements in streamflow quality were muted at all spatial scales considered, with explanations including a limited model of the space-time covariability of the forcing ensembles.

ACKNOWLEDGEMENTS

The authors would like to thank Florian Pappenberger at ECMWF for providing reforecast data. We are also grateful to Tom Hamill, Tom Pagano and an anonymous reviewer for taking the time to review an earlier version of this chapter and for providing constructive comments. Any errors remain ours. Publication of this chapter was partially funded by Deltares research funds. Several open source and freely (as in: without charge) available software and data sources were used for the analyses described in this chapter: Quantum GIS (Quantum GIS Development Team, 2012), R (R Core Team, 2013), the Ensemble Verification System (Brown et al., 2010), the Delft-FEWS forecast production system (Werner et al., 2013), the SRTM digital elevation model (Jarvis et al., 2008), CHR08 precipitation data (Photiadou et al., 2011) and E-OBS temperature data (Haylock et al., 2008).

4

ALTERNATIVE CONFIGURATIONS OF QUANTILE REGRESSION FOR ESTIMATING PREDICTIVE UNCERTAINTY IN WATER LEVEL FORECASTS FOR THE UPPER SEVERN RIVER: A COMPARISON.

ABSTRACT

The present chapter reports an inter-comparison study of different configurations of a statistical post-processor that is used to estimate predictive hydrological uncertainty. It builds on earlier work by Weerts, Winsemius and Verkade (2011; herinafter referred to as WWV2011), who used the Quantile Regression technique to estimate predictive hydrological uncertainty using a deterministic water level forecast as a predictor. The various configurations are designed to address two issues with the WWV2011 implementation: (i) quantile crossing, which causes non-strictly rising cumulative predictive distributions, and (ii) the use of linear quantile models to describe joint distributions that may not be strictly linear. Thus, four configurations were built: (i) 'classical' Quantile Regression, (ii) a configuration that implements a non-crossing quantile technique, (iii) a configuration where quantile models are built in Normal space after application of the Normal Quantile Transform (similar to the implementation used by WWV2011), and (iv) a configuration that builds quantile model separately on separate domains of the predictor. Using each, four re-forecasting series of water levels at fourteen stations in the Upper Severn river were established. The quality of these four series was inter-compared using a set of graphical and numerical verification metrics. Intercomparison showed that reliability and sharpness vary across configurations, but in none of the configurations do these two forecast quality aspects improve simultaneously. Further analysis shows that skills in terms of Brier Skill Score, mean Continuous Ranked Probability Skill Score and Relative Operating Characteristic Score is very similar across the four configurations.

This chapter has been published as López López, P., Verkade, J.S., Weerts, A.H., Solomatine, D.P., 2014. Alternative configurations of Quantile Regression for estimating predictive uncertainty in water level forecasts for the Upper Severn River: a comparison. Hydrology and Earth System Sciences 18, 3411–3428. DOI: 10.5194/HESS-18-3411-2014

4.1 INTRODUCTION

Forecasting may reduce but can never fully eliminate uncertainty about the future. Hydrological forecasts will always be subject to many sources of uncertainty, including those originating in the meteorological forecasts used as inputs to hydrological models (e.g. precipitation and temperature), and in the hydrological models themselves (e.g. model structure, model parameters and human influences). Informed decision making may benefit from estimating the remaining uncertainties. A number of research studies suggest that enclosing predictive uncertainty estimates indeed leads to benefits to end users (Krzysztofowicz, 2001; Collier et al., 2005; Verkade and Werner, 2011; Ramos et al., 2012; Dale et al., 2014).

In the literature, various approaches to estimate predictive uncertainty have been presented. One of those is the use of meteorological ensemble forecasts, where initial atmospheric conditions are perturbed to yield an ensemble of atmospheric forecasts. These can be routed through a hydrological model, thus yielding an ensemble of hydrologic model forecasts which provides insight into the sensitivity of hydrological model results to various possible weather scenarios. Increasingly, hydrologic forecasting systems are including these ensemble predictions in the forecasting routines to capture the meteorological uncertainty. An overview of applications and best practices was given by Cloke and Pappenberger (2009). More recent applications include the Environment Agency's National Flood Forecasting System – NFFS – (Schellekens et al., 2011) and the US National Weather Service's Hydrologic Ensemble Forecast Service HEFS (Demargne et al., 2013). Note that HEFS also includes a statistical post-processor developed by (Seo et al., 2006).

A second approach is statistical post-processing. Estimating predictive uncertainty through statistical post-processing techniques comprises an analysis of past, 'observed' predictive uncertainty to build a model of future predictive uncertainties. It can be used as either an alternative or additional step to hydrological ensemble forecasting. In many hydrological forecasting applications, postprocessing is used in combination with deterministic forecasts (but it can also be applied to ensemble hydrologic forecasts if available; see, for example, Reggiani et al. 2009 and Verkade et al. 2013b). A historical record of past forecasts and their verifying observations is then used to build a model of forecast error. (Note that other configurations are possible, but this one is the most straightforward and common one.) On the assumption that this error will be similar in future cases, the error model is then applied to newly produced deterministic forecasts, thus producing an estimate of predictive hydrological uncertainty. This estimate then includes uncertainties originating in both the atmospheric forecasts as

well as those in the numerical simulation of streamflow generation and routing processes. Post-processing assumes a stationary relation between the predictors and the predictors. It follows that both the forecasts and the observations used for calibration have to be stationary. Also, ideally the calibration record is sufficiently long as to include events that are (relatively) extreme. The reason for this is that the relationship between forecast and observations at extreme events may be different from the relationship in non-extreme hydrological regimes. If the assumption of stationarity cannot be met or if the calibration record is short, the quality of the post-processed forecasts is likely to be reduced. Several hydrologic post-processors have been described in the scientific literature, including the Hydrological Uncertainty Processor (HUP - Krzysztofowicz and Kelly 2000), Bayesian Model Averaging (BMA - Raftery et al. 2005), the Model Conditional Processor (MCP - Todini 2008), UNcertainty Estimation based on local Errors and Clustering (UNEEC - Solomatine and Shrestha 2009), the Hydrologic Model Output Statistics (HMOS - Regonda et al. 2013) and Quantile Regression (QR - Weerts et al. 2011). The present chapter focuses on the latter technique.

Quantile Regression (QR, Koenker and Bassett Jr 1978; Koenker and Hallock 2001; Koenker 2005) aims to describe a full probability distribution of the variable of interest (the predictand), conditional on one or more predictors. Contrary to some of the other post-processors (such as HUP or BMA), QR requires few prior assumptions about the characterization of the model error. While it was originally developed for applications in the economic sciences, it has since been introduced into environmental modelling and climate change impact assessment (e.g. Bremnes, 2004; Nielsen et al., 2006). The technique has been applied in various research studies as a post-processing technique to estimate predictive hydrological uncertainty, including those described by Solomatine and Shrestha (2009), Weerts et al. (2011), Verkade and Werner (2011), and Roscoe et al. (2012). In each of these applications, the quantiles of distribution of the model error are estimated using single valued water level or discharge forecasts as predictors.

Weerts et al. (2011; hereinafter referred to as wwv2011) describe an implementation of QR for the Environment Agency in the United Kingdom. The "Historic Forecast Performance Tool" (HFPT; Sene et al. 2009) makes use of QR to estimate a predictive distribution of future water levels using the deterministic water level forecast as a predictor. The wwv2011 configuration of QR includes a transformation into Gaussian space using the Normal Quantile Transform (Krzysztofowicz and Kelly, 2000; Montanari and Brath, 2004; Bogner and Pappenberger, 2011). In QR, the quantiles are estimated one at a time. Potentially, these quantiles cross, thus yielding implausible predictive distributions. The quantile crossing problem was addressed by omitting the

domain of the predictor where the crossing occurred from the QR procedure and instead, in that domain, imposing a prior assumed distribution of the predictand. The present chapter's study basin was used as a study basin in WWV2011 also.

The results of the WWV2011 analysis were verified for reliability and showed to be satisfactory. However, this verification was unconditional in the sense that only the full available sample of paired (probabilistic) forecasts and observation was assessed for reliability. When the HFPT was further tested (Vaughan, 2012), it was noticed that the probabilistic forecasts did not perform equally well in high flow conditions. One of the contributions of the present chapter consists of a conditional analysis of forecast skill. Forecast skill is assessed for progressively higher flood levels, in terms of commonly used verification metrics and skill scores. These include Brier's probability Score, the Continuous Ranked Probability Score and corresponding skill scores as well as the Relative Operating Characteristic Score.

The configuration of QR in WWV2011 included two elements that, in the present chapter, are explored in additional detail. These steps are (i) the technique for avoiding crossing quantiles and (ii) the derivation of regression quantiles in Normal space using the Normal Quantile Transform (NQT).

In WWV2011, crossing quantiles were avoided by manually imposing a distribution of the predictand in the domain of the predictor where crossing occurred. Since designing and implementing that particular configuration, an alternative technique for estimating non - crossing quantile regression curves has emerged (Bondell et al., 2010b). As the latter technique requires less manual interference by the modelers, the present chapter investigates whether implementation thereof yields estimates of predictive uncertainty that are of equal or higher quality.

In WWV2011, QR was applied using first degree polynomials, i.e. describing the distribution of the predictand as a linear function of the predictor. This, of course, assumes that the joint distribution of predictor and predictand can be described in linear fashion. To facilitate this, both marginal distributions (of forecasts and of observations) were transformed into Normal or Gaussian domain using the NQT. The joint distribution was subsequently described in Normal space using linear regression quantiles, and then back-transformed into original space. The resulting regression quantiles are then no longer linear. While this procedure yielded satisfactory results, there is no requirement on the part of QR of either the marginal or joint distributions to be marginally or jointly Normal distributed. Also, the transformation and especially the back-transformation impose additional assumptions on the marginal distributions and can thus be problematic. Hence a justified question is whether this transformation to and from Normal space actually yields better results. In the present chapter, this is tested

by comparing multiple configurations of QR: derivation of regression quantiles in original space and in Normal space. As an additional step, a piecewise linear configuration is tested, where the domain of the predictor is split into several, mutually exclusive and collectively exhaustive domains, on each of which the regression quantiles are calibrated.

The objective of this work is to thoroughly verify uncertainty estimates using the implementation of QR that was used by WWV2011, and to inter-compare forecast quality and skill in various, differing configurations of QR. The configurations are (i) 'classical' QR, (ii) QR constrained by a requirement that quantiles do not cross, (iii) QR derived on time series that have been transformed into the Normal domain (similar to WWV2011 QR configurations, with the exception of how the quantile crossing problem is addressed), and (iv) a piecewise linear derivation of QR models. A priori, it is expected that imposing a non-crossing requirement yields results that are at least as good as those of the 'classical' implementation of QR, and that derivation in Normal space and piecewise linear derivation each constitute a further improvement in quality and skill compared to derivation in original space.

The novel aspects and new contributions of this work include the thorough verification of an earlier implementation of QR, the application of the non-crossing QR to this particular case study and the exploration of techniques for ensuring that joint distributions can be described using linear QR models.

This chapter first describes the approach, materials and methods, including the study basin, the hindcasting process, the analysed QR configurations and the verification process. Subsequently, results and analysis are presented. The chapter ends with conclusions and discussion.

4.2 APPROACH, MATERIALS AND METHODS

The present study consists of an experiment in which verification results of four differently configured post-processors (each based on the Quantile Regression technique) are inter–compared. By the varying configurations, two potential issues are addressed: quantile crossing and possible non-linearity of the joint distribution of predictor and predictand.

4.2.1 *Study basin: Upper Severn River*

The Upper Severn basin (Figure 31) serves as the study basin for the present study; it was also one of the study basins in WWV2011. Its predominantly hilly catchment extends from the Welsh Hills at Plynlimon to the gauge at Welshbridge in Shrewsbury and is approximately 2,284

Figure 31: The Upper Severn basin including the 14 forecasting locations used in the present study. Note that the smallest river streams are not shown in the stream network. (The digital elevation model is made available by the European Environment Agency on a Creative Commons Attribution License; http://www.eea.europa.eu/data-and-maps/data/digital-elevation-model-of-europe)

km^2 large. Lake Vyrnwy (Vyrnwy River) and Llyn Clywedog (Clywedog River) are two reservoirs located in the headwaters of the catchment. The Upper Severn includes rock formations classified as non-aquifers as well as loamy soils characterised by their high water retention capacity. Annual precipitation varies with topography from 700 to 2,500 mm (EA, 2009). Flooding occurs relatively frequently, with major floods occurring in autumn 2000, February 2002, February 2004, summer 2007, fall 2012 as well as at the time of writing this chapter, early 2014. To manage flood risk, the UK Environment Agency developed the River Severn Catchment Flood Management Plan in 2009. Flood risk management is supported by the Midlands Flood Forecasting System (MFFS), which is based on the Delft-FEWS forecast production system (Werner et al., 2013). The Upper Severn configuration in MFFS consists of a sequence of numerical models for modeling of

Table 7: Hydrometeorological and topographical information of analysed catchments at Upper Severn River (adapted from EA 2013 and Marsh and Hannaford 2008).

Station name	River	Basin area	Elevation	Mean annual rainfall	Mean flow	Highest river level recorded	Basin lag time
		$[km^2]$	[m AOD]	[mm]	$[m^3/s]$	[m]	[h]
Caersws	Severn	205	119	-	-	3.69	8 - 10
Abermule	Severn	580	83	1291	14.58	5.26	13 - 17
Buttington	Severn	653	62	-	-	5.5	8 - 10
Montford	Severn	1784	52	1184	43.3	6.96	10 - 15
Welshbridge	Severn	2025	47	-	-	5.25	15 - 20
Vyrnwy Weir	Vyrnwy	94	226.34	1951	4.24	1.8	2 - 5
Pont Robert	Vyrnwy	417	100	-	-	3.07	5 - 9
Meifod	Vyrnwy	675	81	-	-	3.67	7 - 10
Llanymynech	Vyrnwy	778	62	1358	21.08	5.19	3 - 6
Bryntail	Clywedog	49	212.05	2026	2.4	1.61	2 - 4
Rhos Y Pentref	Dulas	53	178.49	1313	1.45	2.42	1 - 3
Llanerfyl	Banwy	124	151	-	-	3.5	3 - 5
Llanyblodwel	Tanat	229	77.28	1267	6.58	2.68	7 - 10
Yeaton	Perry	109	61.18	767	1.6	1.13	15 - 20

rainfall-runoff (MCRM; Bailey and Dobson 1981), hydrological routing (DODO; Wallingford 1994) and hydrodynamical routing (ISIS; Wallingford 1997) processes as well as an internal MCRM error correction procedure based on the Autoregressive Moving Average (ARMA) technique. The input data for MFFS consists of Real Time Spatial (RTS) data (observed water level data, rain gauge data, air temperature and catchment average rainfall data), Radar Actuals, Radar Forecast, and Numerical Weather Prediction data. This input data is provided by the UK Meteorological Office.

The uncertainty models are used to estimate predictive uncertainty at fourteen hydrological stations on the Upper Severn River, each having different catchments characteristics. Figure 31 shows a map with the forecasting locations and their basins. Table 7 summarizes some key hydrological data.

4.2.2 *Hindcasting process*

The uncertainty models (Section 4.2.3) are derived using a joint historical record of observations and forecasts. The latter is acquired through the process of reforecasting or hindcasting. For this, a standalone version of the forecast production system MFFS is used. Prior to every forecast, the models are run in historical mode over the previous period to produce an estimate of internal states (groundwater level, soil moisture deficit, snow water equivalent, snow density, etc). In this historical mode, models are forced with observed precipitation, evapotranspiration and temperature. The system is subsequently run in forecast mode twice daily, with forecast issue times of 08:00 and 20:00 UTC, with a maximum lead time of 48 hours. The selected reforecasting period is from January 1st, 2006 through March 7th, 2013. Of this period, the period up to March 6th, 2007 is used to 'spin up' the models. The remaining six years are used for the calibration and validation of the uncertainty models.

4.2.3 *Uncertainty models*

In the present study, predictive uncertainty is modelled using Quantile Regression (QR). The basic configuration is simple, and identical across all cases: the predictive distribution of future observed water levels is modeled as a series of quantiles, each estimated as a linear function of a single predictor which is the deterministic water level forecast. Four different configurations are inter-compared. Configuration Zero (QR0) constitutes the most straightforward case, where QR is applied 'as is', i.e. in its most basic form, in which no attempt is made to avoid crossing quantiles and no transformation or piecewise derivation is applied. Configuration One (QR1) addresses the problem of the crossing quantiles using the technique proposed by Bondell et al. (2010b). If quantile crossing problem does not occur, this technique provides the same estimates as in the base scenario. Because of this, it is also applied to the remaining configurations. In some cases, the joint distribution of forecasts and observations is not best modelled using linear quantile regression models across the full domain of the predictor. However, by applying a transformation or by modelling sub-domains of the predictor, linear models may be used nonetheless. This is what is done in Configurations Two (QR2) and Three (QR3) respectively. The configurations are each described in detail in the following four subsections; for reference, they are also listed in Table 8. As the non-crossing quantiles are applied to QR1, QR2 and QR3, the comparison in the present chapter is effectively between these three latter configurations.

The joint distribution of forecasts and their verifying observations is based on the UK Environment Agency archives of water level ob-

Table 8: QR Configurations used in the present study

Identifier	Description
QR0	Classical Quantile Regression - Base scenario
QR1	Quantile Regression constrained by a non–crossing quantiles restriction
QR2	Quantile Regression, constrained by a non–crossing quantiles restriction, on the transformed data into Normal domain through Normal Quantile Transformation (NQT)
QR3	Piecewise linear derivation of Quantile Regression, constrained by a non—crossing quantiles restriction

servations and on the forecasts from the hindcasting procedure. The available record is cross-validated through a leave-one-year-out cross-validation analysis. From the six years' worth of forecasts that are available for calibration and validation, five are used for model calibration and the single remaining year is used for model validation. Subsequently, another year is chosen for validation and the calibration period then comprises the remaining five years. This is repeated until all six years have been used for validation.

Uncertainty models are developed for each combination of lead time and location separately. While the forecasts have a maximum lead time of 48 hours with one-hour intervals, the QR models are derived on a limited number of lead times, namely for 1 hour lead time and then 3 through 48 hours lead time with 3-hour increments. The leave-one-year-out cross validation procedure yields approximately 3,760 observation-forecast pairs for every combination of lead time and location.

4.2.3.1 *QR0: Quantile Regression*

Quantile Regression (QR; Koenker and Bassett Jr, 1978; Koenker and Hallock, 2001; Koenker, 2005). QR is a regression technique for estimating the conditional quantiles of a multivariate distribution. The technique is described in detail in Appendix A. Figures 32, 33 and 34 give a graphical overview of the resulting quantiles. These plots are discussed in the Results and Analysis section.

4.2.3.2 *QR1: Non-Crossing Quantile Regression*

A potential problem with using QR for derivation of multiple conditional quantiles is that quantiles may cross, yielding predictive distributions that are not, as a function of increasing quantiles, monotonously increasing. WWV2011 have addressed this issue by assuming a fixed

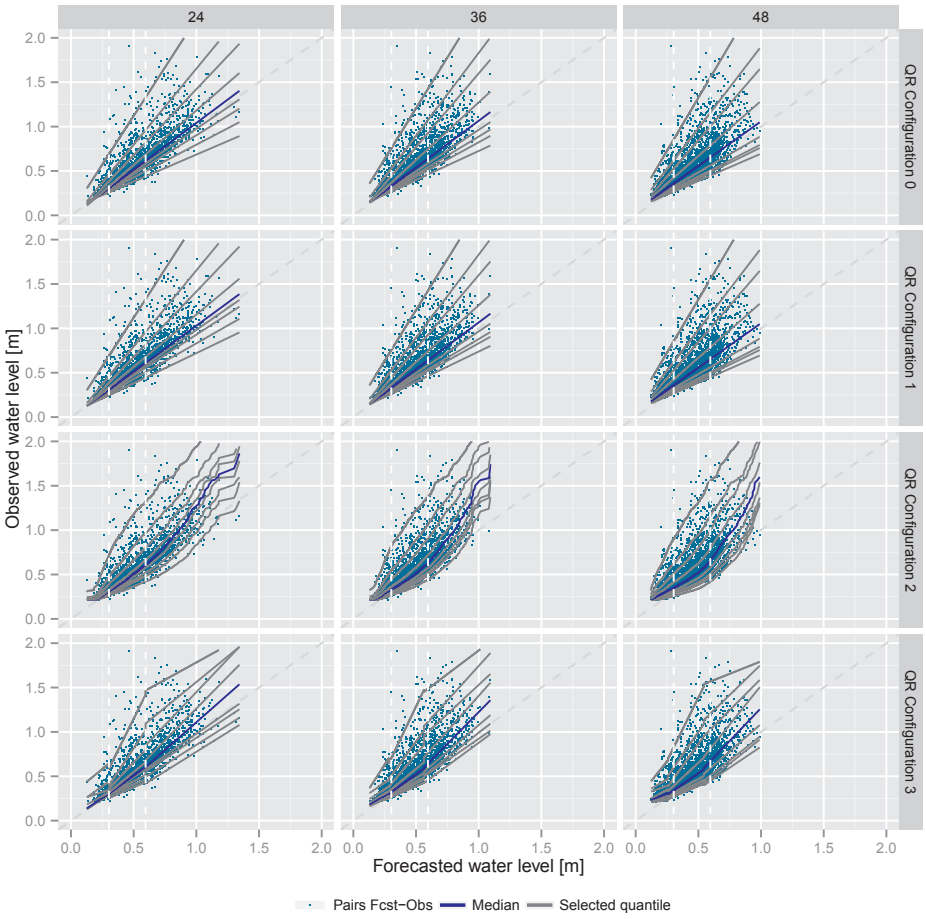


Figure 32: Quantile Regression models for Llanyblodwel. Rows show the four different configurations; columns show different lead times.

error model in the domain of the predictor where there is the danger of quantiles crossing. In the present research study, the technique proposed by (Bondell et al., 2010b) is used. This technique imposes a non-crossing restriction on the solution of Equation A.7. Without this restriction, the solution to the proposed optimization problem is identical to that of classical quantile regression, i.e. to the models derived using QR0. For a more detailed description of the non-crossing quantiles technique, the reader is referred to (Bondell et al., 2010b). The technique is freely available online (Bondell et al., 2010a) and is coded in the statistical computing language R (R Core Team, 2013).

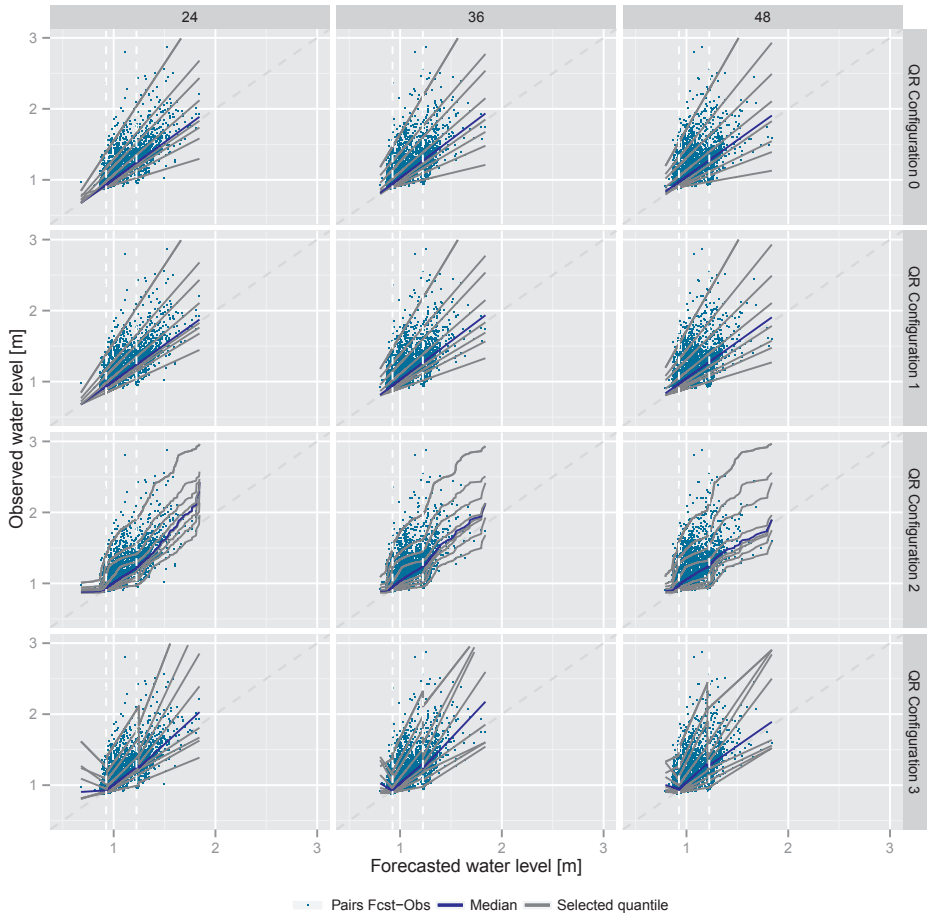


Figure 33: Quantile Regression models for Pont Robert. Rows show the four different configurations; columns show different lead times.

4.2.3.3 *QR2: Quantile Regression in Normal space*

In this configuration, timeseries of water level observations and water level forecasts are first transformed into the Normal domain. This results in timeseries that are marginally Normal distributed. Subsequently, Quantile Regression models are calibrated using the non-crossing quantiles technique. Posterior to the derivation of QR models, the variables are back-transformed into original space. The rationale for using the transformation is that the joint distribution of transformed timeseries appears to be more linear, and can thus be better described by linear conditional quantiles.

