First of all, please let me offer my sincere apologies for being over two months late in submitting this review. In hindsight, I should have maybe declined the review request, but as I found the topic of this paper interesting and relevant to my own work also, I didn't want to pass up this opportunity. At the expense of timeliness however.

The Hoss and Fischbeck paper is an interesting and very worthwhile addition to the existing literature on predictive hydrological uncertainty insofar as it explores the optimal selection of predictors to configure a Quantile Regression based statistical post-processor for estimating predictive uncertainty. However, the paper requires some work prior to its publication. I don't think the the computations underlying the manuscript need to be modified in great extent, but the description and the analysis thereof can be improved substantially. I hope to give some suggestions on how that can be done.

General comments

- The manuscript could benefit from a more substantial "hydrological analysis" of the forecasts made. Post-processors can be used to find statistical relations between predictors and predictands. There needs to be correlation and causality. The paper could benefit from a more in-depth analysis of the latter: what does the 'forecast error' depend on? Here, the authors choose rate of rise and past forecast error: these appear to be more or less randomly chosen, and are subsequently applied to all forecasting locations considered. However, I think that an analysis of the hydrology of the basins considered, in conjunction with the forecasting models for those basins, could reveal important information on how those models are expected to perform. How are the models calibrated? What does this mean for extreme events? Is the relation between predictors and predictand stationary across 'normal flow regimes' and 'extremes'? This likely varies with basin, and therefore one should consider varying post-processing configurations with basin also.
- There is one important assumption underlying the use of statistical post-processors: stationarity of the joint predictor, predictand distributions. The paper would benefit from a discussion thereof, particularly in relation to the results section, and the 'robustness' section contained therein.
- "First US application" is irrelevant to the science and also incorrect, as Wood et al (see reference in Weerts et al, 2011) applied QR previously. This comes back a couple of times in the paper. Also, QR was originally devised by Roger Koenker; not by Weerts et al (I wish!).

- Different users have different needs for uncertainty information; it is not universally true that users benefit most from probabilities of exceedence or non-exceedance. Likewise, not all users are interested in extreme events per sé. This comes back a couple of times in the paper.
- I would recommend to streamline use of terms:
  - ‘predictor’ or ‘independent variable’
  - ‘predictand’ or ‘dependent variable’
  - preferably omit use of ‘variable’ in context of statistical post-processors, as its interpretation can be ambiguous
  - ‘configuration’ rather than ‘model’ (to avoid confusion with underlying hydrological models)
- Please consider removing the footnotes. If the text contained therein is important, include it in the main body of the paper. If not, you may want to consider omitting it altogether.
- Practicalities of data access are not too relevant to the science and I would suggest omitting descriptions of why certain data sources could (not) be accessed and how much effort that would require. Instead, you could turn the argument around and say: “this and this is available and we’re trying to assess if there is any signal that can contribute to better probabilistic forecasts.”

<u>Specific comments</u>

Introduction:

- Some elements can be safely omitted from the introduction:
  - Discussion on QPF forecasts
  - Discussion of RFC produced “outlooks”
- Verifying by means of BSS only is somewhat limited I think, but it does fit with the authors’ wish to verify exceedence probabilities only. Why not, however, use a range of verification metrics? See, for example, some of the recent Brown and Seo papers as well as some of my own work (where the verification approach was inspired on the Brown/Seo papers).
- “Rate of rise” is more commonly used than “rise rate” I think.

2.2 Brier Skill Score:

- The ‘method’ section would benefit from a subsection on verification metrics. That section would then include the current sub-section on BSS, but also some discussion of other metrics now included in the ‘results’ section.
- A decomposition of Brier’s probability score is included; what’s

missing, is a note on how these decompositions are computed in terms of *skill*. See one of the Brown and Seo papers for how that's done. Also, no quantified decompositions are shown in the results/analysis section?

2.3 Proposed addition

- The current title "Proposed addition: more than one independent variable" suggests that it is the *number* of predictors that's important. This is not necessarily so - it's content, not just quantity that's relevant. Please consider retitling this section.
- This section could really benefit from some 'hydrological intelligence': what are the factors determining level of accuracy of model predictions? Are these already included in the model itself somehow? If so, how? If not, why not? To me, it is still an open question: what to include in a model, and what to include in a post-processor? Where is the boundary between statistical modeling and modeling of physical processes? This point is one that the authors should also re-visit in the discussion/conclusions section.
- Table 1: "forecast error 24 hours ago". I understand this to be the difference between the current (i.e. at issue time of the forecast) water level and the forecast that was produced 24/48 hours agao - correct? Maybe good to state this.

2.5 Data:

- First sentence may be omitted, or moved to the introduction.
- The manuscript would benefit from a custom made map showing the forecasting locations and basin delineations.

3.2.2 Best performing combinations

- The forecasts for extreme conditions perform worse when using multiple predictors. Why - overfitting? Some in-depth analysis would be good.

3.3 Robustness

- I think the 'robustness' analysis could, and should, be simplified by using a leave-one-year-out analysis. Length of training set is less relevant than stationarity of joint predictand, predictor distributions. Why not simply use all of the available data most efficiently and then discuss any drops in forecast quality? Also, the current analysis results in a difference in sample size and this would require an analysis of the uncertainty in resulting BSS - which is likely bigger for smaller samples. With a leave-one-year-out analysis, sample size would be equal and the authors would be more easily forgiven for not analysing uncertainty.

- Some hydrologic analysis could contribute to explaining why forecast quality is different between locations.

“Future work”

- Yes, more analysis on which predictors to use could work. Please refer to my earlier comments also on statistical modeling versus numerical modeling of physical processes, and on using knowledge of the hydrology of basins to determine meaningful predictors.

Figures:

- The multi-plot figures contain a lot of white space between plots. As some horizontal and vertical axes are identical across plots within the figure, I would suggest eliminating the in-between space altogether. In figures 10 and 11, this can be done for the vertical axes also. In R: par(mar = c(.5,0,0,0)) and then plot(…, xaxt=”n”) for plots where you can omit horizontal axis.

Additional specific comments are included in attached, annotated PDF.

Hydrol. Earth Syst. Sci. Discuss., 11, 11281–11333, 2014
www.hydrol-earth-syst-sci-discuss.net/11/11281/2014/
doi:10.5194/hessd-11-11281-2014

Hydrology and Earth System Sciences
Open Access
Discussions

Performance and robustness of probabilistic river forecasts computed with quantile regression based on multiple independent variables in the North Central USA

F. Hoss[1,2] and P. S. Fischbeck[1]

[1]Carnegie Mellon University, Department of Engineering & Public Policy, 5000 Forbes Avenue, Pittsburgh, 15213, USA
[2]Harvard Kennedy School, Belfer Center, 79 JFK Street, Cambridge, MA, 02138, USA

Received: 28 August 2014 – Accepted: 20 September 2014 – Published: 14 October 2014

Correspondence to: F. Hoss (fraukehoss@gmail.com)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

This study ~~further~~ develops the method of quantile regression (QR) to predict exceedance probabilities of flood stages by post-processing forecasts. Using data from the 82 river gages, for which the National Weather Service's North Central River Forecast Center issues forecasts daily, this is the first QR application to US ~~American~~ river gages. Archived forecasts for lead times up to six days from 2001–2013 were analyzed. Earlier implementations of QR used the forecast itself as the only independent variable (Weerts et al., 2011; López López et al., 2014). This study adds the rise rate of the river stage in the last 24 and 48 h and the forecast error 24 and 48 h ago to the QR model. Including those four variables significantly improved the forecasts, as measured by the Brier Skill Score (BSS). Mainly, the resolution increases, as the original QR implementation already delivered high reliability. Combining the forecast with the other four variables results in much less favorable BSSs. Lastly, the forecast performance does not depend on the size of the training dataset, but on the year, the river gage, lead time and event threshold that are being forecast. We find that each event threshold requires a separate model configuration or at least calibration.

1 Introduction

River-stage forecasts are inherently uncertain. The past has shown that unfortunate decisions have been made in ignorance of the potential forecast errors (e.g., Pielke, 1999; Morss, 2010). For users, forecasts are most important in extreme situations, such as droughts and floods. Due to their infrequency and the subsequent scarcity of data, forecasts have larger errors where accuracy has the most value. Additionally, users might only experience such an event once or twice in their lifetime, so that they have no experience to what extent they can rely on deterministic forecasts in such situations. Given the many sources and complexity of uncertainty and the lacking user experience, it is easy to see how forecast users find it difficult to estimate the forecast error. Including uncertainty in weather forecasts has been strongly recommended (e.g., National Research Council, 2006).

There are two types of approaches to quantify uncertainty (e.g., Leahy, 2007; Demargne et al., 2013; Regonda et al., 2013): those addressing certain sources of uncertainty in the output, e.g., input uncertainty and hydrological uncertainty, and those taking into account all sources of uncertainty in a lumped fashion. Both approaches have their advantages. Modelling each source separately can take into account that the different sources of uncertainty have different characteristics (e.g., some sources of uncertainty depend on lead time, while others do not). This approach is likely to result in better performing, more parsimonious models. On the downside, it is expensive to develop, maintain and run. As an alternative, the lumped quantification of uncertainty is a less resource-intensive approach (Regonda et al., 2013).

The National Weather Service has chosen for ensemble forecasting to quantify the uncertainty from major sources (Demargne et al., 2013). As of today, the National Weather Service does not routinely publish uncertainty information along with their short-term river-stage forecast (Fig. 1). Until the NWS has implemented probabilistic forecasting for short-term products (next few hours and days), the only way that users can get a sense of the uncertainty is by comparing the quantitative precipitation forecast (QPF) with the non-QPF forecast. The QPF-forecast includes the precipitation predicted for the next 12 h and zero precipitation for the forecasts beyond 12 h.[1] The non-QPF forecast assumes no precipitation. Combined, these two forecasts give an idea of how much difference (a short period of) precipitation would make for the stage height in the river. The non-QPF serves as a reasonable lower bound; however, the QPF forecast is not an upper bound (i.e., precipitation could exceed the forecast values).

[1]This practice differs from RFC to RFC and also over time. For the ABRFC Welles et al. (2007) report: ~ 1993–1994: zero QPF; ~ 1995–2000 24 h QPF for first 24 h, zero QPF beyond 24 h; ~ 2001–2003 12 h QPF for first 12 h, zero QPF beyond 12 h.

As of today, only the "outlooks" produced by the Ensemble Streamflow Prediction part of the NWS River Forecasting System are probabilistic, i.e., quantify uncertainty: an exceedance curve for a period of three month and bar plots for each week of a three months period, see Figs. 2 and 3. These graphs can be used to determine with which probability each river stage will be exceeded in those weeks or three-months period. Although the short-term weather forecasts for the next few days are much used to prepare for flood events, they have remained deterministic, as shown in Fig. 1.[2]

NWS has developed the Hydrologic Ensemble Forecast Service (HEFS) to be able to provide short-term and medium-term probabilistic forecasts. Its implementation at all 13 river forecasts centers is planned to be completed in 2014 (Demargne et al., 2013).

In contrast to the ensemble approach chosen by the NWS, the post-processing method that is further developed in this paper – quantile regression – does not distinguish between sources of uncertainty, but studies the overall uncertainty in a lumped fashion. This choice is motivated by the fact that the total predictive uncertainty, rather than its different sources, are relevant for decision-making (Solomatine and Shrestha, 2009). To further strengthen the main advantage of this method, i.e., requiring relatively little resources, we exclusively use publicly available data to build our models.

Most previously developed post-processors to generate probabilistic forecasts share the overall set-up but differ in their implementation. Explanatory variables such as the forecasted and observed river stage, river flow or precipitation, and previous forecast errors are used to predict the forecast error, conditional probability distribution of the forecast error or other metrics of uncertainty for various lead times (e.g., Kelly and Krzysztofowicz, 1997; Montanari and Brath, 2004; Montanari and Grossi, 2008; Regonda et al., 2013; Seo et al., 2006; Solomatine and Shrestha, 2009; Weerts et al., 2011). Among others, these methods differ in their mathematical methods, their sub-setting of data, and the output metric. Please see Regonda et al. (2013) and Solomatine and Shrestha (2009) for a summary of each method. In a meta-

[2]The deterministic forecasts are also available as text or tables.

analysis of four different post-processing methods to generate confidence intervals, the quantile regression method was one of the two most reliable methods (Solomatine and Shrestha, 2009), while being the mathematically least complicated method and requiring few assumptions.

