Third Revision Article *Hydrology and Earth System Sciences*

Title:

Performance and Robustness of Probabilistic River Forecasts Computed with Quantile Regression based on Multiple Independent Variables in the North Central U.S.A.

Authors:

Frauke Hoss, Paul S. Fischbeck

Affiliation:

Carnegie Mellon University

Department of Engineering & Public Policy

5000 Forbes Avenue

Pittsburgh, PA 15213

Corresponding Author:

Frauke Hoss: fraukehoss@gmail.com

Performance and Robustness of Probabilistic River Forecasts Computed with Quantile Regression based on Multiple Independent Variables in the North Central U.S.A.

Abstract

This study applies Quantile Regression (QR) to predict exceedance probabilities of various water levels, including flood stages, with combinations of deterministic forecasts, past forecast errors and rates of water level rise as independent variables. A computationally cheap technique to estimate forecast uncertainty is valuable, because many national flood forecasting services, such as the National Weather Service (NWS), only publish deterministic single-valued forecasts. The study uses data from the 82 river gages, for which the NWS' North Central River Forecast Center issues forecasts daily. Archived forecasts for lead times up to six days from 2001-2013 were analyzed. Besides the forecast itself, this study uses the rate of rise of the river stage in the last 24 and 48 hours and the forecast error 24 and 48 hours ago as predictors in QR configurations. When compared to just using the forecast as independent variable, adding the latter four predictors significantly improved the forecasts, as measured by the Brier Skill Score and the Continuous Ranked Probability Score. Mainly, the resolution increases, as the forecast-only QR configuration already delivered high reliability. Combining the forecast with the other four predictors results in much less favorable performance. Lastly, the forecast performance does not strongly 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 configuration or at least calibration.

Keywords: River forecasts, quantile regression, probabilistic forecasts, robustness

1 Introduction

River-stage forecasts are no crystal ball; the future remains uncertain. The past has shown that unfortunate decisions have been made, because of users' unawareness of the magnitude of potential forecast errors (Pielke, 1999; Morss, 2010). For many users, such as emergency managers, forecasts are most important in extreme situations, such as droughts and floods. Unfortunately, it is exactly in those situations that forecast are the most uncertain, i.e., forecast errors are the largest, due to the infrequency and the subsequent scarcity of data.

Currently, the National Weather Service does not routinely publish uncertainty information along with their deterministic short-term river-stage forecast (Figure 1). 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. 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 forecasts in such situations. Including uncertainty in river forecast would therefore be valuable, just as has been recommended for weather forecasts in general (e.g., National Research Council, 2006). Hopefully, decision-makers would then consider the whole bandwidth of possible future water levels, rather than focusing on the best estimate that is currently being published.

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://water.weather.gov/ahps2/hydrograph.php?wfo=lsx&gage=hari2.

There are two types of approaches to estimate forecast uncertainty (e.g., Leahy, 2007; Demargne et al., 2013; Regonda et al., 2013): Those addressing major sources of uncertainty individually, 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 and disadvantages. When source of uncertainty are modelled separately, their different characteristics can be taken into account (e.g., some sources of uncertainty depend on lead time, while others do not). Consequently, the approach addressing major source of output uncertainty is likely to result in better performing, more parsimonious model configurations. On the downside, this approach is expensive to develop, maintain and run. The alternative, i.e., the lumped quantification of uncertainties, is a less demanding in development and computation run-time, but glosses over many of the finer details of uncertainties (Regonda et al., 2013).

Most previously developed post-processors to generate probabilistic forecasts share the overall set-up but differ in their implementation. Independent 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 measures 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). These techniques differ in a number of ways, including their sub-setting of data, and the output metric. Please see Regonda et al. (2013) and Solomatine & Shrestha (2009) for a summary of each technique.

The National Weather Service has chosen to quantify the most significant sources of uncertainty using ensemble techniques (Demargne et al., 2013). The NWS has developed the Hydrologic Ensemble Forecast Service (HEFS) to be able to provide short-term and medium-term probabilistic forecasts (Demargne et al., 2013). HEFS includes a post-processor, the Hydrologic Ensemble Post-Processing (EnsPost). It models the hydrological uncertainty by estimating the probability distribution for each of the ensemble members which have been produced with varying input to account for input uncertainty (NWS-OHD, 2013). The

Experimental ensemble forecast service (XEFS) additionally features the more parsimonious Hydrologic Model Output Statistics (HMOS) Streamflow Ensemble Processor, which estimates the total uncertainty (input and hydrological uncertainty) of single-valued streamflow forecasts based on conditional probability distributions (U.S. Department of Commerce/NOAA, 2012).

This paper further develops one of the techniques mentioned above: the Quantile Regression approach to post-process river forecasts first introduced by Wood et al. (2009) and further elaborated by Weerts et al. (2011) and López López et al. (2014). In a comparative analysis of four different post-processing techniques to generate confidence intervals, the quantile regression technique was one of the two most reliable techniques (Solomatine and Shrestha, 2009), while being the mathematically least complicated and requiring few assumptions. After Wood et al. (2009) presented the proof-of-concept for the Lewis River in Washington State at a conference, Weerts et al. (2011) published a formal study of quantile regression to compute confidence intervals for river-stage forecasts. Weerts et al. (2011) 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. When applying QR to river forecasts, Weerts et al. (2011) transformed the deterministic forecasts and the corresponding forecast errors into the Gaussian domain using Normal Quantile Transformation (NQT) to account for heteroscedasticity. Building on Weerts et al. (2011) study, López López et al. (2014) compare different configurations of QR with the forecast as the only independent variable, including configurations without NQT and preventing the crossing of quantiles. They found that no configuration was consistently superior for a range of forecast quality measures (López López et al., 2014).

This paper combines elements of the studies mentioned above. In some aspects, our approach differs from those three studies. We predict the exceedance probabilities of flood stages rather than uncertainty bounds. Additionally, we are fortunate to have a much larger dataset than the three earlier studies, consisting of archived forecasts for 82 river gages covering 11 years. Furthermore, we introduce additional predictors, as was suggested by López López et al. (2014). This study does not add to the mathematical technique of quantile regression itself.

The proposed QR approach is similar to the HMOS approach, but it differs in the following ways. First, HMOS uses ordinary linear regression instead of quantile regression. Second, the QR method uses the single-valued forecast, rates of rise and past forecast errors as independent variables, while HMOS includes recently observed and current flows, and quantitative precipitation forecasts (QPF) as predictors. Third, in this paper QR models are built for a number of event thresholds, whereas HMOS develops models for subsets of forecasted streamflows (Regonda et al., 2013).

Identifying the best-performing set of independent variables is central to this paper. All possible combinations of the following predictors have been studied: forecast, the rate of rise of water levels in past hours, and the past forecast errors. Additionally, the robustness of the resulting QR configurations across different sizes of training datasets, locations, lead times, water levels, and forecast year has been assessed.

The paper is structured as follows. The Data section describes the used data and reviews the overall forecast error for the dataset. The Method section introduces quantile regression and the performance measures, and discusses the performed analyses. The Results describes the results of identifying the best-performing set of independent variables. Additionally, it discusses the robustness of the studied QR configurations. The fourth and last section presents the conclusions and proposes further research ideas.

2 Data

The National Weather Service (NWS)'s daily short-term river forecasts predict the stage height in six-hour intervals for up to six days ahead (20 6-hour intervals). 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 (National Climatic Data Center, 2014; Station ID: KMSR, Bulletin ID: FGUS5). In total, the NCRFC produces forecasts for 525 gages. For 82 of those gages, forecasts have been published daily for at least two 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 (Figure 2). 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). Figure 3 shows an empirical cumulative density function of drainage areas sizes.

Figure 2: River gages for which the North Central River Forecast Centers publishes forecasts daily. Henry (HYNI2) and Hardin (HARI2) are indicated by the upper and lower red arrow respectively. For gages indicated by black dots the basin size is missing. The color scale for basin size in square miles is logarithmic.

Figure 3: Empirical cumulative density function (ecdf) of sizes of drainage area for the river gages that are being forecasted daily by the NCRFC.

Two river gages serve as an illustration for the points made throughout this paper. These two gages were chosen to capture different, but representative conditions. Hardin, IL is just upstream of the confluence of the Illinois River and the Mississippi River (Figure 2). Therefore, it can experience backwatering, when the high water levels in the Mississippi River prevent the Illinois River from draining. Henry, IL is located ~200 miles upstream of Hardin, having a difference in elevation of ~25 feet. The Illinois River is ~330 miles long (Illinois Department of Natural Resources, 2011), draining an area of ~13,500 square miles at Henry (USGS, 2015a) and ~28,700 square miles at Hardin (USGS, 2015b). The number of case studies has been limited to two because of computation time.

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 (Figure 4). 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 (Figure 4a, Figure 5). Extremely low water levels, defined as below the 10th percentile of observed water levels, are also well calibrated (Figure 4b, Figure 5). However, when considering higher water levels the picture changes. When only observations exceeding the 90th percentile of all observations are considered, 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 (Figure 4c, Figure 5). When only looking at observations that exceeded the minor flood stages corresponding to each gage, the underestimation averages 0.45 feet for three days of lead time and 1.51 feet for 6 days of lead time (Figure 4d, Figure 5). 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 12 feet for water levels higher than minor flood stage.

The gages MORI2 and MMOI2 are upstream of a dam. Possibly, the forecasts performed so poorly there, because the dam operators deviated from the schedules that they provide the river forecast centers to base their calculations on. In sum, predicting the forecast error distribution as is done in this paper has much added value for forecast users, because the forecast error can amount to several feet.

Figure 4: 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 5: Empirical cumulative distribution function (ecdf) of forecast error at 82 river gages for six lead times. Vertical lines show the median forecast error of the corresponding subset.