The Normal Quantile Transformation (NQT) is a quantile mapping or cdf-matching technique that matches the (empirical or modeled) cdf of the marginal distributions with a standard normal cdf. Here, the

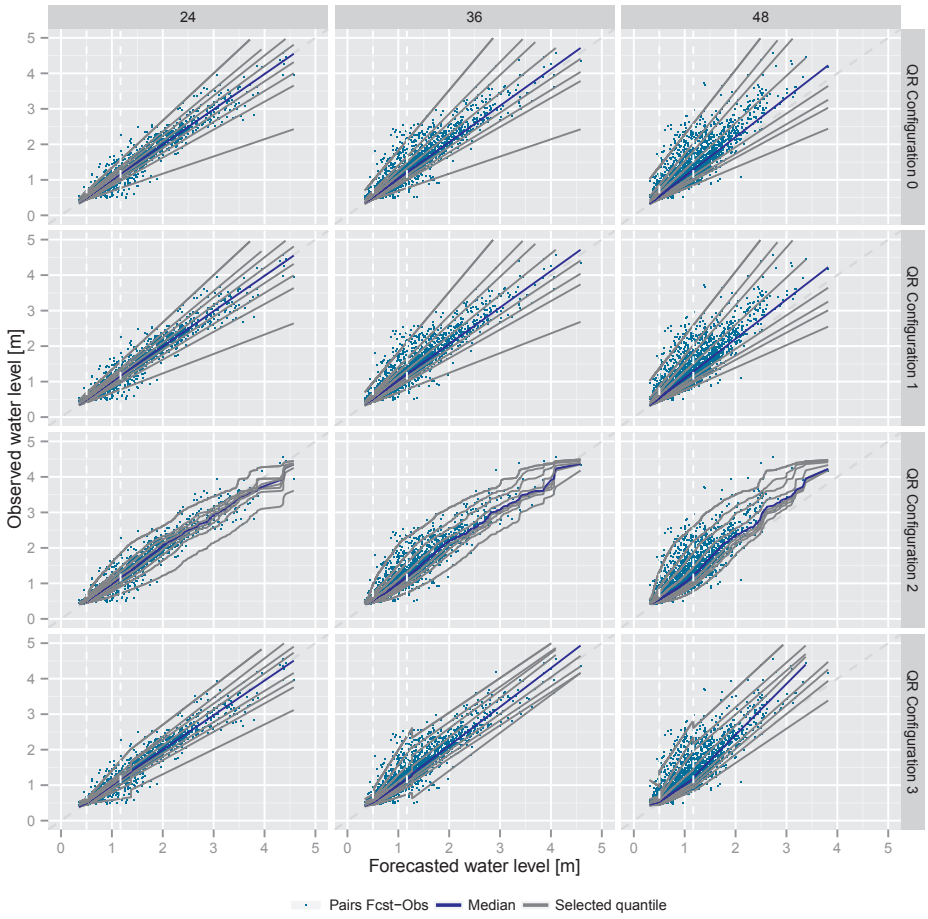


Figure 34: Quantile Regression models for Welshbridge. Rows show the four different configurations; columns show different lead times.

empirical cdf of the marginal distributions is used. Thus, the variables are mapped to a standard normal distribution,

$$\begin{aligned} Y_{nqt} &= Q^{-1}(F(Y)) \\ X_{nqt,n} &= Q^{-1}(F(X_n)) \end{aligned} \tag{12}$$

where $F(\cdot)$ is the Weibull plotting position of the data point considered. The equivalent of Equation A.6 then becomes

$$Y_{nqt,n,\tau} = a_{n,\tau} X_{nqt,n} + b_{n,\tau} \tag{13}$$

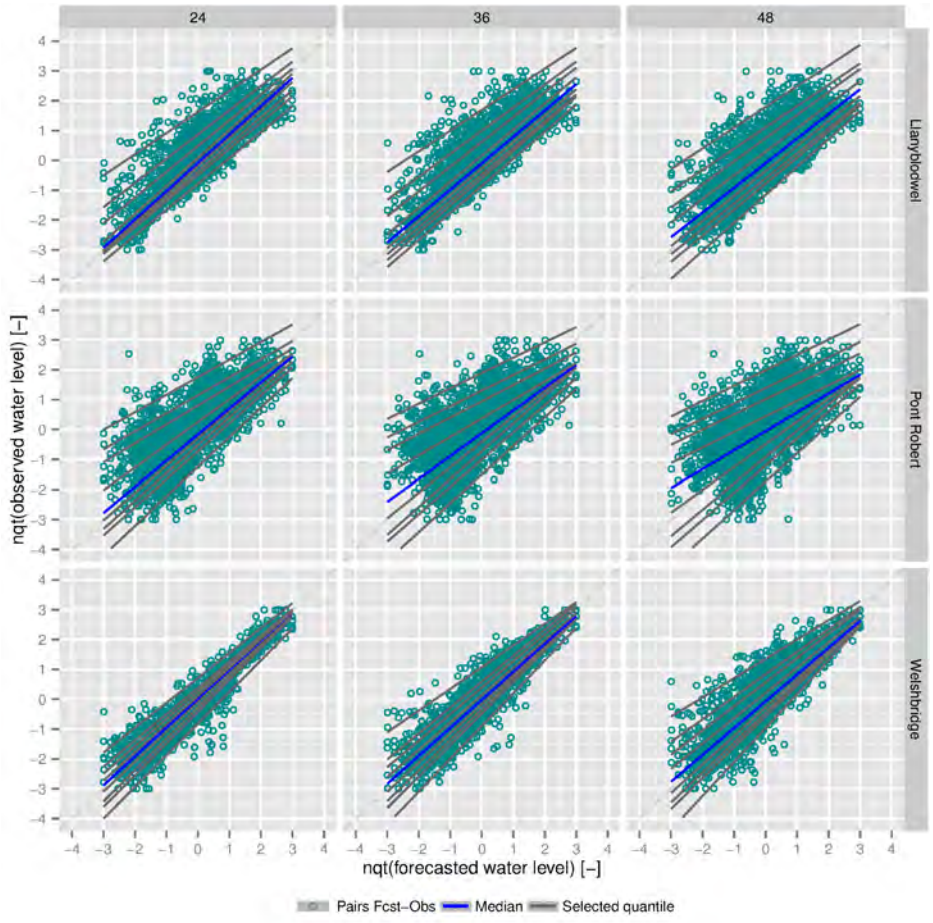


Figure 35: Quantile Regression models for Llanyblodwel, Pont Robert and Welshbridge in Normal space (QR2). Rows show the three different locations; columns show different lead times.

which is solved by minimising the sum of residuals,

$$\min \sum_{j=1}^{J} \rho_{n,\tau} \left(y_{nqt,n,j} - \left(a_{n,\tau} x_{nqt,n,j} + b_{n,\tau} \right) \right) \tag{14}$$

Posterior to the analysis in normal space, the variables are back-transformed to original space using a reversed procedure,

$$\begin{aligned} Y &= Q(F(Y_{nqt})) \\ X_n &= Q(F(X_{nqt,n})) \end{aligned} \tag{15}$$

Back-transformation is problematic if the quantiles of interest lie outside of the range of the empirical distribution of the untransformed

variable in original space. In those cases, assumptions will have to be made on the shape of the tails of the distribution (see Bogner et al. 2012 for a more elaborate discussion). Some authors have chosen to parameterize the distribution of the untransformed variable and use those statistical models for the back-transformation (see, for example, Krzysztofowicz and Kelly 2000). In the present study, this matter is treated through a linear extrapolation on a number of points in the tails of the distribution which was the solution chosen by Montanari (2005) and by WWV2011.

4.2.3.4 *QR3: Piecewise Linear Quantile Regression*

In an effort to try and use linear quantile models to describe a joint distribution that may be slightly nonlinear in nature, Van Steenbergen et al. (2012) applied linear models to partial domains of the predictor. They found the resulting distributions to be both more reliable and sharper than those attained by application of a single linear model to the full domain of the predictor. Multiple, mutually exclusive and collectively exhaustive domains were identified based on a visual inspection of the data and taking into account the requirement that each sub group will have to contain a sufficiently sized data sample. As this selection more or less coincided with two splits at the 20^{th} and 80^{th} percentile, thus three sub-domains were defined, comprising 20%, 60% and 20% of the data respectively.

4.2.4 *Verification strategy*

To understand and inter-compare the performance of different QR configurations, an extensive verification of forecast quality was carried out. The post-processing procedure separated calibration from validation hence the verification can be considered to be independent. The old(-ish) adage has it that probabilistic forecasts should strive for sharpness subject to reliability (Gneiting et al., 2005): an improvement in sharpness at the expense of reliability is not desirable. In addition, decision makers may be interested in event discrimination skill for specific flood thresholds, for example. Forecasts were therefore assessed for reliability, sharpness and event discrimination, and a number of metrics were calculated.

These verification metrics include the Brier Score (BS), the mean Continuous Ranked Probability Score (CRPS) and area beneath the Relative Operating Characteristic (ROCA). Reliability was assessed using reliability diagrams, that plot the relative frequency of occurrence of an event versus the predicted probability of event occurrence. Proximity to the 1:1 diagonal, where observed frequency equals predicted probability, indicates higher reliability. Sharpness was explored by determin-

ing the width of the centred 80% interval of the predictive distributions; the full sample of these widths is shown by means of an empirical cumulative distribution function (ecdf). The Brier Score (Brier, 1950) is defined as the mean squared error of a probabilistic forecast of a binary event. The mean CRPS (Brown, 1974; Matheson and Winkler, 1976) is a measure of the squared probabilistic error in the forecasts across all possible discrete events. The area beneath the Relative Operating Characteristic is a measure of the forecasts' ability to discriminate between the exceedence and non-exceedence of a threshold, for example a flood threshold. A detailed description of these measures with their mathematical formulation can be found in Appendix B.

To allow for comparison across different locations, BS, CRPS and ROCA are expressed as skill, thus becoming Brier Skill Score (BSS), Continuous Ranked Probability Skill Score (CRPSS) and the Relative Operating Characteristic Score (ROCS),

$$\text{skill} = \frac{\text{score} - \text{score}_{\text{ref}}}{\text{score}_{\text{perfect}} - \text{score}_{\text{ref}}} \tag{16}$$

where score is the score of the system considered, $\text{score}_{\text{ref}}$ is the score of a reference system and $\text{score}_{\text{perfect}}$ is the highest possible score. Skill scores range from $-\infty$ to 1. The highest possible value is 1. If skill $= 0$, the system's score is as good as that of the reference system. If skill < 0 then the system's score is less than that of the reference system. In the case of BSS and CRPSS, the reference score comprises that of the sample unconditional climatology; in case of the ROCS, the reference score is the ROCA associated with an unskilled forecast which states that the probability of event occurrence is equal to the probability of non-event occurrence.

As the post-processor is intended to be used in flood forecasting, forecast skill is not only assessed for the full available sample of forecast, observation pairs, but also for subsets of high and extreme events. These subsets are defined by the climatological probability of non-exceedence P of the observation. For example, P = .95 denotes the sub–sample of forecast, observation pairs where the observation falls in the top 5% of observations. Increasing the value of P from 0 (i.e. the full available sample) to a value close to 1 thus gives an indication of forecast performance for high events. Note that for event metrics such as BS and ROC and associated skill scores, the value of P denotes the threshold that defines the event. This procedure is identical to the procedure outlined in Section 3.2.5.

By construction, sample size for the computation of every verification metric varies with the climatological probability of non-exceedance P considered (Figure 36). Increasing the value of P means lower sample size. Sampling uncertainties of the verification metrics were explored by bootstrapping. The stationary block bootstrapping

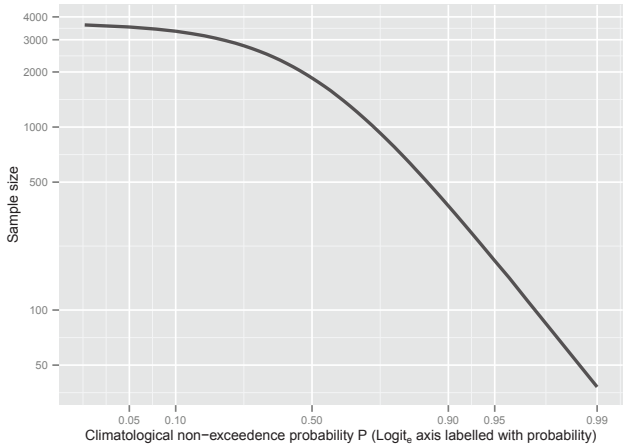


Figure 36: Sample size as a function of the climatological probability of non-exceedence P.

techniques was applied. This method constructs resample blocks of observations to form a pseudo time series, so that the statistic of interest may be recalculated based on the resampled data set (Politis and Romano, 1994). The minimum sample size was set to 50 and the number of bootstrap samples to use in computing the confidence intervals was set to 1000. The applied resampling method estimates the sampling distribution of each verification score. Here, the 5^{th} and 95^{th} percentiles of those distributions are shown. These thus constitute the centred 90% confidence intervals.

Verification metrics were calculated using the Ensemble Verification System (Brown et al., 2010).

4.3 RESULTS AND ANALYSIS

Results were produced for each of the fourteen locations listed in Table 7 and all of the lead times were considered. For practical reasons, the present section includes results for a limited number of lead times and locations only: 24-hour, 36-hour and 48-hour lead times at Llanyblodwel, Pont Robert and Welshbridge. This combination thus comprises forecasting locations with varying sizes of contributing area. Pont Robert is located upstream, Llanyblodwel somewhere in the middle, and Welshbridge at the very outlet of the Upper Severn basin.

4.3.1 *Uncertainty models*

Uncertainty models for the three locations are shown in Figures 32, 33 and 34. All scatter plots show observed water levels on the vertical

axis versus water level forecasts on the horizontal axis. Each of the figures consists of a matrix of multiple panels, with rows showing the four configurations considered and columns showing various lead times. Note that across configurations, the scattered pairs are identical. On the scatter plots, a summary of the estimated uncertainty models is superimposed, consisting of a selection of quantiles only: $\tau \in \{0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.99\}$. Note that these quantiles were derived for plotting purposes only, and do not necessarily coincide with the quantiles derived for verification. In the analysis, a more elaborate set of quantile is used. The latter quantiles are derived using a leave-one-year-out procedure (see Section 4.2.3 for details), whereas this was not the case for the example quantiles in Figures 32 through 35. However, the derived models do not differ markedly. In the plots, the QR-estimated quantiles are shown in grey with the exception of the median quantile which is shown in blue.

From Figures 32, 33 an 34, a few general observations can be made. All scatter plots show that there is an obvious correlation of forecasted and observed water levels, although in none of the combinations of location and lead time, all forecasts are equal to the observations. Spread of the forecast, observation pairs increases with increasing lead time. At zero lead time, the error correction technique ensures that modeled (i.e. simulated or forecasted) water levels are equal to the water level observation, hence at issue time there is no forecasting uncertainty. With increasing lead time, this uncertainty increases. The location with largest lag time (Welshbridge) shows spread that is more concentrated around the 1:1 diagonal than the other locations that have smaller contributing areas and shorter lag times. The location and slope of the quantiles show that in most cases, spread is modeled to be very small at low predicted values of the forecast, and increases with increasing value of the forecast.

The figures show how the uncertainty models, each based on a different configuration of Quantile Regression, differ from one another. Configurations 0 and 1 appear to be very similar. They differ only in those instances where the former configuration would lead to quantile crossing but are identical otherwise, which was indeed anticipated. Configurations 2 (derived using NQT transform) and 3 (piecewise linear approach) are quite different from the first two configuration, but not dissimilar to one another. In these configurations, the quantiles are not a linear function of the water level forecast, that is, not along the full domain. Note that this non-linearity constituted the very reason why these configurations were included in the analysis. Both models often - but not always - show a very small spread at lowest water level forecasts, followed by an increasing spread. At high water level forecasts however, spread no longer increases and sometimes decreases. This means that sharpness of the resulting probability forecasts then

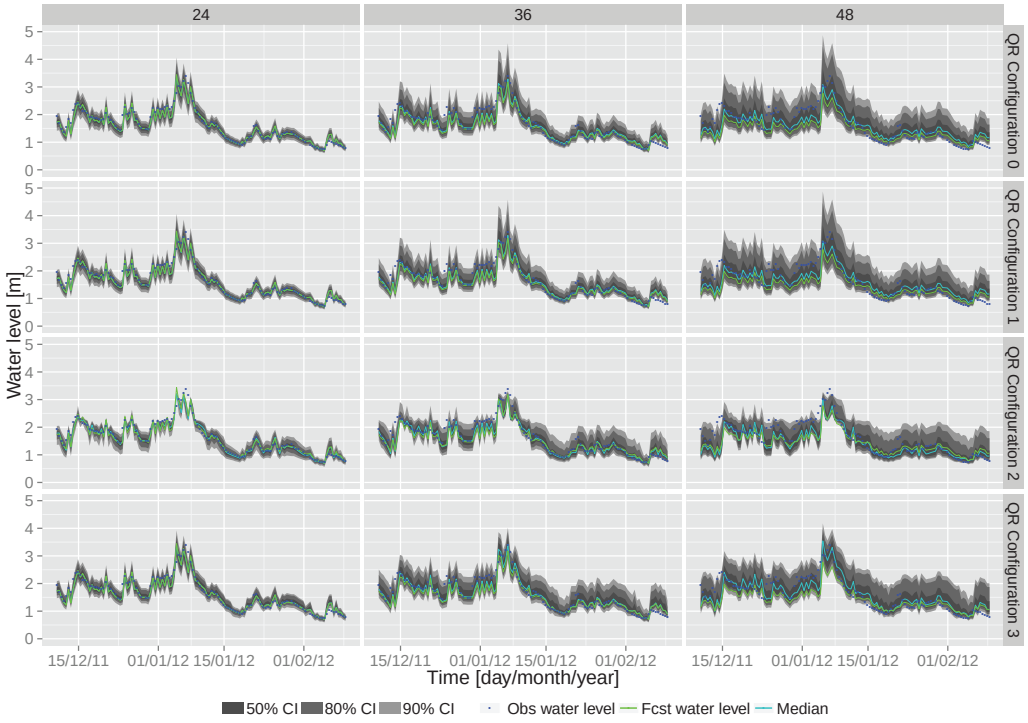


Figure 37: Hydrographs of late 2011 and early 2012 events at Welshbridge.

no longer reduces with increasing values of the water level prediction; sometimes it even increases.

Figure 35 gives some additional background to the QR2 scenario and shows the estimated quantiles in Normal space, i.e. prior to back-transformation to original space. Similar to the other configurations, the estimated quantiles are linear. The strong non-linearity that is shown in Figures 2 through 4 is a result of the back-transformation from Normal to original space.

From the pairs and the models, we can see that at both Llanyblodwel and Pont Robert, the deterministic forecast has a tendency towards underforecasting, i.e. to underestimate future water levels. This underforecasting is corrected for by the uncertainty models, that thus include a bias correction by resulting in a median forecast that is higher than the deterministic forecast. The joint forecast, observation distribution for Welshbridge shows that there is much less obvious underforecasting, or overforecasting for that matter.

4.3.2 *Hydrographs*

Hydrographs are shown in Figure 37 at Welshbridge for a flood event that took place late 2011 and early 2012. The multiplot panel is composed by three columns representing three different lead times; 24-hour, 36-hour and 48-hour, and four rows for each of the four QR configurations. Each of these plots shows time in the horizontal axis, approximately 3 months and water level in the vertical axis. Deterministic forecast water level (green line), observations (blue dots), median quantile (light blue) and centered 50 %, 80 % and 90 % confidence intervals are included (in shades of grey). Across the configurations for a particular lead time, water level observations and deterministic forecasts are identical.

From the plots, a number of observations can be made, each consistent with what was to be expected based on the QR models. Uncertainty increases with lead time, as is shown by the widest intervals at highest lead times, and vice versa. The deterministic forecast tends to underestimate water level observations. With increasing lead time, underforecasting increases. At 48-hour lead time for high water levels the deterministic forecast provides a higher underestimation than for low and medium water levels, which is consistent with QR models shown in Figure 34.

The probabilistic forecasts resulting from configurations 0 and 1 are quite similar to one another. They both show highest uncertainty at higher deterministic water level forecasts. Configuration 2 does not show this behaviour; at higher deterministic forecasts, probabilistic forecasts are sharper. Again, this is consistent with the QR model plots in Figure 34. Configuration 3 results in forecasts whose width in the top 20 % of forecasts varies only slightly (at 24-hour lead time) or almost not at all (at 36-hour and 48-hour lead times) with the value of the predictor.

From a visual inspection, it appears that the median quantile obtained with the four QR Configurations improves the deterministic forecast. QR Configurations 0 and 1 provide a median quantile with a minor improvement. Differences between the median quantile of QR Configuration 2 and the deterministic forecast are the lowest ones. QR Configuration 3 median quantile reproduces with the highest accuracy water level observations, including high, medium and low values.

4.3.3 *Verification*

4.3.3.1 *Reliability and sharpness*

Figures 38 and 39 show reliability diagrams for the full data sample and for the forecasts whose verifying observation falls in the top 10% of

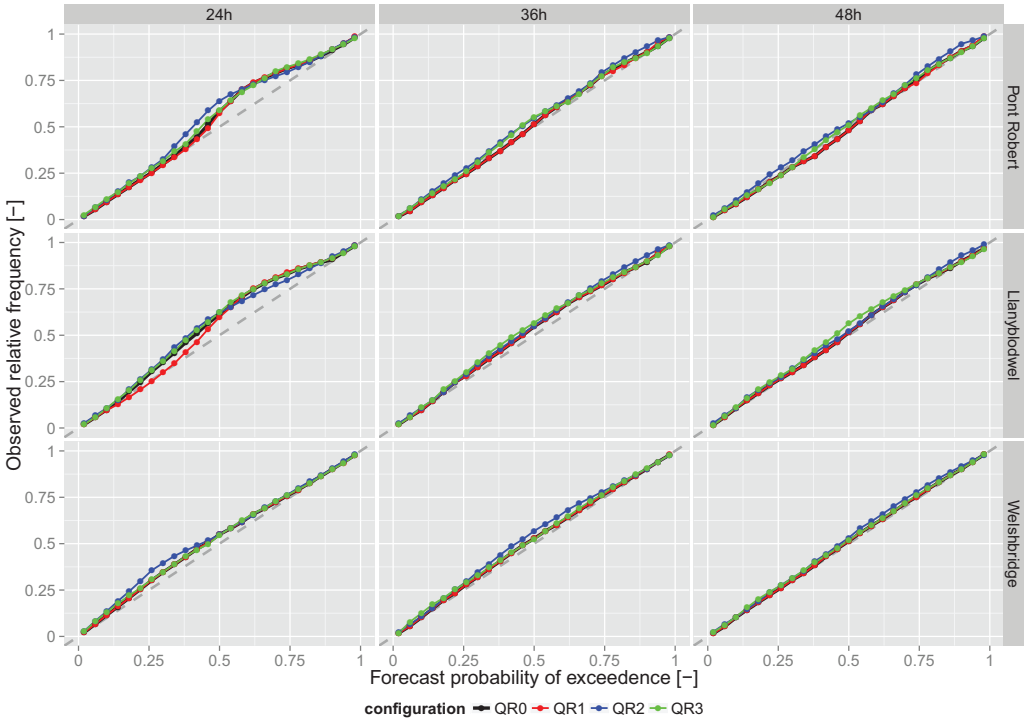


Figure 38: Full sample reliability plots.

observations (P = 0.90), respectively. When looking at the full available sample, the diagrams show reasonably high reliability: most plotting points are very near, or on the 1:1 diagonal. At 24-hour leadtime, there is some underforecasting but this is no longer the case at the longer leadtimes shown.

At P = 0.90, forecasts are considerably less reliable. At all locations and at all leadtimes, there is considerable underforecasting at all but the tails of the predictive distributions. This underforecasting is more pronounced for the smaller basins, and vice versa. Forecasts from QR0 and QR1 are equally (un-)reliable. When comparing these to forecasts from QR2 and QR3, there is no configuration that yields more, or less, reliable forecasts across all cases. QR3 forecasts are nearly always among the least unreliable forecasts, although in many cases this is a shared position with varying other configurations.

Figures 40 and 41 show the distribution of width of the centred 90% predictive intervals for the full available sample (P = 0) and the top 10% of observations only (P = 0.90), respectively. The figures show that sharpness reduces with increasing lead-time as well as with increasing basin lag time. Inter-comparison of sharpness between the different cases shows that for the full sample (Figure 40) there is little if any dif-

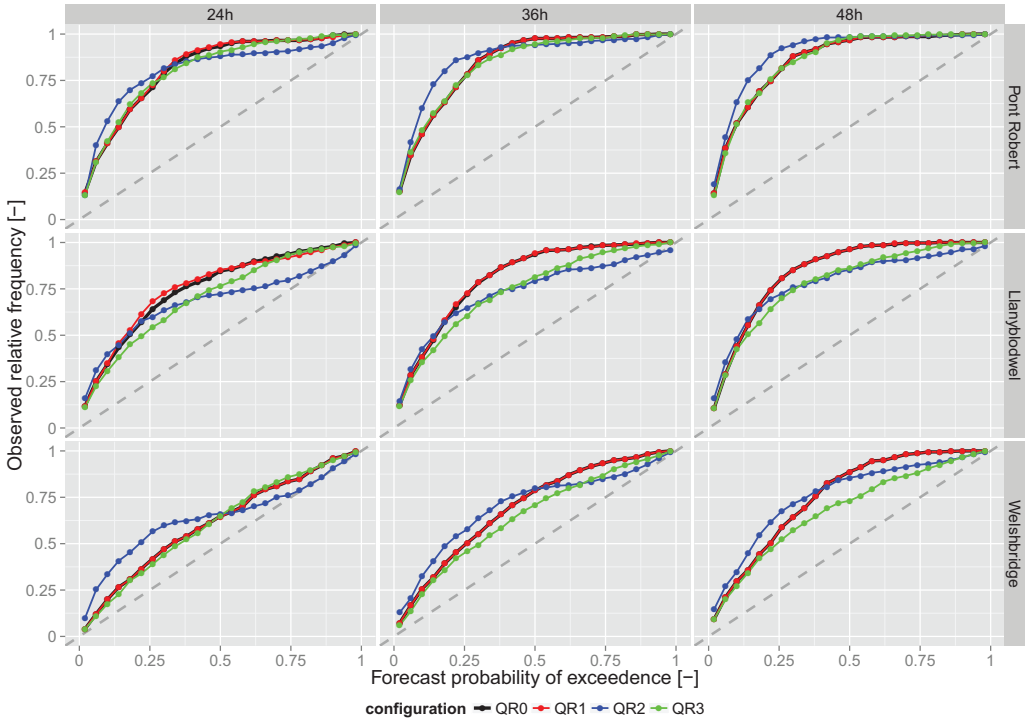


Figure 39: Reliability plots for the forecasts associated with the top 10% observations (P = 0.90).

ference between the four configurations, and virtually none between QR0 and QR1. Forecasts for events that are more extreme (P = 0.90) show larger differences. Again, QR0 and QR1 yield forecasts of more or less equal width, but there are some differences between these configurations and QR2 and QR3. These differences increase with increasing lead time and increasing basin lag time. At Welshbridge, QR2 yields sharpest forecasts, followed by QR3.