This paper further develops one of the methods mentioned above: the Quantile Regression method to post-process river forecasts introduced by Weerts et al. (2011). That study achieved impressive results in estimating the 50 and 90 % confidence interval of river-stage forecasts for three case studies in England and Wales using QR with calibration and validation datasets spanning two years each. In some aspects, our approach differs from the original approach by Weerts et al. (2011) and López López et al. (2014). We predict the probabilities that flood stages are exceeded rather than uncertainty bounds, because the former are more relevant to decision-making. In an attempt to balance missed alarms and false alarms, decision-makers are likely to resort to the best estimate (i.e., the deterministic forecast) rather than basing actions on the 50 or 90 % confidence interval. Additionally, predicting the probability of an event corresponds with other forecasts with which users have much experience, e.g., the probability of precipitation. Morss et al. (2010) found in a survey of the general US public that most people are able to base decisions on those forecasts. Additionally, we are fortunate to have a much larger dataset, consisting of archived forecasts for 82 river gages covering 11 years available.

In this paper, the QR method is applied to the 82 river gages of the North Central River Forecast Center (NCRFC) encompassing (parts of) Illinois, Michigan, Wisconsin, Minnesota, Indiana, North Dakota, Iowa, and Missouri.[3] To our knowledge, this paper is the first application of the QR method to the US American context.

The method is further developed by demonstrating the benefit – measured by an increase in Brier Skill Score (BSS) – of including the rise rates of water levels in past hours and the past forecast errors as independent variables into the quantile regression. For extremely high water levels the variable combination has to be

[3] As of spring 2014, the NCRFC does not publish any sort of probabilistic forecasts.

customized for each river gage. For those, sets of few independent variables work best. Variable combinations for other event thresholds should include as many dependent variables as possible. Using the same combination for all of them works satisfactorily. Furthermore, it is found that the forecast – the only independent variable in the original QR method – is difficult to combine with the other dependent variables. Last, the method is shown to be robust to the size of the training dataset. However, the forecast performance does vary significantly across locations, lead times, water levels, and forecast year.

The paper is structured as follows. The Method section summarizes the additions that this paper makes to the quantile regression method introduced by Weerts et al. (2011). It reviews the method, explains the additions, introduces the performance metric, and discusses the computations and data. The Results section first reviews the overall forecast error for the dataset. It then compares the proposed method to the original quantile regression as demonstrated for river gages in Wales and England (Weerts et al., 2011). Finally, it discusses the robustness of the proposed method. The fourth and last section presents the conclusions and proposes further research ideas.

2 Method

The use of quantile regression to quantify the error distribution of river-stage forecasts has first been presented by Weerts et al. (2011) for river catchments in the England and Wales. In this paper, we further develop Weerts' original method in three ways: (a) by including additional variables instead of using only the forecast itself as an independent variable, (b) by testing the robustness of the method across locations, lead times, event thresholds, forecast years, and the size of training dataset, (c) by estimating the more decision-relevant probability of exceeding flood stages rather than confidence bounds. To develop the different configurations of quantile regression and to compare their performance, the Brier Skill Score (BSS) is used.

In the following, the quantile regression itself, the proposed addition to the method, and the undertaken computations are explained.

2.1 Quantile Regression

In the context of river forecasts, linear quantile regression has been used to estimate the distribution of forecast errors as a function of the forecast itself. Weerts et al. (2011) summarize this stochastic approach as follows:

"[It] estimates effective uncertainty due to all uncertainty sources. The approach is implemented as a post-processor on a deterministic forecast. [It] estimates the probability distribution of the forecast error at different lead times, by conditioning the forecast error on the predicted value itself. Once this distribution is known, it can be efficiently imposed on forecast values."

Quantile Regression was first introduced by Koenker (2005, 1978). It is different from ordinary least square regression in that it predicts percentiles rather than the mean of a dataset. Koenker and Machado (1999, p. 1305) and Alexander et al. (2011) demonstrate that studying the coefficients and their uncertainty for different percentiles generates new insights, especially for non-normally distributed data. For example, using quantile regression to analyze the drivers of international economic growths, Koenker and Machado (1999) find that benefits of improving the terms of trade show a monotonously increasing trend across percentiles, thus benefitting faster-growing countries proportionally more.

In its original application to river forecasts by Weerts et al. (2011), the forecast values and the corresponding forecast errors are transformed into the Gaussian domain using Normal Quantile Transformation (NQT), as instructed by Bogner et al. (2012) to account for heteroscedasticity. Building on this study, López López et al. (2014) compare different configurations of QR with the forecast as the only independent variable, including configurations omitting NQT. They find that no configuration was consistently superior for a range of forecast quality metrics (López López et al., 2014). To be able to combine variables of different nature, we build a model based on untransformed

variables. The reason to do so will be discussed and illustrated later (see Figs. 11 and 12).

Using the transformed data, a quantile regression is run for each lead time and desired percentile with the forecast error as the dependent variable and the forecast and other variables as the independent variables.[4] To prevent the quantile regression lines from crossing each other, a fixed effects model is implemented below a certain forecast value. Weerts et al. (2011) give a detailed mathematical description for applying QR to river forecasts. Mathematically, the approach is formulated as follows:

Equation (1): Original QR implementation with NQT, with percentiles of the forecast error as the dependent variable and the only independent variable being the forecast itself, bot transformed into the normal domain.

$$F_\tau(t) = f(t) + \mathrm{NQT}^{-1}[a_\tau \cdot V_{\mathrm{NQT}}(t) + b_\tau] \tag{1}$$

Equation (2): QR implementation without NQT, with percentiles of the forecast error as the dependent variable and multiple independent variables.

$$F_\tau(t) = f(t) + \sum_i^I a_{i,\tau} \cdot V_i(t) + b_\tau \tag{2}$$

with

$F_\tau(t)$ – estimated forecast associated with percentile τ and time t

$f(t)$ – original forecast at time t

$V_i(t)$ – the independent variable i (e.g., the original forecast) at time t

$V_{i;\mathrm{NQT}}(t)$ – the independent variable I transformed by NQT at time t

$a_{i,\tau}, b_\tau$ – model coefficients

[4] As mentioned in Weerts et al. (2011), our quantile regression models have likewise a higher predictive capacity, if the forecast error rather than the forecast itself is used as the dependent variable.

The second part of the equations stands for the error estimate based on the quantile regression model for each percentile τ and lead time. In Eq. (1), that was used in the original QR method proposed by Weerts et al. (2011), this estimation was executed in the Gaussian domain using only the forecast as independent variable.[5]

2.2 Brier Skill Score

The original QR implementation by Weerts et al. (2011) was evaluated by determining the fraction of observations that fell into the confidence intervals predicted by the QR model; i.e., ideally, 90 % of the observations should be larger than the predicted 10th percentile for that day, and smaller than the predicted 90th percentile. López López et al. (2014) used a number of metrics to assess model performance, e.g., the Brier Skill Score (BSS), the mean continuous ranked probability (skill) score (RPSS), the relative operating characteristic (ROC), and reliability diagrams to compare QR configurations.

We use the Brier Skill Score to compare the different versions of the QR model proposed in this paper. We chose to optimize our QR models based on the BSS, first introduced by Brier (1950), for two reasons. First, for decision-making the probability with which a certain water level, e.g., a flood stage, is exceeded is more useful than confidence intervals. Second, the Brier Score can be decomposed into two different measures of forecast quality (see Eq. 3): reliability and resolution. The third component is uncertainty, which is a hydrological characteristic inherent to the river gage. Thus, it is not subject to the forecast quality. Equation (3) gives the definition of the (de-composed) Brier Score (e.g., Jolliffe and Stephenson, 2012; Anon, 2014; WWRP/WGNE, 2009).[6]

[5]All quantile regressions were done using the command *rq ()* in the R-package “quantreg” (Koenker, 2013).

[6]Bröcker (2012) showed that the conventional decomposition of the Brier Score is biased for finite sample sizes. It systematically overestimates reliability, under- or

Equation (3): Brier Score; de-composed into three terms: reliability, resolution and uncertainty.

$$\mathrm{BS} = \frac{1}{N}\sum_{k=1}^{K} n_k (f_k - \overline{o}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k (\overline{o}_k - \overline{o})^2 + \overline{o}(1-\overline{o}) = \frac{1}{N}\sum_{t=1}^{N} (f_t - o_t)^2 \tag{3}$$

with

BS – Brier Score

N – number of forecasts

K – the number of bins for forecast probability of binary event occurring on each day

n_k – the number of forecasts falling into each bin

$\overline{o}_k$ – the frequency of binary event occurring on days in which forecast falls into bin k

f_k – forecast probability

$\overline{o}$ – frequency of binary event occurring

f_t – forecast probability at time t

o_t – observed event at time t (binary: 0 – event did not happen, 1 – event happened)

The Brier Score pertains to binary events, e.g., the exceedance of a certain river stage or flood stage. Reliability compares the estimated probability of such an event

overestimates resolution, and underestimates uncertainty. Several authors proposed less biased decompositions (e.g., Bröcker, 2012; Ferro and Fricker, 2012). Additionally, Stephenson et al. (2008) proved that the Brier Score has two additional components when it is computed based on bins, as is usually done. Nonetheless, we chose to stick to the conventional decomposition and using bins, as implemented in the R-package “verification” (NCAR-Research Applications Laboratory, 2014; Wilks, 1995) to ensure that our results can be readily compared to other studies like López López et al. (2014). After all, the Score is mainly used to compare model configurations, rather than establishing the absolute performance of each model.

with its actual frequency. For example, perfect reliability means that on 60 % of all days for which it was predicted that the water level would exceed flood stage with a 60 % probability, it actually does so. A forecast with perfect reliability would follow the diagonal in Fig. 4, i.e., the area in Fig. 4a representing reliability would equal zero (e.g., Jolliffe and Stephenson, 2012; Anon, 2014; WWRP/WGNE, 2009). The configuration by López López et al. (2014) performs well in terms of reliability. When estimating confidence intervals, Weerts et al. (2011) achieved good results especially for the more extreme percentiles (i.e., 10th and 90th).

Resolution pertains to how much better the forecast performs than taking the historical frequency (climatology) as a forecast. For example, for a gage where flood stage is exceeded on 5 % of the days in a year, simply using the historical frequency as the forecast would mean forecasting that the probability of the water level exceeding flood stage is 5 % on any given day (e.g., Jolliffe and Stephenson, 2012; Anon, 2014; WWRP/WGNE, 2009). In Fig. 4, a forecast with good resolution would be steeper than the dashed line that represents climatology, i.e., the area in Fig. 4a representing resolution would be maximized. In absolute terms, the resolution can never exceed the third term in Eq. (3) representing the uncertainty inherent to the river gage. Through the resolution component, the Brier Score is related to the area under the relative operating characteristic (ROC) curve (for more details, see Ikeda et al., 2002). The latter likewise quantifies how much better a forecast is than random guessing in detecting a binary event; though unlike the Brier Score it focuses on the ratios of false and missed alarms (e.g., Jolliffe and Stephenson, 2012; Anon, 2014; WWRP/WGNE, 2009).

A forecast possesses skill, i.e., performs better than random guessing or climatology, if it is inside the shaded area in Fig. 4b. The Brier *Skill* Score (BSS) equals the Brier Score normalized by climatology to make the score comparable across gages with different frequencies of a binary event.[7] The BSS can range from minus infinity to

[7]All measures of forecast quality were computed using the R-package "verification" (NCAR, 2014).

one. A BSS below zero indicates no skill; the perfect score is one (e.g., Jolliffe and Stephenson, 2012; Anon, 2014; WWRP/WGNE, 2009).

2.3 Proposed addition: more than one independent variable

Intuitively, more information should lead to better prediction of the distribution of the forecast error, because the regression models would be based on more data. The most obvious variables to include besides the forecast itself are the observed water level 24 and 48 h ago, the observed rise in water level in the last 24 and 48 h (called rise rate hereafter), the forecast error 24 and 48 h ago, or the time of the year, e.g., month or season. Other potential variables are the water levels observed up- and downstream at various times, the precipitation upstream of the catchment area, and the precipitation forecast. However, these latter variables are much more difficult to gather because of the way in which data is archived at the National Climatic Data Center (NCDC).[8]

In preliminary trials on two case studies (gages HARI2 and HYNI2), it was found that season and months are not significant in quantile regression models to predict the quantiles of the forecast error. It was also found that the rise rates and the forecast errors are better predictors than the water levels observed in previous days. After all, the observed water levels are used to compute the rise rates and forecast errors so that these latter variables include the information of the former variable.