3 Method

Quantile Regression (QR) is used to estimate the distribution of river-stage forecasts for each forecast point in time and location. This information can be published in a number of formats to suit the needs of the forecast users. Wood et al. (2009) and Weerts et al. (2011) chose to study confidence intervals. A confidence interval is the range between two points on the estimated forecast distribution, e.g., between the 10th and 90th percentile. Our paper differs in that our output is the probability of exceeding a flood stage. A flood stage and the corresponding probability of it being exceeded are represented by a single point on the estimated forecast distribution. Assessing forecast performance for a single point rather than for two points on the estimated distribution allows for scrutinizing forecast performance more closely, not least because the method is not necessarily equally successful in both tails of the distribution.

In the following, quantile regression itself and the analysis to identify the best-performing set of independent variables are explained.

3.1 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 (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.

López López et al. (2014) did not find that the quantile regression method produces better forecasts if the variables are subject to NQT beforehand, as was practiced by Weerts et al. (2011). We chose not to apply NQT, because four of five of our independent variables are already approximately normally distributed; only the forecast itself is not.

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 independent variables. 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. Detailed instructions to perform NQT can be found in Bogner et al. (2012). Mathematically, the approach is formulated as follows (with and without NQT):

Equation 1: QR configuration *with* NQT , with percentiles of the forecast error as the dependent variable and the one independent variable, bot transformed into the normal domain.

$$F_\tau(t) = fcst(t) + NQT^{-1}[\sum_{i}^{I} a_{i,\tau} * V_{NQT,i}(t) + b_\tau]$$

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

$$F_\tau(t) = fcst(t) + \sum_{i}^{I} a_{i,\tau} * V_i(t) + b_\tau$$

with	$F_\tau(t)$	– estimated forecast associated with percentile τ and time t
	fcst(t)	– original forecast at time t
	$V_i(t)$	– the independent variable i (e.g., the original forecast) at time t
	$V_{i;NQT}(t)$	– the independent variable I transformed by NQT at time t
	$a_{i,\tau}$, b_τ	– configuration coefficients

The second part of the equations stands for the error estimate based on the quantile regression configuration for each error percentile τ and lead time. In Equation 1, that was used by Weerts et al. (2011), this estimation was executed in the Gaussian domain using only the forecast as independent variable. Our study mainly uses Equation 2, i.e., it does not transform the predictors and the predictand. All quantile regressions were done using the command *rq()* in the R-package “quantreg” (Koenker, 2013).

3.2 Verification Measures

The QR configuration by Weerts et al. (2011) was evaluated by determining the fraction of observations that fell into the confidence intervals predicted by the QR configuration; i.e., ideally, 80% 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 measures to assess configuration performance, e.g., the Brier Skill Score (BSS), the mean continuous ranked probability (skill) score (CRPSS), the relative operating characteristic (ROC), and reliability diagrams to compare QR configurations.

We focus on the Brier Skill Score (BSS) – first introduced by Brier (1950) – to assess QR configurations for two reasons. First, to be able to determine the best set of predictors it is easiest to choose a single measure. Second, the BSS allows us to study forecast performance at individual event thresholds. Third, out of the available measures the Brier Score is attractive, because it can be decomposed into two different measures of forecast quality (see Equation 3): Reliability and resolution. The third component is uncertainty. This type of uncertainty describes the uncertainty inherent in an event caused by natural variability. It is narrower than forecast uncertainty, because the latter additionally includes the uncertainty that is caused by imperfections of the forecast model, i.e., the variables that could explain some of the uncertainty have not been identified or correctly parameterized yet. In sum, the BS' uncertainty term is not subject to the forecast quality. Equation 3 gives the definition of the (de-composed) Brier Score (e.g., Jolliffe and Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009).

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

$$BS = \underbrace{\frac{1}{N}\sum_{k=1}^{K} n_k(f_k - \bar{o}_k)^2}_{\text{Reliability}} - \underbrace{\frac{1}{N}\sum_{k=1}^{K} n_k(\bar{o}_k - \bar{o})^2}_{\text{Resolution}} + \underbrace{\bar{o}(1-\bar{o})}_{\text{Uncertainty}} = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

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
$\bar{o}_k$ – the frequency of binary event occurring on days in which forecast falls into bin k
f_k – forecast probability
$\bar{o}$ – 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 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. The reliability curve for the forecast representing perfect reliability would follow the diagonal in Figure 6, i.e., the area in Figure 6a representing reliability would equal zero (Jolliffe and Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009).

Figure 6: Theory behind Brier Skill Score illustrated for an imaginary forecast (red line): (a) reliability and resolution; (b) skill. In figure 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 the reference forecast, if it is inside the shaded area in the figure b. Ideally, the forecast would follow the diagonal (BSS=1). (Adapted from Hsu and Murphy, 1986; Wilson, n.d.).

Resolution measures the difference between the predicted probability of an event on a given day and the historically observed average probability. For example, imagine a gage where flood stage has historically been exceeded on 5% of the days in a year. If every day at that gage the probability of exceeding flood stage is forecasted to be 5%, the resolution of those forecasts would be zero. After all, the difference between the predicted frequency and the historical average is zero. So a forecast with higher resolution is better. (e.g., Jolliffe and Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009). In Figure 6, the curve for a forecast with good resolution would be steeper than the dashed line that represents the historically observed frequency (climatology). It follows that forecasters should strive to maximize the area in Figure 6a representing resolution. In absolute terms, the resolution can never exceed the uncertainty inherent to the river gage, as represented by the third term in Equation 3. (e.g., Jolliffe and Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009).

A forecast performs better than the reference forecast (in this case the historically observed frequency), if it (the red line) is inside the shaded area in Figure 6b. Then the forecast is said to have “skill”. The Brier *Skill* Score (BSS) equals the Brier Score normalized by the historically observed frequency, i.e., the resolution and reliability terms are being divided by the uncertainty term (Equation 4). In contrast to the Brier Score, this makes the Brier Skill Score comparable across gages with different frequencies of a binary event. The BSS can range from minus infinity to one. A BSS below zero indicates no skill; the perfect score is one (e.g., Jolliffe and Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009).

Equation 4: Decomposition of Brier Skill Score

$$BSS = 1 - \frac{BS}{\bar{o}(1-\bar{o})} = \frac{RES}{\bar{o}(1-\bar{o})} - \frac{REL}{\bar{o}(1-\bar{o})}$$

with BSS – Brier Skill Score
BS – Brier Score
RES – Resolution
REL – Reliability
ō – Frequency of binary event occurring
ō(1−ō) – Climatological variance

To verify that the results hold up for verification measures other than the BSS, we additionally use the Continuous Ranked Probability Score (CRPS). The BSS assesses forecast performance for one point on the forecast distribution, i.e., one event threshold. In contrast, the

CRPS, defined by Equation 5, measures the forecast performance for the forecast distribution as the whole. Therefore, the CRPS cannot detect whether the forecast does better or worse in the tails. Instead, it is a measure of the forecast's overall performance. The CRPS' perfect score equals zero (e.g., Jolliffe and Stephenson, 2012; WWRP/WGNE, 2009).

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

Equation 5:

$$CRPS = \frac{1}{N}\sum_{n=1}^{N}\int_{-\infty}^{\infty}(F_n^f(x) - F_n^o(x))^2 dx$$

with CRPS – Continuous Ranked Probability Score

F_n^f (x) – Forecast probability distribution (cdf) for the n-th forecast case

F_n^o (x) – Observation for n-th forecast case (feet)

N – Number of forecast cases, i.e., length of time series

3.3 Choice of independent variables

The challenge is to identify a well-performing QR model with a set of predictors that is both parsimonious and comprehensive. Wood et al. (2009) found rate of rise and lead time to be informative independent variables. Weerts et al. (2011) achieved good results using only the forecast itself as predictor. Besides these variables, the most obvious predictors to include are the current water levels and those observed 24 and 48 hours ago, and the forecast error 24 and 48 hours ago (i.e., the difference between the current water level at issue time of the forecast that the error distribution is being predicted for, and the forecasts that were produced 24 and 48 hours earlier to predict the current water level). Additional potential independent variables are the water levels observed at gages up- and downstream at various times, the precipitation upstream of the catchment area, and the precipitation forecast.

Rates of rise and forecast errors were chosen to complement the forecast as independent variables for the following reasons. So instead of using it as an independent variable, separate

QR models have been built for each lead time. After all, the best choice of independent variables might depend on lead time. Precipitation and precipitation forecast were not available for this study, because without direct access to the database at the National Climatic Data Center (NCDC) requesting that data is a very lengthy effort.

Forecasts and observed water levels were readily accessible from NCDC databases. Rates of rise and forecast errors can be derived from those two. As will be shown in section 4.3, it is mathematically challenging to combine independent variables with different distributions into a joint predictor. Forecast and observed water levels have a skewed distribution, because low water levels occur more frequently than extremely high water levels, while rates of rise and forecast error are approximately normally distributed. Accordingly, either forecasts and observations can easily be combined into a joint predictor, or rates of rise and forecast errors. For this study the latter option was chosen for the following reasons. Observed water levels are systematically included in the NWS forecast model. Assuming a well-defined NWS forecast model, there should not be statistical relationship between forecast error and observed water levels. In comparison, rates of rise and forecast error are only included in the NWS model at the discretion of the individual forecaster. Therefore, these latter two variables are likely to contribute more information to predicting the distribution of forecast errors than the forecasts and observed water levels. Nonetheless, forecasts were included as predictor in this study to demonstrate the difficulty of combining variables with a skewed distribution with normally distributed variables into a joint predictor, and because it served as the only independent variable in previous studies (Weerts et al., 2011; López López et al., 2014).