Unconditionally, both sharpness and reliability are more or less similar across four configurations. At P = 0.90 however, some forecasts are sharper than others but at the expense of reliability. On balance, usefulness of these forecasts may be equal. The trade-off between probability of detection and probability of false detection can be seen as a measure of this; the derived ROCS is analysed in the next section.

4.3.3.2 *Skill scores*

Figures 42, 43 and 44 present the skill scores computed for probabilistic forecast verification. These plots show BSS, CRPSS and ROCS (vertical axes; each score on a new row) versus the magnitude of the verifying observation, as a function of the observation which is expressed by

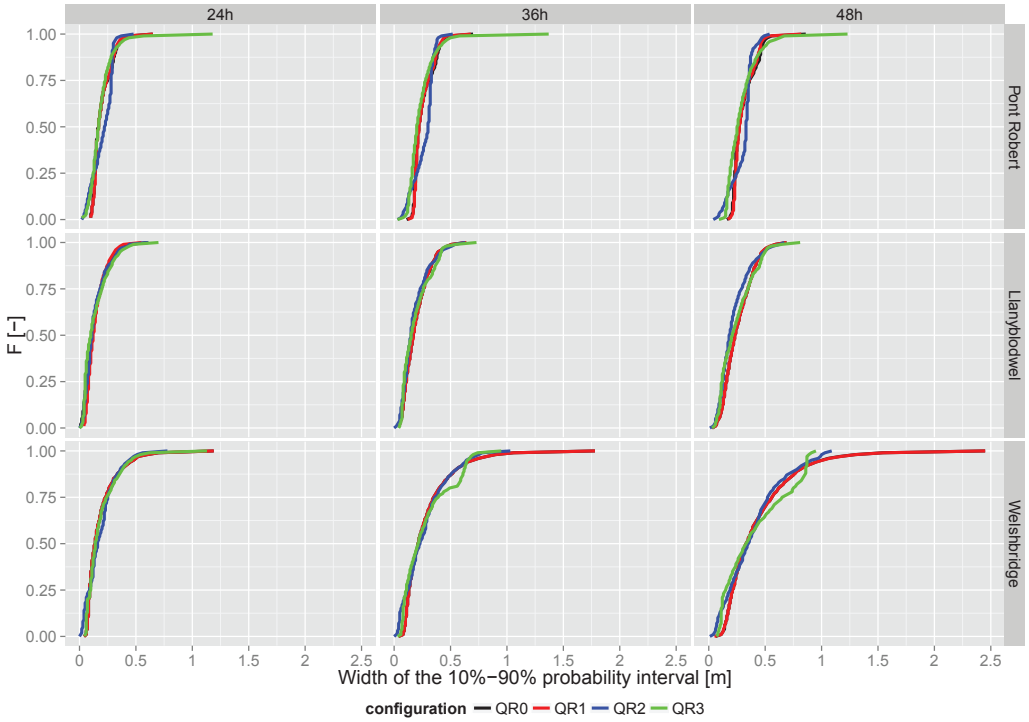


Figure 40: Empirical cumulative distribution function of the centred 80% confidence interval of the predictive distributions.

its climatological probability of non-exceedance P (horizontal axes) for various lead times (columns). In each of the plots, results are shown for four QR configurations considered. To give an indication of the uncertainty in the estimation of metrics, median as well as 10% and 90% estimates are shown.

From the figures, some general observations can be made. First of all, skills are mostly positive, with the exception of BSS and ROCS at the tails of P. Furthermore, skills deteriorate with increasing lead time, increase with increasing basin size and vary with the observation. Many of the plotted results are very similar in that the distribution of verification metrics is very similar - both in terms of the median as well as the confidence bounds shown - across all leadtimes (columns) and values of P (horizontal axes). As the distributions are approximations – the verification pairs used are not strictly independent – a formal statistical hypothesis testing procedure cannot be used. Hence the interpretation is necessarily largely subjective.

The Brier Skill Score (BSS) as a function of P has a concave, inverse U shape curve. BSS is lowest — sometimes even negative — at the tails of P and highest near median P. This is because BSS is calculated

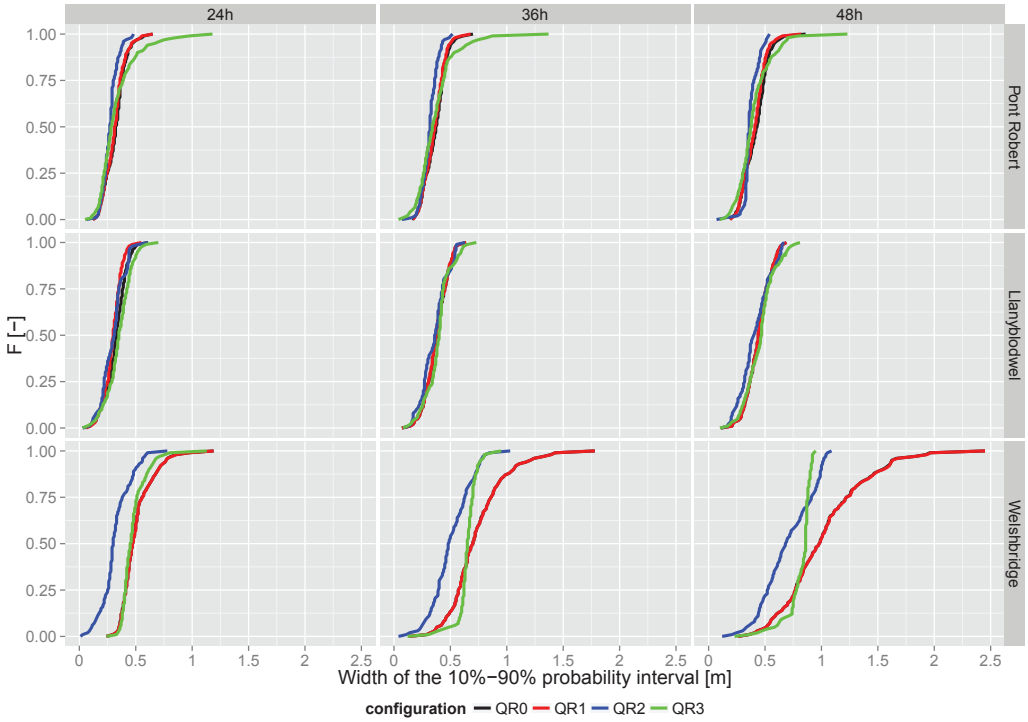


Figure 41: Empirical cumulative distribution function of the centred 80% confidence interval of the predictive distributions associated with the top 10% observations.

using event probabilities; and extreme events, whether low or high, are more difficult to correctly predict than non-extreme events. In terms of difference across the configurations: these are very limited. Only at the low tail do these become apparent, but often the differences are not significant.

Contrary to BSS and ROCS, CRPSS is a smooth, continuous measure that factors skill across all possible thresholds for each paired sample. This different formulation is reflected in its behaviour with increasing value of the observation. For short lead times, CRPSS is approximately constant. With increasing lead time, a small dip in CRPSS values is detected close to the median P. At nearly all lead times, the four QR configurations show very similar skill. The only exception is the highest lead times (48 hours), in which QR Configuration 3 outperforms the remaining cases.

ROCS is a binary event skill with a similar formulation to BSS. However, ROCS values do not show the same pattern than BSS. ROCS is largely constant for the whole climatological distribution of the observations, as it can be seen at Welshbridge in Figure 34. Pont Robert

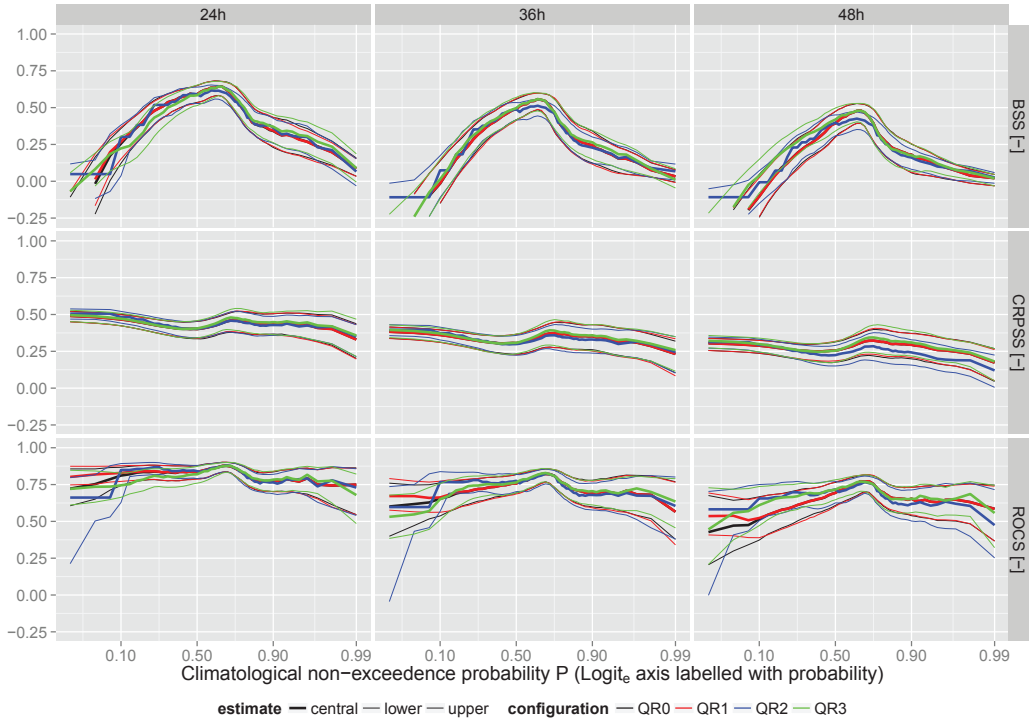


Figure 42: Verification results for water level forecasts at Pont Robert station (5 – 9 h lag time). In the rows, three different metrics are shown; from top to bottom these are the Brier Skill Score (BSS), the mean Continuous Ranked Probability Skill Score (CRPSS) and the Relative Operative Characteristic Score (ROCS). Columns show various leadtimes: 24, 36 and 48 h.

(Figure 42) and Llanyblodwel (Figure 43) present lower skill for the top half of the observations. Forecast quality decreases with increasing lead time, as it happens with BSS and CRPSS. No significant differences can be pointed out among the analysed QR Configurations.

4.4 SUMMARY, CONCLUSIONS AND DISCUSSION

The research described in this chapter had two objectives: (i) to extensively verify the estimates of predictive uncertainty for Upper Severn basins that were produced using the Quantile Regression post-processing technique as described by WWV2011; (ii) to address two issues with the 'as is' implementation of linear models of QR: (a) invalid predictive distributions due to the crossing quantile problem; (b) the description of slightly non-linear joint distributions by a linear QR model.

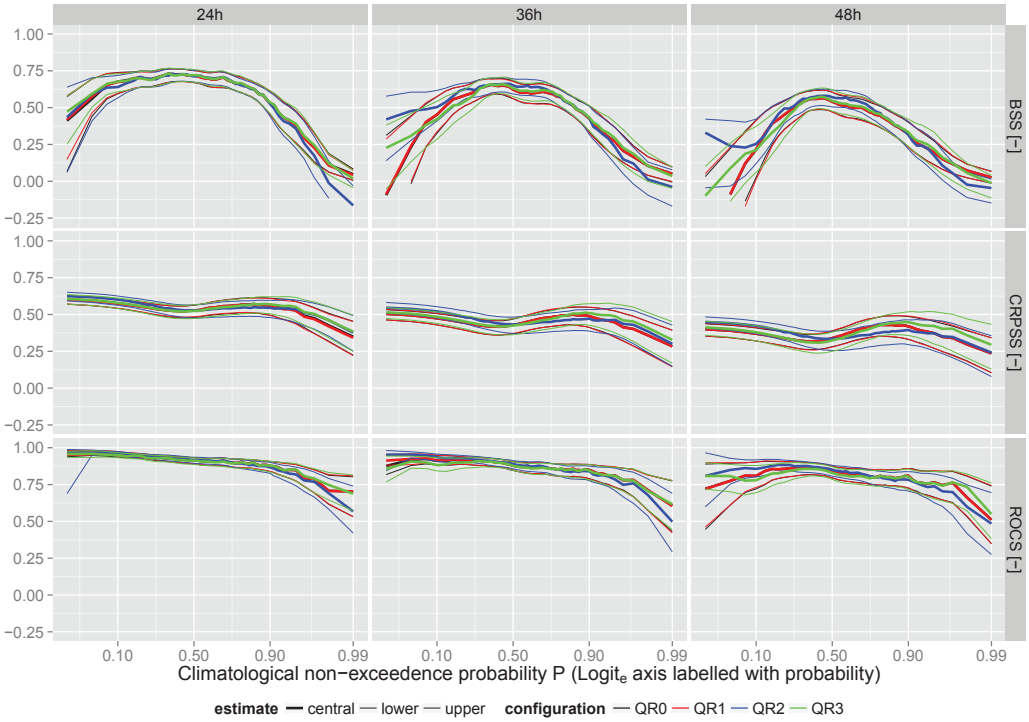


Figure 43: Verification results for water level forecasts at Llanyblodwel station (7 – 10 h lag time). In the rows, three different metrics are shown; from top to bottom these are the Brier Skill Score (BSS), the mean Continuous Ranked Probability Skill Score (CRPSS) and the Relative Operative Characteristic Score (ROCS). Columns show various leadtimes: 24, 36 and 48 h.

The verification of forecast quality builds on the verification that was carried out in WWV2011. In the present chapter, multiple metrics and skill scores are presented. Also, a 'conditional verification' was carried out, i.e. verification was done for a large number of sub-sets of available data, each representative for increasingly higher events. Verification showed that, unconditinally, in terms of all skills and metrics, forecast quality is positive. However, the analysis also shows that forecast quality and skill decreases with increasing value of the event.

The two issues described above were addressed by implementing several techniques, thus arriving at four configurations of Quantile Regression. The problem of crossing quantiles was addressed by adopting the non-crossing quantiles technique that was proposed by Bondell et al. (2010b). This resulted in near-identical sharpness, reliability and skill. From a forecaster's point of view, the technique constitutes a methodological improvement as the post-processor will no longer produce invalid predictive distributions as a result of crossing quantiles,

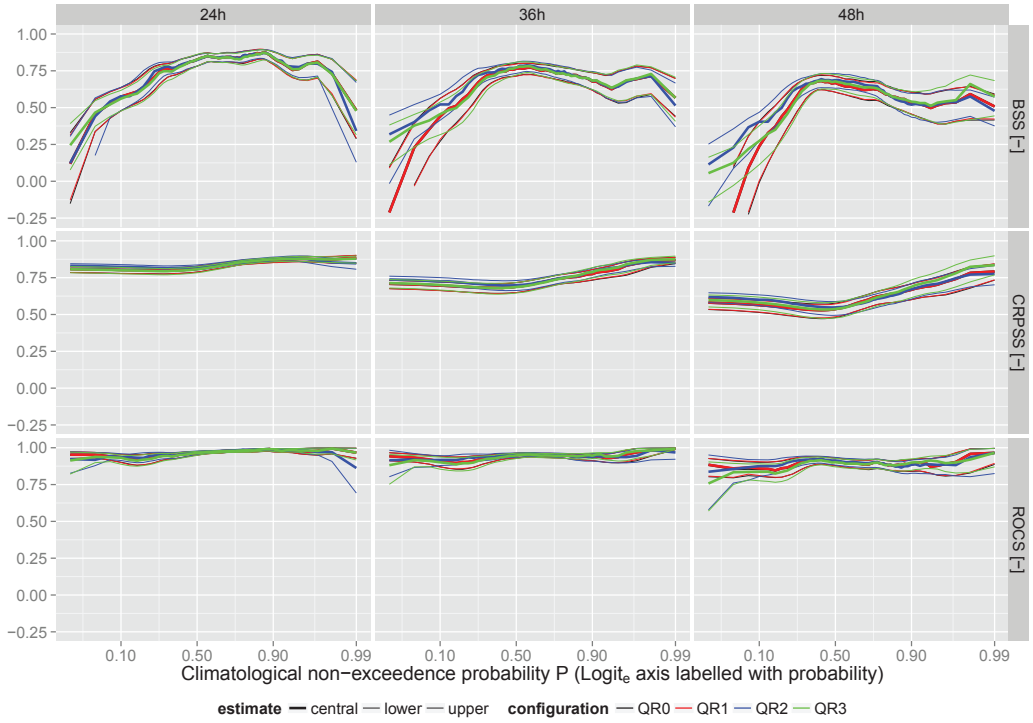


Figure 44: Verification results for water level forecasts at Welshbridge station (5 – 9 h lag time). In the rows, three different metrics are shown; from top to bottom these are the Brier Skill Score (BSS), the mean Continuous Ranked Probability Skill Score (CRPSS) and the Relative Operative Characteristic Score (ROCS). Columns show various leadtimes: 24, 36 and 48 h.

at no noticeable extra computational expense. The problem of linearly describing joint distributions of forecasts and observations that may not be linear in nature was addressed by two different approaches. The transformation to the Normal space attempted to produce a joint distribution that is 'more linear'. The piecewise linear derivation approach constitutes dividing the data into sub-samples on which the joint distribution is linear.

The intercomparison shows that none of the four Quantile Regression configurations consistently outperforms the others. Sharpness and reliability may vary across configurations, but none results in a more favourable combination of the two. In terms of BSS, CRPSS and ROCS, the four configurations yield comparable forecast quality.

Addressing the problem of the non-linearity of the joint distributions by the solutions proposed in the present chapter has not resulted in higher skill. Either the data was sufficiently linear for the techniques not to be required, or the techniques have not performed to expecta-

tion. In any case, a skill improvement does not provide a rationale for derivation of Quantile Regression models in Normal space as was done by WWV2011.

While none of the configurations has a proven higher skill, there may be alternative reasons for choosing one over the other. If the post-processors are used in operational forecasting systems, the forecasters will have to be able to explain to an end user how predictive uncertainty was estimated. Hence more complicated configurations will be less likely to be used. Also, forecasts have to be consistent with forecasters' beliefs (Murphy, 1993), hence the post-processor will have to fit with the forecasters' perceptual model of forecasting error.

Like all post-processing techniques, QR requires a long calibration and validation data set containing several extreme events. If the magnitude of the forecasted water level is outside of the calibration sample range, then any estimate of hydrological predictive uncertainty is not supported by data in that range. In an operational setting, it is important for the forecaster to be aware that this issue may surface. A suggestion to overcome this issue may be to "flag" the uncertainty estimate if it is based on extrapolation outside of the calibration range. Possibly, in those cases the uncertainty estimate can be replaced by an assumed estimate that the forecasters are comfortable with.

What would be a promising route to try and improve the skill of the estimates of predictive uncertainty that are produced by Quantile Regression? There are multiple possible answers here. First of all, there may be merit in adding predictors, i.e. by further conditioning forecast error on additional available variables. These could, for example, include the internal state variables of a model (dry or wet) and/or available observations at upstream locations. This route was taken by Solomatine and Shrestha (2009) in their UNEEC approach, and by Dogulu et al. (2014). Both compare a more complex UNEEC approach to QR and found improvement in skill. Stratification of the post-processing depending on different seasons or water level ranges could represent another alternative configuration. Both the addition of predictors as well as stratification, however, introduce additional data requirements that may not be met, and in the absence of which the quality of post-processed forecasts may be reduced. Alternative techniques may be considered, a recent article by Van Andel et al. (2013) discusses various techniques in the context of the HEPEX intercomparison experiment. Another option would be to fully investigate additional configurations of the piecewise linear approach. For example, c-means or K-means clustering would allow for partitioning data to be used to build several regression models.

All the configurations inter-compared in the present work are parametric Quantile Regression estimations. Non-parametric or semi-parametric Quantile Regression approaches, based on local smoothing

could also be considered in future studies. For example, a comparison between here presented parametric QR configurations and the non-parametric estimation of the water level or discharge conditional distribution with copulas proposed by Smith et al. (2014), would be of interest.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Environment Agency for provision of the data and use of the standalone NFFS system required to do the analyses described in the present chapter. The authors would also like to thank Julie Demargne, Paul Smith and Florian Pappenberger who, in their roles as reviewer (JD, PS) and editor (FP), helped to improve this chapter considerably.

5

ESTIMATING PREDICTIVE HYDROLOGICAL UNCERTAINTY BY DRESSING DETERMINISTIC AND ENSEMBLE FORECASTS; A COMPARISON, WITH APPLICATION TO MEUSE AND RHINE

ABSTRACT

Two techniques for estimation of predictive hydrological uncertainty are compared: post-processing of deterministic forecasts and 'dressing' of ensemble streamflow forecasts by adding estimates of hydrological uncertainties to individual streamflow ensemble members. Both techniques 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.

This chapter has been submitted for publication in Journal of Hydrology as Verkade, J.S., J.D. Brown, F. Davids, P. Reggiani, and A.H. Weerts. Estimating predictive hydrological uncertainty by dressing deterministic and ensemble forecasts; a comparison, with application to River Meuse

5.1 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).

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 (2008a) 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), UNEEC (Solomatine and Shrestha, 2009) and 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 forecassts 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 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 chapter is organised as follows. In the next section, the study approach is detailed, followed by a description of the study basins in section 5.3. In section 5.4 the results of the experiments are presented and analysed. In section 5.5, conclusions are drawn and discussed.

5.2 APPROACH

5.2.1 *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.

5.2.2 *'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),

$$\Phi_n\left(y_n|x_{n,1}, x_{n,2}, \ldots, x_{n,E}\right) = \frac{1}{E}\sum_{e=1}^{E} \phi_n\left(y_n|x_{n,e}\right), \tag{17}$$

where Φ is the aggregated distribution of observed streamflow y at lead time n, conditional on the raw streamflow forecast x 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

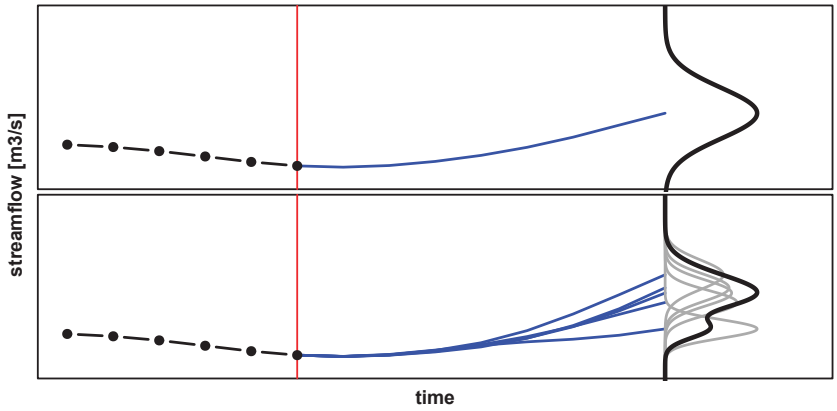


Figure 45: Schematic representation of the dressing procedures for the deterministic (top) and ensemble forecasts (bottom), respectively. The vertical red line denotes the issue time of the forecast, with the observations (black dots) in the past and the raw forecasts (blue lines) in the future.

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. However, in principle, our approach could be extended to account for under-dispersion of the meteorological ensemble via the streamflow. In this context, any under-dispersion would be reflected in the residual of the best forcing ensemble member after propagation through the streamflow model and the simulated streamflow. This would benefit from the mediating effects of basin hydrology on precipitation when compared to directly modelling the under-dispersion of the precipitation ensemble forecasts. Finally, an 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).

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,

$$\Phi_n\left(y_n|x_{n,1}\right) = \phi_n\left(y_n|x_{n,1}\right). \tag{18}$$

The dressing procedures are schematically visualised in Figure 45.

5.2.3 *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 forecasts are dressed with predictive distributions that comprise both hydrological and meteorological uncertainties. In the ensemble case, each of the ensemble members is dressed with a predictive distribution that comprises the hydrological uncertainties only.

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. The latter 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 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 use the latest available observation as a predictor in the uncertainty model; this approach was taken by, among others, Seo et al. (2006).

5.2.4 *Quantile Regression*

In the present chapter, in both scenarios, uncertainty is estimated using Quantile Regression (QR; Koenker and Bassett Jr, 1978; Koenker and Hallock, 2001; Koenker, 2005). QR is a regression technique for estimating the quantiles of a conditional distribution. The technique is described in detail in Appendix A. Figure 49 and Figure 50 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.

5.2.5 *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 B.