To determine which set of variables preforms best in generating probabilistic forecasts, all 31 possible combinations of the forecast (fcst), the rise rate in the last 24 and 48 h (rr24, rr48), and the forecast error 24 and 48 h ago (err24, err48) were tested for 82 gages that the NCRFC issues forecasts for every morning (Table 1). Based on the Bier Skill Score, a metric of forecast quality explained below, it was determined

[8]For the NCRFC, the river forecast and the observed water levels are saved in the same text product available at: http://cdo.ncdc.noaa.gov/pls/plhas/HAS.FileAppSelectfidatasetname=9957ANX (last access: July 2014). (Station ID: KMSR, Bulletin ID: FGUS5). Requesting the corresponding precipitation and precipitation forecast requires an extensive effort or direct access to the database.

which variable combination on average and most often leads to the best out-of-sample results for various lead times and water levels.

2.4 Computations

The output of our QR application to river forecasts is the probability that a certain water level in the river or flood stage is exceeded on a given day, e.g., "On the day after tomorrow, the probability that the river exceeds 15 feet is 60 %." This is done in two steps. First, a training dataset (first half of the data) is used to build one quantile regression model for each of the following percentiles: [0.05, 0.1, 0.15, ..., 0.85, 0.90, 0.95]. The dependent variable is the water level. As described above, the forecast itself, the rise rates and forecast errors serve as independent variables.

In the second step, these QR models are used to predict the water levels corresponding with each model's percentile on each day in the verification dataset (the second half of the dataset). Effectively, for each day in the verification dataset, a discrete probability distribution of water levels is predicted. Each QR model contributes one point to that distribution.

In our opinion, this probability distribution of water levels is too much information to efficiently make decisions. The model performance should be assessed for a decision-relevant output. Therefore, we calculate the probability with which various water levels (called event thresholds hereafter) will be exceeded. The probability of exceeding each water level is computed by linearly interpolating between the points of the discrete probability distribution that was computed in the previous step.[9]

To be able to compare various model configurations, the Brier Skill Score is determined across all the days in the verification dataset. As explained above, the BSS is based on the difference between the predicted exceedance probability and the observed exceedance (binary), averaged across all days in the verification dataset.

[9]Using the command "approx(x, y, xout, yleft = 1, yright = 0, ties = mean)" in the R-package "stats" (R-Core Team, 2014).

To study whether the various combinations of variables perform equally well for high and low thresholds, these last computational steps (i.e., interpolating to determine the exceedance probability for a certain water level and calculating the BSS) were done for the 10th, 25th, 75th, and 90th percentile of observed water levels and the decision-relevant four flood stages (action stage, and minor, moderate, and major flood stage) of each gage.

To determine the optimal set of independent variables, the entire procedure is repeated for each of the 31 variable combinations in Table 1, thus using a different set of independent variables each time. To test the robustness of this approach, the procedure was also repeated for each river gage and for several lead times. The result is 31 BSSs for 82 river gages for four different lead times (one to four days) and for different event thresholds (i.e., flood stages or percentiles of the observed water level).

2.5 Data

The National Weather Service (NWS) issues river-stage forecasts for ~ 4000 river gages every day. Such daily published forecasts predict the stage height in 6 h intervals for up to five days ahead (20 6 h intervals).[10] When floods occur and increased information is needed, the local river forecast center (RFC) can decide to publish river-stage forecasts more frequently and for more locations. Welles et al. (2007) provides a detailed description of the forecasting process.

For this paper, all forecasts published by the North Central River Forecast Center (NCRFC) between 1 May 2001 and 31 December 2013 were requested from the NCDC's HDSS Access System.[11] In total, the NCRFC produces forecasts for 525

[10]The river-stage forecasts are produced by one of NWS' thirteen river forecasts centers (RFCs). Every morning the forecasts are forwarded to one of NWS's 122 local weather forecast offices (WFOs), who then disseminate the information to the public through a variety of media channels or by issuing warnings.

[11]URL (last accessed July 2014): http://cdo.ncdc.noaa.gov/pls/plhas/HAS.FileAppSelect fidatasetname=9957ANX; Station ID: KMSR, Bulletin ID: FGUS5.

gages (Fig. 5). For 82 of those gages, forecasts have been published daily for a sufficient number of years, and are not inflow forecasts. The latter have been excluded from the forecast error analysis because they forecast discharge rather than water level. About half of the analyzed gages are along the Mississippi River. The Illinois River and the Des Moines River are two other prominent rivers in the region. The drainage areas of the 82 river gages average 61 500 square miles (minimum 200 sq.miles; maximum 708 600 sq.miles).

Two river gages serve as an illustration for the points made throughout this paper. Hardin, IL is just upstream the confluence of the Illinois River and the Mississippi River (Fig. 5). Therefore, it probably experiences high water levels through backwatering, when the high water levels in the Mississippi River prevent the Illinois River from draining. Henry, IL is located ~ 200 miles (~ 320 km) upstream of Hardin, having a difference in elevation of ~ 25 feet (~ 7.6 m). The Illinois River is ~ 330 miles (~ 530 km) long,[12] draining an area of ~ 13 500 square miles (~ 35 000 km^2) at Henry[13] and ~ 28 700 square miles (~ 72 000 km^2) at Hardin.[14]

3 Results

3.1 Forecast error at NCRFC's gages

In general, the NCRFC's forecasts are well calibrated across the entire dataset. The average error, defined as observation minus the forecast, is zero for most gages.

[12] Illinois Environmental Protection Agency: "Illinois River and Lakes Fact Sheets", URL (accessed 24 April 2014): http://dnr.state.il.us/education/aquatic/aquaticillinoisrivlakefactshts.pdf.

[13] Source: http://waterdata.usgs.gov/nwis/nwisman/fisite_no=05558300&agency_cd=USGS.

[14] Source: http://waterdata.usgs.gov/nwis/nwisman/fisite_no=05587060&agency_cd=USGS.

For lead times longer than three days, a slight underestimation by the forecast is noticeable. By a lead time of 6 days this underestimation averages 0.41 feet only (Fig. 6a, Table 2a). Extremely low water levels, defined as below the 10th percentile of observed water levels, are also well calibrated (Fig. 6b, Table 2b). However, when considering higher water levels the picture changes.[15] The underestimation becomes more pronounced, averaging 0.29 feet for three days of lead time and 1.14 feet for six days of lead time, when only observations exceeding the 90th percentile of all observations are considered (Fig. 6c, Table 2c). When only looking at observations that exceeded the minor flood stages corresponding to each gage,[16] the underestimation averages 0.45 feet for three days of lead time and 1.51 feet for 6 days of lead time (Fig. 6d, Table 2d). However, some gages, such as Morris (MORI2), Marseilles Lock/Dam (MMOI2) – both on the Illinois River – and Marshall Town on the Iowa River (MIWI4) experience *average* errors of 5 to 1[illegible] feet for water levels higher than minor flood stage.

3.2 Including more variables

In total, the Brier Skill Score (BSS) for 31 variable combinations (Table 1) across various lead times and event threshold have been compared. Across 82 river gages, it has been analyzed (a) which combinations perform best and worst most often, and (b) which sets of variables deliver the best BSSs on average.

[15] The gages MORI2 and MMOI2 are upstream of a dam. It is likely that the forecasts performed so poorly there, because the dam operators deviated from the schedules that they provide the river forecast centers to base their calculation on.

[16] Flood stages are based on the damage done by previous floods. It depends on the context, e.g., the shape of the river bed and the development of the river shores, which water levels cause damage. Therefore, it depends on the river gage which percentiles of observed water levels the flood stages correspond with.

3.2.1 Frequency analysis

For each lead time (i.e., one to four days) and various event thresholds (i.e., 10th, 25th, 75th, 90th percentiles as well as the four flood stages), we counted how often each variable combination resulted in the highest and the lowest BSSs across the 82 river gages. Figure 7 shows that for water levels below the 50th percentile variable combinations with four or more variables return the best BSSs most often, while those with one and two variables perform worst most often. For thresholds higher than the 50th percentile the distributions gradually become more flat. For the 90th percentile, a clear trend is no longer detectable. The same set of histograms for the four flood stages (i.e., action, minor, moderate, and major) confirms this (Fig. 8). Across lead times, there is a slight trend noticeable that single variables tend to be the worst combination more often for longer lead times. Thus, the further out one is forecasting, the more important it becomes to include more data in the model.

3.2.2 Best performing combinations on average

For each river gage, the combinations have been ranked by BSSs. It was found that the more variables are included in a set, the higher that set of variables will rank on average (Fig. 9). However, for extremely high water levels, this trend gradually reverses (Fig. 10). For action stage[17] and minor flood stage,[18] a slightly increasing trend is still

[17]Across the 82 stations, action stage corresponds with water levels between the 60th and 100th percentile.

[18]Across the 82 stations, minor flood stage corresponds with water levels between the 70th and 100th percentile.

visible. For moderate[19] and major flood stage,[20] combinations with fewer variables rank higher on average.

Considering these findings and those of the frequency analysis earlier, the models for the various river gages can generally be based on the same variable combinations of four or more variables. But for extremely high water levels, a model with variable combinations specific to each river gage has to be built in order to achieve high BSSs.

The combinations including the forecast (indicated by gray vertical lines in Figs. 9 and 10) perform less well than those that exclude it. Plotting the independent variables against the forecast error as the dependent variable makes the reason visible (Figs. 11 and 12). Without a transformation into the normal domain, the forecast does not provide a lot of information for the QR model. In contrast, the other four variables do not lend themselves for linear quantile regression after performing NQT. Further research is necessary to reconcile these two types of variables. A possible solution could be to build QR models for subsets of the transformed dependent and independent variable.

3.2.3 Brier Skill Score

Including the rise rate and forecasts errors as independent variables into the QR model improves the Brier Skill Score (BSS) significantly. Figure 13 illustrates the BSS when using the model as originally introduced by Weerts et al. (2011). Using the best performing variable combination instead, gives an upper bound of the BSSs that can be achieved at best. This configuration increases the mean and decreases the standard deviation (Table 3, Fig. 14). The performance improves most where all model

[19]Across the 82 stations, moderate flood stage corresponds with water levels between the 80th and 100th percentile.

[20]Across the 82 stations, major flood stage corresponds with water levels between the 90th and 100th percentile.

configurations perform worst: at the 10th percentile.[21] The decrease of the BSSs with lead time also becomes considerably less with this configuration. Additionally, a one-size-fits-all approach was tested to investigate, whether customizing the QR model to each river gage would be worth it. In this configuration, the rise rates in the past 24 and 48 h and the forecast errors 24 and 48 h ago serve as the independent variables (combination 30). It was found that this approach returns only slightly worse results than working with the best performing configuration for each river gage (Table 3; Fig. 15). Accordingly, the same variable combination can be used for all river gages.

As shown in Fig. 8, this last conclusion is not true for extremely high water levels. Including more variables does improve the BSSs considerably (Figs. 16 and 17, Table 3). However, for each river gage the best combination of variables needs to be identified separately. Because data to build models is scarce for extreme levels, the QR models all perform less well for each increase in flood stage.

The fact that the Brier Score can be de-composed into reliability, resolution and uncertainty allows a closer look at which improvements are being achieved by including more variables. Figure 18 shows that the original QR model configuration by Weerts et al. (2011) has high reliability (i.e., the reliability is close to zero). The Brier Score and the Brier Skill Score mainly improve when using rise rates and forecast errors as independent variables, because the resolution increases. The forecast quality improves along other dimensions as well, i.e., the areas under the ROC curves and the ranked probability skill score (RPSS) increase. The first weighs missed alarms against false alarms and has a perfect score equal to one. The latter is a version of the Brier Skill Score. While the Brier Skill Score pertains to a binary event, the RPSS can take into account various event categories. Its perfect score equals one (e.g., WWRP/WGNE, 2009).