To determine which set of predictors performs best in generating probabilistic forecasts, all 31 possible combinations of the forecast (fcst), the rate of rise in the last 24 and 48 hours (rr24, rr48), and the forecast error 24 and 48 hours ago (err24, err48) – see Equation 5 – were tested for 82 gages that the NCRFC issues forecasts for every morning (Table 1). Based on the Bier Skill Score, it was determined which joint predictor delivers on average the best out-of-sample forecast performance for various lead times and water levels.

Equation 5: QR configuration without NQT, with percentiles of the forecast error as the dependent variable and varying combinations of the five independent variables. This equation was used to predict the water level distribution for each day at 82 gages with different lead times.

$$F_\tau(t) = fcst(t) + a_{fcst,\tau} * fcst(t) + a_{rr24,\tau} * rr24(t) + a_{rr48,\tau} * rr48(t) \\ + a_{err24,\tau} * err24(t) + a_{err48,\tau} * err48(t) + b_\tau$$

with	$F_\tau(t)$	– estimated forecast associated with percentile τ and time t
	fcst(t)	– original forecast at time t
	rr24(t), rr48(t)	– rates of rise in the last 24 and 48 hours at time t
	err24(t), err48(t)	– forecast errors 24 and 48 hours ago (e.g., the original forecast) at time t
	$a_{xx,\tau}$, b_τ	– configuration coefficients; forced to be zero if the predictor is excluded from the joint predictor that is being studied.

Table 1: Joint predictors

3.4 Computational process

The final output of the computational process 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 at location X is 60%." This is done in two steps. First, a training dataset (first half of the data) is used to define one quantile regression configuration for each percentile of the error distribution π = [0.05, 0.1, 0.15, … , 0.85, 0.90, 0.95] and each lead time. The dependent variable is the forecast error, i.e. the difference between forecast and observed water level. To recap, depending on configuration (Table 1) the forecast itself, the rates of rise and forecast errors serve as independent variables.

In the second step, these QR configurations are used to predict percentile by percentile the distribution of forecast error for 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 forecast errors is predicted. Adding the single-valued forecast to the forecast error distribution results in a distribution of predicted water levels. Each estimated percentile π contributes one point to that distribution.

Then, 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. Next, the Brier Skill Score is determined based on predicted exceedance probability for all days in the verification dataset.

To study whether the various combinations of predictors 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 repeated for eight event thresholds: the 10th, 25th, 75th, and 90th percentile of observed water levels and the four decision-relevant flood stages (action stage, and minor, moderate, and major flood stage) of each gage. Flood stages indicated when material damage or substantial hinder is caused by high water levels. Therefore, the flood stages correspond with different percentiles at different river gages. To determine the best-performing set of independent variables, the entire procedure is repeated for each of the 31 joint predictors in Table 1, thus using a different set of independent variables each time. The robustness of the technique was tested analyzing its performance for 82 gage locations using different lengths of data sets for five different lead times.

4 Results

In total, the Brier Skill Score (BSS) for 31 joint predictors (Table 1) across various lead times and event threshold have been compared. Across 82 river gages, it has been analyzed which joint predictor delivers the best BSSs on average. When informative, the CRPS has been used as an additional measure of forecast performance.

4.1 Identifying best performing joint predictors on average

For each river gage, the combinations have been ranked by BSSs. The best performing combination was ranked first, the worst performing 31st. It was found that the more independent variables are included in a joint predictor, the higher that set of predictors will rank on average (Figure 7, Table 2a). Apparently, every additional independent variable does add information. In other words, the future forecast error is a function of rates of rise and past forecast errors. Rising water levels are difficult to anticipate and therefore a common source of forecast error, because precipitation is a major source of input uncertainty. For example, it is never completely certain into which river basin the rain will fall. Additionally, only the expected precipitation for the coming 12 hours is currently included in forecasts, regardless of lead time. The past forecast errors are a measure of the magnitude of impact those unanticipated developments are likely to have.

For extremely high water levels, this trend favoring larger joint predictors gradually reverses (Figure 8). The trend remains statistically significant, but its coefficient decreases for higher event thresholds (Table 2a) until it changes signs for major flood stages (Table 2b). A possible explanation is that combinations with more variables suffer from overfitting for extreme event thresholds characterized by data scarcity.

Table 2: Results of regression analyses to determine the impact of including more variables and the forecast into the joint predictor

Figure 7: Average rank for each joint predictor for one to four days of lead time and two percentiles of observed water levels. Vertical gray lines correspond to the configurations that include forecast as one of the predictors. The y-axis is reversed, so that an increasing trend indicates increasing performance.

Figure 8: Average rank for each joint predictor for one to four days of lead time and the two highest flood stages. Vertical gray lines correspond to the configurations that include forecast as one of the predictors. The y-axis is reversed, so that an increasing trend indicates increasing performance.

The results hold up when CRPS instead of BSS is used as a measure of forecast performance. The average rank of joint predictors based on CRPS is proportional to the average rank as measured by the BSS previously (Figure 9). However, scores themselves are not proportional (Figure 10), because the BSS assesses one point on the estimated distribution, while the CRPS measures the forecast performance for the distribution as a whole. Figure 10 shows that BSS and CRPs correspond well for event thresholds Q25 and Q75. However, the BSS indicates that in the tails (Q10, Q90) the forecast does not perform as well, i.e., despite equally good CRPS scores the BSS varies widely.

Figure 9: Comparing average rank across 82 gages based on Brier Skill Score and CRPS.

Figure 10: Comparing the performance of combination 30 [err24, err48, rr24, rr48] as measured Brier Skill Score and as measured by the Continuous Ranked Probability Score. Each data point corresponds with a gage at a certain lead time. Since the CRPS' perfect score equals zero, the y-axis has been reversed.

4.2 Combining differently distributed variables into a joint predictor

The combinations including the forecast (indicated by gray vertical lines in Figure 7 and Figure 8) perform significantly better than those that exclude it (Table 2). This disadvantageous impact of forecast as an independent variable is less pronounced for very high or low event thresholds

(Table 2a). Including the forecast into the joint predictor is even beneficial for major flood stages (Table 2b), when joint predictors with less rather than more variables perform better.

The forecast is difficult to combine with the other four predictors (err24/48, rr24/48), because their statistical distributions are different. Unlike the dependent variable (forecast error), the forecasts are highly skewed towards the left, because low water levels occur more frequently. Due to its skewed distribution, the forecast becomes a better predictor in a quantile regression predicting a normally distributed dependent variable after a NQT transformation, as successfully used by Weerts et al. (2011). Without a transformation into the normal domain, the scatterplot of forecast and forecast error does not show obvious quantile trends (Figure 11a). After NQT, the percentiles show distinct quantile trends laid out like a fan (Figure 12a).

In contrast, errors and rise rates are already approximately normally distributed. There are no quantile trends visually detectable anymore after the other four predictors have been subjected to NQT (Figure 11 b-e). In sum, forecast performance in this study is better without NQT, because four of the five independent variables were approximately normally distributed already. Further research is necessary to reconcile predictors with different distributions. Possible solutions could be to define QR configurations for subsets of the transformed dependent and independent variables or to experiment with subjecting only some, but not all predictors to NQT.

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: rates of rise.

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: rates of rise.

4.3 Improvement in forecast performance

Using the best performing joint predictor at each river gage gives an upper bound of the BSSs that can be achieved at best. Confirming Wood et al.'s findings (2009), additionally including the rates of rise and forecasts errors as independent variables into the QR configuration improves the Brier Skill Score (BSS) significantly. Figure 13 illustrates the BSS when using the forecast as the only predictor as studied by Weerts et al. (2011), while Figure 14 shows the performance for the best joint predictor at each gage.

Figure 13: Brier Skill Scores (BSS) for forecast-only configuration for different lead times and event thresholds. The BSS' perfect score equals one. A BSS of zero indicates a forecast without skill.

Figure 14: Brier Skill Scores (BSS) for best performing the joint predictor at each gage for different lead times and event thresholds. The BSS' perfect score equals one. A BSS of zero indicates a forecast without skill.

Figure 15: Empirical cumulative density functions of three QR configurations predicting exceedance probabilities of the Action, Minor, Moderate, and Major Flood Stage: the configuration using the transformed forecast as the only independent variable [NQT fcst]; the best performing combination for each river gage (upper performance limit) [Best combis]

Figures 13 to 15 indicate that the QR method performs better for higher than for lower water levels. Due to the skewed distribution of water levels, the ranges between percentiles in the left tail (lower water levels) correspond with much smaller ranges of water levels (feet) than in the right tail. Therefore, achieving good performance in forecasting exceedance probabilities of low event thresholds requires much better prediction of forecast error in feet than for higher event thresholds.

Additionally, Figures 13 to 15 show that forecast performance also decreases with increasing lead time, because variables such as rates of rise and past forecast error become proportionally less representative with lead time.

Paired T-tests for each combination of lead time and event threshold indicate that using the best joint predictor at each gage increased average BSS across all gages statistically significantly (Table 3). The performance improves most where forecasts tend to perform worst. The average increase in BSS is largest for extreme water levels, most notably moderate and major flood stages and the 10th percentile of water levels (Table 3). The average increase of BSS for major flood stage is even larger than one, meaning that the method did frequently not have skill before, i.e., negative BSSs. Additionally, predictions with longer lead times experience larger increases in BSS. Compared to using only the forecast as an independent variable, using the best combinations of forecast, rates of rise and past forecast errors as predictors at each gage not only increases the mean BSS, but also decreases the standard deviation of skill scores across gages, i.e., performance becomes more consistent (Figures 13 and 14).