Reliability is the degree to which predicted probabilities coincide with the observed relative frequencies, given those forecast probabilities. Here, we consider reliability for subsets of verification pairs that exceed a quantile of the predictive distribution, as well as the overall paired samples. Reliability is measured in terms of the dispersion of the observations within the forecast distribution (see Appendix B), which is akin to the Probability Integral Transform. Proximity to the 1:1 diagonal, where observed frequency equals predicted probability, indicates higher reliability. Sharpness plots show the empirical cumulative distribution of the width of the 10^{th}–90^{th} 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, relative mean error (RME), mean absolute error (MAE) and the root mean squared error (RMSE). The correlation

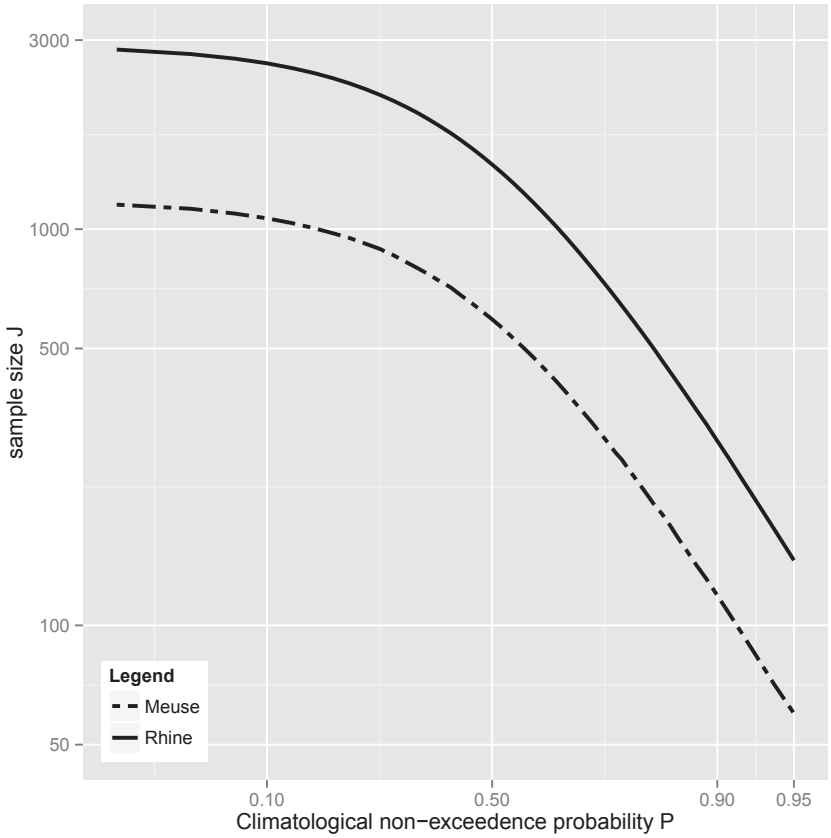


Figure 46: Sample and subsample size as a function of the climatological probability P of non-exceedence of observation.

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 chapter, 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. $2\mathrm{AUC} - 1$. 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 1). 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. This procedure is identical to that in Chapters 3 and 4. Essentially, $P = 0$ constitutes an unconditional verification for continuous measures, such as the CRPSS, as all available data pairs are considered (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 (Figure 46).

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 chapter, the comparison between scenarios is (necessarily) based on a qualitative description of the uncertainty intervals.

5.3 STUDY BASINS AND DATA USED

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

5.3.1 *Meuse*

The river Meuse (Figure 47) 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 soil layers. 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 900mm. The Meuse is a typically rain-fed river; long lasting, extensive snowpacks do not occur. Figure 48 shows the distribution of streamflow at the forecasting locations considered in this study. Near Maastricht, average runoff equals approx. 200 m^3/s. Temporal variability can be large as, during summer, streamflow can be less than 10 m^3/s, while the design flood, associated with an average return period of 1,250 years, has been established at approx. 3,000 m^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 the Meuse, just upstream of the Dutch-Belgian border, are heavily reg-

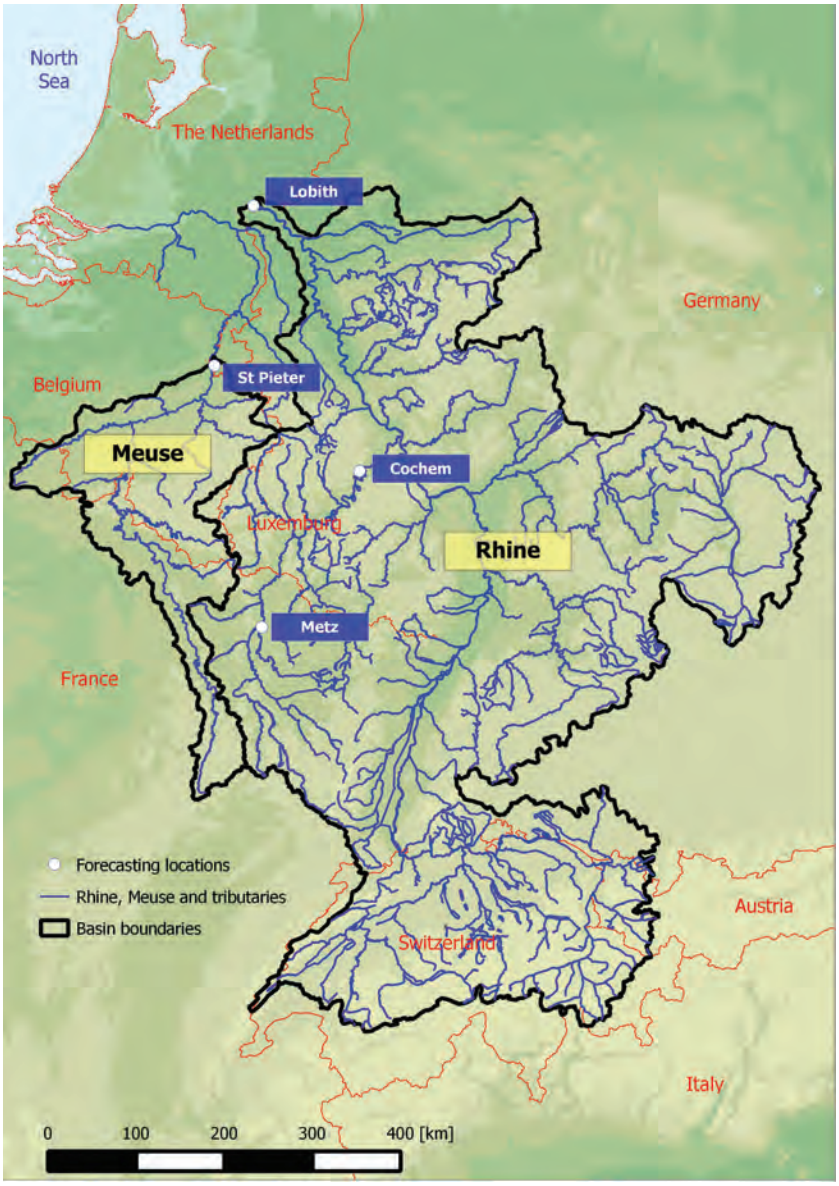


Figure 47: Map of the Meuse and Rhine basins and the forecasting locations that are considered in this chapter.

ulated 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.

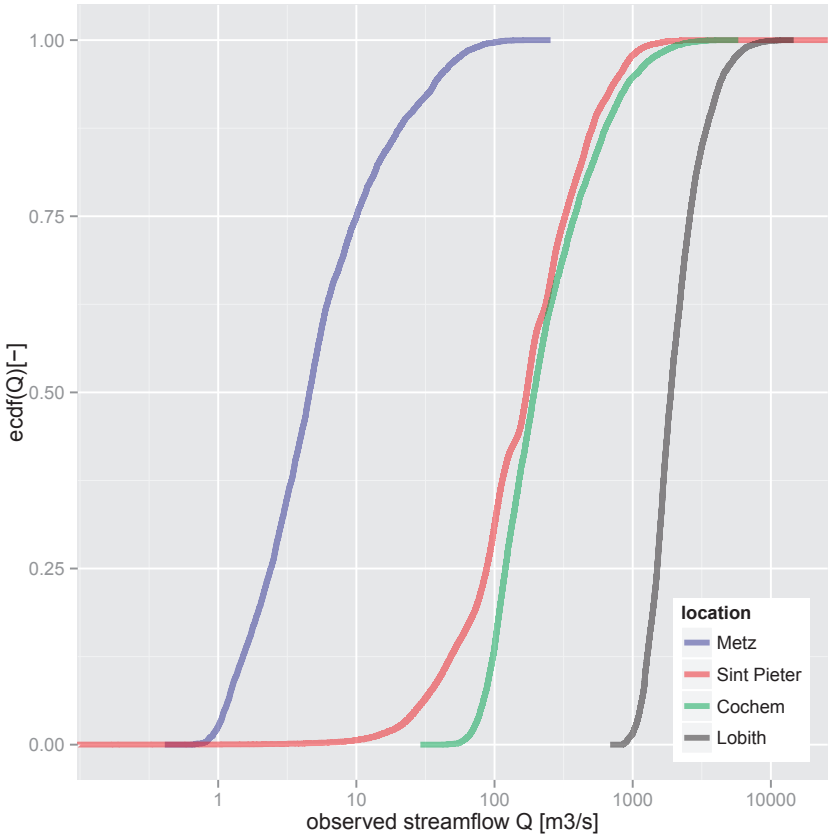


Figure 48: Distribution of streamflow observations at the forecasting locations considered in this study.

Hindcasts for the river Meuse 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. The forecasting system contains an implementation of the HBV rainfall-runoff model (Bergström and Singh, 1995). This is a semi-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.

Hindcasting is a two-step process: first, the hydrological model is forced with observed temperature and precipitation for a period up to the forecast initialisation time. Thus, the internal model state reflects the basin's actual initial conditions as closely as possible. The initial state is used as the starting point for forecast runs, where the model is forced with COSMO-LEPS precipitation and temperature ensemble forecasts.

COSMO-LEPS (Marsigli et al., 2005) 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 10km grid spacing and 40 vertical layers. It covers large parts of continental Europe including the Meuse basin. For the present experiment, approx. 1,400 historical COSMO-LEPS forecasts were available (Figure 46): 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 flood volume originates from the basin upstream from Chooz hence the 'naive' forecast is expected to be skillful. 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 chapter.

In the hindcasting procedure, an ARMA error correction procedure (Broersen and Weerts, 2005) was used. Error correction is applied to the streamflow forecast at St Pieter but not for water level forecasts or streamflow forecasts for other forecasting locations. The effect of error correction will therefore diminish with increasing lead time as well as with increasing distance from St Pieter. Hourly streamflow observations for hydrological stations along the stream network as well as temperature and precipitation observations within the Meuse basin were obtained from the Water Management Centre of The Netherlands.

5.3.2 *Rhine*

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 delt, 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. Figure 47 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.

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 and Singh, 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.

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 (Figure 46), 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

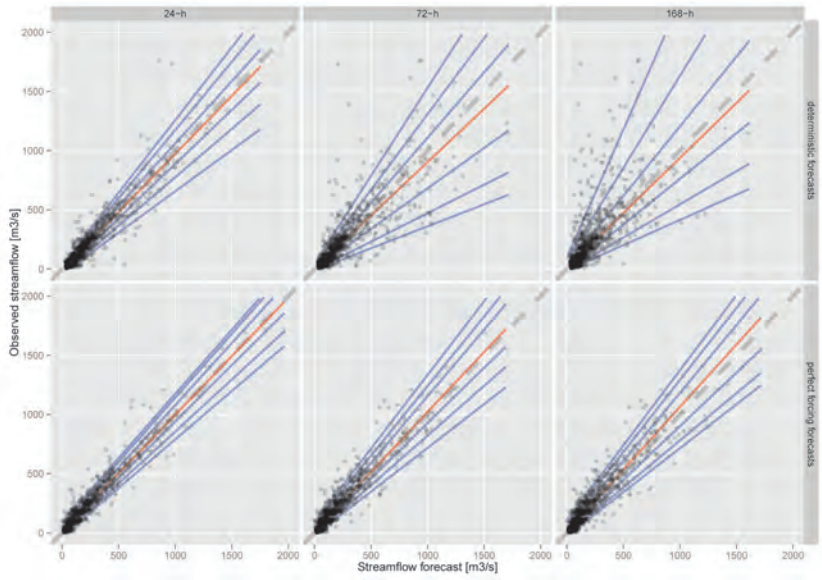


Figure 49: Quantile Regression plots for St Pieter for both deterministic (top row) and perfect forcing (bottom row) forecasts with 24-h, 72-h and 168-h forecasts (columns).

a deterministic forecast which was subsequently dressed and analysed. However, results for one of these forecasts is presented only.

5.4 RESULTS AND ANALYSIS

5.4.1 *Post-processing of single valued forecasts*

Scatter plots of single–valued forecasts and observations are shown in Figures 49 and 50 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 5^{th}, 10^{th}, 25^{th}, 75^{th}, 90^{th} and 95^{th} percentiles in blue.

The pairs and the estimated quantiles in the St Pieter figure (Figure 49) 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

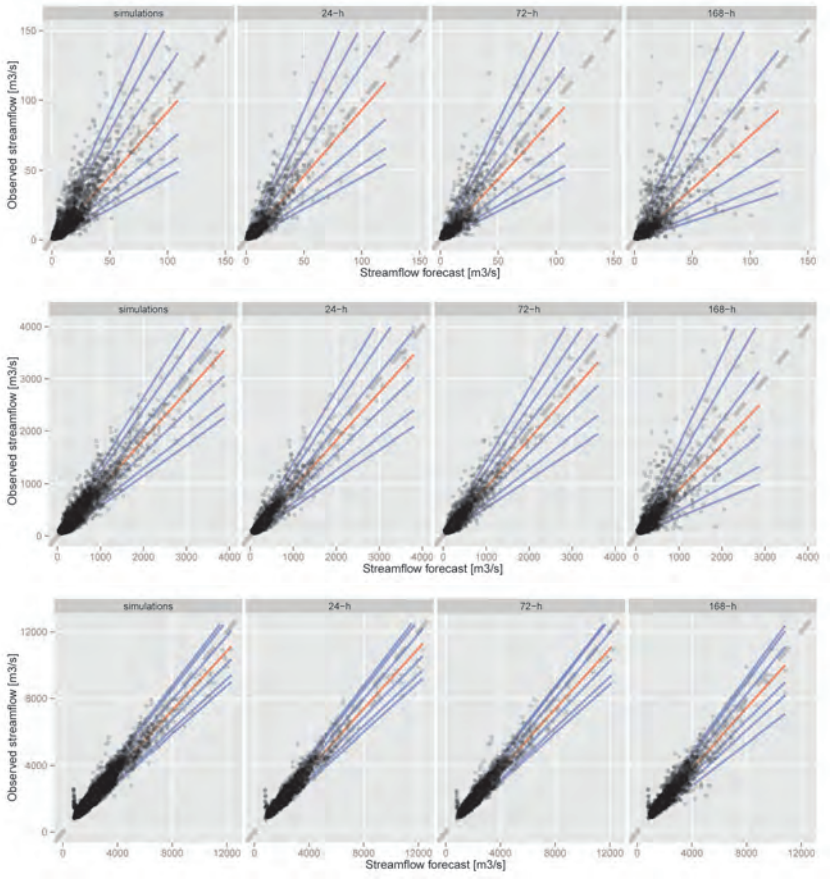


Figure 50: Quantile Regression plots for Metz (top), Cochem (middle) and Lobith (bottom) for both simulations (leftmost column) and deterministic forecasts with 24-h, 72-h and 168-h forecasts (rightmost three columns).

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 (Figure 50), 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

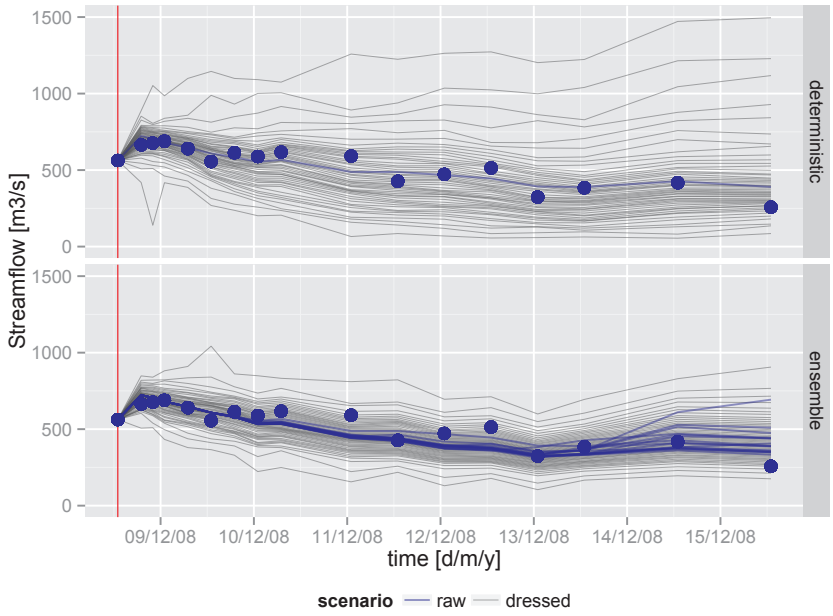


Figure 51: Sample forecasts of the scenarios: deterministic and ensemble based forecasts in top and bottom row, respectively. Observed values are indicated by blue dots. Note that the (blue) lines for the raw forecasts represent one or multiple traces, whereas the (black) lines for the dressed forecasts represent quantiles. The red vertical line denotes the issue time of the forecast.

exception, the median forecast is located below the diagonal which indicates an overforecasting bias.

5.4.2 *Forecast hydrographs*

Sample forecasts or predictive distributions for both scenarios are shown in Figure 51. 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.

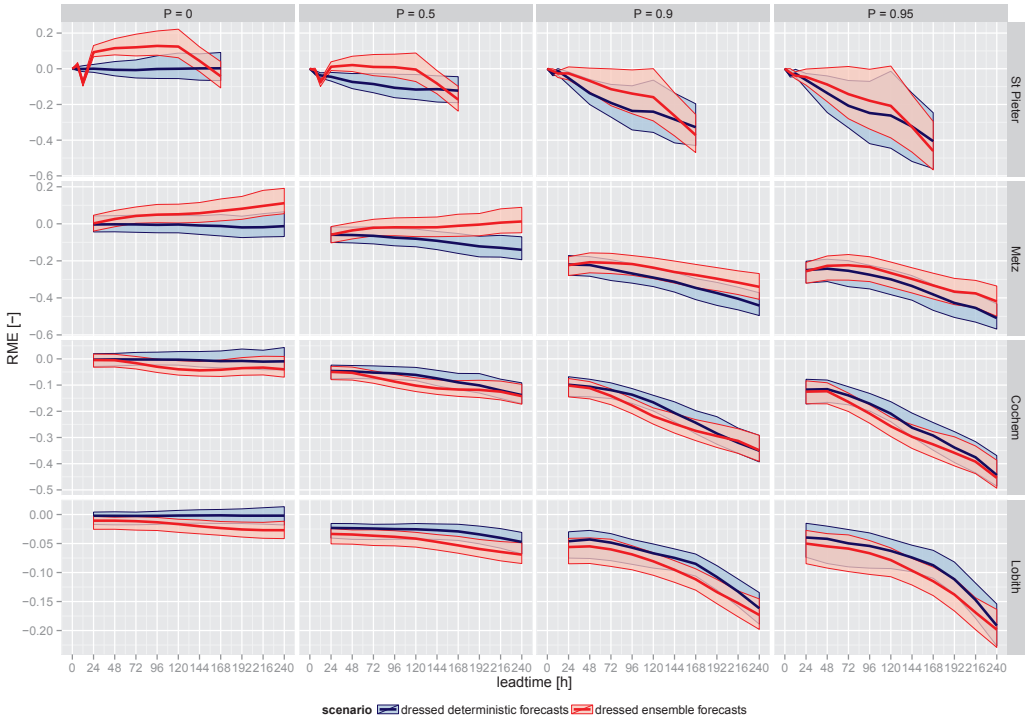


Figure 52: Relative Mean Error (RME) as a function of lead time for several subsamples of the verification pairs (columns) and several locations (rows). Note that the vertical scales used in the various columns differ.

The 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.

5.4.3 *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 chapter, only the RME is shown (Figure 52).

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

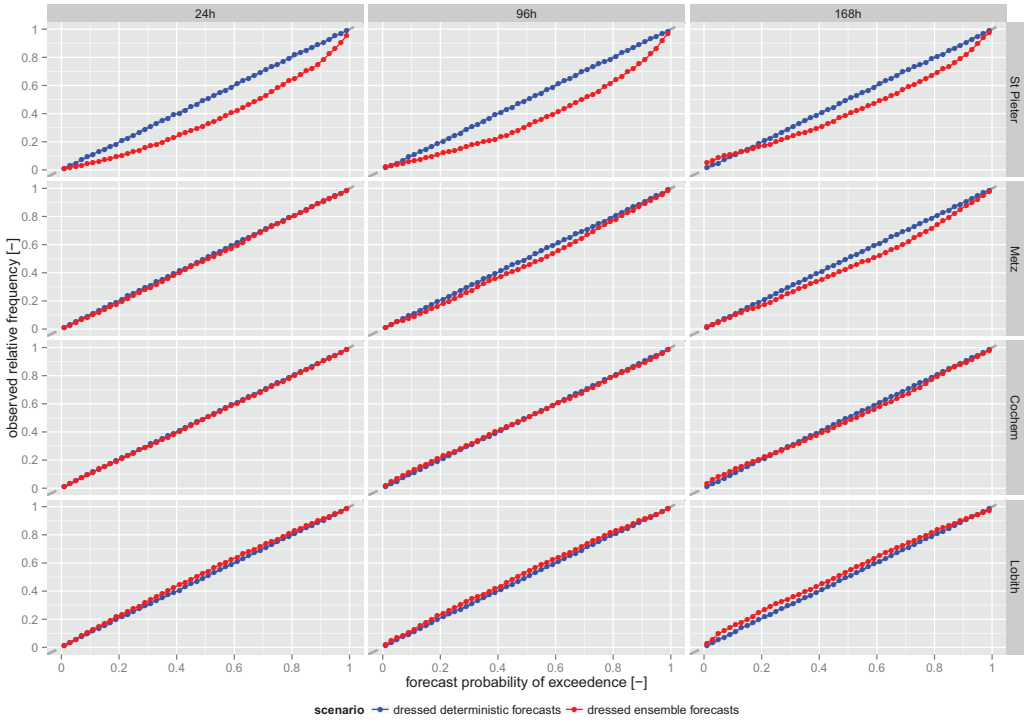


Figure 53: Reliability plots for various lead times (columns) for several locations (rows). The plot is unconditional, i.e. for the full data sample, (i.e. $P = 0$).

those cases, the dressed ensembles outperform the dressed deterministic forecasts.

The RME plots (Figure 52) 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 worsens (i.e., increases) with increasing forecast lead time, increasing threshold amount, and 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

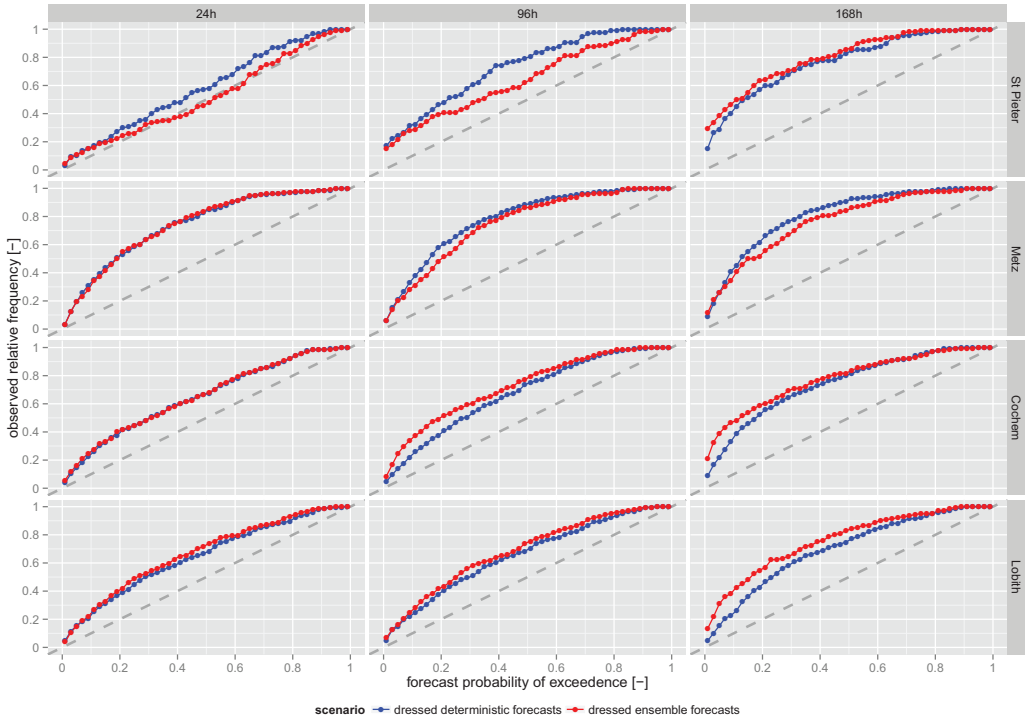


Figure 54: Reliability plots for various lead times (columns) for several locations (rows). The plot is conditional on the observations exceeding the 90^{th} percentile of the climatological exceedence probability (i.e., $P = 0.90$).

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.