[21]Possibly, the models do not perform well for low percentiles, because the dependent variable – the forecast error – exhibits very little variance at those water levels, i.e., the average error is very small (Table 2).

3.3 Robustness

The impact of the length of the training dataset on the model's performance measured by the Brier Skill Score (BSS) was assessed for the one-size-fits-all QR model (i.e., rise rates and forecast errors as independent variables for all gages) for Hardin and Henry on the Illinois River. Each year between 2003 and 2013 was forecast by models trained on one year up to however many years of archived forecasts were available. Figures 19 and 20 show that for those gages, it does not matter for the BSS how many years are included in the training dataset. That is good news, if stationarity cannot be assumed (Milly et al., 2008), a step-change in river regime has occurred, or forecast data have not been archived in the past. In those cases, only short training datasets are available. However, the BSS varies considerably for what year is being forecast. The forecast performance varies greatly, especially for the 10th and 25th percentile of observed water levels. It is likely that a very large dataset, including more infrequent events, would improve these results. However, most river forecast centers only recently started archiving forecasts in a text-format, so that even having ten years' worth of data is an exception.[22]

To generalize the result, the same analysis as for Hardin and Henry was done for all 82 gages. Following that, a regression analysis was executed with the BSS score as the dependent variable and the river gages and forecast years as factorial independent variables and the lead time, event thresholds, and number of training years as numerical independent variables. The forecast performance was found to vary significantly across all those dimensions except the number of training years. This results in a very wide range of Brier Skill Scores (Fig. 22). Accordingly, for the user, it is particularly difficult to know how much to trust a forecast, if the performance depends so much on context. Likewise, this is case for the original QR configuration (not shown).

[22]To illustrate that point, the National Climatic Data Center has archived data from 2001 onwards available in their HDSS Access System.

For low event thresholds, the BSSs are much worse than for high thresholds, and the BSSs slightly decrease with lead time (Table 4). The regression is slightly biased regarding the forecast quality for each forecast year. The earlier years are included less often in the dataset with on average less years' worth of data in their training dataset, because, for example, unlike for the year 2013, ten years of training data were not available for the year 2006. Nonetheless, the regression indicates that 2008 was particularly difficult to forecast and 2012 relatively easy, i.e. they are associated with relatively low and high coefficients respectively (Table 4). The performance of the forecast additionally depends on the river gage. The coefficients of the river gages, included as factors in the regression, have been excluded from Table 4 for the sake of brevity. Instead, Fig. 21 maps the geographic position of the river gages with the color code indicating each gage's regression coefficient. The coefficient is lower, and therefore the Brier Skill Scores are lower, for gages far upstream a river and those close to confluences. The latter is particularly visible where the Illinois River and the Mississippi River join. At least for the gages at confluences, the QR model could probably be improved by including the rise rates at the river gages on the other joining river into the regression.

4 Conclusions

In this study, quantile regression (QR) has been applied to estimate the probability of the river water level exceeding various event thresholds (i.e., 10th, 25th, 75th, 90th percentiles of observed water levels as well as the four flood stages of each river gage). This is the first study applying this method to the US American context. Additionally, it further develops the method by including more independent variables and testing the method's robustness across locations, lead times, event thresholds, forecast years and sizes of training dataset.

Most importantly, it was found that including rise rates in the past 24 and 48 h and the forecast errors of 24 and 48 h ago as independent variables improves the performance

of the QR model, as measured by the Brier Skill Score. Since the reliability was already high with the original QR method as proposed by Weerts et al. (2011), the new configuration mainly increases the resolution.

For extremely high water levels, the combinations of independent variables that perform best vary across stations. On those days, combinations of fewer variables perform better than those that include more. In contrast to these extremely high event thresholds, larger sets of variables work better than smaller ones for non-extreme and low event thresholds. Additionally, a one-size-fits-all approach (i.e. the rise rates and forecasts errors as independent variables) performs satisfactorily for those cases.

The new independent variables – rise rates and forecast errors – do not combine well with forecast itself. The latter was the only variable included in the original QR configuration as studied by Weerts et al. (2011) and López López et al. (2014). To account for heteroscedasticity, the forecast was transformed into the Gaussian domain. However, the rise rates and the forecast errors do not lend themselves for linear quantile regression after such a transformation. Therefore, it is difficult to combine these two variables. A possible solution could be to build regression models for subsets of the transformed data. However, such an approach drastically decreases the amount of data available for each model.

The proposed QR method is robust to the size of training dataset, which is convenient if stationarity cannot be assumed (Milly et al., 2008), a step-change in the river regime has occurred, or – as is the case for most river forecast centers – only recent forecast data have been archived. However, the performance of the method does depend on the river gage, the lead time, event threshold and year that are being forecast. This results in a very wide range of Brier Skill Scores. This means that the danger remains that forecast users make good experiences with a forecast in one year or at one location and assume it is equally reliable in other locations and every year. As is the case with most other forecasts, an indication of uncertainty needs to be communicated alongside the exceedance probabilities generated by our approach.

The proposed approach performs less well for longer lead times, for gages far upstream a river or close to confluences, for low event thresholds and extremely high ones. The model might be performing less well for low event thresholds, because the variance in the dependent variable – the forecast error – is smaller. After all, river forecasts have much smaller errors for lower water levels. In turn, for extremely high water levels, the scarcity of data decreases the model performance.

Future work

The methods can be further developed in several ways to achieve higher Brier Skill Scores and more robustness. First, more independent variables can be added. Trials with a different method, classification trees, showed that the observed precipitation, the precipitation forecast (i.e., POP – probability of precipitation) and the upstream water levels significantly improve models. Presumably, this is the case, because the QPF-forecast includes the precipitation forecast only for the next 12 h. However, currently, the precipitation data and forecasts can only be requested in chunks of a month, three chunks per day, from the NCDC's HDSS Access System.[23] For a period of 12 years, requesting such data for several weather stations[24] is obviously time-consuming. Upstream water levels can easily be included after manually determining the upstream gage(s) for each of the 82 NCRFC gages. To improve model performance at gages close to river confluences, the upstream water level of the gages on the joining river should be included as well.

Different approaches of sub-setting the data to improve models results also warrant consideration. Particularly, clustering the data by variability seems promising. However, early trials indicated that this method is very sensitive to the training dataset.

[23]URL: http://cdo.ncdc.noaa.gov/pls/plhas/HAS.FileAppSelectfidata setname=9957ANX, last access: July 2014.

[24]The geographical units of the weather forecasts bulletins do not correspond with those of the river forecast bulletins.

As mentioned above, the QR method works less well for low than for high event thresholds. Further study should investigate, why that is the case, and identify possible solutions. The current study focused on extremely high event thresholds, i.e., flood stages, but not on lower ones, i.e., below the 50th percentile of observed water levels.

Last, the proposed method would need to be verified for gages for which the NCRFC does not publish daily forecasts. Ignorance of the uncertainty inherent in river forecasts have had some of the most unfortunate impacts on decision-making in Grand Forks, ND and Fargo, ND (Pielke, 1999; Morss, 2010). Both of those stages are discontinuously forecast NCRFC gages.

Acknowledgements. To ensure anonymity, this section will be added after the review process.

References

Alexander, M., Harding, M., and Lamarche, C.: Quantile regression for time-series-cross-section-data, International Journal of Statistics and Management System, 4, 47–72, 2011.

Anon.: Brier Score, Wikipedia Free Encycl., available at: http://en.wikipedia.org/wiki/Brier_score (last access: 27 August 2014), 2014.

Bogner, K., Pappenberger, F., and Cloke, H. L.: Technical Note: The normal quantile transformation and its application in a flood forecasting system, Hydrol. Earth Syst. Sci., 16, 1085–1094, doi:10.5194/hess-16-1085-2012, 2012.

Brier, G. W.: Verification of forecasts expressed in terms of probability, Mon. Weather Rev., 78, 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2, 1950.

Bröcker, J.: Estimating reliability and resolution of probability forecasts through decomposition of the empirical score, Clim. Dynam., 39, 655–667, doi:10.1007/s00382-011-1191-1, 2012.

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., 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, B. Am. Meteorol. Soc., 95, 79–98, doi:10.1175/BAMS-D-12-00081.1, 2013.

Ferro, C. A. T. and Fricker, T. E.: A bias-corrected decomposition of the Brier Score, Q. J. Roy. Meteor. Soc., 138, 1954–1960, doi:10.1002/qj.1924, 2012.

Hsu, W. and Murphy, A. H.: The attributes diagram a geometrical framework for assessing the quality of probability forecasts, Int. J. Forecasting, 2, 285–293, doi:10.1016/0169-2070(86)90048-8, 1986.
Ikeda, M., Ishigaki, T., and Yamauchi, K.: Relationship between Brier Score and area under the binormal ROC curve, Comput. Meth. Prog. Bio., 67, 187–194, doi:10.1016/S0169-2607(01)00157-2, 2002.
Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification: A Practitioner's Guide in Atmospheric Science, John Wiley & Sons, 2012.
Kelly, K. S. and Krzysztofowicz, R.: A bivariate meta-Gaussian density for use in hydrology, Stoch. Hydrol. Hydraul., 11, 17–31, doi:10.1007/BF02428423, 1997.
Koenker, R.: Quantile Regression, Cambridge University Press, New York, 2005.
Koenker, R.: quantreg: Quantile Regression, R Package Version 505, available at: http://CRAN.R-project.org/package=quantreg (last access: 27 August 2014), 2013.
Koenker, R. and Bassett, G.: Regression quantiles, Econometrica, 46, 33–50, doi:10.2307/1913643, 1978.
Koenker, R. and Machado, J. A. F.: Goodness of fit and related inference processes for quantile regression, J. Am. Stat. Assoc., 94, 1296–1310, doi:10.1080/01621459.1999.10473882, 1999.
Leahy, C. P., Sri Srikanthan, G. Amirthanathan, Soori Sooriyakumaran, and Hydrology Unit: Objective Assessment and Communication of Uncertainty in Flood Warnings, in: 5 th Flood Management Conference Warrnamboll, 9–12 October 2007, 1–6, 2007.
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, Hydrol. Earth Syst. Sci., 18, 3411–3428, doi:10.5194/hess-18-3411-2014, 2014.
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.
Montanari, A. and Brath, A.: A stochastic approach for assessing the uncertainty of rainfall–runoff simulations, Water Resour. Res., 40, W01106, doi:10.1029/2003WR002540, 2004.
Montanari, A. and Grossi, G.: Estimating the uncertainty of hydrological forecasts: a statistical approach, Water Resour. Res., 44, W00B08, doi:10.1029/2008WR006897, 2008.

Morss, R. E.: Interactions among flood predictions, decisions, and outcomes: synthesis of three cases, Natural Hazards Review, 11, 83–96, doi:10.1061/(ASCE)NH.1527-6996.0000011, 2010.
Morss, R. E., Lazo, J. K., and Demuth, J. L.: Examining the use of weather forecasts in decision scenarios: results from a US survey with implications for uncertainty communication, Meteorol. Appl., 17, 149–162, doi:10.1002/met.196, 2010.
National Research Council: Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts, National Academies Press, Washingtion, DC, available at: http://www.nap.edu/catalog.php?record_id=11699&utm_expid=4418042-5.krRTDpXJQISoXLpdo-1Ynw.0 (last access: 18 September 2014), 2006.
NCAR-Research Applications Laboratory, N.-R. A.: verification: Weather Forecast Verification Utilities, available at: http://cran.r-project.org/web/packages/verification/index.html (last access: 27 August 2014), 2014.
Pielke, R. A.: Who decides? Forecasts and responsibilities in the 1997 Red River Flood, Applied Behavioral Science Review, 7, 83–101, 1999.
R-Core Team: R: A language and Environment For Statistical Computing, available at: http://ww.R-project.org/ (last access: 27 August 2014), 2014.
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, J. Hydrol., 497, 80–96, doi:10.1016/j.jhydrol.2013.05.028, 2013.
Seo, D.-J., Herr, H. D., and Schaake, J. C.: A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction, Hydrol. Earth Syst. Sci. Discuss., 3, 1987–2035, doi:10.5194/hessd-3-1987-2006, 2006.
Solomatine, D. P. and Shrestha, D. L.: A novel method to estimate model uncertainty using machine learning techniques, Water Resour. Res., 45, W00B11, doi:10.1029/2008WR006839, 2009.
Stephenson, D. B., Coelho, C. A. S., and Jolliffe, I. T.: Two extra components in the Brier Score decomposition, Weather Forecast., 23, 752–757, doi:10.1175/2007WAF2006116.1, 2008.
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), Hydrol. Earth Syst. Sci., 15, 255–265, doi:10.5194/hess-15-255-2011, 2011.
Welles, E., Sorooshian, S., Carter, G., and Olsen, B.: Hydrologic verification: a call for action and collaboration, B. Am. Meteorol. Soc., 88, 503–511, doi:10.1175/BAMS-88-4-503, 2007.
Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press, San Diego, 1995.
Wilson, L. J.: Verification of Probability and Ensemble Forecasts, available at: http://www.swpc.noaa.gov/forecast_verification/Assets/Tutorials/Ensemble Forecast Verification.pdf, last access: 27 August 2014.
WWRP/WGNE: Methods For Probabilistic Forecasts, Forecast Verification – Issues, Methods and FAQ, available at: http://www.cawcr.gov.au/projects/verification/verif_web_page.html#BSS (last access: 27 August 2014), 2009.