Table 3: Results of paired t-tests comparing the QR method`s performance with only forecast as predictor and the best-performing combination of five predictors for each river gage

As expected, the CRPS improves as well when using the best joint predictor at each gage instead of forecast as the only predictor. The average CRPS and its standard deviation decrease. The improvement is more pronounced for longer lead times (Figure 16). Moving away from average CRPS, Table 4 reveals that the best joint predictors for high event thresholds (Q75, Q90) do not benefit the average CRPS. The fact that the average CRPS does not improve implies that the best joint predictors for high event thresholds increase forecast performance less for high event thresholds than it worsens performance for low event thresholds. The best joint predictors for low event thresholds (Q10, Q25) do improve average CRPS. So they must be improving the forecast so substantially that the average CRPS increases, even though those best predictors might not perform well for high event thresholds. This is congruent with the finding that average BSS increases much more for percentiles Q10 and Q25 than for Q75 and Q90, as shown in Table

3. This reinforces the finding that separate QR models should be configured for individual event thresholds based on the BSS, rather than for the whole distribution based the CRPS.

Figure 16: Continuous Ranked Probability Score (CRPS) for the forecast-only configuration and for the best performing the joint predictor at each gage for different lead times and event thresholds. The CRPS' perfect score equals zero.

Table 4: Results of paired t-tests comparing the QR method`s performance with only forecast as predictor and the best-performing combination of five predictors for each river gage for the Brier score.

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 predictors than just the forecast. Table 4 summarizes the results of paired t-tests comparing the forecast-only and the best performing joint predictor for each gage for the components of the BSS as well as the CRPS.

The Brier Score and the Brier Skill Score mainly improve, because the resolution increases when using the best-performing set of independent variables at each gage (Table 4). Visualizing the improvement in forecast performance for a lead time of three days and the 75th percentile threshold (Q75), Figure 17 illustrates that the forecast-only QR configuration as studied by Weerts et al. (2011) has high reliability (i.e., the reliability is close to zero). So reliability improves statistically significantly for lower water levels (Q10, Q25), but the magnitude of improvement in reliability is by one order smaller than the improvement in resolution (Table 4).

Figure 17: Comparison of the forecast-only QR configuration (i.e., only transformed forecast as independent variables) and using the best-performing joint predictor at each gage along various measures of forecast quality: Brier Score (BS), Brier Skill Score (BSS), Reliability (Rel), Resolution (Res), and continuous ranked probability score (CRPS). Lead time: 3 days; 75th percentile of observation levels as threshold.

4.4 One-size-fits-all approach – Brier Skill Score

Combing these findings, the configurations for the various river gages can generally be based on the same joint predictor of the four independent variables excluding the forecast itself (combination 30). But for extremely high water levels, a configuration specific to each river gage has to be built in order to achieve high BSSs.

Verifying this finding, a one size-fits-all approach was tested to investigate, whether customizing the QR configuration to each river gage would be worth it. The rates of rise in the past 24 and 48 hours and the forecast errors 24 and 48 hours ago (combination 30 in Table 1) serve as independent variables for this approach. This combination of predictors has been chosen, because it performed well for most gages (see section 4.1). Furthermore, less important predictors in the combination will get small coefficients in the quantile regression. So additional variables are unlikely to do harm, but can improve the estimates at various stages. The price of opting for a joint predictor with more variables is an increase of the risk of overfitting.

Paired t-tests have been executed to investigate whether this one-size-fits all approach performs statistically significantly worse than using the best combination of predictors for each gage. It was found that this approach on average performs statistically significantly not as well as using the best-performing combination of predictors. But the difference in average BSS is small, ranging between 0.003 and 0.075 (Table 5).

However, using the best joint predictors results in much better performance for major flood stages than the one-size-fits-all approach. The average difference between average BSSs amounts to 0.21 to 0.38 (Table 5). Given that a BSS for a forecast with skill ranges between one and zero, this is a substantial difference. In sum, the same joint predictor can be used for all river gages without much loss in performance, except for extremely high water levels.

Table 5: Results of paired t-test comparing best combinations of predictors with one-size-fits-all approach.

4.5 Robustness

4.5.1 Minimum length of training dataset

Stationarity cannot always be assumed (Milly et al., 2008). River regimes can change through natural processes like sedimentation or human intervention. Those changes can occur gradually or as step-changes. This analysis of robustness is meant to determine the minimum length of the training dataset to be able to produce skillful forecasts again after a step-change using the QR method. Additionally, the analysis is meant to find out to which length the forecaster should limit the training dataset when gradual change is occurring. After all, in such a case each year further in the past is less representative of the year ahead, so that training dataset should be as short as possible.

The impact of the length of the training dataset on the configuration's performance measured by the BSS was assessed for the best joint predictor (i.e., rates of rise 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 QR configurations trained on however many years of archived forecasts were available in that year, i.e., the forecasts for 2005 is produced by a model trained on less data than those for 2013. Then, the BSS for that year (e.g., 2005 or 2013) was computed.

Figure 18 and Figure 19 show that at Henry and Hardin it barely matters for the BSS how many years are included in the training dataset. This finding is congruent with the fact that Weerts et al. (2011) were able to achieve outstanding results with the QR method using training datasets that were only two years long. Only needing short time series to define a skillful QR configuration implies (i) skillful forecasts can be produced not long after a step-change, and that

(ii) the configuration parameters can be updated regularly so that gradually changing relationships between predictors etc. can be taken into account.

Figure 18: 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 19: 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.

4.5.2 Sensitivity Analysis

Furthermore, we aim to identify the factors that impact forecast skill as quantified by the Brier Skill Score (BSS) and to generalize the result regarding training data length described for Hardin and Henry above. To do so, the same analysis as for Hardin and Henry was repeated for all 82 gages. Following that, a regression analysis was executed with the BSS as the dependent variable and event thresholds (Q10, Q25, Q75, Q90), the river gages and forecast years as independent nominal variables, and the lead time (one to four days) and number of training years as independent ratio variables. This regression is meant to identify the factors to which the forecast performance as measured by the BSS is sensitive to, i.e., which factors statistically significantly impact forecast performance.

The forecast performance was found to vary statistically significantly across all tested dimensions, except the number of training years (Table 6). This results in a very wide range of BSSs (Figure 13 and 14). 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 QR configuration based on the forecast only (not shown).

Table 6: Regression results sensitivity analysis

A closer look at the regression coefficients (Table 6) provides interesting insights. For low event thresholds, the BSSs are much worse than for high thresholds. As mentioned above, for such low event thresholds the forecast has to predict the water levels much more accurately to achieve similar forecast performance than for higher water levels due to the skewed distribution of water levels. In the lower tail, each percentile corresponds with a much shorter span of water levels than in the upper tail. Using higher resolution in the lower tail is therefore advisable.

As expected, the BSSs slightly decrease with lead time, because independent variables such as rates of rise and past forecast error gradually become less representative of the days to be forecasted.

Regarding the forecast quality for each forecast year, the regression is slightly biased. 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 6).

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 6 for the sake of brevity. Instead, Figure 20 maps the geographic position of the river gages with the color code indicating each gage's regression coefficient. The coefficient indicates the method's performance at the particular gage as compared to the average performance. The coefficients are lower, and therefore the Brier Skill Scores are lower, for gages far upstream a river, off the main stream, and those close to confluences.

Precipitation is one of the major sources of uncertainty in river forecasting. For example, if rainfall shifts by a few miles it might be raining down in a different river basin. This makes rises in water level difficult to anticipate, making rates of rise such a successful predictor of the distribution of forecast errors. However, upstream and close to confluences rates of rise and past forecast errors perform less well as predictors than elsewhere. This suggests that uncertain expected rainfall constitutes a smaller part of the overall uncertainty.

Close to confluences the joining second river adds a major part of that additional uncertainty. The interaction between the rivers increases uncertainty, in addition to the uncertainty associated with the joining river itself, e.g., the uncertain expected rainfall along its course. At upstream gages, the rates of rise possibly provide less information, because due to smaller basin sizes concentration times are shorter, i.e., water levels rise quicker. In that case, the rise in water level of the past 24 and 48 hours may not sufficiently capture rises occurring with shorter notice. The argument holds for forecast errors as well. If concentration times are short, the forecast error of 48 hours ago is not representative of those in the near future.

Figure 20: Geographical position of rivers. Colors indicate the regression coefficient of each station with the Brier Skill Score as dependent variable.

5 Conclusion

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). It further develops the application of QR to estimating river forecast uncertainty (a) comparing different sets of independent variables, (b) and testing the technique’s robustness across locations, lead times, event thresholds, forecast years and sizes of training dataset.

When compared to the configuration using only the forecast, it was found that including rates of rise in the past 24 and 48 hours and the forecast errors of 24 and 48 hours ago as independent variables improves the performance of the QR configuration, as measured by the Brier Skill Score. This confirms Wood et al.’s (2009) finding that rate of rise is a valuable predictor for QR error models. The configuration with the forecast as the only independent variable, as studied by Weerts et al. (2011), produced estimates with high reliability. Including the other four predictors mentioned above 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 independent variables perform better than those that include more. The most likely explanation is that QR configurations based on large joint predictors result in overfitting the data. In contrast to these extremely high event thresholds, larger sets of predictors work better than smaller ones for non-extreme and low event thresholds. Additionally, customizing the set of predictors to the event thresholds does not improve the BSS much, except for extremely high event thresholds, i.e. major flood stage.

When forming a joint predictor, the independent variables rates of rise and forecast errors do not combine well with the forecast itself, because the forecast has a skewed distribution, while the other predictors are approximately normally distributed. The forecast becomes an excellent predictor for linear quantile regression after NQT. However, the other four variables lose their value as predictors when subjected to NQT, because their original distribution is already approximately normal. Therefore, it is difficult to combine predictors with different distributions.

A possible solution could be to define QR configurations for subsets of the transformed data or to experiment with only subjecting some of the predictors to NQT.