5.4.4 *Reliability and sharpness*

Reliability plots for the unconditional sample at $P = 0$ (Figure 53) show that the dressed deterministic forecasts are extremely reliable at all forecast lead times. In contrast, the dressed ensemble forecasts are not consistently reliable. In the Rhine basins, the dressed ensemble forecasts are reliable at the shortest forecast lead time but less reliable at longer forecast lead times. At St Pieter, the dressed ensembles are much less reliable. Here, the dressed ensemble forecasts overestimate the ob-

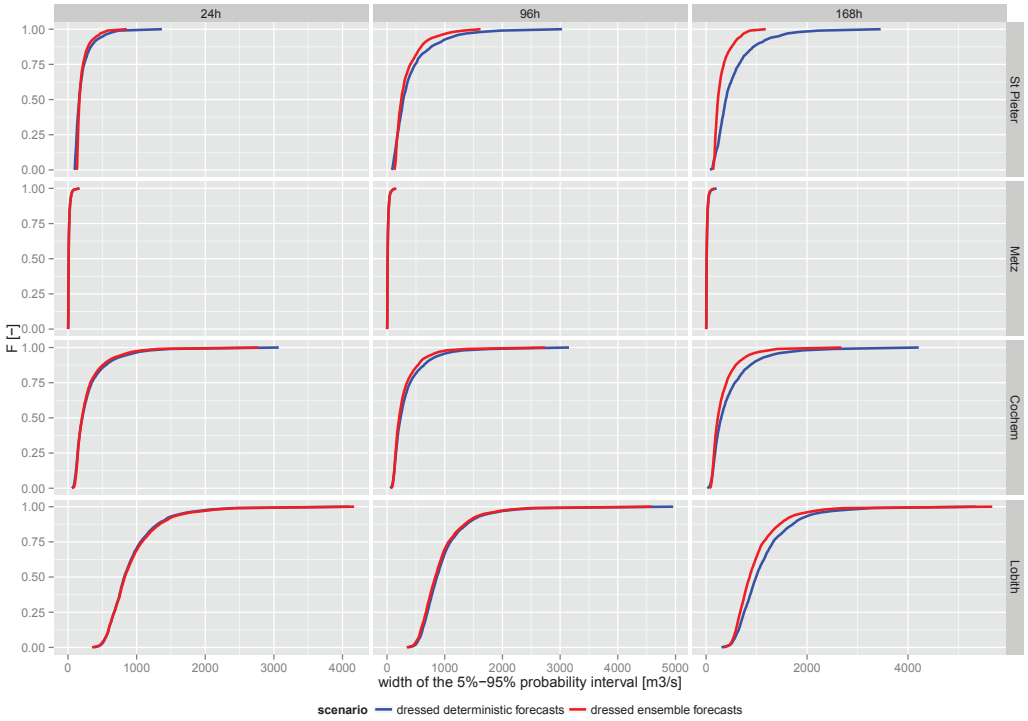


Figure 55: Sharpness plots for various lead times (columns) for several locations (rows). The plot is unconditional, i.e. for the full data sample (i.e., $P = 0$).

served probability of exceeding a given threshold, i.e. they show a wet bias (cf. Wilks 2011).

For the $P = 0.90$ subsample (Figure 54) show that most forecasts are a lot less reliable compared to the unconditional sample. The only exception is St Pieter, where the dressed ensembles show higher reliability at $P = 0.90$ than at $P = 0$. At St Pieter and Metz, the dressed ensembles are more reliable than the dressed deterministic forecasts; the converse is true for Cochem and Lobith.

Sharpness (Figures 55 and 56) reduces with increasing lead time, with increasing basin size and with increasing value of P. At lead times longer than 24-h, the differences in sharpness between scenarios becomes noticeable. In all cases, the dressed ensembles result in sharper predictive distributions than the dressed deterministic forecasts. These differences are more pronounced at higher values of P. However, sharpness is only valuable in a decision making context if the forecasts are also reliable, i.e. if the spread is sufficient to make reliable predictions.

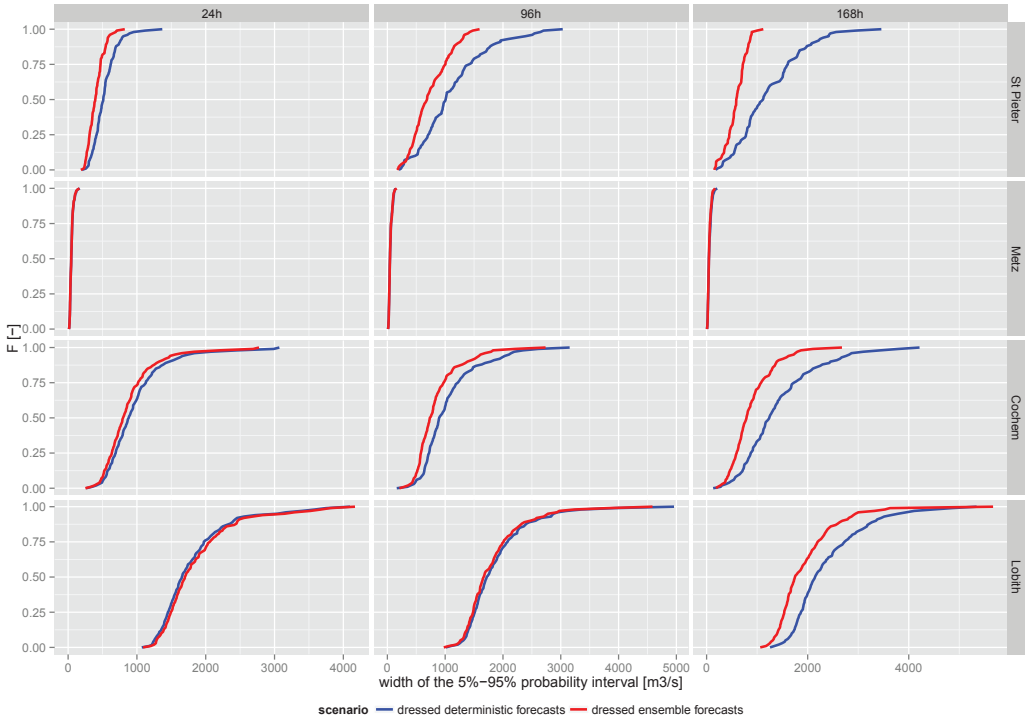


Figure 56: Sharpness plots for various lead times (columns) for several locations (rows). The plot is conditional on the observations exceeding the 90^{th} percentile of the climatological exceedence probability (i.e., $P = 0.90$).

5.4.5 *Probabilistic Skill Scores*

For the skill scores (Figure 57), 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 (Figure 58) 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

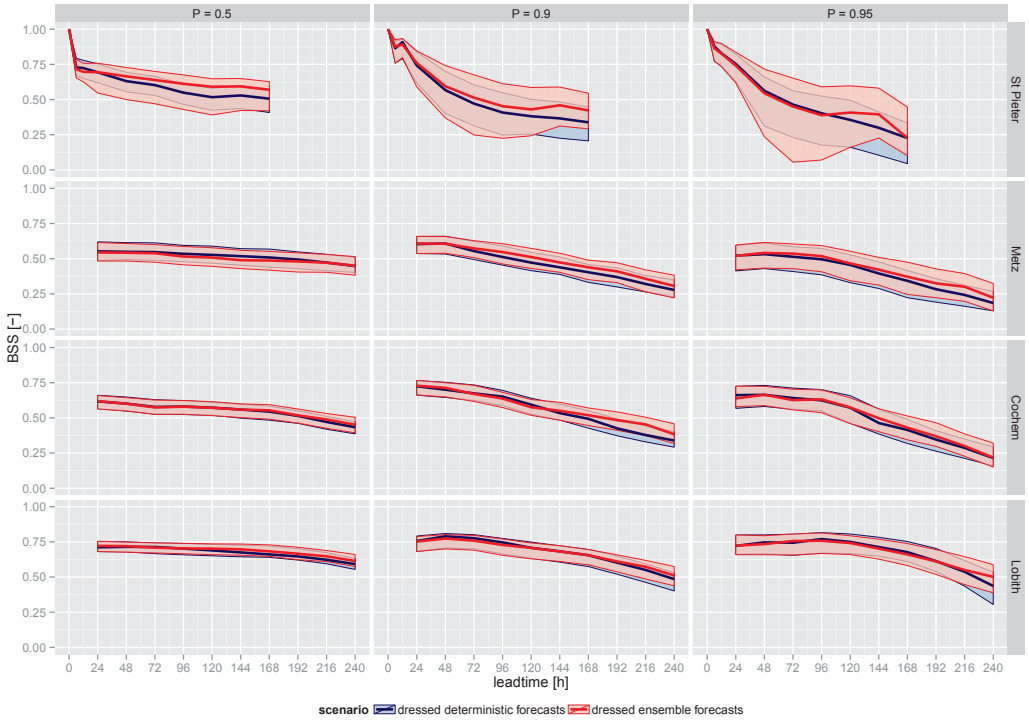


Figure 57: Brier Skill Score (BSS) as a function of lead time for several events (columns) and several locations (rows).

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).

5.4.6 *Forecast value*

Relative Operating Characteristic plots for the event defined by the exceedence of the 90th percentile of the observational record are shown in Figure 59. 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 ensemble forecasts are slightly more discriminatory than the dressed deterministic forecasts.

The associated ROC scores (Figure 60) are very similar at most locations and forecast lead times and generally decline with increasing forecast lead time and increase with threshold amount.

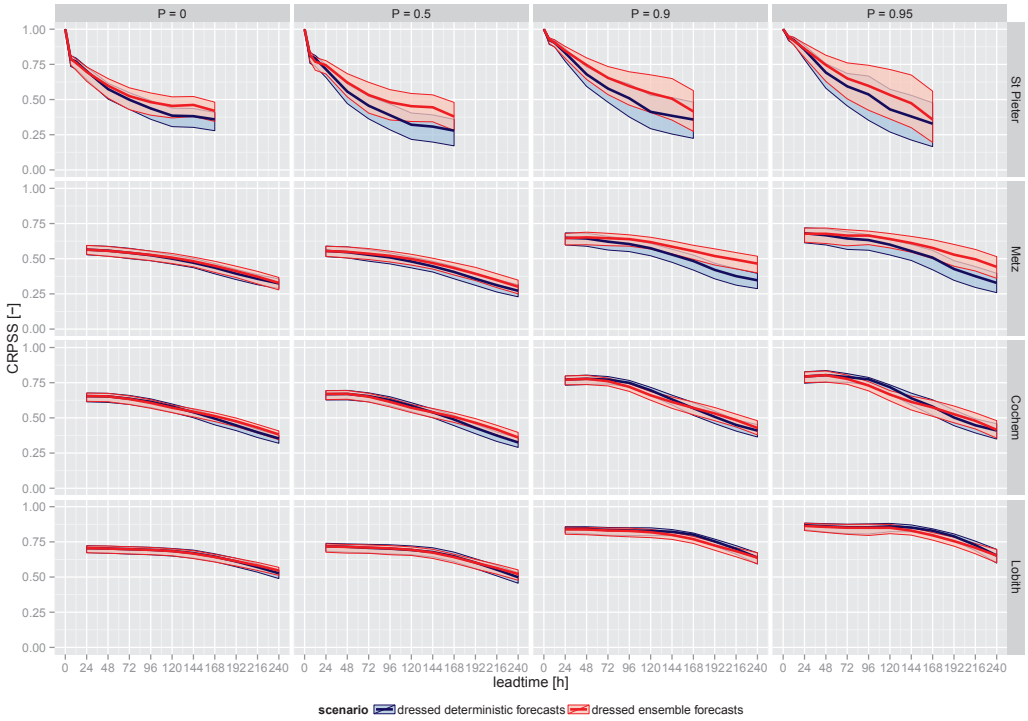


Figure 58: Mean Continuous Ranked Probability Skill Score (CRPSS) as a function of lead time for several subsamples of the verification pairs (columns) and several locations (rows).

The 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.

5.4.7 *Analysis*

The results show that, at $P = 0$, the dressed deterministic forecasts improve on the dressed ensemble forecasts in terms of reliability and RME. However, the dressed ensemble forecasts are sharper. On balance, the dressed ensemble forecasts have 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

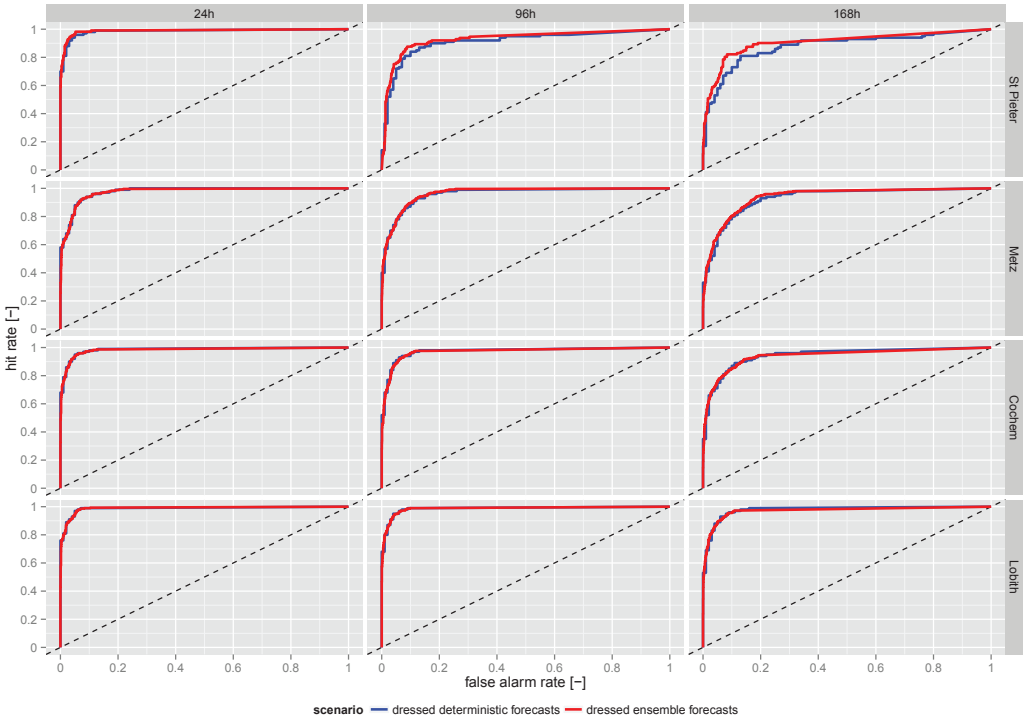


Figure 59: ROC plots for various lead times (columns) for several locations (rows). The plot is for the event that the posterior water level exceeds the 90^{th} percentile of the climatological exceedence probability (i.e., $P = 0.90$).

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.

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.

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

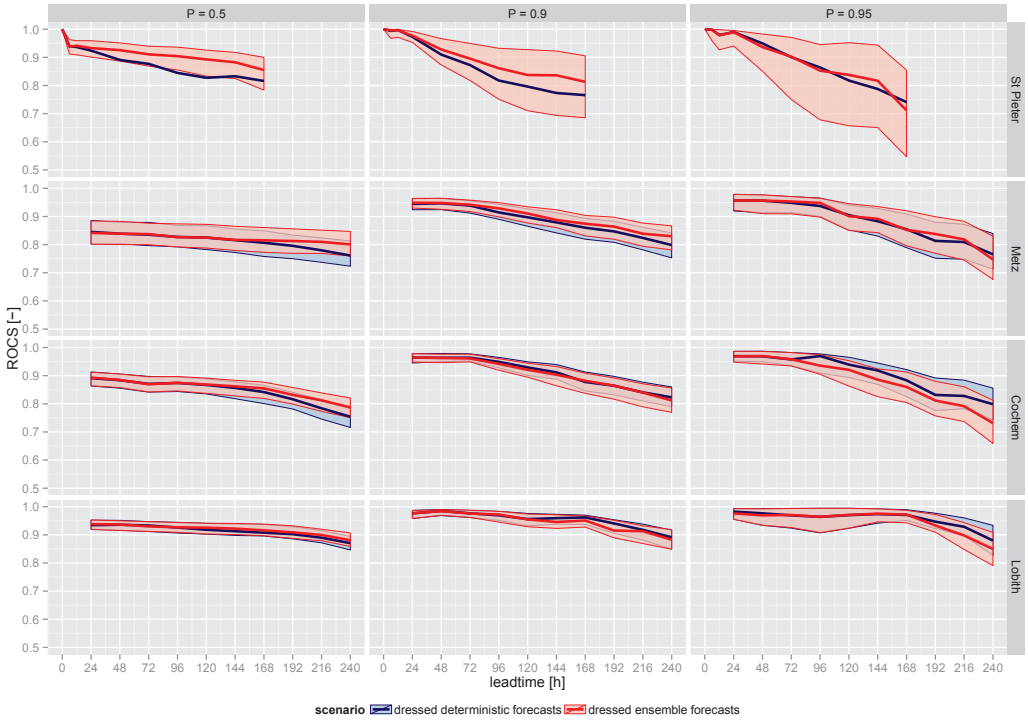


Figure 60: Relative Operating Characteristic Score (ROCS) as a function of lead time for several subsamples of the verification pairs (columns) and several locations (rows).

is offset by an unconditional wet bias, leading to reasonably reliable forecasts at higher thresholds and early forecast lead times.

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., 2013b). 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,

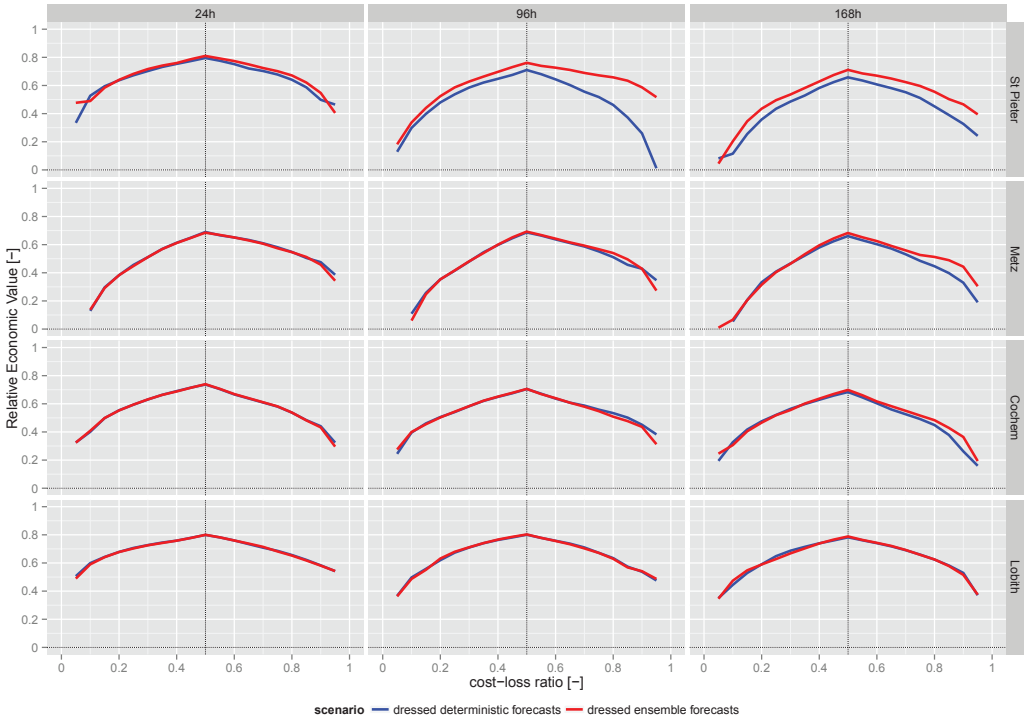


Figure 61: Value plots for various lead times (columns) for several locations (rows). The plot is conditional on the observations exceeding the 50^{th} percentile of the climatological exceedence probability (i.e., $P = 0.50$).

the dressed ensembles were both sharper and more reliable than the dressed deterministic forecasts.

5.5 CONCLUSIONS

Estimates of the predictive uncertainty in hydrological forecasts should capture all major sources of uncertainty. This can be achieved with a source-based approach or a lumped approach. We compare these two approaches 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.

While the analysis revealed quite similar results between the scenarios, further studies or different approaches to quantifying the various

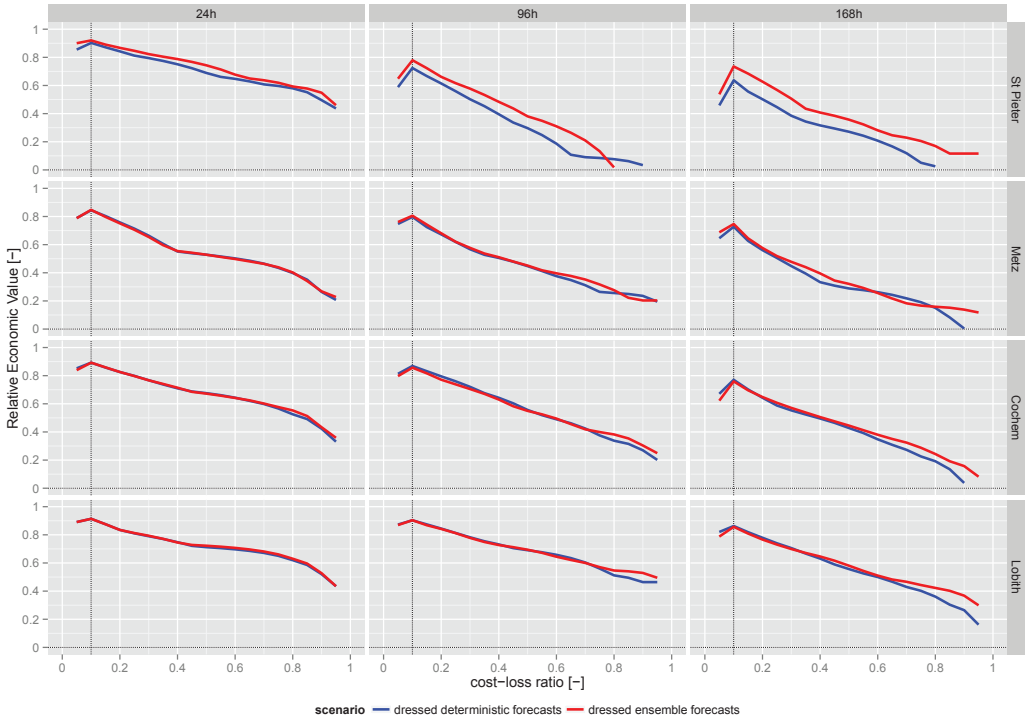


Figure 62: Value plots for various lead times (columns) for several locations (rows). The plot is conditional on the observations exceeding the 90th percentile of the climatological exceedence probability (i.e., P = 0.90).

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. Also, the meteorological biases could be addressed through meteorological post-processing or by accounting for the effects of under-dispersion on the streamflow forecasts. For example, the spread contributed by the residuals between the best streamflow ensemble member (before accounting for any hydrological uncertainties) and the simulated streamflow could be included in the ensemble dressing technique.

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 (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 Figure 47 was made available by the European Environment Agency on a Creative Commons Attribution License (WWW.EEA.EUROPA.EU).

6 SYNTHESIS

The objective of this research project was defined in Section 1.3 as *to contribute — in two distinct ways — to the use of probabilistic hydrologic forecasts in flood early warning systems: (i) by providing a valuing technique for estimating the value of probabilistic flood forecasts in terms of flood risk so that the value of flood early warning systems can be compared to the value of other risk reduction measures; and (ii) by the development of various post-processing approaches for improving the skill of probabilistic hydrological forecasts.*

The objective was addressed by means of research (sub-)questions that are addressed in the chapters 2 through 5. In the present chapter, these are revisited and conclusions are presented. Subsequently, the implications for forecasting system managers as well as decision makers are presented and remaining challenges for both researchers and practitioners are described. In the final section, some final thoughts on the topic of estimating and using predictive hydrological uncertainty are presented.

6.1 CONCLUSIONS

6.1.1 *How can the value of probabilistic forecasts be expressed in terms of flood risk?*

A theoretical framework is developed to express the value of forecasts in terms of the reduction of flood risk through flood early warning. This framework comprises several steps. First, a technique for estimating flood risk is presented, with flood risk comprising the expected (in a statistical sense) annual flood damage. Expected annual flood damage is estimated using the 'hydro-economic EAD model'. This model allows for inter-comparing flood risk management strategies regardless of whether these are structural or non-structural in nature. Thus, it also allows for estimating flood risk in a scenario where an early warning system is present, and a scenario in which such a system is absent. Then, the costs of the adverse effects of forecasting uncertainty are estimated. This is achieved by combining the estimated EAD with the 'relative economic value' (REV) of a forecasting system. This REV is essentially a function of the elements of a contingency table that, for a given decision threshold, tallies the number of hits, misses, false alarms and correct negatives. REV is used to scale the difference between expected annual damage in two scenarios. The first scenario comprises the presence of a forecasting system that is 'perfect', i.e. not adversely

affected by forecasting uncertainty. In the second scenario, there is no forecasting system present hence flood damage remains unmitigated.

The technique is used to estimate the value of flood early warning systems in a flood prone area in the headwaters of the Scottish White Cart basin. This analysis reveals that probabilistic forecasts have more value than their deterministic equivalents. The reason is that the former allow for balancing the level of certainty about the future with the required investment of a flood damage mitigation measure. The technique also allows for choosing an optimal lead time at which flood mitigating action should be taken. At this optimal lead time, the combined cost of mitigating action and both mitigated and unmitigated flood damage is lower than at other lead times. It is shown that the optimal lead time is not necessarily the longest available lead time. While increasing the lead time of a forecast can result in longer mitigation time and thus for more damage mitigation, the increased lead time also results in increased predictive uncertainty and the costs of that may be higher than the benefit — the reduction in flood damage.

6.1.2 *Can the skill of ensemble streamflow forecasts be improved by post-processing ensemble NWP temperature and precipitation forecasts?*

Precipitation and temperature ensemble forecasts are often biased, for example in mean, spread or both. In Chapter 3, the ECMWF-EPS precipitation and temperature reforecast ensembles (Hagedorn, 2008) are analysed for biases and subsequently post-processed in various ways in order to try and minimize these biases as much as possible. Post-processing techniques include the unconditional quantile-to-quantile transform, Gaussian regression (for temperature forecasts) and logistic regression (for precipitation forecasts). Subsequently, the Schaake Shuffle (Clark et al., 2004) is applied in order to restore spatial and temporal correlation structures as much as possible. Both raw and post-processed ensembles are then used to force a hydrological model of the Rhine basin, resulting in ensemble streamflow ensembles. The raw and post-processed forcing ensembles as well as the streamflow ensembles are subsequently verified and the metrics and skills are used to intercompare forecast quality.