Table 1. Variable combinations.

Combi	fcst	err24	err48	rr24	rr48	Combi	fcst	err24	err48	rr24	rr48
1	●					16	●	●	●		
2		●				17	●	●		●	
3			●			18	●	●			●
4				●		19	●		●	●	
5					●	20	●		●		●
6	○	○				21	●			●	●
7	○		○			22		●	●	●	
8	○			○		23		●	●		●
9	○				○	24		●		●	●
10		○	○			25			●	●	●
11		○		○		26	○	○	○	○	
12		○			○	27	○	○	○		○
13			○	○		28	○	○		○	○
14			○		○	29	○		○	○	○
15				○	○	30		○	○	○	○
						31	●	●	●	●	●

fcst = forecast; rr24, rr48 = rise rate in the past 24 and 48 h;
err24, err 48 = forecast error 24 and 48 h ago

Table 2. Error statistics for the forecast error (a) of the whole dataset, (b) on days that the water level did not exceed the 10th percentile of observations, (c) on days that the water level exceeded the 90th percentile of observations, (d) on days that the water level exceeded minor flood stage. For easier reading, the mean values are in bold.

Average errors of 82 gages	Lead Time Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
(a) All Observations						
Minimum	−0.21	−0.08	−0.09	−0.07	−0.04	0.02
Median	0.01	0.02	0.06	0.13	0.22	0.30
Mean	**0.01**	**0.04**	**0.10**	**0.18**	**0.30**	**0.41**
Maximum	0.19	0.21	0.76	1.65	2.62	3.47
(b) Observations < 10th Percentile						
Minimum	−1.2	−0.35	−0.38	−0.41	−0.38	−0.39
Median	−0.03	−0.04	−0.05	−0.05	−0.04	−0.04
Mean	**-0.06**	**-0.06**	**-0.06**	**-0.06**	**-0.05**	**-0.04**
Maximum	0.03	0.04	0.05	0.12	0.17	0.25
(c) Observations > 90th Percentile						
Minimum	−0.11	−0.23	−0.31	−0.38	−0.38	−0.27
Median	−0.01	0.02	0.15	0.32	0.55	0.81
Mean	**0.01**	**0.09**	**0.29**	**0.55**	**0.82**	**1.14**
Maximum	0.34	1.01	3.12	5.13	6.81	8.56
(d) Observations > Flood Stage						
Minimum	−0.20	−0.30	−0.44	−0.63	−0.78	−0.80
Median	−0.02	−0.03	0.22	0.45	0.78	1.10
Mean	**0.01**	**0.17**	**0.45**	**0.80**	**1.14**	**1.51**
Maximum	0.65	2.44	5.70	8.37	10.40	11.74

Table 3. Means and standard deviations of Brier Skill Scores resulting from three QR configurations: the original using the transformed forecast only as independent variable; the best performing combination for each river gage (upper performance limit); rise rates in the past 24 and 48 h and the forecast errors 24 and 48 h ago as independent variables (one-size-fits-all solution).

	Q10	Q25	Q75	Q90	Q10	Q25	Q75	Q90
	Day 1				Day 2			
NQT-fcst	0.34 (0.52)	0.65 (0.36)	0.90 (0.07)	0.88 (0.08)	0.24 (0.57)	0.59 (0.35)	0.85 (0.10)	0.82 (0.12)
Best combi.s	0.54 (0.34)	0.78 (0.18)	0.93 (0.05)	0.91 (0.06)	0.49 (0.36)	0.74 (0.19)	0.90 (0.05)	0.87 (0.07)
Rise rate 24/48 + error 24/48*	0.49 (0.41)	0.77 (0.18)	0.92 (0.05)	0.93 (0.06)	0.42 (0.44)	0.73 (0.19)	0.90 (0.06)	0.86 (0.09)
	Day 3				Day 4			
NQT-fcst	0.20 (0.61)	0.56 (0.33)	0.81 (0.10)	0.75 (0.15)	0.19 (0.55)	0.55 (0.31)	0.77 (0.13)	0.69 (0.18)
Best combi.s	0.47 (0.37)	0.74 (0.17)	0.89 (0.05)	0.85 (0.09)	0.46 (0.37)	0.73 (0.18)	0.89 (0.05)	0.84 (0.09)
Rise rate 24/48 + error 24/48*	0.40 (0.44)	0.72 (0.19)	0.88 (0.06)	0.84 (0.11)	0.39 (0.43)	0.71 (0.20)	0.88 (0.05)	0.82 (0.20)
	Action	Minor	Moderate	Major	Action	Minor	Moderate	Major
	Day 1				Day 2			
NQT-fcst	0.81 (0.27)	0.42 (1.12)	0.38 (1.02)	−0.80 (2.07)	0.68 (0.59)	0.41 (0.90)	0.25 (1.2)	−1.30 (1.96)
Best combi.s	0.86 (0.26)	0.78 (0.27)	0.73 (0.24)	0.36 (0.66)	0.82 (0.29)	0.73 (0.28)	0.68 (0.24)	0.26 (0.67)
	Day 3				Day 4			
NQT-fcst	0.67 (0.37)	0.37 (0.87)	−0.09 (1.42)	−1.69 (2.24)	0.62 (0.35)	0.22 (1.00)	−0.07 (1.05)	−1.52 (1.96)
Best combi.s	0.81 (0.26)	0.71 (0.31)	0.64 (0.23)	0.19 (0.76)	0.79 (0.26)	0.69 (0.30)	0.60 (0.23)	0.13 (0.72)

* Combination 30.

Table 4. Regression results.

	Coef.	SD	
Intercept	−0.206	0.031	***
Event thresholds	0.265	0.003	***
Lead Times	−0.021	0.003	***
Forecast Years			
2004	−0.266	0.020	***
2005	−0.081	0.018	***
2006	−0.125	0.017	***
2007	−0.129	0.017	***
2008	−0.203	0.017	***
2009	−0.125	0.016	***
2010	−0.140	0.017	***
2011	−0.128	0.016	***
2012	0.056	0.017	***
2013	−0.054	0.016	***
Number of Years in Training Dataset	0.001	0.001	
River Gages			***
For the sake of brevity, the 82 river gages included in the regression as factors are omitted here.			
R^2		0.26	
Adjusted R^2		0.25	
P values: *** – < 0.001; ** – 0.01; * – 0.05; . – 0.1			

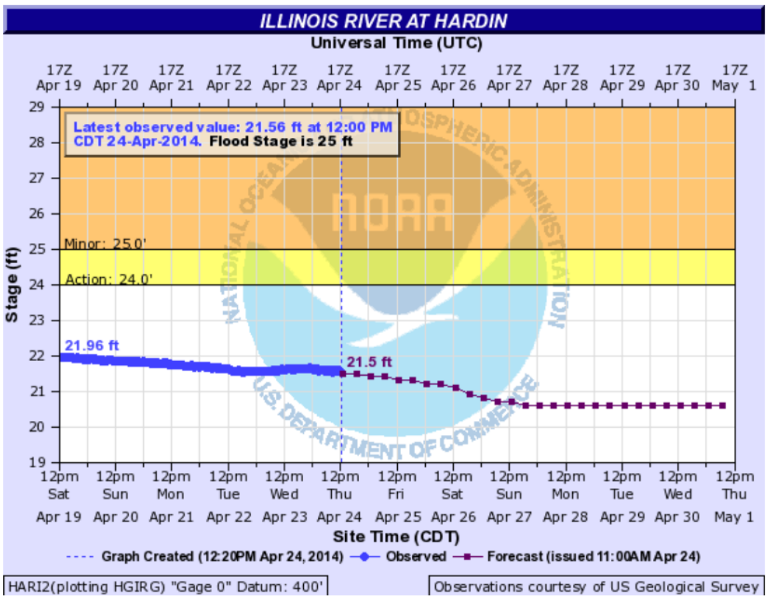

Figure 1. Deterministic short-term weather forecast in six hour intervals as published by the NWS for Hardin, IL on 24 April 2014. Source: http:// r.weather.gov/ahps2/hydrograph.php?wfo=lsx&gage=hari2 (last access: 1 October 2014).

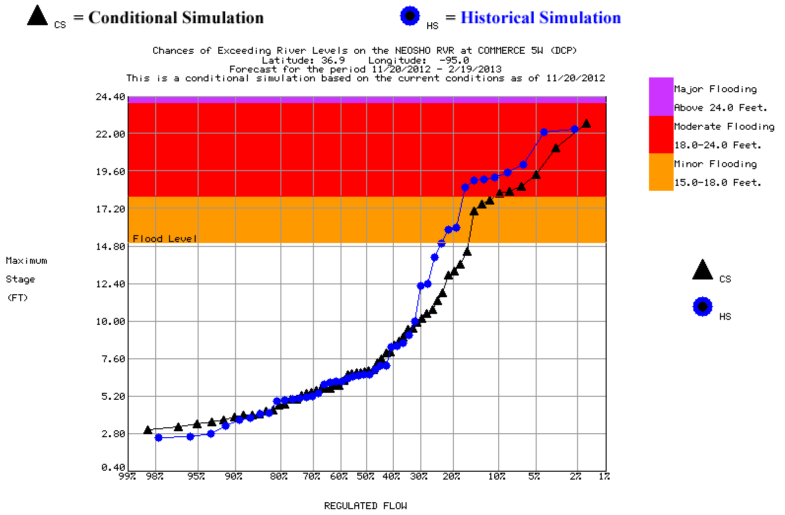

Figure 2. Probabilistic long-term forecast as published by the NWS for Commerce, OK on 14 December 2012: exceedance curve for three months period. (Not available for Hardin, IL.) Source: http://water.weather.gov/ahps2/probability_information.php?wfo=tsa&gage=COMO2&graph_id=2 (last access: 1 October 2014).

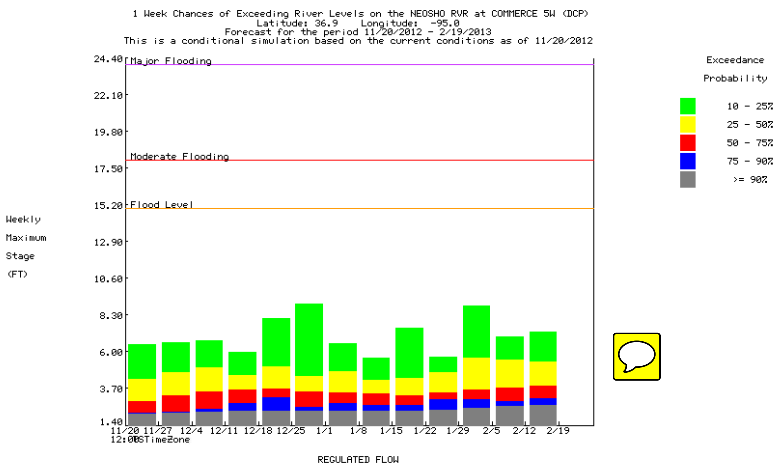

Figure 3. Probabilistic long-term forecast as published by the NWS for Commerce, OK on 14 December 2012: bar plot for each week of a three months period. (Not available for Hardin, IL.) Source: http://water.weather.gov/ahps2/probability_information.php?wfo=tsa&gage=COMO2&graph_id=0 (last access: 1 October 2014).