This study shows the importance of configuring QR models for individual event thresholds, rather than using one configuration to estimate the whole forecast distribution. The tails are too different to use the same joint predictors and parametrization.

The studied QR configurations are relatively 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 technique depends heavily 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 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 forecast uncertainty needs to be communicated alongside the exceedance probabilities generated by our approach.

As is the case for many forecasting methods, the studied QR configurations perform less well for longer lead times, extreme event thresholds that are characterized by data scarcity, and for gages far upstream a river, off the main stream or close to confluences where different factors interact with each other. Additionally, QR configurations underperform for low event thresholds. Due to the skewed distribution of water levels, forecasts have to perform better in estimating low water levels to achieve the same BSSs as for high event thresholds, because in the lower tail each percentile spans a smaller range of water levels. Using higher resolution in the lower tail would probably improve forecast performance for low event thresholds.

Future Work

This technique can be further developed in several ways to achieve higher Brier Skill Scores and more robustness. First, more independent variables can be added. Observed precipitation, the precipitation forecast (i.e., POP – probability of precipitation) and the upstream water levels are promising candidates, because the forecast used in this study includes the precipitation forecast for only the next 12 hours. 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. For a period of 12 years, requesting such data for several weather stations is obviously time-consuming; not least, because the geographical units of the weather forecasts bulletins do not correspond with those of the river forecast bulletins. Upstream water levels can easily be included after manually determining the upstream gage(s) for each of the 82 NCRFC gages. To improve performance at gages close to river confluences, off the main stream, and the upstream water level of the gages on the joining river should be included as well.

Note though that many hydrological variables have a skewed distribution, so that they cannot readily be combined into a joint predictor with normally distributed variables such as rates of rise and past forecast errors as used in this study. Future work should focus on reconciling predictors with different distributions.

Different approaches of sub-setting the data to improve performance also warrant consideration to boost performance of the QR method. Particularly, clustering the data by variability seems promising.

Additionally, the studied technique would need to be verified for gages for which the NCRFC does not publish daily forecasts. Ignorance of the uncertainty inherent in river forecasts has 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.

Finally, this paper uses a brute force approach by simply calculating and comparing all possible combinations of independent variables. Mathematically more challenging stepwise quantile regression would not only be more elegant, but also provide better safeguards against overfitting the data.

Acknowledgements:

Many thanks to Grant Weller who suggested looking into quantile regression to predict forecast errors. We would like to thank the three reviewers for their insightful comments. The paper greatly benefitted from their comments. As to funding, Frauke Hoss is supported by an ERP fellowship of the German National Academic Foundation and by the Center of Climate and Energy Decision Making (SES-0949710), through a cooperative agreement between the National Science Foundation and Carnegie Mellon University (CMU).

References

Alexander, M., Harding, M. and Lamarche, C.: Quantile Regression for Time-Series-Cross-Section-Data, Int. J. Stat. Manag. Syst., 4(1-2), 47–72, 2011.

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(4), 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), 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2, 1950.

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

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, Bull. Am. Meteorol. Soc., 95(1), 79–98, doi:10.1175/BAMS-D-12-00081.1, 2013.

Hsu, W. and Murphy, A. H.: The attributes diagram A geometrical framework for assessing the quality of probability forecasts, Int. J. Forecast., 2(3), 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. Methods Programs Biomed., 67(3), 187–194, doi:10.1016/S0169-2607(01)00157-2, 2002.

Illinois Department of Natural Resources: Aquatic Illinois - Illinois Rivers and Lakes Fact Sheets, [online] Available from: http://dnr.state.il.us/education/aquatic/aquaticillinoisrivlakefactshts.pdf (Accessed 3 February 2015), 2011.

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(1), 17–31, doi:10.1007/BF02428423, 1997.

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

Koenker, R.: quantreg: Quantile Regression, R Package Version 505 [online] Available from: http://CRAN.R-project.org/package=quantreg (Accessed 27 August 2014), 2013.

Koenker, R. and Bassett, G.: Regression Quantiles, Econometrica, 46(1), 33, 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(448), 1296–1310, doi:10.1080/01621459.1999.10473882, 1999.

Leahy, C. P.: Objective Assessment and Communication of Uncertainty in Flood Warnings., 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. Discuss., 11(4), 3811–3855, 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(5863), 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(1), W01106, doi:10.1029/2003WR002540, 2004.

Montanari, A. and Grossi, G.: Estimating the uncertainty of hydrological forecasts: A statistical approach, Water Resour. Res., 44(12), W00B08, doi:10.1029/2008WR006897, 2008.

Morss, R. E.: Interactions among Flood Predictions, Decisions, and Outcomes: Synthesis of Three Cases, Nat. Hazards Rev., 11(3), 83–96, doi:10.1061/(ASCE)NH.1527-6996.0000011, 2010.

National Climatic Data Center: HDSS Access System, [online] Available from: http://cdo.ncdc.noaa.gov/pls/plhas/HAS.FileAppSelect?datasetname=9957ANX; (Accessed 15 July 2014), 2014.

National Research Council: Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts, National Academies Press, Washingtion, DC. [online] Available from: http://www.nap.edu/catalog.php?record_id=11699 (Accessed 18 September 2014), 2006.

National Weather Service, Office of Hydrologic Development: Ensemble Postprocessor (EnsPost) User's Manual. HEFS Release 0.3.2. [online] Available from: http://www.nws.noaa.gov/oh/hrl/general/HEFS_doc/HEFS-0.3.2_EnsPost_Users_Manual.pdf (Accessed 22 July 2015), 2013.

Pielke, R. A.: Who Decides? Forecasts and Responsibilities in the 1997 Red River Flood, Appl. Behav. Sci. Rev., 7(2), 83–101, 1999.

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.: Hydrologic Ensemble Processing Overview, [online] Available from: http://www.nws.noaa.gov/oh/hrl/hsmb/docs/hep/events_announce/Hydro_Ens_Overview_DJ.pdf (Accessed 29 January 2015), 2008.

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(4), 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, doi:10.1029/2008WR006839, 2009.

U.S. Department of Commerce, NOAA: NOAA/NWS Hydrologic Ensemble Forecasting, [online] Available from: http://www.nws.noaa.gov/ohd/XEFS/ (Accessed 22 July 2015), 2012.

USGS: Stream Site - USGS 05558300 Illinois River at Henry, IL, [online] Available from: http://waterdata.usgs.gov/nwis/inventory/?site_no=05558300&agency_cd=USGS (Accessed 2 February 2015a), 2015.

USGS: Stream Site - USGS 05587060 Illinois River at Hardin, IL, [online] Available from: http://waterdata.usgs.gov/il/nwis/inventory/?site_no=05587060& (Accessed 3 February 2015b), 2015.

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(1), 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, Bull. Am. Meteorol. Soc., 88(4), 503–511, doi:10.1175/BAMS-88-4-503, 2007.

Wikipedia: Brier score, [online] Available from: http://en.wikipedia.org/w/index.php?title=Brier_score&oldid=619686224 (Accessed 27 August 2014), 2014.

Wilson, L. J.: Verification of probability and ensemble forecasts, [online] Available from: http://www.swpc.noaa.gov/forecast_verification/Assets/Tutorials/Ensemble%20Forecast%20Verification.pdf (Accessed 27 August 2014), n.d.

Wood, A. W., Wiley, M. and Nijssen, B.: Use of quantile regression for calibration of hydrologic forecasts, [online] Available from: http://ams.confex.com/ams/89annual/wrfredirect.cgi?id=10049, 2009.

WWRP/WGNE: Methods for probabilistic forecasts. Forecast Verification – Issues, Methods and FAQ, [online] Available from: http://www.cawcr.gov.au/projects/verification/verif_web_page.html#BSS (Accessed 27 August 2014), 2009.

Tables

Table 1: Joint predictors

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 hours;
err24, err 48 = forecast error 24 and 48 hours ago

Table 2: Results of regression analyses to determine the impact of including more variables and the forecast into the joint predictor

(a) PERCENTILES of observed water levels

Independent Variable: **Rank (1 to 31)**	**Q10** **Coef (St.Err.)**	**Q25** **Coef (St.Err.)**	**Q50** **Coef (St.Err.)**	**Q75** **Coef (St.Err.)**
Intercept	26.49 (.21) ***	27.54 (.19) ***	24.47 (.19) ***	20.09 (.22) ***
Number of variables	-4.47 (.08) ***	-5.59 (.08) ***	-4.98 (.08) ***	-3.02 (.09) ***
Forecast included? (binary)	2.01 (.17) ***	5.15 (.16) ***	8.51 (.16) ***	7.18 (.18) ***
R^2	0.23	0.34	0.33	0.17
Adjusted R^2	0.23	0.34	0.33	0.17

P-Values: *** – <0.001; ** – 0.01; * – 0.05; **.** – 0.1

(b) FLOOD STAGES

Independent Variable: **Rank (1 to 31)**	**Action FS** **Coef (St.Err.)**	**Minor FS** **Coef (St.Err.)**	**Moderate FS** **Coef (St.Err.)**	**Major FS** **Coef (St.Err.)**
Intercept	20.92 (.22) ***	18.76 (.23) ***	15.49 (.27) ***	12.58 (.29) ***
Number of variables	-3.33 (.09) ***	-2.40 (.09) ***	-0.22 (.11) *	1.59 (-12) ***
Forecast included? (binary)	7.11 (.18) ***	6.68 (.19) ***	2.02 (.22) ***	-1.30 (.24) ***
R^2	0.18	0.13	0.01	0.03
Adjusted R^2	0.18	0.13	0.01	0.03

P-Values: *** – <0.001; ** – 0.01; * – 0.05; **.** – 0.1

Table 3: Results of paired t-tests comparing the QR method`s performance with only forecast as predictor and the best-performing combination of five predictors for each river gage