The raw temperature and precipitation ensemble forecasts are found to be biased in both mean and spread, possibly as a result of the relatively low number of members in the ensemble. However, they are skilful with respect to sample climatology. Post-processing results in a modest increase in skill – more so for the temperature ensembles than for the precipitation ensembles. However, this skill increase does not proportionally translate to the skill of the streamflow ensembles.

The implication of these findings is that commonly used techniques for post-processing of temperature and precipitation forecasts are not

necessarily suitable for hydrological applications. Possibly, more elaborate implementations may result in larger improvements in skill of the ensemble NWP predictions. These include, for example, using additional predictors or a stratification: separately post-processing multiple sub-samples defined by certain conditions. However, the use of additional predictors may lead to overfitting and the introduction of a stratification further reduces sample size.

Physically plausible spatio-temporal relations in temperature and precipitation ensemble forecasts are important for hydrological applications. However, these are not necessarily preserved posterior to the application of a statistical post-processing technique, even when the Schaake Shuffle is applied. This reduces the quality of the resulting streamflow forecasts. Larger improvements of skill in streamflow predictions may be attained by post-processing techniques that specifically take into account spatio-temporal relations (Wilks, 2014).

Statistical post-processing requires long, consistent timeseries of forecasts and observations. While, often, reasonably long observational records are available, the same is not true for forecasts. Forecasting models improve continuously and there is a tendency of meteorological forecasting centres to include these improvements in their operationally used models. While the 'new' forecasts are likely better than before, the scope for post-processing will be reduced as the forecast record is no longer of consistent quality. Hence there is merit in 'freezing' versions of forecasting models and thus create long, consistent forecast records. for the purpose of post-processing. Similarly, there is merit in the production of reforecasts - retrospectively forecasting using a current version of a forecast model.

In the absence of possibilities of creating long, consistent timeseries, post-processing may have to be replaced by ensemble techniques. This, however, requires that any source of uncertainty can be described by a probability distribution from which multiple plausible values can be drawn. These are then routed through the models and thus form ensemble members. Currently, not all sources of uncertainty can be described in this way.

6.1.3 *Can the estimates of predictive hydrological uncertainty be improved by changing the configuration of a post-processor?*

Weerts et al. (2011, WWV2011) describe a statistical post-processor based on the Quantile Regression technique. The post-processor is applied to water level forecasts in the Upper Severn basin. The WWV2011 application is relatively straightforward: a posterior predictive distribution of water levels was derived using deterministic water levels as predictors, i.e. inputs to the statistical prediction model. This analysis was performed in Gaussian space to ensure that the joint distribution of

forecasts and observations could be described linearly. The problem of quantile crossing – resulting in non strictly rising cumulative distributions –is addressed in a very pragmatic manner, namely by manually imposing predictive distributions where crossing occurs.

Since the WWV2011 analysis was carried out, a new technique for avoiding crossing quantiles has been published. The non-crossing quantile technique can be applied for linear quantile regression only. In Chapter 4, this technique is applied to the very same study basin and data that was used for the WWV2011 analysis. At the same time, the necessity of using the Normal Quantile Transform – for transforming timeseries into Gaussian space – is questioned by testing additional configurations. These comprise an application in the original space and by an application on multiple domains of the predictor, each of which is chosen as to contain a more or less linear joint distribution of forecasts and observations. The four configurations are verified against observations and verification results are intercompared.

In the Severn case study, the simplest possible configuration of Quantile Regression yields forecasts of similar quality to those attained by more complex configurations. When trying to further improve forecast quality and skill, alternative routes may be tried. These include the use of a post-processor with alternative or multiple predictors, or both. Additional alternative implementations include the use of non-linear quantile regression models or the use of other smoothing techniques.

The uncertainty models that are derived in the four scenarios are sometimes visually quite different, yet — as reviewer Paul Smith pointed out — this is not reflected in the resulting metrics and skill scores (Smith, 2014). One reason is that the most striking visual differences are in the extreme quantiles, i.e. those that have very small or very high exceedence probabilities. While these may be highly relevant in early identification (or ruling out) of possible events, these quantiles have a very limited effect on the resulting metrics.

6.1.4 *Can the skill of raw ensemble streamflow forecasts be improved by 'dressing' the ensemble members with distributions that describe the hydrologic uncertainties?*

In Chapter 5, two approaches for estimating the 'total predictive uncertainty' are intercompared. This total uncertainty comprises uncertainty originating in both the modelling of the future state of the atmosphere as well as in the modelling of the rainfall to runoff and the streamflow propagation processes. In the 'lumped approach', the joint distribution of deterministic forecasts and their verifying observations is characterised and subsequently used to estimate predictive uncertainty about future forecasts. In the 'source specific approach', the joint distribution of hydrological simulations or perfect forecasts and their verifying

observations is used to characterise the hydrologic uncertainties. The resulting distributions are subsequently used to 'dress' raw ensemble streamflow predictions that are indicative of uncertainties in the atmospheric forecasts only. The forecasts from both approaches are verified against observations and forecast quality and skill are intercompared.

The combined Meuse and Rhine case study shows that the "source-specific approach" yields forecasts that are at least as good as those from the "lumped approach". The former, however, has more scope for further improvement through post-processing of the raw meteorological ensembles. Possibly, skill can be further increased if additional sources of uncertainty are isolated and described using ensemble techniques.

The goal of probabilistic forecasting is to maximize sharpness, given reliability (Gneiting and Katzfuss, 2014). The two scenarios that are analysed show that to some degree, there can be a trade-off between these two properties, depending on which metric is used to express quality or skill. Quality metrics that are highly sensitive to reliability — or lack thereof — will show reduced quality. Metrics that take both reliability and sharpness into account, however, may not show such a reduction. The latter is true for Relative Economic Value, for example.

Analyses described in the literature show that the skill of the dressed ensembles is significantly larger than that of the raw ensembles (e.g. Pagano et al. 2013). While these particular results are not shown, this is confirmed by the analysis. This leads to the conclusion that the hydrologic uncertainties matter and should not be ignored which is the case in many operational ensemble streamflow forecasting systems (Pappenberger, 2013).

6.1.5 *Can statistical post-processing further improve the skill of probabilistic forecasts?*

This is the overall research question that comprises sub-questions 2a — 2c. Table 9 summarizes the approaches taken.

Of the three approaches taken here, none resulted in significant changes in the skill of hydrologic probabilistic forecasts. Does that mean that forecast skill cannot be improved through statistical post-processing? Not necessarily. While the range of approaches taken here is reasonably wide, there are still ample additional approaches that may be tried and tested. Some of the results obtained here may guide those developments. For example, the post-processing of temperature and precipitation forecasts did not result in significant skill improvements, mainly — or so we believe — because of the change in spatio-temporal relations due to the post-processing techniques used. Application of a technique that is better able to preserve those spatio-temporal relations may well result in skill improvements. In addition,

Table 9: Summary conclusions of the three chapters comprising post-processing experiments

Chapter	3	4	5
Study basin	Rhine	Severn	Rhine and Meuse
Post-processed variables	Temperature and Precipitation	River stage	Streamflow
Baseline for verification	Temperature and precipitation: observations Streamflow: simulations	Observations	Observations
Techniques	Temperature: Quantile–to–quantile transform and linear regression; Precipitation: Quantile–to–quantile transform and logistic regression	Various configurations of Quantile Regression in combination with deterministic forecasts	Quantile regression in combination with (i) deterministic forecasts and (ii) ensemble forecasts
Conclusions	Forcing forecasts increase in skill following post-processing. This increase does not proportionally propagate to streamflow forecasts.	Raw QR forecasts are skilful; alternative configurations do not further add skill.	Dressed ensemble forecasts are at least as skilful as the dressed deterministic forecasts.

the approaches taken to post-process meteorological forecasts and hydrological forecasts in Chapters 3 and 5 made use of one single predictor: a single valued forecast. The use of additional predictors was not tested. In the case of the meteorological forecasts, the use of additional variables such as atmospheric pressure may well carry an additional signal that can be used to improve on the estimates of future precipitation, for example. The same is true for hydrological forecasts: the hydrologic behavior of a basin likely depends on that basin's initial conditions, hence conditioning a post-processor on that additional predictor could possible improve skill of the resulting forecast. One question – implicitly addressed to some degree in Chapter 5 – that is still open is when to use lumped and when to use source-specific approaches. The results show that the skill of a combination of these two approaches is similar to that of lumping - even when there are obvious ways to try and improve the reliability of the forecasts that are combined.

These examples show that there are still promising routes for improving forecast skill using post-processors. However, in all cases one needs to keep in mind the assumptions underlying the use of post-processing techniques. These mainly include the assumption of stationarity. Joint distributions of predictors and predictands that are observed in the past are deemed to be similar in the future also. Additional points of attention include the homogeneity of the joint sample, that is: do all observations in the relevant domain adhere to the mathematical description? Often, this is not the case of extreme values.

6.2 IMPLICATIONS

The conclusions presented in this dissertation have a number of potential implications for forecasting system managers and decision makers.

The analysis of potential value shows that probabilistic forecasts have higher value than their deterministic equivalents. Essentially, the probabilities comprise additional information that has value. The probabilistic forecasts themselves are not any less certain, but the degree of certainty is shown rather than obscured. Using the probabilities then constitutes, cf Pielke (2011), a choice for decision making under uncertainty over decision making under ignorance.

The value analysis also shows that the most optimal lead time for decision-making is not necessarily the longest lead time available. While longer lead times go hand in hand with longer mitigation times and thus allows for more damage mitigating actions, longer lead times also mean that the forecasts are more uncertain. An analysis of the cost of uncertainty versus the benefits of longer mitigation time can help decision makers in choosing an optimal balance.

The review of flood management practices following the 2007 summer floods in England and Wales suggested that flood warnings

should be issued against lower thresholds of probability (Pitt, 2008). This may indeed benefit *some* users but certainly not all. Any decision maker should be aware of her own particular situation in terms of costs and benefits of initiating flood damage mitigation. It is reasonable to assume that many decision makers are not aware of their own cost-to-loss ratio, or of how to incorporate predictive uncertainty in decision making for that matter. Some guidance by the scientific community may be required in this respect.

Increasingly often, forecasting systems constitute Hydrological Ensemble Prediction Systems (HEPSs), where ensemble NWP models are propagated through hydrological models to arrive at a streamflow ensemble. While this constitutes an improvement with respect to the production of deterministic forecasts only, the raw NWP ensembles are likely to be biased in mean, spread or both, and these biases will propagate to the streamflow ensembles. If anything, decision makers should be (made) aware of the presence of these biases. Better even, an attempt should be made to remove or reduce them – even though, as was shown in Chapter 3, this may prove to be difficult.

While, almost by construction, HEPSs take into account uncertainty originating in the predicted forcings (i.e. in the weather forecast), many do not account for hydrological uncertainties, i.e. those originating in the modelling of rainfall to runoff processes as well as in streamflow propagation. These hydrological uncertainties are not negligible and should be incorporated in the estimate of total uncertainty to attain more reliable forecasts. In the absence thereof, decision makers should be (made) aware of this lack of reliability due to the incomplete modelling of uncertainties.

In operational practice, many estimates of predictive uncertainty are – at least in part – based on statistical post-processing techniques. While it has been shown that using these techniques can result in skilful forecasts, there are some underlying assumptions that should be taken into account when using the forecasts in operational practice. One major assumption is that past forecast performance is assumed to remain unchanged in the future. This is especially important when this past performance is identified on the basis of relatively few data points. That is almost always the case for extreme events – that by definition do not feature often in the available records – but also for very high or very low (non-)exceedence probabilities. In these domains, care should be made when interpreting the forecasts.

Given these cautions, there may be merit in moving towards ensemble techniques, thus phasing out the use of post-processing techniques for uncertainty estimation. In principle, ensemble techniques should be better able to estimate predictive uncertainty in extreme situations – assuming that the underlying process models capture the behaviour of the physical system in extreme situations reasonably well. Also, the

output from ensemble models constitute 'plausible traces' in that it its temporal and spatial correlation is physically plausible. This is not necessarily so in the case of post-processsing techniques.

The analysis of alternating configurations of Quantile Regression suggested that the simplest possible configuration yielded forecasts with similar quality to those produced by more complicated configurations. Given the requirement that forecasts have to be consistent with the beliefs of the forecaster, the implication of this may be that the simplest configurations are not only as good as other configurations, but better. More complicated configurations may be less well understood by a forecaster, hence it is unlikely that the outcomes will be consistent with her beliefs.

6.3 REMAINING CHALLENGES

Addressing the research questions has resulted in some worthwhile insights. Obviously – and thankfully – additional challenges remain, both for scientists and practitioners. Some of these are described in the present section.

6.3.1 *Remaining challenges pertaining to value estimation*

The benefits of flood forecasting are estimated using a framework that comprises a flood risk estimation technique and the cost-to-loss framework. The flood risk estimation model is deterministic in nature and could benefit from an analysis of the uncertainties contained therein. Also, the estimated flood damage comprises direct, tangible damage also, whereas indirect and intangible damage may be considerable. Finally, flood damage was modelled to be a function of inundation depth only, whereas in reality other flood hazard characteristics – such as flow velocity, sediment content and rate of rise – are likely to contribute to flood damage also. Hence the remaining challenge comprises extension of the framework to include additional damage types as well as damage estimates resulting from additional hazard characteristics.

The risk-based cost-to-loss decision framework is conceptually very simple yet its use in operational practice is not trivial. It requires that decision makers are familiar with the framework and know how to use it. The framework rests on the assumption of rational decision making, which is known to be problematic. We know, for example, that humans can be risk averse or risk seeking depending on how consequences of actions are presented (Kahneman and Tversky, 1979). It is also unknown if decision makers would be willing to apply rational decision theory to decisions that they know will have to be taken very infrequently – which is the case in extreme events. Indeed, the principle of risk-based decision making is that some optimum will be obtained

over a high number of decisions only. Finally, the framework can only be applied if both costs and damage reduction are known, whereas in reality these are likely to be unknown or in any case highly uncertain. The implication of this is twofold: (i) incorporation of uncertainties in the estimation of costs and benefits requires additional research; and (ii) possibly statistical decision making will have to make way for decision frameworks that address different types of uncertainties altogether. These could be borrowed from approaches used for long-term planning such as robust and flexible decision making (see Haasnoot, 2013, for examples). Both implications require additional research.

6.3.2 *Remaining challenges pertaining to uncertainty estimation*

Chapters 3, 4 and 5 discuss techniques for statistically post-processing forecasts. These techniques assume that past forecast behaviour – with respect to their verifying observations – is representative of future behaviour. This assumption may be violated in a two ways:

First, the joint distribution of forecasts and observations is unlikely to be stationary if any of the marginal distributions change. While water management practice is largely built on the assumption of stationarity – meaning that the marginal distribution of observations is time invariant – it is becoming increasingly clear that this assumption is invalid (Milly et al., 2008). Rivers and flood risk are affected by both human interference and shifting climates, in ways that are not always clear. In addition, climate change may affect the hydrological response to changing forcings. It is unlikely that the forecasting models will keep up and ensure an unaltered joint distribution. Milly's (2008) proposition that "Stationarity is dead" is now widely accepted, yet the use of statistical post-processing techniques has yet to find a way to manage this paradigm shift.

A second issue in the use of statistical post-processing is the small number of extreme events that are present in the observational record. This makes it difficult to extrapolate the observed behaviour (of a forecasting system) to the future with a high degree of certainty. This may potentially be resolved by artificially extending records of observation, for example by pooling data from additional, similar forecasting locations.

Chapter 3 showed that the inability to preserve spatio-temporal forecast correlations may render post-processed forcing forecasts less useful for hydrological applications. The Schaake Shuffle (Clark et al., 2004) goes some way towards restoring these correlations. However, there may be merit in the development of a post-processing technique that maintains the correlations, thus obviating the necessity to try and restore these after the fact.

Statistical post-processing for uncertainty estimation does not produce plausible traces, i.e. forecast hydrographs that show temporal dynamics consistent with physical behaviour. This prevents these techniques from being used if the outcomes have to be re-used in 'downstream' models. Some authors have proposed techniques for this (see, for example, Regonda et al. 2013) and testing the validity of these approaches on basins with varying characteristics constitutes a worthwhile challenge.

The application of Monte Carlo (ensemble) techniques for uncertainty estimation does preserve plausible traces. However, describing all sources of uncertainty this way requires that every source of uncertainty can be described with a probability distribution from which samples can be drawn. This is reasonably well possible for uncertainties originating in future atmospheric forcings (Cloke and Pappenberger, 2009) and sometimes also for uncertainties originating in the estimation of model parameter values (De Wit and Buishand, 2007). For other sources of uncertainty, however, the suitability of Monte Carlo is less obvious. How, for instance, can the entire space of possible models be described? Or the human behaviour that affects streamflow patterns in the presence of weirs and dams? And even if this were possible, the "curse of dimensionality" would make the computational burden possibly too big to be resolved in an operational setting.

In hydrological forecasting, predictive uncertainty may be estimated by lumped and by source-specific approaches. These are essentially the ends of a scale; in-between approaches may exist, such as the 'ensemble dressing' technique that was used in chapter 4. The question of when lumped approaches and when source-specific approaches are justified is not fully addressed.

Scientists have thought of many ways to address the estimation of predictive hydrological uncertainty. Some of these are relatively simple, others can be very complex. If any of these approaches are to be used operationally, the forecasters must fully understand the techniques to take ownership of it. This move from science to operations constitutes a challenge in itself.

6.4 CLOSURE

Decisions pertaining to the future benefit from an estimate of what that future will hold. The future, however, is inherently uncertain. Forecasts contribute to reducing that uncertainty. Consequently, there is a large demand for forecasts in fields ranging from military planning to business planning to earth sciences such as meteorology and hydrology. In these earth sciences disciplines, this demand has been met by enormous progress in the science of forecasting as well as by the de-

velopment of the computational resources that are required to produce forecasts in real-time.

For many years, forecasts have been deterministic in nature: for any location and point in time, a single estimate of future conditions was given. Possibly, forecasters had a high degree in confidence in the science underlying the forecasts. Equally likely is that 'determinism' was fueled by many forecast users' requests (not to say demands) for certainty. The reason may also have been purely pragmatic in nature as computational resources only allowed for a single forecast run.

Whatever the reason, the forecasting communities may have promised too much by issuing deterministic forecasts only. The implicit promise of certainty has been widely challenged since Lorenz (1969) and Epstein (1969) recognized that that the atmosphere constitutes a chaotic system. Forecasting the atmosphere is essentially an initial value problem (Inness and Dorling, 2013) that is very sensitive to the initial conditions used in the forecast run. As it is impossible to accurately determine the initial conditions of the global atmosphere at any point in time, meteorological forecasts cannot be completely accurate. By construction, neither can the hydrologic forecasts for which the meteorological forecasts constitute inputs. Also by construction, the models used in earth sciences are simplifications of reality and will therefore be able to eliminate uncertainty altogether even if forcings were certain. Hence the notion that forecasts should be accompanied by a measure of the remaining uncertainty – already voiced by the Australian "government astronomer" W. Ernest Cooke as early as 1906 – has become increasingly accepted in communities of scientists and practitioners alike.

Since the seminal works of Lorenz and Epstein, ample resources have been devoted to the development of uncertainty estimation through a mixture of ensemble prediction systems and statistical post-processing techniques. Especially the former have been made possible through the massive advances in computational speed. Currently, most of the ensemble prediction systems that are in use around the globe are skilful (Buizza, 2014). Much of the techniques developed in the atmospheric sciences have found their way to the hydrologic sciences and there has been a significant move towards the use of hydrologic ensemble predictions systems (Cloke and Pappenberger, 2009).

There is still ample room for improvement of forecast skill. However, the time may have come to start thinking about how to transform this skill into value, that is, a benefit for the forecast user.

> "THERE IS VALUE IN IMPROVING THE QUALITY OF PROBABILITY FORECASTS, BUT THE VALUE OF IMPROVING THE ACTUAL USAGE THEREOF IS AN ORDER OF MAGNITUDE HIGHER."
>
> Robert Hartman, Hydrologist-in-Charge, US NWS California-Nevada River Forecast Centre, October 2012

Value is realised in the response to a forecasted hydrological hazard, i.e. at the very end of the forecast – decision – response chain. As the chain is only as strong as its weakest link, each of the elements of that chain will have to be prepared for using probabilistic forecasts.

> "ADDITIONAL EFFORT IS REQUIRED IN THE COMMUNICATION, VISUALIZATION, AND EVALUATION OF PROBABILISTIC FORECASTS ... TO AVOID THE RISK OF PERFECTING ENSEMBLE METHODOLOGIES WITHOUT A CLEAR AIM."
>
> Gneiting and Raftery (2005)

This requires expertise from a diverse range of disciplines including — but probably not limited to — communication and visualization, cognitive processing, decision-making and behavioural sciences. Such an effort is most likely to be successful if expert communities commit themselves to it. Indeed, the HEPEX community is gearing up towards that next step. Once that step has been taken, the full value of probabilistic forecasts will be showcased and this will further strengthen the role of uncertainty estimation within operational systems around the world.

Hence it seems appropriate to end this dissertation with a slightly adapted quote from a recent presentation (Buizza, 2014) by one of the lead model developers at ECMWF,

> "PROBABILISTIC FORECASTS ARE HERE TO STAY".

A

POST-PROCESSING TECHNIQUES

QUANTILE-TO-QUANTILE TRANSFORM

The quantile-to-quantile transform, also known as quantile mapping or cdf matching, is given by

$$x_{qqt,c} = \bar{F}_Y^{-1}\left(\bar{F}_{X_c}(x_c)\right), \tag{A.1}$$

where $\bar{F}_Y$ denotes the sample climatology of the predictand Y, or the empirical distribution of observations, $\bar{F}_{X_c}$ denotes the sample climatology of the predictor X_c and $x_{qqt,c}$ represents the quantile-to-quantile transformed prediction for the c^{th} member of the C–member forcing ensemble. Thus, the transform is applied to each of the C members and their C separate, but practically identical, climatologies. In general, $x_{qqt,c}$ will not map linearly to x_c, because the curvatures of $\bar{F}_Y$ and $\bar{F}_{X_c}$ are different.

LINEAR REGRESSION

Given a training data set, a simple linear regression relation is assumed to exist between observed temperature and the mean of the ensemble prediction (Wilks, 2011),

$$Y = \beta_0 + \beta_1 \bar{X} + \varepsilon, \tag{A.2}$$

where β_0 and β_1 are regression parameters to estimate and ε is a stochastic residual. This relation is sought for each location and lead time separately but subscripts denoting these are omitted from Equation A.2. The regression coefficients are found by minimising the expected square difference between the temperatures predicted by the model and observed. The regression constants are determined for each lead time and location separately.

The residuals are assumed to be Normally distributed with zero mean, μ,

$$\varepsilon = N(\mu = 0, \sigma), \tag{A.3}$$

and σ given by the sample standard deviation of errors.

From this regression equation, probabilistic temperature forecasts are produced for a given value of the raw ensemble mean, $\bar{x}$, by sampling from $N(\beta_0 + \beta_1 \bar{x}, \sigma)$.

LOGISTIC REGRESSION

The conditional probability that the future amount of precipitation, Y, does not exceed a discrete threshold, y, given the raw ensemble mean, $\bar{x}$, is

$$P\left(Y \leqslant y | \bar{X} = \bar{x}\right) = \frac{1.0}{1.0 + \exp^{-\left(\beta_0 + \beta_1 \bar{x}^{1/3}\right)}}, \tag{A.4}$$

where β_0 and β_1 are the parameters of the linear model to estimate through maximum likelihood. The power transformation has the effect of allowing the precipitation forecast data to be more normally distributed (Hamill et al., 2008). Similar to the experiment described in Sloughter et al. (2007), a one-third power transformation is used. In Chapter 3, 200 thresholds are considered. The thresholds are then interpolated using a spline constrained to be a valid cumulative distribution function using the method described by He and Ng (1999).

QUANTILE REGRESSION

Quantile Regression (QR; Koenker and Bassett Jr, 1978; Koenker and Hallock, 2001; Koenker, 2005) 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 dissertation, Quantile Regression is used to estimate lead time n specific conditional distributions of either streamflow or water level; in any case the dependent variable is indicated by Y,

$$\phi_n = \left\{Y_{n,\tau_1}, Y_{n,\tau_2}, \ldots, Y_{n,\tau_T}\right\} \tag{A.5}$$

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.

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 (independent) hydrologic forecast X and its (dependent) verifying observation Y,

$$Y_{n,\tau} = a_{n,\tau} X_n + b_{n,\tau} \tag{A.6}$$

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,

$$\min \sum_{j=1}^{J} \rho_{n,\tau} \left(y_{n,j} - \left(a_{n,\tau} x_{n,j} + b_{n,\tau} \right) \right) \tag{A.7}$$

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

In the research described in the present dissertation, solving equation A.7 was done using the `quantreg` package (Koenker, 2013) in the R software environment (R Core Team, 2013).

B

VERIFICATION METRICS

For ease of reference, the probabilistic verification metrics used in this dissertation are briefly explained; this description is based on Brown and Seo (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).

RELATIVE MEAN ERROR

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

$$\mathrm{RME} = \frac{\sum_{i=1}^{J} \left(\bar{X}_i - Y_i\right)}{\sum_{i=1}^{J} Y_i}, \tag{B.1}$$

where Y is the observation and $\bar{X}$ 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.

BRIER SCORE AND BRIER SKILL SCORE

For a given binary event, such as the exceedence of a flood threshold, the (half) Brier score (BS, Brier 1950) measures the mean square error of J predicted probabilities that X exceeds a threshold x,

$$\mathrm{BS} = \frac{1}{J} \sum_{i=1}^{J} \left\{F_{X_i}(x) - F_{Y_i}(x)\right\}^2, \tag{B.2}$$

where $F_{X_i}(x) = \Pr[X_i > x]$ and $F_{Y_i}(x) = \begin{cases} 1 & \text{if } Y_i > x; \\ 0 & \text{otherwise} \end{cases}$.