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

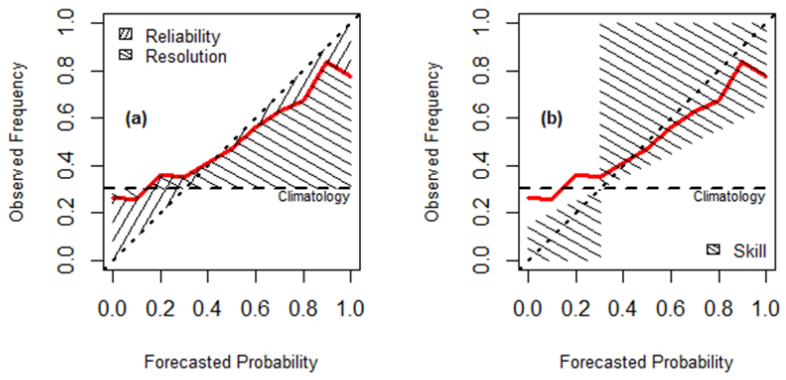

Figure 4. Theory behind Brier Skill Score illustrated for an imaginary forecast (red line): **(a)** reliability and resolution; **(b)** skill. In **(a)**, the area representing reliability should be as small, and for resolution as large as possible. The forecast has skill (BSS > 0), i.e. performs better than random guessing, if it is inside the shaded area in **(b)**. Ideally, the forecast would follow the diagonal (BSS = 1). (Adapted from Hsu and Murphy, 1986; Wilson, n.d.)

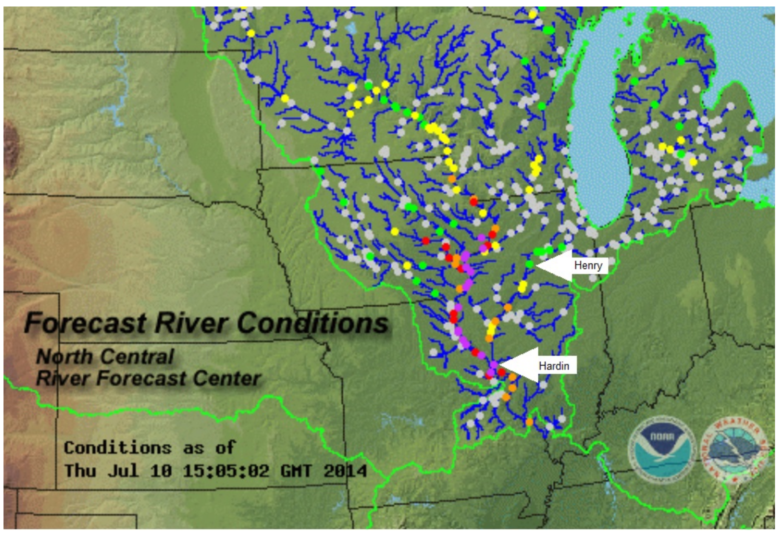

Figure 5. Portion of the North Central River Forecast Centers river gages with Henry (HYNI2) and Hardin (HARI2) indicated by the upper and lower red arrow respectively. Source: http://www.crh.noaa.gov/ncrfc/.

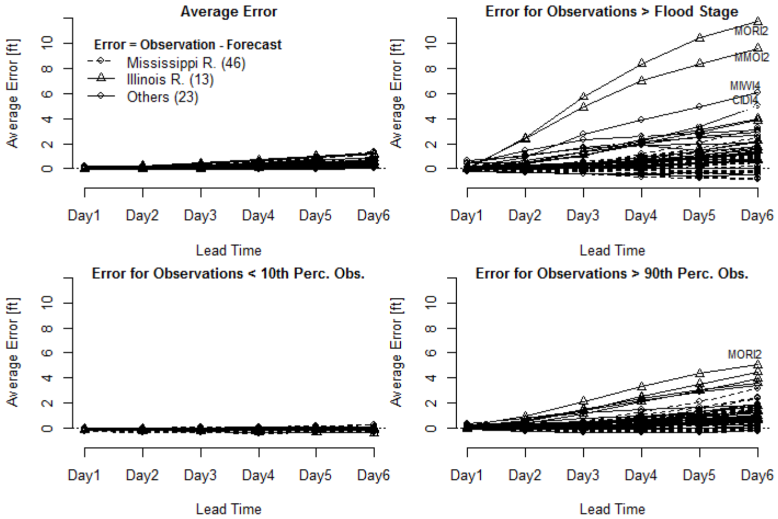

Figure 6. Forecast error for 82 river gages that the NCRFC publishes daily forecasts for. In anti-clockwise direction starting at the top left: **(a)** average error; **(b)** error on days that the water level did not exceed the 10th percentile of observations; **(c)** error on days that the water level exceeded the 90th percentile of observations; **(d)** error on days that the water level exceeded minor flood stage.

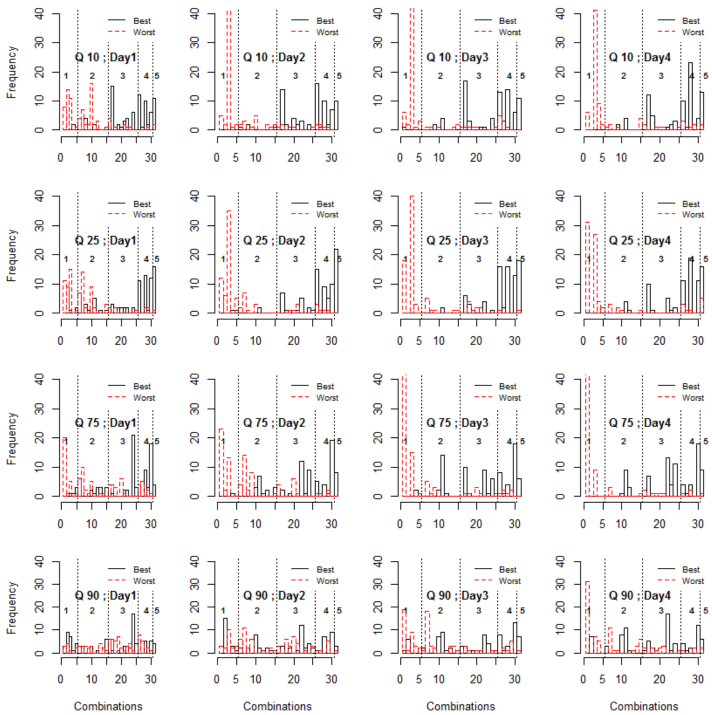

Figure 7. Histograms of variable combinations returning the best and worst Brier Skill Scores across 82 river gages. Each row of histograms refers to an event threshold defined as a percentile of the observed water levels, and each column to a lead time. The dotted vertical lines in the histograms distinguish variable combinations with different numbers of variables.

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

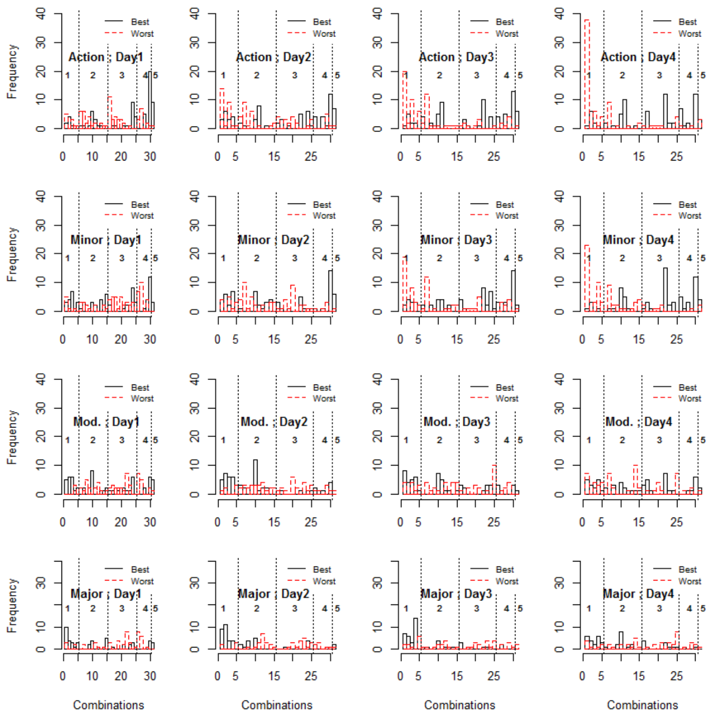

Figure 8. Histograms of variable combinations returning the best and worst Brier Skill Scores across 82 river gages. Each row of histograms refers to a flood stage, and each column to a lead time. The dotted vertical lines in the histograms distinguish variable combinations with different numbers of variables.

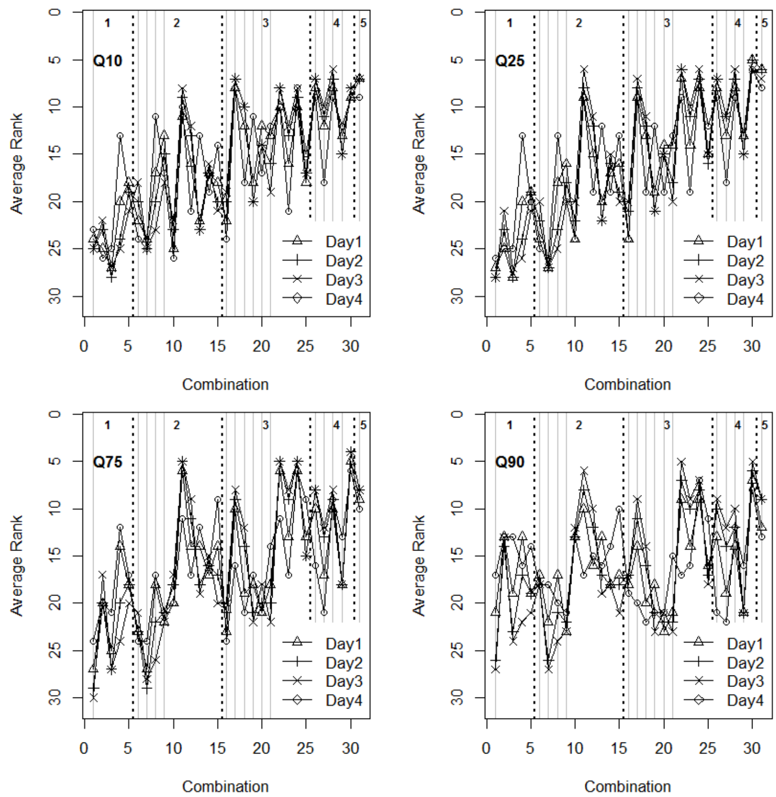

Figure 9. Average rank for each variable combination for one to four days of lead time and four percentiles of observed water levels. Vertical gray lines indicate variable combinations including the forecast.

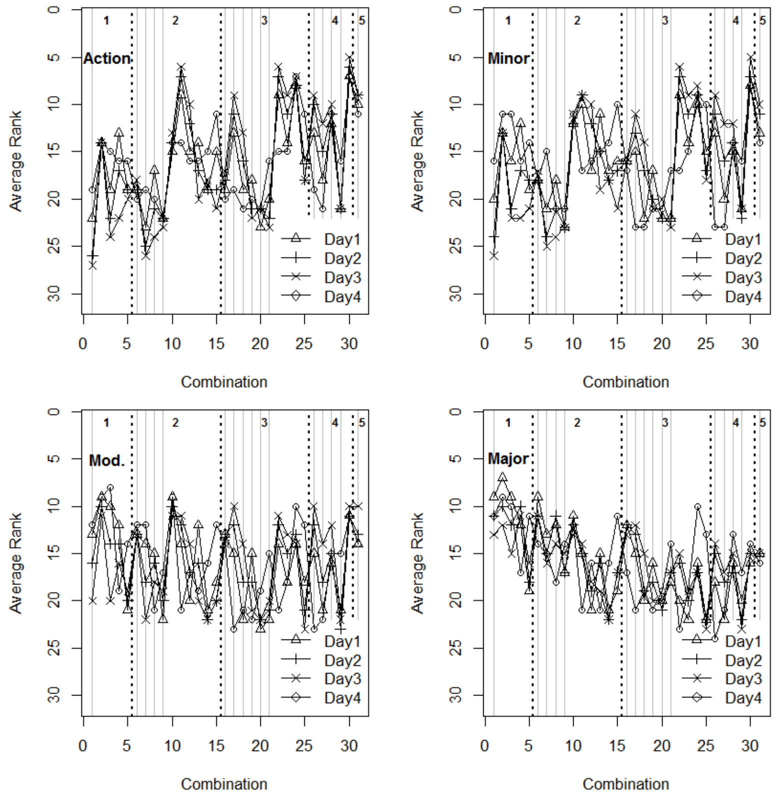

Figure 10. Average rank for each variable combination for one to four days of lead time and four flood stages. Vertical gray lines indicate variable combinations including the forecast.