	1 Day				2 Days				3 Days				4 Days			
	Diff.	**T-stat.**	**Df**	**p-val.**	**Diff.**	**T-stat.**	**Df**	**p-val.**	**Diff.**	**T-stat.**	**Df**	**p-val.**	**Diff.**	**T-stat.**	**Df**	**p-val.**
Q10	0.20	8.68	80	.000	0.25	8.98	79	.000	0.28	8.53	79	.000	0.27	10.08	79	.000
Q25	0.13	6.06	81	.000	0.15	7.10	81	.000	0.18	9.00	80	.000	0.20	11.35	80	.000
Q75	0.03	10.19	81	.000	0.05	9.58	81	.000	0.08	11.00	81	.000	0.12	10.80	81	.000
Q90	0.03	8.38	81	.000	0.06	9.33	81	.000	0.10	10.54	81	.000	0.15	11.95	81	.000
Action	0.05	7.76	72	.000	0.14	2.37	73	.010	0.14	5.39	73	.000	0.18	7.30	73	.000
Minor	0.40	2.98	60	.002	0.35	3.37	60	.001	0.37	3.70	60	.000	0.51	4.35	62	.000
Mod.	0.44	2.93	41	.003	0.52	2.94	42	.003	0.81	3.97	45	.000	0.74	5.08	47	.000
Major	1.36	3.00	19	.004	1.84	4.27	22	.000	2.14	4.85	26	.000	1.80	6.01	34	.000

Table 4: Results of paired t-tests comparing the QR method`s performance with only forecast as predictor and the best-performing combination of five predictors for each river gage for the Brier score.

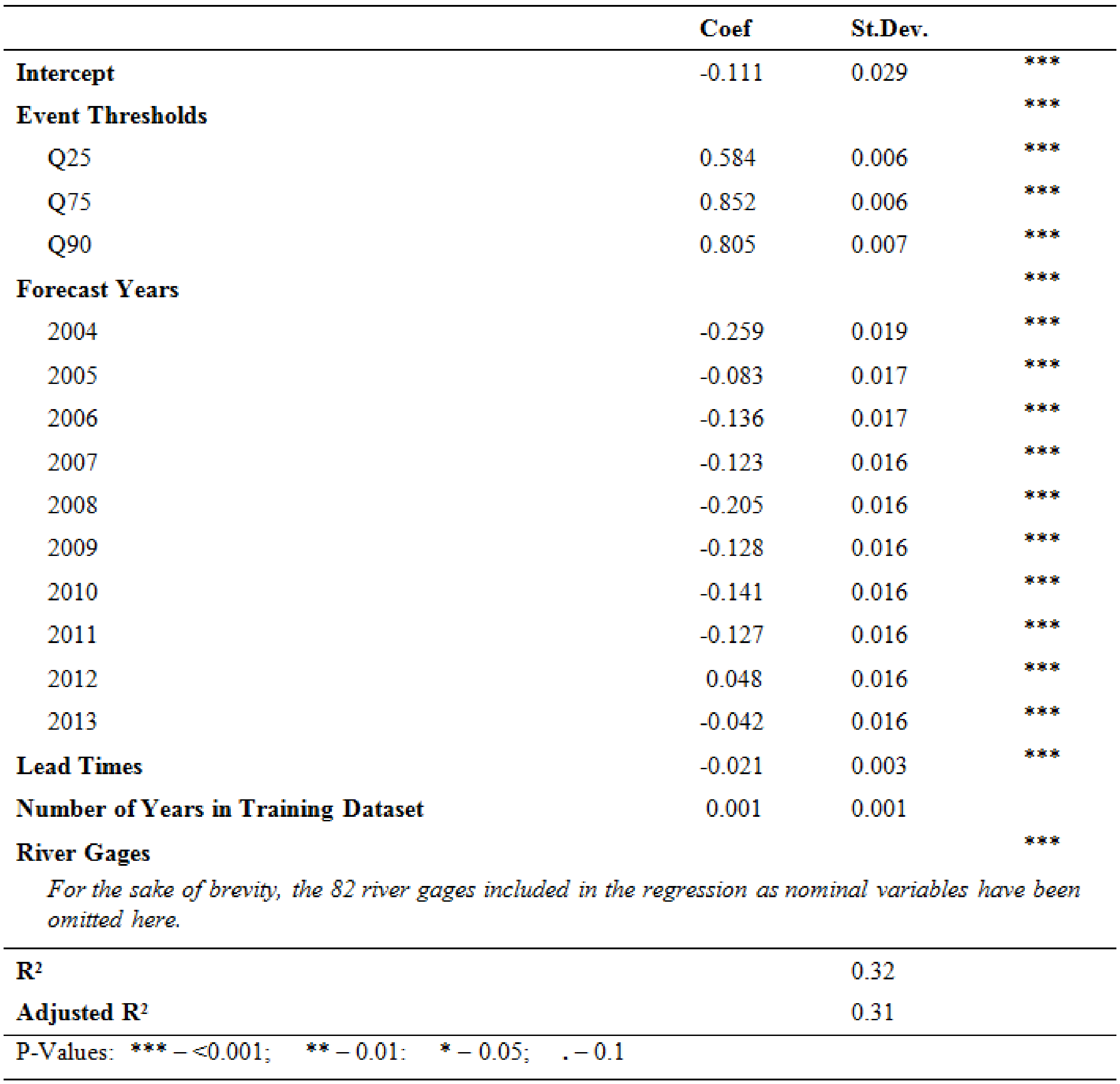
Event Thresh.	**Lead Time**	**Brier Score**	**Brier Skill Sc.**	**Reliabil.**	**Resol.**	**CRPS**
Q10	**1 Day**	-.012***	.20***	-.002***	.008***	-.026**
	2 Days	-.014***	.25***	-.002***	.010***	-.082**
	3 Days	-.016***	.28***	-.002***	.012***	-.121***
	4 Days	-0.17***	.27***	-.001*	.013***	-.054
Q25	**1 Day**	-.018***	.13***	-.003***	.013***	-.028**
	2 Days	-.023***	.16***	-.002***	.018***	-.088**
	3 Days	-.027***	.18***	-.003***	.021***	-.097**
	4 Days	-.031***	.20***	-.002***	.025***	-.475 **.**
Q75	**1 Day**	-.005***	.03***	.000	.011***	.342
	2 Days	-.011***	.05***	-.000 **.**	.015***	.009
	3 Days	-.016***	.08***	-.000	.021***	.188
	4 Days	-.025***	.12***	-.000	.028***	-.064
Q90	**1 Day**	-.003***	.03***	-.000**	.013***	.159
	2 Days	-.005***	.06***	-.000*	.015***	-.086**
	3 Days	-.010***	.10***	-.000	.019***	.163
	4 Days	-.015***	.15***	-.000*	.025***	-.075

P-Values: *** – <0.001; ** – 0.01; * – 0.05; **.** – 0.1

Table 5: Results of paired t-test comparing best combinations of predictors with one-size-fits-all approach.

	1 Day				**2 Days**				**3 Days**				**4 Days**			
	Diff.	**T-stat.**	**Df**	**p-val.**	**Diff.**	**T-stat.**	**Df**	**p-val.**	**Diff.**	**T-stat.**	**Df**	**p-val.**	**Diff.**	**T-stat.**	**Df**	**p-val.**
Q10	.054	4.61	79	.000	.071	5.56	79	.000	.075	6.36	79	.000	.071	7.54	79	.000
Q25	.010	5.73	80	.000	.016	4.17	80	.000	.016	5.11	80	.000	.019	3.76	80	.000
Q75	.003	6.56	81	.000	.004	7.25	81	.000	.005	4.63	81	.000	.004	6.42	81	.000
Q90	.008	7.10	81	.000	.015	4.37	81	.000	.012	5.16	81	.000	.021	1.84	81	.035
Action	.024	1.94	72	.028	.031	1.97	73	.026	.039	1.96	73	.027	.022	2.20	73	.016
Minor	.023	3.14	60	.001	.028	3.52	60	.000	.021	4.89	60	.000	.023	3.89	62	.000
Mod.	.039	4.79	41	.000	.052	6.18	42	.000	.063	4.98	45	.000	.060	4.40	47	.000
Major	.245	2.09	19	.025	.212	2.34	22	.014	.234	2.66	26	.007	.375	3.25	34	.001

Table 6: Regression results sensitivity analysis



	Coef	St.Dev.	
Intercept	-0.111	0.029	***
Event Thresholds			***
Q25	0.584	0.006	***
Q75	0.852	0.006	***
Q90	0.805	0.007	***
Forecast Years			***
2004	-0.259	0.019	***
2005	-0.083	0.017	***
2006	-0.136	0.017	***
2007	-0.123	0.016	***
2008	-0.205	0.016	***
2009	-0.128	0.016	***
2010	-0.141	0.016	***
2011	-0.127	0.016	***
2012	0.048	0.016	***
2013	-0.042	0.016	***
Lead Times	-0.021	0.003	***
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 nominal variables have been omitted here.			
R^2		0.32	
Adjusted R^2		0.31	