Brier Skill Score

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 $BS_{perfect}$ (which is equal to 0) and to a reference system BS_{ref},

$$\begin{aligned} BSS &= \frac{BS - BS_{ref}}{BS_{perfect} - BS_{ref}} \\ &= \frac{BS - BS_{ref}}{0 - BS_{ref}} = \frac{BS_{ref} - BS}{BS_{ref}} \\ &= 1 - \frac{BS}{BS_{ref}} \end{aligned} \tag{B.3}$$

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

Decomposition

By conditioning on the predicted probability, and partitioning over K discrete categories, the BS is decomposed into the calibration-refinement (CR) measures of Type-I conditional bias or reliability (REL), resolution (RES) and uncertainty (UNC),

$$\begin{aligned} BS = & \underbrace{\frac{1}{J}\sum_{k=1}^{K} N_k\{F_{X_k}(x) - \bar{F}_{Y_k}(x)\}^2}_{REL} \\ & - \underbrace{\frac{1}{J}\sum_{k=1}^{K} N_k\{F_{Y_k}(x) - \bar{F}_{Y}(x)\}^2}_{RES} \\ & + \underbrace{\sigma_Y^2(x)}_{UNC}. \end{aligned} \tag{B.4}$$

Here, $\bar{F}_Y(x)$ represents the average relative frequency (ARF) with which the observation exceeds the threshold x. The term $F_{Y_k}(x)$ represents the conditional observed ARF, given that the predicted probability falls within the k^{th} of K probability categories, which happens N_k times. Normalizing by the climatological variance UNC, $\sigma_Y^2(x)$, leads to the Brier Skill Score (BSS),

$$BSS = 1 - \frac{BS}{UNC} = \frac{RES}{UNC} - \frac{REL}{UNC}. \tag{B.5}$$

By conditioning on the $O = 2$ two possible observed outcomes, $\{0, 1\}$, the BS is decomposed into the likelihood-base-rate (LBR) measures of Type-II conditional bias (TP2), discrimination (DIS), and sharpness (SHA),

$$\begin{aligned} \mathrm{BS} &= \underbrace{\frac{1}{J}\sum_{o=1}^{O} N_o \left\{\bar{F}_{X_o}(x) - \bar{F}_{Y_o}(x)\right\}^2}_{\mathrm{TP2}} \\ &\quad \underbrace{- \frac{1}{J}\sum_{o=1}^{O} N_o \left\{F_{X_o}(x) - \bar{F}_X(x)\right\}^2}_{\mathrm{DIS}} \\ &\quad + \underbrace{\sigma_X^2(x)}_{\mathrm{SHA}}. \end{aligned} \tag{B.6}$$

Here, $\bar{F}_{X_o}(x)$ represents the average probability with which X is predicted to exceed x, given that Y exceeds x ($o = 1$) or does not exceed x ($o = 2$), where N_o is the conditional sample size for each case. The BSS is then given by

$$\begin{aligned} \mathrm{BSS} &= 1 - \frac{\mathrm{BS}}{\mathrm{UNC}} \\ &= 1 - \frac{\mathrm{TP2}}{\mathrm{UNC}} + \frac{\mathrm{DIS}}{\mathrm{UNC}} - \frac{\mathrm{SHA}}{\mathrm{UNC}}. \end{aligned} \tag{B.7}$$

MEAN CONTINUOUS RANKED PROBABILITY SCORE AND SKILL SCORE

The mean Continuous Ranked Probability Score (CRPS) measures the integral square difference between the cumulative distribution function (cdf) of the forecast $F_X(q)$, and the corresponding cdf of the observed variable $F_Y(q)$, averaged across J pairs of forecasts and observations,

$$\overline{\mathrm{CRPS}} = \frac{1}{J}\int_{-\infty}^{\infty} \{F_X(q) - F_Y(q)\}\,dq. \tag{B.8}$$

The mean Continuous Ranked Probability Skill Score (CRPSS) is a scaled representation of forecast quality that relates the quality of a particular system $\overline{\mathrm{CRPS}}$ to that of a perfect system $\overline{\mathrm{CRPS}}_{\mathrm{perfect}}$ (which is equal to 0) and to a reference system $\overline{\mathrm{CRPS}}_{\mathrm{ref}}$,

$$\begin{aligned} \text{CRPSS} &= \frac{\overline{\text{CRPS}} - \overline{\text{CRPS}}_{\text{ref}}}{\overline{\text{CRPS}}_{\text{perfect}} - \overline{\text{CRPS}}_{\text{ref}}} \\ &= \frac{\overline{\text{CRPS}} - \overline{\text{CRPS}}_{\text{ref}}}{0 - \overline{\text{CRPS}}_{\text{ref}}} = \frac{\overline{\text{CRPS}}_{\text{ref}} - \overline{\text{CRPS}}}{\overline{\text{CRPS}}_{\text{ref}}} \\ &= 1 - \frac{\overline{\text{CRPS}}}{\overline{\text{CRPS}}_{\text{ref}}} \end{aligned} \tag{B.9}$$

RELATIVE ECONOMIC VALUE

In the absence of a flood warning system (FWS), a user's flood losses will be determined by the climatological frequency of flooding and consist of unmitigated losses, which is the sum of the losses avoided through warning response L_a, and the losses that cannot be avoided L_u for every flood event e,

$$V_{\text{noFWS}} = e\ (\ L_a + L_u\ ). \tag{B.10}$$

If the FWS generates perfect forecasts, a flood event is always preceded by a warning and flood damage can always be reduced by mitigating action. False alarms and missed events do not occur. The expected damage then consists of the sum of cost for warning response and unavoidable losses for every flood event:

$$V_{\text{perfect}} = e\ (C + L_u). \tag{B.11}$$

The FWS performance based on imperfect forecasts can be assessed using a contingency table. Missed events result in unmitigated flood losses, which equal the sum of avoidable and unavoidable losses $L_a + L_u$. Loss mitigation through warning response can only be achieved at a cost C. In case of false warnings, these are the only costs incurred by a user. A user's expected costs and losses consist of those associated with hits, misses and false alarms:

$$\begin{aligned} V_{\text{FWS}} &= h\ (C + L_u) + f\ C + m\ (L_a + L_u) \\ &= e\ L_u + (h + f)\ C + m\ L_a. \end{aligned} \tag{B.12}$$

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 ($V = 1$):

$$V = \frac{V_{\text{noFWS}} - V_{\text{FWS}}}{V_{\text{noFWS}} - V_{\text{perfect}}}. \tag{B.13}$$

Note that REV can be less than 0 if the cost of false alarms is higher than the benefits attained by the warning system.

Substituting Eqs. (B.10), (B.11) and (B.12) in (B.13), subsequent division by L_a and substitution of C/L_a by the cost-loss ratio r yields:

$$\begin{aligned} V &= \frac{e\,L_a - (h + f)\,C - m\,L_a}{e\,L_a - e\,C} \\ &= \frac{e - (h + f)\,r - m}{e - e\,r} \\ &= \frac{e - (h + f)\,r - m}{e\,(1 - r)}, \end{aligned} \tag{B.14}$$

which allows for expressing V as a function of r.

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 (for example, Table 1), which is valid for a single probabilistic decision rule, and are defined as follows

$$\begin{aligned} \text{hit rate} &= \frac{\#\text{ hits}}{\#\text{ observed events}} = \frac{h}{e} \\ \text{false alarm rate} &= \frac{\#\text{ false alarms}}{\#\text{ events not observed}} = \frac{f}{e'}, \end{aligned} \tag{B.15}$$

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

$$\text{ROCS} = 2 \times (\text{AUC} - 0.5)\,. \tag{B.16}$$

BIBLIOGRAPHY

Bailey, R. and Dobson, C.: Forecasting for floods in the Severn catchment, J. Inst. Water Engrs Sci., 35, 168–178, 1981.

Bergström, S. and Singh, V. P.: The HBV model, in: Computer models of watershed hydrology, edited by Singh, V., pp. 443–476, Water Resources Publications, Highlands Ranch, Colorado, United States, 1995.

Bogner, K. and Pappenberger, F.: Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system, Water Resources Research, 47, W07 524, DOI:10.1029/2010WR009137, 2011.

Bogner, K., Pappenberger, F., and Cloke, H.: Technical Note: The normal quantile transformation and its application in a flood forecasting system, Hydrology and Earth System Sciences, 16, 1085–1094, 2012.

Bondell, H. D., Reich, B. J., and Wang, H.: R–code for Non-crossing quantile regression curve estimation, North Carolina State University, Department of Statistics, 2010a.

Bondell, H. D., Reich, B. J., and Wang, H.: Noncrossing quantile regression curve estimation, Biometrika, 97, 825 –838, DOI:10.1093/BIOMET/ASQ048, 2010b.

Boucher, M.-A., Tremblay, D., Delorme, L., Perreault, L., and Anctil, F.: Hydro-economic assessment of hydrological forecasting systems, Journal of Hydrology, 416-417, 133–144, DOI:10.1016/J.JHYDROL.2011.11.042, 2012.

Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B., and Beare, S. E.: The MOGREPS short-range ensemble prediction system, Quarterly Journal of the Royal Meteorological Society, 134, 703–722, DOI:10.1002/QJ.234, 2008.

Bradley, A. A. and Schwartz, S. S.: Summary Verification Measures and Their Interpretation for Ensemble Forecasts, Monthly Weather Review, 139, 3075–3089, DOI:10.1175/2010MWR3305.1, 2011.

Bremnes, J. B.: Probabilistic Forecasts of Precipitation in Terms of Quantiles Using NWP Model Output, Monthly Weather Review, 132, 338–347, DOI:10.1175/1520-0493(2004)132%3C0338:PFOPIT%3E2.0.CO;2, 2004.

Brier, G.: Verification of forecasts expressed in terms of probability, Monthly weather review, 78, 1–3, 1950.

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

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

Brown, J. D. and Seo, D.-J.: A Nonparametric Postprocessor for Bias Correction of Hydrometeorological and Hydrologic Ensemble Forecasts, Journal of Hydrometeorology, 11, 642–665, DOI:10.1175/2009JHM1188.1, 2010.

Brown, J. D. and Seo, D.-J.: Evaluation of a nonparametric post-processor for bias correction and uncertainty estimation of hydrologic predictions, Hydrological Processes, 27, 83–105, DOI:10.1002/HYP.9263, 2013.

Brown, J. D., Demargne, J., Seo, D.-J., and Liu, Y.: The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations, Environmental Modelling & Software, 25, 854 – 872, 2010.

Brown, T. A.: Admissible Scoring Systems for Continuous Distributions., 1974.

Buizza, R.: The value of probabilistic prediction, Atmospheric Science Letters, 9, 36–42, DOI:10.1002/ASL.170, 2008.

Buizza, R.: The rise of ensemble forecasting at ECMWF in the 10 years of HEPEX and applications, `http://hepex.irstea.fr/tenth-ann-workshop/program/#day2`, 2014.

Buizza, R., Bidlot, J.-R., Wedi, N., Fuentes, M., Hamrud, M., Holt, G., and Vitart, F.: The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System), Quarterly Journal of the Royal Meteorological Society, 133, 681–695, DOI:10.1002/QJ.75, 2007.

Carsell, K. M., Pingel, N. D., and Ford, D. T.: Quantifying the benefit of a flood warning system, Natural Hazards Review, 5, 131, 2004.

Chen, S. and Yu, P.: Real-time probabilistic forecasting of flood stages, Journal of Hydrology, 340, 63–77, 2007.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., and Wilby, R.: The Schaake Shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields, Journal of Hydrometeorology, 5, 243–262, 2004.

Cloke, H. and Pappenberger, F.: Ensemble flood forecasting: a review, Journal of Hydrology, 375, 613–626, 2009.

Coccia, G. and Todini, E.: Recent developments in predictive uncertainty assessment based on the model conditional processor approach, Hydrology and Earth System Sciences, 15, 3253–3274, 2011.

Collier, C. G., Cross, R., Khatibi, R., Levizzani, V., Solheim, I., and Todini, E.: ACTIF best practice paper – the requirements of flood forecasters for the preparation of specific types of warnings, in: ACTIF international conference on innovation advances and implementation of flood forecasting technology, Tromsø, Norway, pp. 17–19, 2005.

Cooke, E. W.: Forecasts and verifications in Western Australia, Monthly Weather Review, 34, 23–24, 1906.

Cranston, M., Werner, M. G. F., Janssen, A., Hollebrandse, F., Lardet, P., Oxbrow, J., and Piedra, M.: Flood Early Warning System (FEWS) Scotland: An example of real time system and forecasting model development and delivery best practice, in: DEFRA Conf. Flood and Coastal Management York, UK, pp. 02–3, 2007.

Dale, M., Wicks, J., Mylne, K., Pappenberger, F., Laeger, S., and Taylor, S.: Probabilistic flood forecasting and decision-making: an innovative risk-based approach, Natural Hazards, 70, 159–172, 2014.

Davis, D., Faber, B., and Stedinger, J.: USACE Experience in Implementing Risk Analysis for Flood Damage Reduction Projects, Journal of Contemporary Water Research & Education, 140, 3–14, 2008.

De Bruijn, K.: Resilience and flood risk management: a systems approach applied to lowland rivers, Ph.D. dissertation, Delft University of Technology, 2005.

De Moel, H.: Uncertainty in Flood Risk, Ph.D. dissertation, Vrije Universiteit Amsterdam, 2012.

De Wit, M. J. M. and Buishand, T. A.: Generator of rainfall and discharge extremes (GRADE) for the Rhine and Meuse basins, Rijkswaterstaat/RIZA, 2007.

DEFRA: Flood warning and forecasting, Tech. rep., Department for Environment, Food and Rural Affairs, 2004.

Deltares: The World of Deltares 2013-2016, Tech. rep., Deltares, Delft, The Netherlands, `http://www.deltares.nl/en/knowledge-and-innovation`, 2013.

Demargne, J., Wu, L., Regonda, S., Brown, J., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H. D., Fresch, M., Schaake, J., and Zhu, Y.: The Science of NOAA's Operational Hydrologic Ensemble Forecast Service, Bulletin of the American Meteorological Society, p. 130611111953000, DOI:10.1175/BAMS-D-12-00081.1, 2013.

Dingman, S.: Physical Hydrology, Prentice-Hall, Upper Saddle River, NJ, 2002.

Dogulu, N., López López, P., Solomatine, D. P., Weerts, A. H., and Shrestha, D. L.: Predicting model uncertainty using quantile regression and UNEEC methods and their comparison on contrasting catchments, Hydrology and Earth System Sciences Dis, 11, 10 179–10 233, DOI:10.5194/HESSD-11-10179-2014, 2014.

EA: Environment Agency. River levels: Midlands., `http://apps.environment-agency.gov.uk/river-and-sea-levels`, 2013.

Epstein, E. S.: Stochastic dynamic prediction, Tellus, 21, 739–759, DOI: 10.1111/J.2153-3490.1969.TB00483.X, 1969.

Fleming, G., Frost, L., Huntingtin, S., Knight, D., Law, F., and Ard, C.: Learning to live with rivers, 2001.

Fortin, V., Favre, A.-C., and Said, M.: Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member, Quarterly Journal of the Royal Meteorological Society, 132, 1349–1369, 2006.

Gigerenzer, G., Hertwig, R., Van Den Broek, E., Fasolo, B., and Katsikopoulos, K.: "30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts?, Risk analysis, 25, 623–629, 2005.

Gilleland, M.: Forecast Value Added Analysis: Step-by-Step, Tech. rep., SAS Institute, Inc, `www.sas.com/reg/web/corp/4385`, 2013.

Glahn, H. R. and Lowry, D. A.: The use of model output statistics (MOS) in objective weather forecasting, Journal of Applied Meteorology, 11, 1203–1211, 1972.

Gneiting, T. and Katzfuss, M.: Probabilistic Forecasting, Annual Review of Statistics and Its Application, 1, 125–151, DOI:10.1146/ANNUREV-STATISTICS-062713-085831, 2014.

Gneiting, T. and Raftery, A.: Weather forecasting with ensemble methods, Science, 310, 248–249, 2005.

Gneiting, T., Raftery, A., Westveld, A., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Monthly Weather Review, 133, 1098–1118, 2005.

Gouldby, B. and Samuels, P.: Language of Risk - project definitions, Tech. rep., FLOODsite consortium, www.floodsite.net, 2005.

Green, C. and Herschy, R.: Assessing the benefits of streamflow gauging, Tech. rep., Flood Hazard Research Centre, Middlesex University, 1994.

Green, D. M. and Swets, J. A.: Signal detection theory and psychophysics, John Wiley & Sons, Inc., New York, 1966.

Haasnoot, M.: Anticipating Change: Sustainable Water Policy Pathways for an Uncertain Future, Ph.D. dissertation, University of Twente, Enschede, The Netherlands, 2013.

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

Hagedorn, R., Hamill, T., and Whitaker, J.: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures, Monthly Weather Review, 136, 2608–2619, 2008.

Hamill, T. and Whitaker, J.: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application, Monthly weather review, 134, 3209–3229, 2006.

Hamill, T., Whitaker, J., and Mullen, S.: Reforecasts: An important dataset for improving weather predictions, Bull. Amer. Meteor. Soc, 87, 33–46, 2006.

Hamill, T., Hagedorn, R., and Whitaker, J.: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation, Monthly Weather Review, 136, 2620–2632, 2008.

Hamill, T. M.: Verification of TIGGE Multimodel and ECMWF Reforecast-Calibrated Probabilistic Precipitation Forecasts over the Contiguous United States*, Monthly Weather Review, 140, 2232–2252, DOI:10.1175/MWR-D-11-00220.1, 2012.

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau Jr, T. J., Zhu, Y., and Lapenta, W.: NOAA's second-generation global medium-range ensemble reforecast dataset, Bulletin of the American Meteorological Society, 94, 1553–1565, 2013.

Hantush, M. M. and Kalin, L.: Stochastic residual-error analysis for estimating hydrologic model predictive uncertainty, Journal of Hydrologic Engineering, 13, 585–596, 2008.

Hashino, T., Bradley, A., and Schwartz, S.: Evaluation of bias-correction methods for ensemble streamflow volume forecasts, Hydrology and Earth System Sciences, 11, 939–950, 2007.

Haylock, M., Hofstra, N., Klein Tank, A., Klok, E., Jones, P., and New, M.: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, J. Geophys. Res, 113, D20 119, 2008.

He, X. and Ng, P.: COBS: qualitatively constrained smoothing via linear programming, Computational Statistics, 14, 315–338, 1999.

HEPEX community: HEPEX: the hydrologic ensemble prediction experiment | community portal, www.hepex.org, 2013.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather and Forecasting, 15, 559–570, DOI:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Hopson, T. M.: Operational Flood-Forecasting for Bangladesh, Ph.D. dissertation, University of Colorado, Department of Astrophysical, Planetary, and Atmospheric Science, 2005.

Hyndman, R. J.: Backcasting in R, http://robjhyndman.com/hyndsight/backcasting/, 2014.

IFRC: World Disaster Report 2013, International Federation of Red Cross and Red Crescent Societies, Geneva, Switzerland, 2013.

Inness, P. and Dorling, S.: Operational weather forecasting, John Wiley & Sons, Hoboken, NJ, 2013.

Jarvis, A., Reuter, H., Nelson, A., and Guevara, E.: Hole-filled seamless SRTM data V4, http://srtm.csi.cgiar.org, http://srtm.csi.cgiar.org, 2008.

Jha, A. K., Bloch, R., and Lamond, J.: Cities and Flooding: A Guide to Integrated Urban Flood Risk Management for the 21st Century, The World Bank, 2012.

Jolliffe, I. T. and Stephenson, D. B., eds.: Forecast Verification: A Practitioner's Guide in Atmospheric Science, Second Edition, John Wiley & Sons, 2012.

Kahneman, D. and Tversky, A.: Prospect theory: An analysis of decision under risk, Econometrica: Journal of the Econometric Society, pp. 263–291, 1979.

Kang, T.-H., Kim, Y.-O., and Hong, I.-P.: Comparison of pre- and post-processors for ensemble streamflow prediction, Atmospheric Science Letters, 11, 153–159, DOI:10.1002/ASL.276, 2010.

Katz, R. W. and Murphy, A. H.: Economic value of weather and climate forecasts, Cambridge University Press, Cambridge, UK; New York, 1997.

Kelly, K. and Krzysztofowicz, R.: Precipitation uncertainty processor for probabilistic river stage forecasting, Water resources research, 36, 2000.

Kleiber, W., Raftery, A. E., Baars, J., Gneiting, T., Mass, C. F., and Grimit, E.: Locally Calibrated Probabilistic Temperature Forecasting Using Geostatistical Model Averaging and Local Bayesian Model Averaging, Monthly Weather Review, 139, 2630–2649, DOI:10.1175/2010MWR3511.1, 2011.

Koenker, R.: Quantile regression, Cambridge University Press, 2005.

Koenker, R.: `quantreg`: Quantile Regression, r package version 5.05, 2013.

Koenker, R. and Bassett Jr, G.: Regression quantiles, Econometrica: journal of the Econometric Society, pp. 33–50, 1978.

Koenker, R. and Hallock, K.: Quantile regression, The Journal of Economic Perspectives, 15, 143–156, 2001.

Kron, W.: Keynote lecture: Flood risk= hazard x exposure x vulnerability, in: Proceedings of the International Symposium on Flood Defence, edited by Wu, pp. 82 – 97, Science Press, New York, 2002.

Krzysztofowicz, R.: Bayesian forecasting via deterministic model, Risk Analysis, 19, 739–749, 1999.

Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, Journal of Hydrology, 249, 2–9, 2001.

Krzysztofowicz, R.: Bayesian system for probabilistic river stage forecasting, Journal of hydrology, 268, 16–40, 2002.

Krzysztofowicz, R. and Kelly, K.: Hydrologic uncertainty processor for probabilistic river stage forecasting, Water resources research, 36, 2000.

Lahiri, P.: On the impact of bootstrap in survey sampling and small-area estimation, Statistical Science, 18, 199–210, 2003.

López López, P., Verkade, J. S., Weerts, A. H., and Solomatine, D. P.: Alternative configurations of Quantile Regression for estimating predictive uncertainty in water level forecasts for the Upper Severn River: a comparison., Hydrology and Earth System Sciences, 18, 3411–3428, DOI:10.5194/HESS-18-3411-2014, 2014.

Lorenz, E. N.: The predictability of a flow which possesses many scales of motion, Tellus, 21, 289–307, DOI:10.1111/J.2153-3490.1969.TB00444.X, 1969.

Loucks, D., Van Beek, E., Stedinger, J., Dijkman, J., and Villars, M.: Water resources systems planning and management: an introduction to methods, models and applications, Paris: UNESCO, 2005.

Madadgar, S., Moradkhani, H., and Garen, D.: Towards Improved Post-processing of Hydrologic Forecast Ensembles, Hydrological Processes, 2012.

Marsh, T. and Hannaford, J.: UK hydrometric register, Hydrological data UK series. Centre for Ecology and Hydrology, Wallingford, UK, pp. 1–210, 2008.

Marsigli, C., Boccanera, F., Montani, A., Paccagnella, T., et al.: The COSMO-LEPS mesoscale ensemble system: validation of the methodology and verification, Nonlinear Processes in Geophysics, 12, 527–536, 2005.

Matheson, J. E. and Winkler, R. L.: Scoring rules for continuous probability distributions, Management science, 22, 1087–1096, 1976.

Merz, B., Kreibich, H., Schwarze, R., and Thieken, A.: Assessment of economic flood damage, Nat. Hazards Earth Syst. Sci, 10, 1697–1724, 2010.

Messner, F., Penning-Rowsell, E., Green, C., Meyer, V., Tunstall, S., and van der Veen, A.: Evaluating flood damages: guidance and recommendations on principles and methods, Final Report T09-06-01, FLOODsite, 2007.

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., and Stouffer, R. J.: Stationarity Is Dead: Whither Water Management?, Science, 319, 573–574, DOI:10.1126/SCIENCE.1151915, 2008.

Molinari, D.: Flood forecast verification to support emergency management, in: Proceedings of the 34th IAHR World Congress, Brisbane, June 26 - July 1, 2011, 2011.

Molinari, D. and Handmer, J.: A behavioural model for quantifying flood warning effectiveness, Journal of Flood Risk Management, 4:1, 23–32, 2011.

Montanari, A.: Deseasonalisation of hydrological time series through the normal quantile transform, Journal of Hydrology, 313, 274–282, 2005.

Montanari, A. and Brath, A.: A stochastic approach for assessing the uncertainty of rainfall-runoff simulations., Water Resources Research, 40, W01 106, 2004.

Montanari, A. and Grossi, G.: Estimating the uncertainty of hydrological forecasts: A statistical approach, Water resources research, 44, 2008.

Moore, R.: The probability-distributed principle and runoff production at point and basin scales, Hydrological Sciences Journal, 30, 273–297, 1985.

Moore, R., Jones, D., Bird, P., and Cottingham, M.: A basin-wide flow forecasting system for real time flood warning, river control and water management, in: International Conference on River Flood Hydraulics, edited by W.R., W., pp. 21–30, John Wiley & Sons, Ltd., UK, 1990.

Murphy, A.: A New Vector Partition of the Probability Score, Journal of Applied Meteorology, 12, 595–600, 1973.

Murphy, A.: Decision Making and the Value of Forecasts in a Generalized Model of the Cost-Loss Ratio Situation, Monthly Weather Review, 113, 362–369, 1985.

Murphy, A.: What is a good forecast? An essay on the nature of goodness in weather forecasting, Weather and Forecasting, 8, 281–293, 1993.

Murphy, A. and Ehrendorfer, M.: On the relationship between the accuracy and value of forecasts in the cost-loss ratio situation, Weather and forecasting, 2, 243–251, 1987.