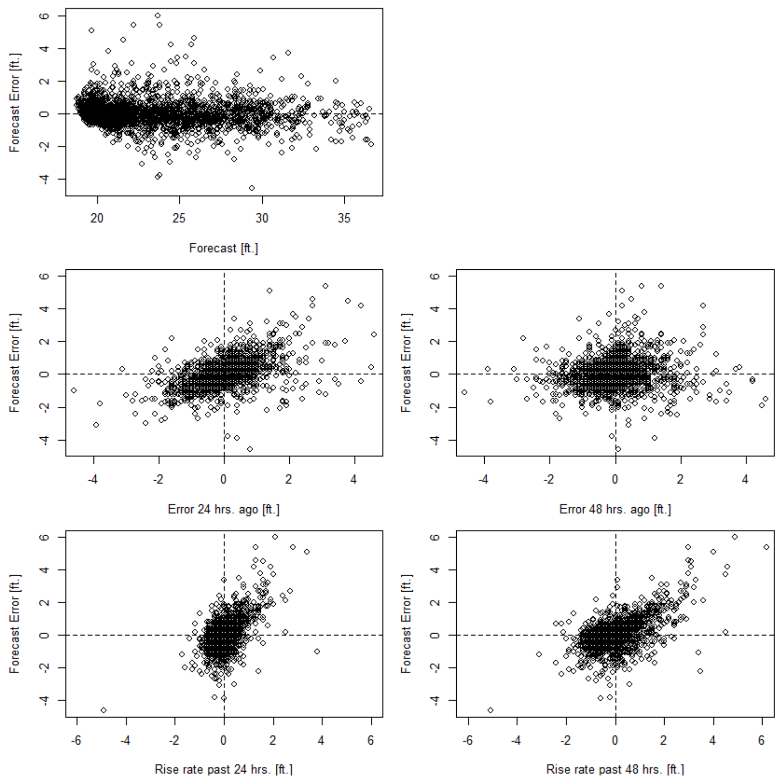

Figure 11. Independent variables plotted against the forecast error for Hardin IL with 3 days of lead time. First row: forecast; second row: past forecast errors; third row: rise rates.

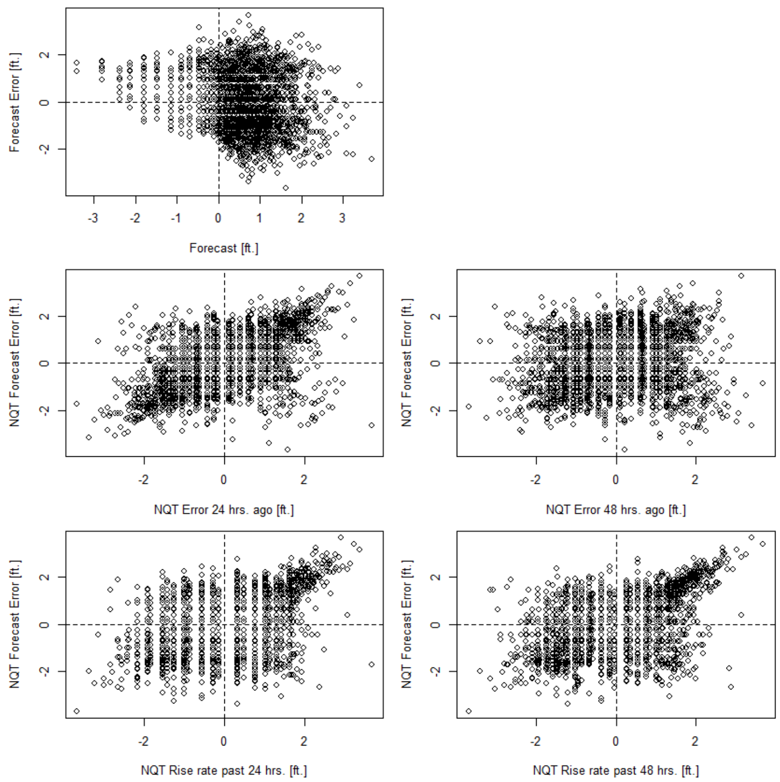

Figure 12. Independent variables after transforming into the Gaussian domain plotted against the forecast error for Hardin IL with 3 days of lead time. First row: forecast; second row: past forecast errors; third row: rise rates.

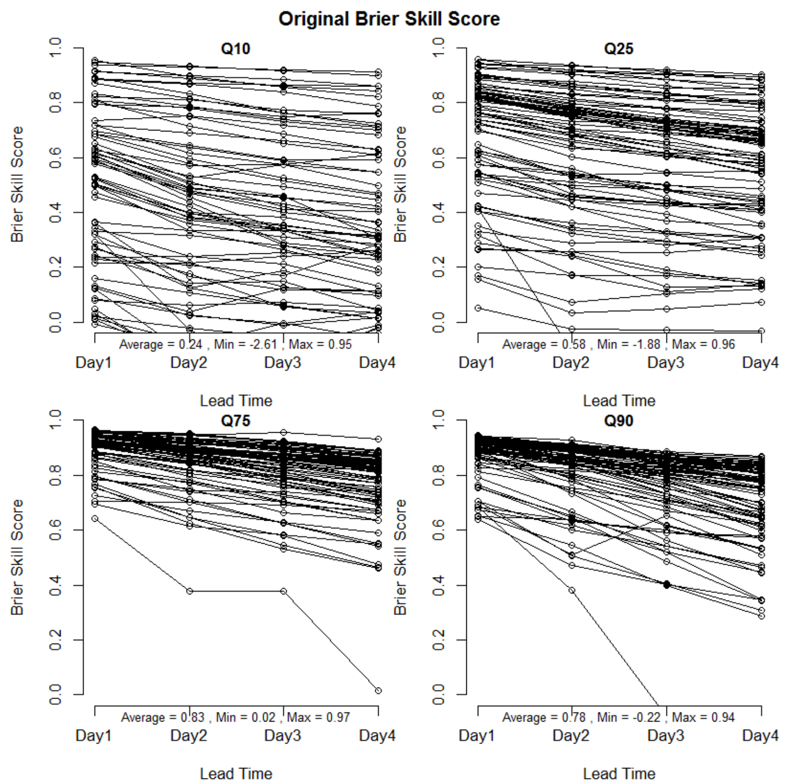

Figure 13. Brier Skill Scores of the original QR model (i.e., using the transformed forecast as the only independent variable) for four lead times and percentiles of observed water levels.

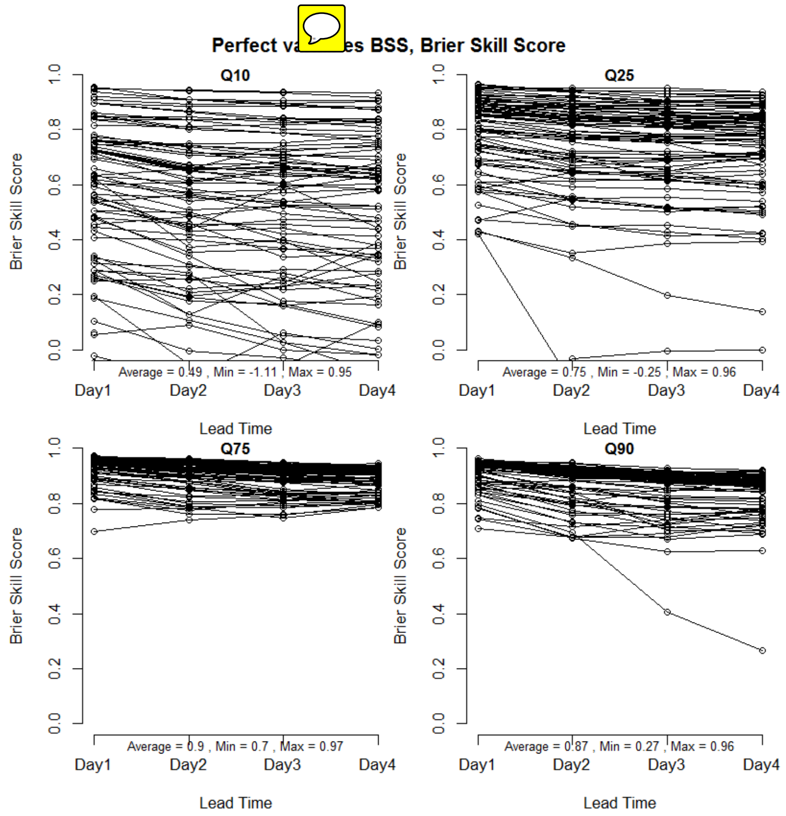

Figure 14. Brier Skill Scores for four lead times and percentiles of observed water levels using the best variable combination for each river gage as independent variables in the QR model.

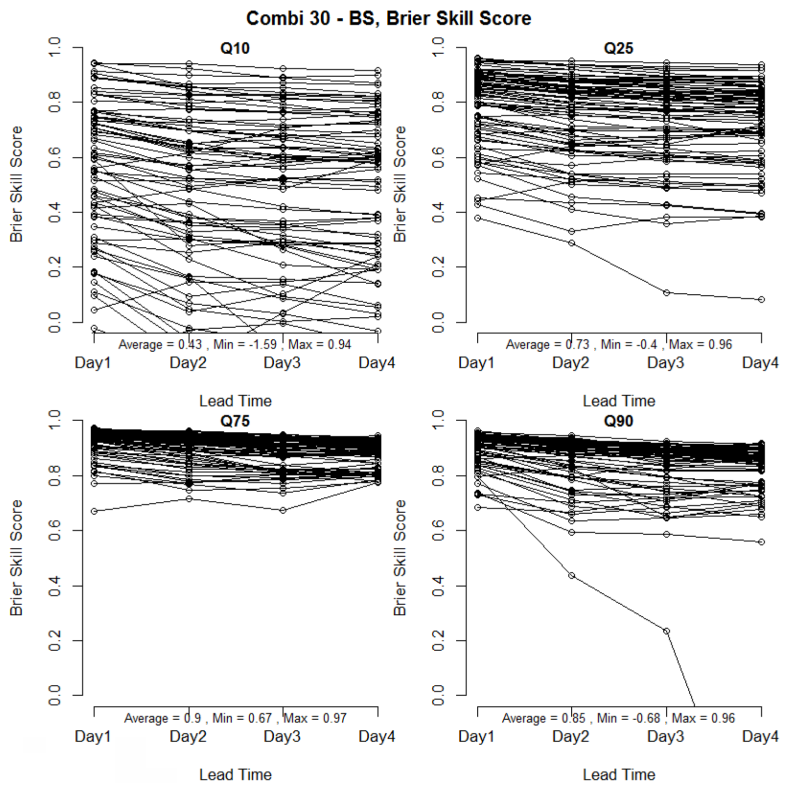

Figure 15. Brier Skill Scores for four lead times and percentiles of observed water levels using a one-size-fits-all approach (i.e., rr24, rr48, err24, err48) for the independent variables in the QR model.