P-Values: *** – <0.001; ** – 0.01: * – 0.05; . – 0.1

Figures

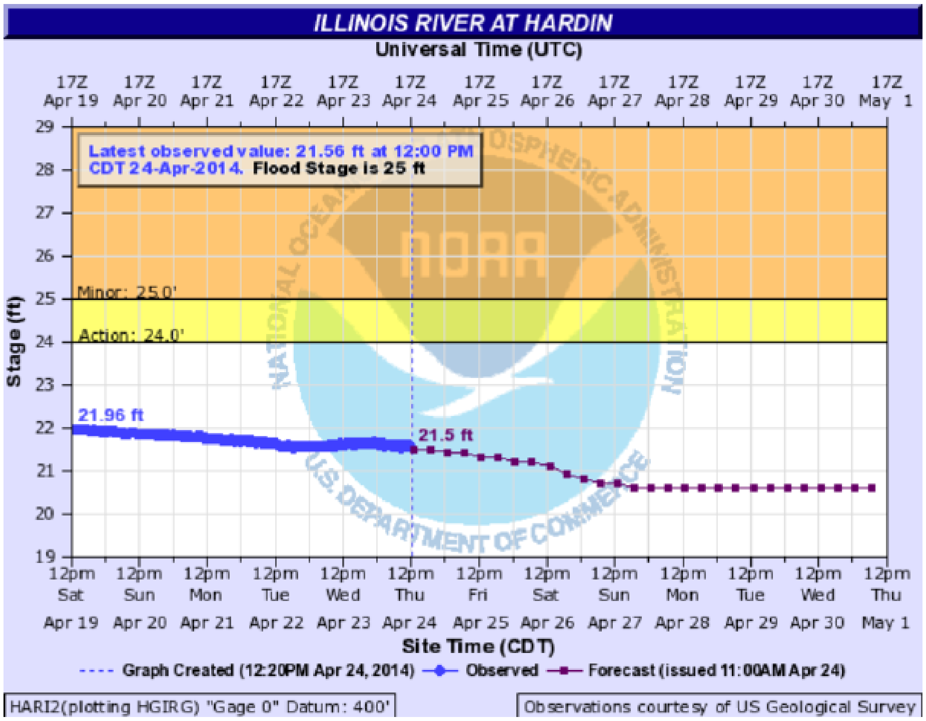


Figure 1: Deterministic short-term weather forecast in six hour intervals as published by the NWS for Hardin, IL on 24 April 2014.

Source:http://water.weather.gov/ahps2/hydrograph.php?wfo=lsx&gage=hari2.

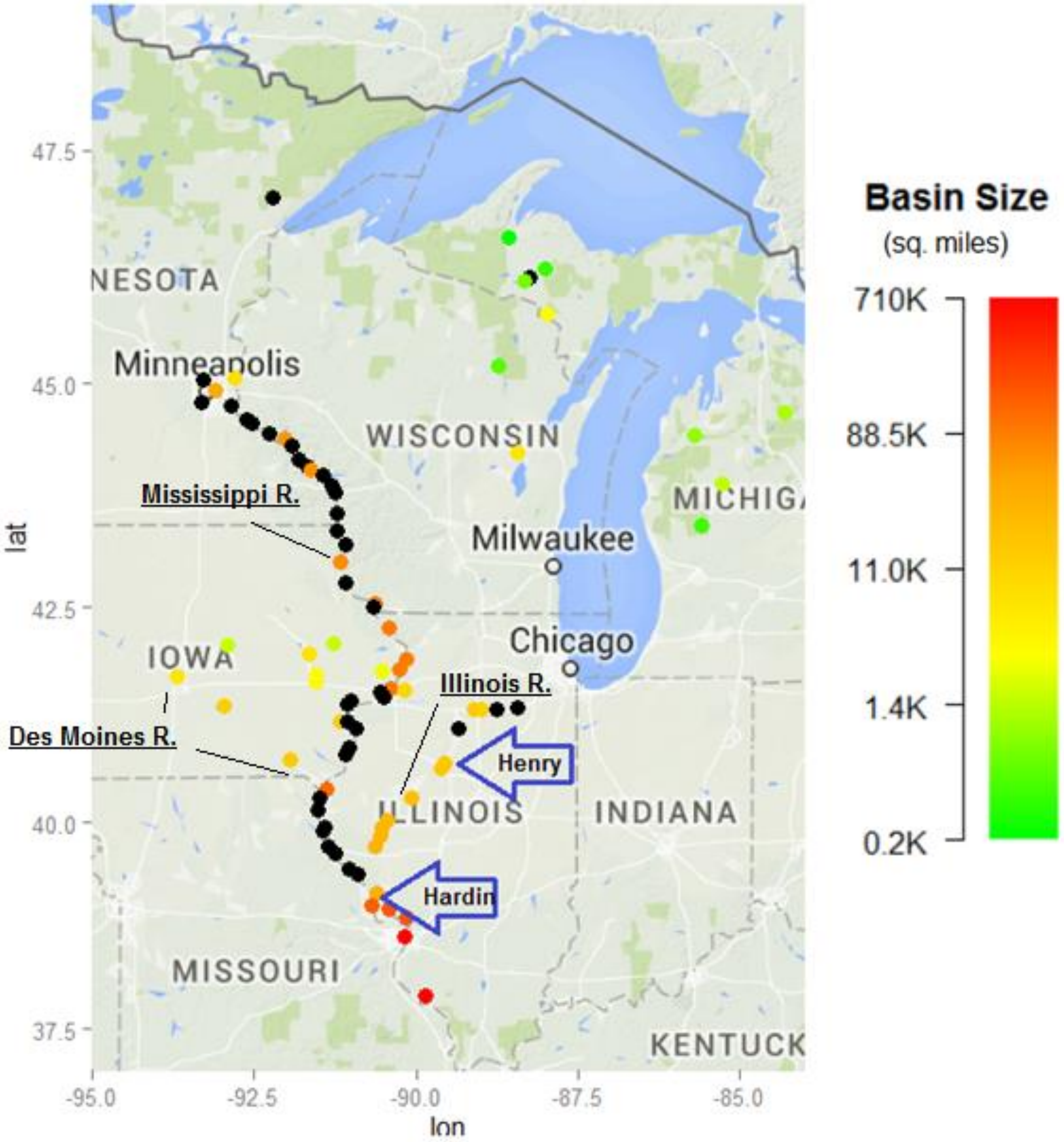


Figure 2: River gages for which the North Central River Forecast Centers publishes forecasts daily. Henry (HYNI2) and Hardin (HARI2) are indicated by the upper and lower red arrow respectively. For gages indicated by black dots the basin size is missing. The color scale for basin size in square miles is logarithmic.

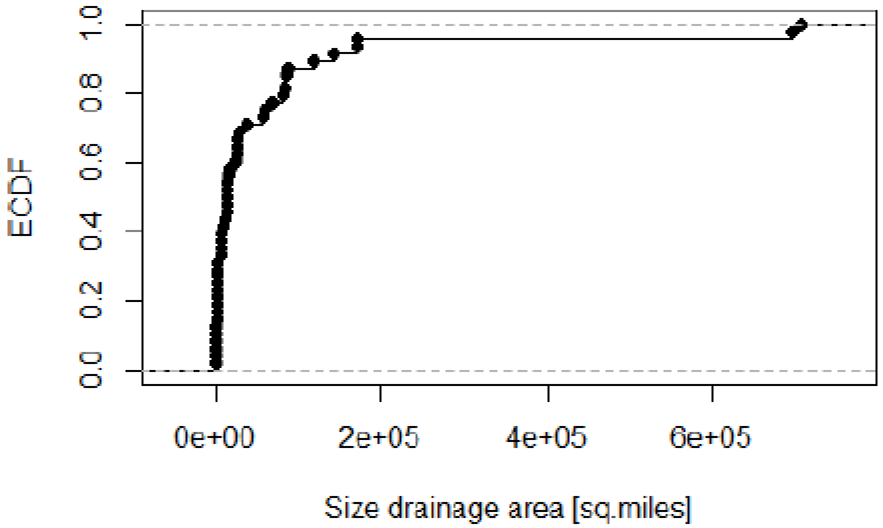


Figure 3: Empirical cumulative density function (ecdf) of sizes of drainage area for the river gages that are being forecasted daily by the NCRFC.