Murphy, A., Lichtenstein, S., Fischhoff, B., and Winkler, R.: Misinterpretations of precipitation probability forecasts., Bulletin of the American Meteorological Society, 61, 695–701, 1980.

NHRC, R. F.: The business of warning, Risk Frontiers Quarterly Newsletter, 1:3, 2002.

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

O'Connor, J. E. and Costa, J. E.: The world's largest floods, past and present: their causes and magnitudes, US Geological Survey, Reston, Virginia, 2004.

Oxford University Press: Oxford Dictionaries, http://www.oxforddictionaries.com/, 2014.

Pagano, T. C., Shrestha, D. L., Wang, Q. J., Robertson, D., and Hapuarachchi, P.: Ensemble dressing for hydrological applications, Hydrological Processes, 263, 106–116, DOI:10.1002/HYP.9313, 2013.

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., and Verkade, J. S.: Challenges of Operational River Forecasting, Journal of Hydrometeorology, DOI:10.1175/JHM-D-13-0188.1, 2014.

Pappenberger, F.: Operational HEPS systems around the globe | HEPEX, http://hepex.irstea.fr/operational-heps-systems-around-the-globe/, 2013.

Pappenberger, F., Thielen, J., and Del Medico, M.: The impact of weather forecast improvements on large scale hydrology: analysing a decade of forecasts of the European Flood Alert System, Hydrological Processes, 25, 1091–1113, 2011.

Parker, D.: The damage-reducing effects of flood warnings, Tech. rep., Flood Hazard Research Centre, Middlesex University, 1991.

Parker, D. and Fordham, M.: An evaluation of flood forecasting, warning and response systems in the European Union, Water Resources Management, 10, 279–302, 1996.

Parker, D., Priest, S., Schildt, A., and Handmer, J.: Modelling the damage reducing effects of flood warnings, Final Report T10-07-12, FLOODsite, 2008.

Penning-Rowsell, E., Johnson, C., Tunstall, S., Tapsell, S., Morris, J., Chatterton, J., Green, C., Wilson, T., Koussela, K., and Fernandez-Bilbao, A.: The benefits of flood and coastal risk management: a manual of assessment techniques, Middlesex University Press, London, 2005.

Photiadou, C. S., Weerts, A. H., and van den Hurk, B. J. J. M.: Evaluation of two precipitation data sets for the Rhine River using streamflow simulations, Hydrology and Earth System Sciences, 15, 3355–3366, DOI:10.5194/HESS-15-3355-2011, 2011.

Pielke, R.: Certain Ignorance versus Uncertain Uncertainty, http://rogerpielkejr.blogspot.nl/2011/08/certain-ignornace-versus-uncertain.html, 2011.

Pielke Jr, R. A. and Downton, M. W.: Precipitation and Damaging Floods: Trends in the United States, 1932—97, Journal of Climate, 13, 3625–3637, 2000.

Pitt, M.: Learning lessons from the 2007 floods (The Pitt review), London: Cabinet Office, 2008.

Politis, D. N. and Romano, J. P.: The stationary bootstrap, Journal of the American Statistical Association, 89, 1303–1313, 1994.

Quantum GIS Development Team: Quantum GIS Geographic Information System, 2012.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013.

Raftery, A., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, Monthly Weather Review, 133, 1155–1174, 2005.

Raiffa, H. and Schlaifer, R.: Applied Statistical Decision Theory, Harvard University Press, 1961.

Ramos, M. H., van Andel, S. J., and Pappenberger, F.: Do probabilistic forecasts lead to better decisions?, Hydrology and Earth System Sciences Discussions, 9, 13 569–13 607, DOI:10.5194/HESSD-9-13569-2012, 2012.

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

Reggiani, P. and Weerts, A. H.: A Bayesian approach to decision-making under uncertainty: An application to real-time forecasting in the river Rhine, Journal of Hydrology, 356, 56–69, 2008a.

Reggiani, P. and Weerts, A. H.: Probabilistic quantitative precipitation forecast for flood prediction: An application, Journal of Hydrometeorology, 9, 76–95, 2008b.

Reggiani, P., Renner, M., Weerts, A. H., and Van Gelder, P. H. A. J. M.: Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system, Water Resources Research, 45, 2009.

Regonda, S. K., Seo, D.-J., Lawrence, B., Brown, J. D., and Demargne, J.: Short-term ensemble streamflow forecasting using operationally-produced single-valued streamflow forecasts – A Hydrologic Model Output Statistics (HMOS) approach, Journal of Hydrology, 497, 80–96, DOI:10.1016/J.JHYDROL.2013.05.028, 2013.

Roscoe, K. L., Weerts, A. H., and Schroevers, M.: Estimation of the uncertainty in water level forecasts at ungauged river locations using quantile regression, International Journal of River Basin Management, 10, 383–394, 2012.

Roulin, E.: Skill and relative economic value of medium-range hydrologicalensemble predictions, Hydrology and Earth System Sciences, 11, 725–737, DOI:10.5194/HESS-11-725-2007, 2007.

Roulston, M. S. and Smith, L. A.: Combining dynamical and statistical ensembles, Tellus A, 55, 16–30, DOI:10.1034/J.1600-0870.2003.201378.X, 2003.

Schaake, J., Hamill, T., Buizza, R., and Clark, M.: HEPEX: the hydrological ensemble prediction experiment, Bulletin of the American Meteorological Society, 88, 1541–1547, 2007.

Schellekens, J., Weerts, A. H., Moore, R. J., Pierce, C. E., and Hildon, S.: The use of MOGREPS ensemble rainfall forecasts in operational flood forecasting systems across England and Wales, Advances in Geosciences, 29, 77–84, 2011.

Schmeits, M. J. and Kok, K. J.: A Comparison between Raw Ensemble Output, (Modified) Bayesian Model Averaging, and Extended Logistic Regression Using ECMWF Ensemble Precipitation Reforecasts, Monthly Weather Review, 138, 4199–4211, DOI:10.1175/2010MWR3285.1, 2010.

Sene, K., Weerts, A., Beven, K., Moore, R., Whitlow, C., Beckers, P., Minett, A., Winsemius, H., Verkade, J. S., Young, P., et al.: Risk-based Probabilistic Fluvial Flood Forecasting for Integrated Catchment Models: Phase 2-final. Science Report–SR SC080030, 2009.

Seo, D., Herr, H., and Schaake, J.: A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction, Hydrology and Earth System Sciences Discussions, 3, 1987–2035, 2006.

Seo, D. J. and Demargne, J.: Use of Ensembles in Operational Hydrologic Forecasting in NWS, in: 4th Ensemble User Workshop, NCEP, College Park, Maryland, 2008.

Sloughter, J., Raftery, A., Gneiting, T., and Fraley, C.: Probabilistic quantitative precipitation forecasting using Bayesian model averaging, Monthly Weather Review, 135, 3209–3220, 2007.

Slovic, P.: Perception of risk, Science, 236, 280–285, 1987.

Smith, P., Panziera, L., and Beven, K.: Forecasting flash floods using data-based mechanistic models and NORA radar rainfall forecasts, Hydrological Sciences Journal, pp. 1–15, 2014.

Smith, P. J.: HESSD - Interactive Discussion - Alternative configurations of Quantile Regression for estimating predictive uncertainty in water level forecasts for the Upper Severn River:

a comparison, http://www.hydrol-earth-syst-sci-discuss.net/11/3811/2014/hessd-11-3811-2014-discussion.html, 2014.

Solomatine, D. P. and Shrestha, D. L.: A novel method to estimate model uncertainty using machine learning techniques, Water Resources Research, 45, 2009.

Spiegelhalter, D., Pearson, M., and Short, I.: Visualizing uncertainty about the future, Science, 333, 1393–1400, 2011.

StackExchange: What's the difference between probability and statistics?, accessed November 12, 2014, http://stats.stackexchange.com/questions/665/whats-the-difference-between-probability-and-statistics, 2010.

Stanski, H., Wilson, L., and Burrows, W.: Survey of common verification methods in meteorology, World Meteorological Organization Geneva, 1989.

Todini, E.: Role and treatment of uncertainty in real-time flood forecasting, Hydrological Processes, 18, 2743–2746, 2004.

Todini, E.: A model conditional processor to assess predictive uncertainty in flood forecasting, International Journal of River Basin Management, 6, 123–138, 2008.

Unger, D., van den Dool, H., O'Lenic, E., and Collins, D.: Ensemble regression, Monthly Weather Review, 137, 2365–2379, 2009.

UNISDR: Guidelines for Reducing Flood Losses, 2004.

USACE: Framework for estimating national economic benefits and other beneficial effects of flood warning and preparedness systems, Tech. rep., United States Army Corps of Engineers, Institute for Water Resources, Alexandria, Virginia, United States, 1994.

Van Andel, S. J., Weerts, A. H., Schaake, J., and Bogner, K.: Post-processing hydrological ensemble predictions intercomparison experiment, Hydrological Processes, 27, 158–161, 2013.

Van Asselt, M.: Perspectives on uncertainty and risk, Dordrecht: Kluwer, 2000.

Van Dantzig, D. and Kriens, J.: Het economisch beslissingsprobleem inzake de beveiliging van Nederland tegen stormvloeden, in: Rapport Deltacommissie: Eindverslag en Interimadviezen, edited by Maris, A., De Blocq van Kuffeler, V., Harmsen, W., Jansen, P., Nijhoff, G., Thijsse, J., Verloren van Themaat, R., De Vries, J., and Van der Wal, L., Deltacommissie, 1960.

Van Steenbergen, N., Ronsyn, J., and Willems, P.: A non-parametric data-based approach for probabilistic flood forecasting in support of uncertainty communication, Environmental Modelling & Software, 33, 92–105, 2012.

Vaughan, M.: Probabilistic flood forecasting, in: 85th European Study Group with Industry (ESGI), University of East Anglia, Norwich, United Kingdom, 2012.

Verkade, J. S. and Werner, M. G. F.: Estimating the benefits of single value and probability forecasting for flood warning, Hydrology and Earth System Sciences, 15, 3751–3765, DOI:10.5194/HESS-15-3751-2011, 2011.

Verkade, J. S., Brown, J. D., Reggiani, P., and Weerts, A. H.: Dataset for Verkade et al., 2013. Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. Journal of Hydrology 501, 73–91, DOI:10.4121/UUID:56637037-8197-472B-B143-2F87ADF49ABC, 2013a.

Verkade, J. S., Brown, J. D., Reggiani, P., and Weerts, A. H.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, Journal of Hydrology, 501, 73–91, DOI:10.1016/J.JHYDROL.2013.07.039, 2013b.

Wallingford: Wallingford Water. A flood forecasting and warning system for the river Soar, Wallingford Water, 1994.

Wallingford: HR Wallingford. User Manual, vol.2. Hydraulic Unit. Reference, Halcrow/HR, HR Wallingford, 1997.

Weerts, A. H., Winsemius, H. C., and Verkade, J. S.: Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales), Hydrology and Earth System Sciences, 15, 255–265, 2011.

Werner, K., Verkade, J. S., Pagano, T. C., and Welles, E.: Handbook of Hydrometeorological Ensemble Forecasting, chap. Application and Communication of Verification Information, pp. –, Springer, 2015.

Werner, M., Cranston, M., Harrison, T., Whitfield, D., and Schellekens, J.: Recent developments in operational flood forecasting in England, Wales and Scotland, Meteorological Applications, 16, 13–22, 2009.

Werner, M. G. F.: Spatial flood extent modelling. A performance based comparison, Ph.D. dissertation, Delft University of Technology, 2004.

Werner, M. G. F. and Cranston, M.: Understanding the Value of Radar Rainfall Nowcasts in Flood Forecasting and Warning in Flashy Catchments, Meteorological Applications, 16, 41–55, 2009.

Werner, M. G. F., Schellekens, J., Gijsbers, P. J. A., Van Dijk, M. J., Van den Akker, O., and Heynert, K. V.: The Delft-FEWS flow forecasting system, Environmental Modelling & Software, 40, 65–77, DOI: 10.1016/J.ENVSOFT.2012.07.010, 2013.

White, G. F.: Human Adjustement to Floods: A Geographical Approach to the Flood Problem in the United States, Ph.D. dissertation, The University of Chicago, 1942.

Wilks, D. S.: A skill score based on economic value for probability forecasts, Meteorological Applications, 8, 209–219, DOI:10.1017/S1350482701002092, 2001.

Wilks, D. S.: Extending logistic regression to provide full-probability-distribution MOS forecasts, Meteorological Applications, 16, 361–368, DOI:10.1002/MET.134, 2009.

Wilks, D. S.: Statistical methods in the atmospheric sciences, Academic Press, Oxford, United Kingdom, 2011.

Wilks, D. S.: Multivariate Ensemble-MOS using Empirical Copulas, Quarterly Journal of the Royal Meteorological Society, pp. n/a–n/a, DOI:10.1002/QJ.2414, 2014.

Wilks, D. S. and Hamill, T. M.: Comparison of Ensemble-MOS Methods Using GFS Reforecasts, Monthly Weather Review, 135, 2379–2390, DOI:10.1175/MWR3402.1, 2007.

Wood, A. W., Maurer, E. P., Kumar, A., and Lettenmaier, D. P.: Long-range experimental hydrologic forecasting for the eastern United States, J. Geophys. Res, 107, 4429, 2002.

World Meteorological Organization: Nowcasting, `http://www.wmo.int/pages/prog/amp/pwsp/Nowcasting.htm`, 2014.

Wu, L., Seo, D.-J., Demargne, J., Brown, J. D., Cong, S., and Schaake, J.: Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction, Journal of Hydrology, 399, 281–298, DOI:10.1016/J.JHYDROL.2011.01.013, 2011.

Yuan, X. and Wood, E. F.: Downscaling precipitation or bias-correcting streamflow? Some implications for coupled general circulation model (CGCM)-based ensemble seasonal hydrologic forecast, Water Resources Research, 48, 2012.

Zalachori, I., Ramos, M.-H., Garçon, R., Mathevet, T., and Gailhard, J.: Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies, Advances in Science and Research, 8, 135–141, DOI:10.5194/ASR-8-135-2012, 2012.

Zhao, L., Duan, Q., Schaake, J., Ye, A., and Xia, J.: A hydrologic post-processor for ensemble streamflow predictions, Adv. Geosci., 29, 51–59, DOI:10.5194/ADGEO-29-51-2011, 2011.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., and Mylne, K.: The economic value of ensemble-based weather forecasts, Bulletin of the American Meteorological Society, 83, 73–83, 2002.

ACKNOWLEDGEMENTS

Now at the close of a chapter in my life, I'll take a moment and reflect on how I got here, and whose invaluable support helped me navigate to this point[1].

One afternoon, late 1998, I walked into the offices of then Delft Hydraulics to apply for a temporary job as department secretary. The job turned out to be a start of a career in water management: a discipline that, until then, I had not given serious thought or consideration. However, some of the people I met showed me a compelling alternative to the business and marketing career I had in mind at the time. These notably included Karel Heynert, Hans Wesseling, Mindert de Vries, Gerben Boot, Marcel Ververs, Erik Ruijgh, Marc van Dijk, Ferdinand Diermanse and Berend Bakker - I am thankful to you all, and privileged to still be working with many of you after all those years. Eelco van Beek and Jos Dijkman deserve a special mention here: Eelco for hiring me for that first job and for continuing to provide inspiration to this very day and Jos for convincing me that commencing a Master's programme (in Civil Engineering / Water Resources Management) at 27 was maybe not such a bad idea after all.

I am very grateful to Deltares for allowing me to spend some of my time on a PhD project. 'The nod' was given by Toon Segeren and Kees Bons, whom I reported to when I started the research. Hanneke van der Klis was the perfect manager: supportive all along, critical when required and never short on good advice. As office mates, Jaap Kwadijk, Jaap Schellekens, Laurène Bouaziz and Patricia López witnessed (if not suffered) the PhD project from up close; you've all been very patient with me, for which I am thankful. Ursula van de Pol, later joined by Marjolein Schaafsma, thank you for the many helping hands along the way, and for always taking time for a chat. You play a major part in oiling the wheels of our organization – keep up the good work!

Thankfully, this project was not a solitary enterprise. Pieter van Gelder, Albrecht Weerts, Paolo Reggiani and Hanneke van der Klis comprised the 'daily supervisory committee'. Your advice has been very helpful. Han Vrijling: thank you for offering your unique perspective. I thoroughly enjoyed discussing our — sometimes differing — views on forecasting and flood risk management. Pieter van Gelder, thank you for your encouragement and kindness, always. Albrecht Weerts and Paolo Reggiani: I consider myself lucky to be able to work with such knowledgeable people. Much looking forward to continuing that. Micha Werner and James Brown, it was a tremendous pleasure

[1] I 'borrowed' this opening line from Tom Hopson's dissertation (Hopson, 2005)

to co-author papers with you. The co-operation did not only teach me a lot (this is probably an understatement) about real-time forecasting and predictive uncertainty, but also about critical thinking and writing clearly and concisely. I hope you'll find some of that reflected in the present dissertation.

Many thanks also to the members of the examination committee, for the time taken to review the dissertation and be present at the defence. I hope you have found and will find it worthwhile.

The HEPEX community of practitioners and researchers in hydrometeorological forecasting has offered inspiration and expertise in addition to fun and friendships. Florian Pappenberger, Maria–Helena Ramos, Jutta Thielen, Fredrik Wetterhall and Andy Wood: I'm looking forward to many more sessions of science and socializing.

This PhD project has spilled over into private life. I am fortunate enough to have some very good friends who have always offered kind words about my seemingly never-ending PhD project. Emma Dornan, Markus Masche, Benjamin Fischer, Pieter van Berkum, Léon Hijweege, Ruud van der Hout and Mark Godfrey: thank you for sticking around.

It's great to have a few neighbours to share the occasional drink with, and thus occasionally distract oneself from the research: Leendert, Marion, Menno, Maria, Katrijn, Rutger, Henri, and Bianca – looking forward to the next occasion! And if none is planned, one can always decide to organize a *straatfeest* — with Leendert, Menno, Heera and Jeroen — for additional fun and distraction!

To my extended family and family–in–law: thank you for your support and kind words along the way.

Pa en Ma, your support and encouragement did not start when I started this PhD project, but almost forty years ago. You've been fantastic parents all along, and a wonderful example now that I am a parent myself. I cannot thank you enough.

Most importantly, I am deeply indebted to Marjolijn, Nils and Mette. Lijn, this project has brought additional challenges to an already mad busy period in our lives. Thank you so much for putting up with me. I'm looking forward to calmer times and spending more time together. Nils and Mette, you have never known your father without this PhD project to worry about. Probably without realizing it, you have been a wonderful source of inspiration to me. You are my pride and joy.

Finally[2], to all those who I have not mentioned, but who have helped, contributed, encouraged or even discouraged. Maximum respect.

Jan Verkade
Delft, March 2015

[2] This closing sentence was 'borrowed' from Micha Werner's dissertation (Werner, 2004)

ABOUT THE AUTHOR

Jan Verkade was born on July 23rd, 1975 in Maassluis, The Netherlands. He holds a Bachelor's degree in Business Administration, awarded in 1997 by The Hague University of Applied Sciences, a Master of Arts degree in International Relations, awarded 'with distinction' in 2000 by Dublin City University and a Master of Science degree in Water Resources Management, aawarded 'cum laude' in 2008 by Delft University of Technology.

Prior to joining Delft Hydraulics in 1998, Jan has worked in various roles and professional environments. At Delft Hydraulics, which has merged into Deltares in 2008, he has worked as a hydrologist and has specialized in real–time forecasting. Part of the work at Deltares constitutes membership of the River Forecasting Service at Rijkswaterstaat's Water Management Centre of The Netherlands. The service is responsible for both low flow forecasting and flood forecasting in rivers Meuse and Rhine, as well as interpretation and dissemination of fluvial flood forecasts produced by the European Flood Awareness System EFAS. Deltares has allowed Jan to work — on a part–time basis — on a PhD research project on estimating real-time predictive hydrological uncertainty from 2008 through 2014.

At Deltares, Jan was one of the founding members of the 'Statistics Community' which aims to share experience and expertise in the use of statistics and uncertainty analyses. He is also an active member of the HEPEX community, where he initiated, co-implemented and co-manages the current HEPEX portal (at `www.hepex.org`). The portal is instrumental in engaging the HEPEX community in between meetings and has lead to a considerable growth of community membership and activity. Jan has acted as a convener and as co–convener of conference sessions at several General Assemblies of the European Geosciences Union as well as at the 2014 GEWEX conference. These sessions all pertained to the generation and use of hydrological predictions and forecasts.

After completion of the PhD project, Jan will continue to work at Deltares. He currently lives in Delft with his partner Marjolijn and their beautiful children Nils (2010) and Mette (2014).

PUBLICATIONS BY THE AUTHOR

Some ideas and figures have appeared previously in the following publications:

Peer-reviewed literature

López López, P., Verkade, J. S., Weerts, A. H., and Solomatine, D. P.: Alternative configurations of Quantile Regression for estimating predictive uncertainty in water level forecasts for the Upper Severn River: a comparison., Hydrology and Earth System Sciences, 18, 3411–3428, DOI:10.5194/HESS-18-3411-2014, 2014.

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., and Verkade, J. S.: Challenges of Operational River Forecasting, Journal of Hydrometeorology, DOI:10.1175/JHM-D-13-0188.1, 2014.

Verkade, J. S. and Werner, M. G. F.: Estimating the benefits of single value and probability forecasting for flood warning, Hydrology and Earth System Sciences, 15, 3751–3765, DOI:10.5194/HESS-15-3751-2011, 2011.

Verkade, J. S., Brown, J. D., Reggiani, P., and Weerts, A. H.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, Journal of Hydrology, 501, 73–91, DOI:10.1016/J.JHYDROL.2013.07.039, 2013.

Verkade, J. S., Brown, J. D., Davids, F., Reggiani, P., and Weerts, A. H.: Estimating predictive hydrological uncertainty by dressing deterministic and ensemble forecasts; a comparison, with application to River Meuse, Journal of Hydrology, submitted.

Weerts, A. H., Winsemius, H. C., and Verkade, J. S.: Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales), Hydrology and Earth System Sciences, 15, 255–265, 2011.

Conference papers

De Kleermaeker, S. and Verkade, J. S.: A decision support system for effective use of probability forecasts, in: Proceedings of the 10th International ISCRAM Conference, Baden – Baden, Germany, 2013.

Haasnoot, M., Verkade, J. S., and De Bruijn, K. M.: HABITAT, a spatial analysis tool for environmental impact and damage assessment, in: HydroInformatics, Concepción, Chile, 2009.

Leahy, C. P., Pagano, T. C., Elliott, J. ., Sooriyakumaran, S., Schellekens, J., and Verkade, J. S.: Seven day flow forecasting using hydrological models and Numerical Weather Prediction Rainfall Forecasts, in: Proceedings of the 34th World Congress of the International Association for Hydro-Environment Research and Engineering: 33rd Hydrology and Water Resources Symposium and 10th Conference on Hydraulics in Water Engineering., p. 248, Engineers Australia, Adelaide, Australia, 2010.

Other publications

Verkade, J. S., Van Loenen, A., Beckers, J., Weerts, A. H., and Van Leeuwen, P. E. R. M.: Kansverwachtingen in het regionaal-waterbeheer, H2O, 16, 20–21, 2011.

In addition, there have been contributions to conference sessions at the 2009 through 2014 General Assemblies of the European Geosciences Union, at the 2011 Fall Meeting of the American Geophysical Union, at the 2011, 2012 and 2014 HEPEX meetings in Delft, Beijing and Washington, respectively and at the 2014 GEWEX conference in The Hague. Also, Jan has authored and co-authored blog posts that can be accessed at `www.hepex.org` and `http://janverkade.wordpress.com`

In development

Pappenberger, F., Pagano, T. C., Alfieri, L., Berthet, L., Brown, J. D., Cloke, H. L., Cranston, M., Demargne, J., de Saint-Aubin, C., Feikema, P., Fresch, M., Garçon, R., Gelfan, A., He, Y., Hu, Y., Janet, B., Jurdy, N., Javelle, P., Kuchment, L., Le Lay, M., Li, Z., Marty, R., Meissner, D., Manful, D., Organde, D., Rademacher, S., Ramos, M., Reinbold, D., Ricciardi, G., Salamon, P., Shin, D., Sprokkereef, E., Thielen, J., Thiemig, V., Tuteja, N., Van Andel, S.-J., Verkade, J. S., Wetterhall, F., and Zappa, M.: Handbook of Hydrometeorological Ensemble Forecasting, chap. Hydrological Ensemble Prediction Systems Around the Globe, Springer, 2015.

Verkade, J. S., Pappenberger, F., Ramos, M.-H., Buan, S., Restrepo, P., Spångmyr, H., Hopson, T., Price, D., Welles, E., McManamon, A., and Stephens, E.: A Staged Development Model for the Effective Use of Probabilistic Forecasts in Flood Forecasting, Warning and Response Systems, TBD, 2015.

Werner, K., Verkade, J. S., Pagano, T. C., and Welles, E.: Handbook of Hydrometeorological Ensemble Forecasting, chap. Application and Communication of Verification Information, pp. –, Springer, 2015.

ESTIMATING REAL-TIME PREDICTIVE HYDROLOGICAL UNCERTAINTY

JAN VERKADE

Water management decisions may benefit from real-time hydrological forecasts. These forecasts reduce but cannot eliminate uncertainty about the future. Probabilistic forecasting aims to estimate the remaining predictive hydrological uncertainty.

The present dissertation explores the value of probabilistic forecasting as well as various statistical techniques for improving the estimates of predictive hydrological uncertainty.

Jan Verkade has worked at Delft Hydraulics and Deltares for over fifteen years. His research into predictive hydrological uncertainty has resulted in the present PhD dissertation. Jan continues to research various aspects of predictive uncertainty, including the use thereof in decision-making as well as how to improve forecast-decision-response systems through forecast verification.

ISBN 978-94-6186-446-8