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

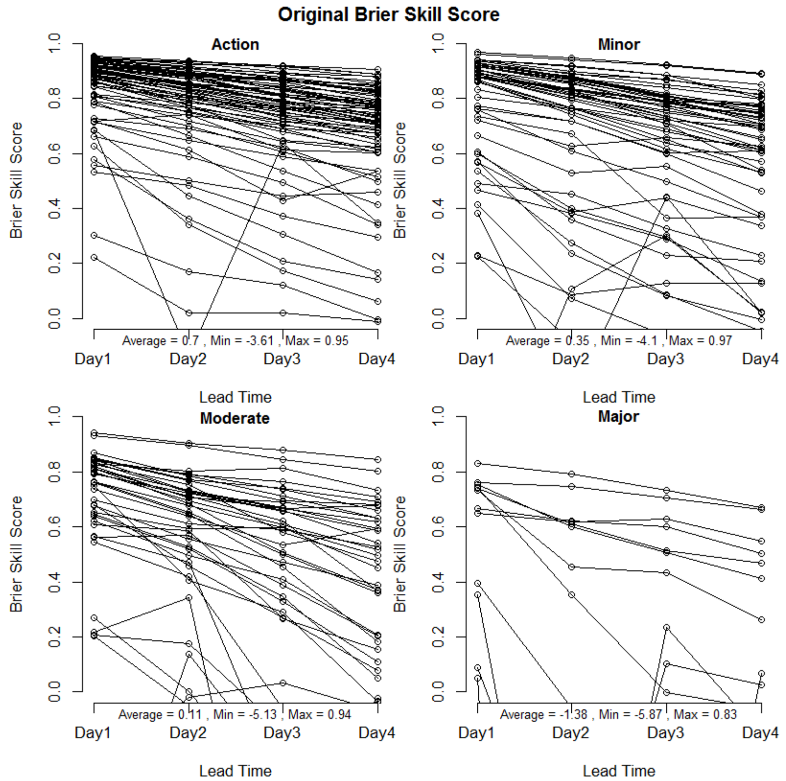

Figure 16. Brier Skill Scores of the original QR model (i.e., using the transformed forecast as the only independent variable) for four lead times and flood stages.

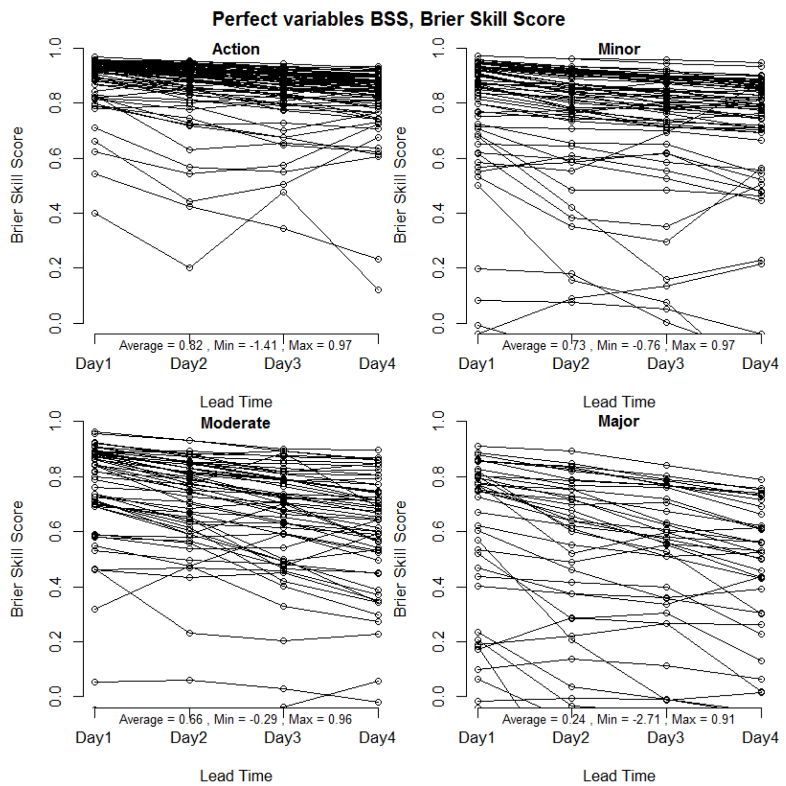

Figure 17. Brier Skill Scores for four lead times and flood stages of observed water levels using the best variable combination for each river gage as independent variables in the QR model.

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

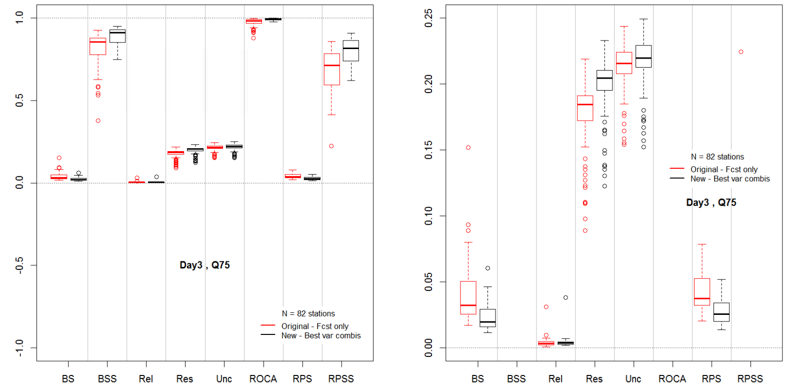

Figure 18. Comparison of the original QR model (i.e., only transformed forecast as independent variables) and the one-size-fits-all approach (i.e., rise rates and forecast errors as independent variables) using various measures of forecast quality: Brier Score (BS), Brier Skill Score (BSS), Reliability (Rel), Resolution (Res), Uncertainty (Unc), Area under the ROC curve (ROCA), ranked probability score (RPS), ranked probability skill score (RPSS). Lead time: 3 days; 75th percentile of observation levels as threshold. The left figure zooms in on the right figure to make changes in reliability and resolution better visible.

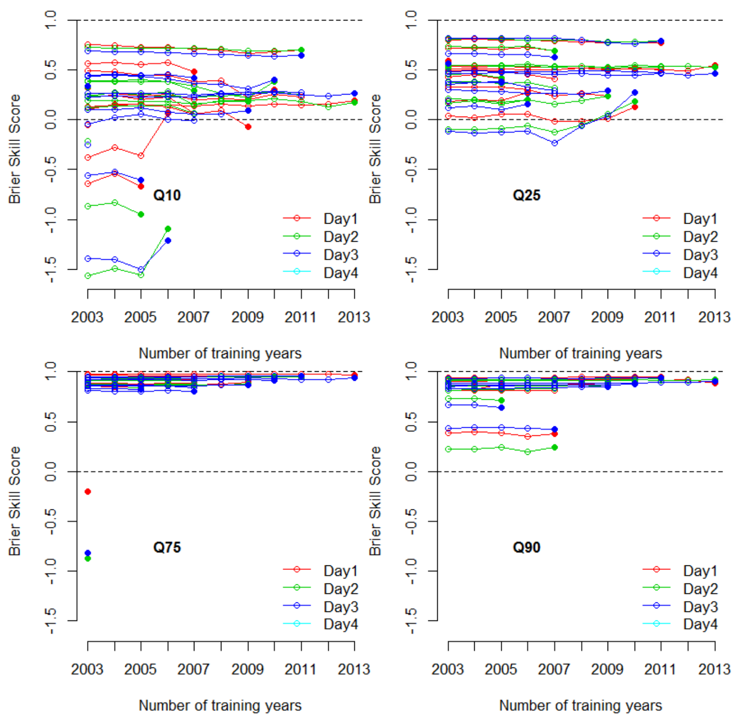

Figure 19. Brier Skill Score for various forecast years and various sizes of training dataset across different lead times (colors) and event thresholds (plots) for Hardin, IL (HARI2). The filled-in end point of each line indicates the BSS for the forecast year on the x axis with one year in the training dataset. Each point further to the left stands for one additional training year for that same forecast year.

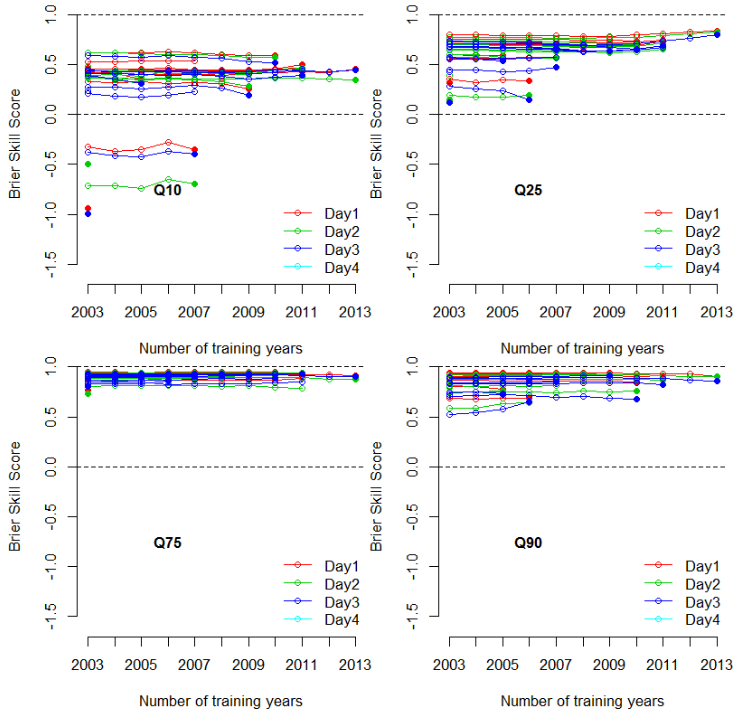

Figure 20. Brier Skill Score for various forecast years and various sizes of training dataset across different lead times (colors) and event thresholds (plots) for Henry, IL (HNYI2). The filled-in end point of each line indicates the BSS for the forecast year on the *x* axis with one year in the training dataset. Each point further to the left stands for one additional training year for that same forecast year.

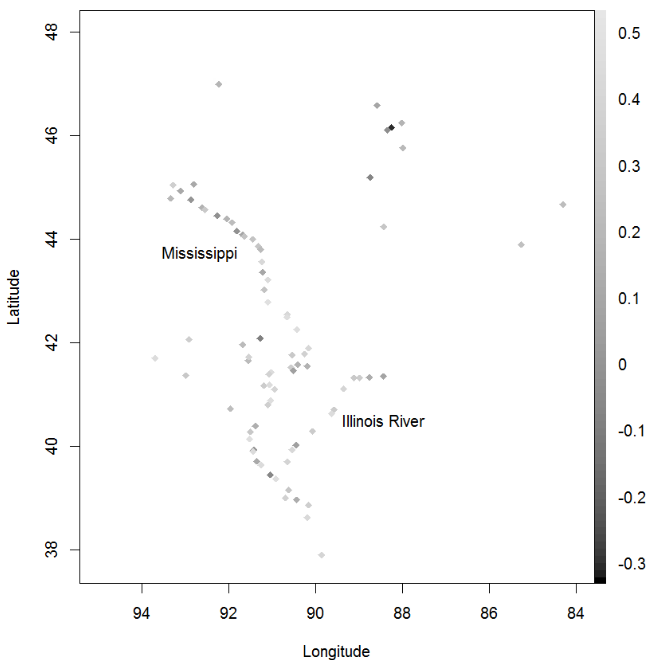

Figure 21. Geographical position of rivers. Colors indicate the regression coefficient of each station with the Brier Skill Score as dependent variable.

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

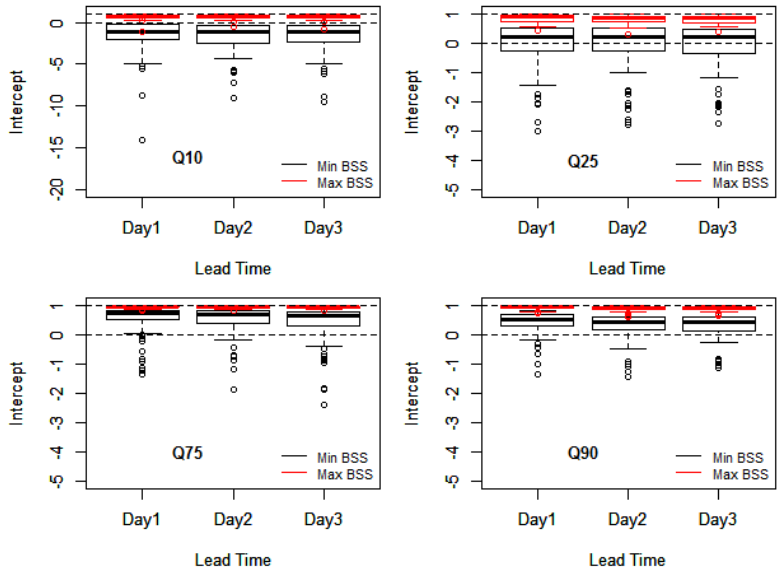

Figure 22. Minimum (black) and maximum (red) Brier Skill Scores for various lead times and event thresholds across locations, size of training dataset and forecast years.