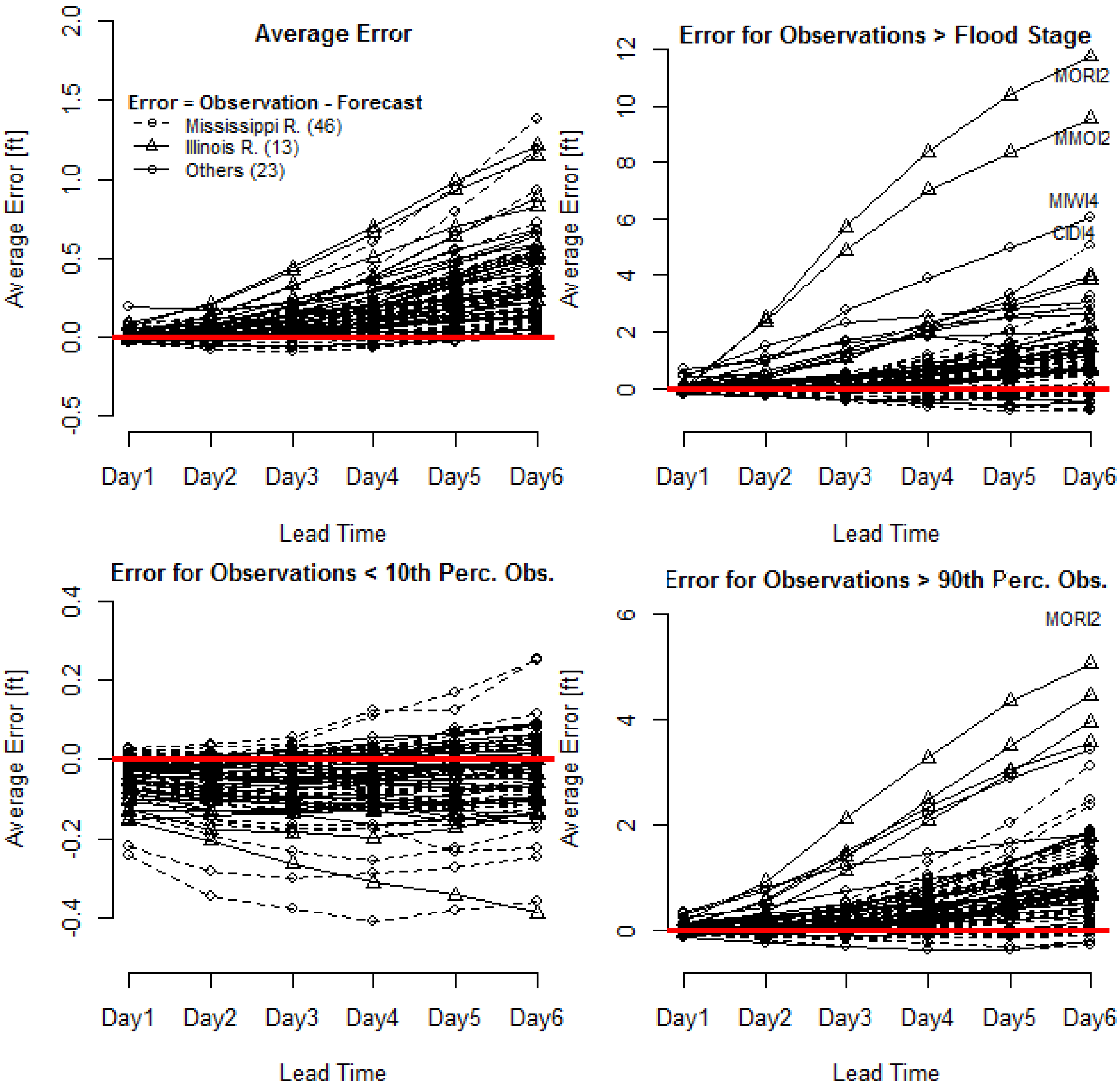


Figure 4: Forecast error for 82 river gages that the NCRFC publishes daily forecasts for. In anti-clockwise direction starting at the top left: (a) Average error; (b) error on days that the water level did not exceed the 10th percentile of observations; (c) error on days that the water level exceeded the 90th percentile of observations; (d) error on days that the water level exceeded minor flood stage

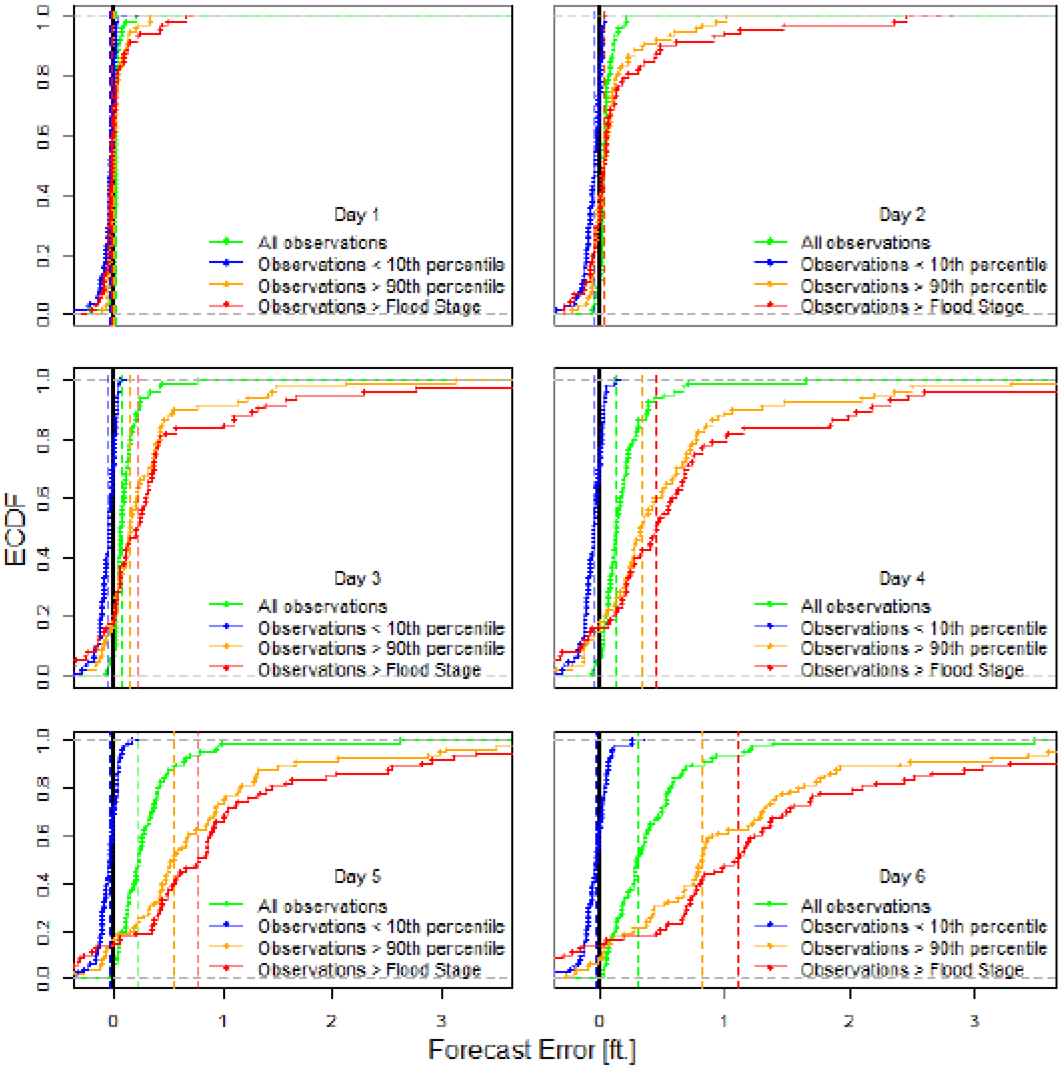


Figure 5: Empirical cumulative distribution function (ecdf) of forecast error at 82 river gages for six lead times. Vertical lines show the median forecast error of the corresponding subset.

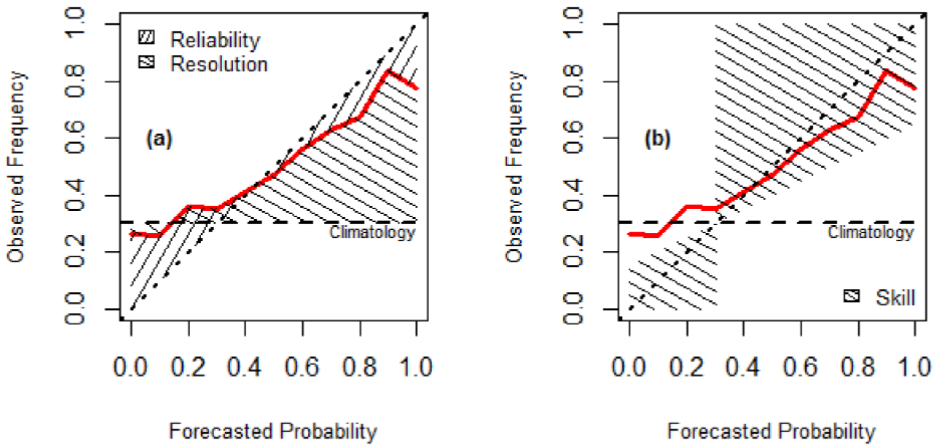


Figure 6: Theory behind Brier Skill Score illustrated for an imaginary forecast (red line): (a) reliability and resolution; (b) skill. In figure a, the area representing reliability should be as small, and for resolution as large as possible. The forecast has skill (BSS > 0), i.e., performs better than the reference forecast, if it is inside the shaded area in the figure b. Ideally, the forecast would follow the diagonal (BSS=1). (Adapted from Hsu and Murphy, 1986; Wilson, n.d.).

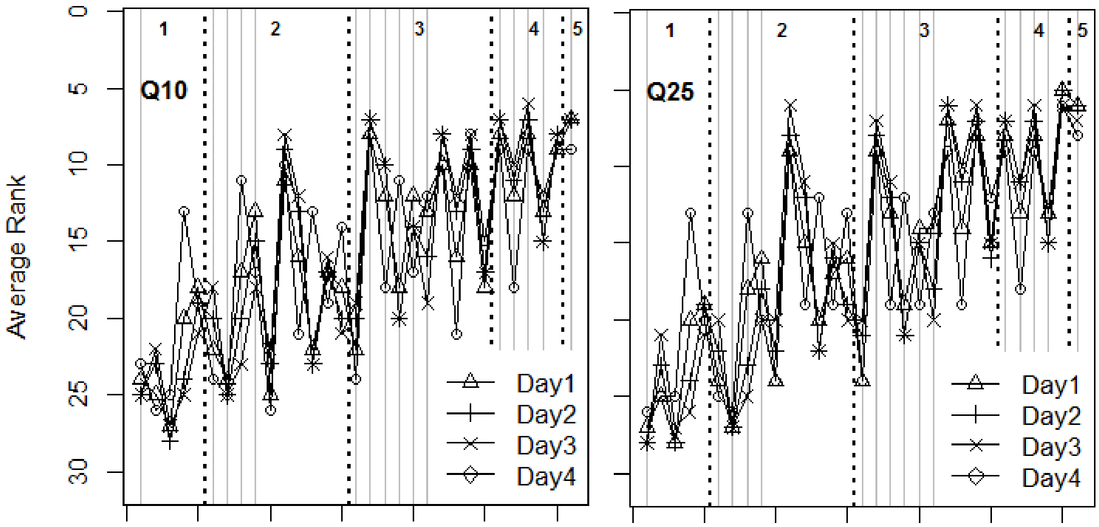


Figure 7: Average rank for each joint predictor for one to four days of lead time and two percentiles of observed water levels. Vertical gray lines correspond to the configurations that include forecast as one of the predictors. The y-axis is reversed, so that an increasing trend indicates increasing performance.

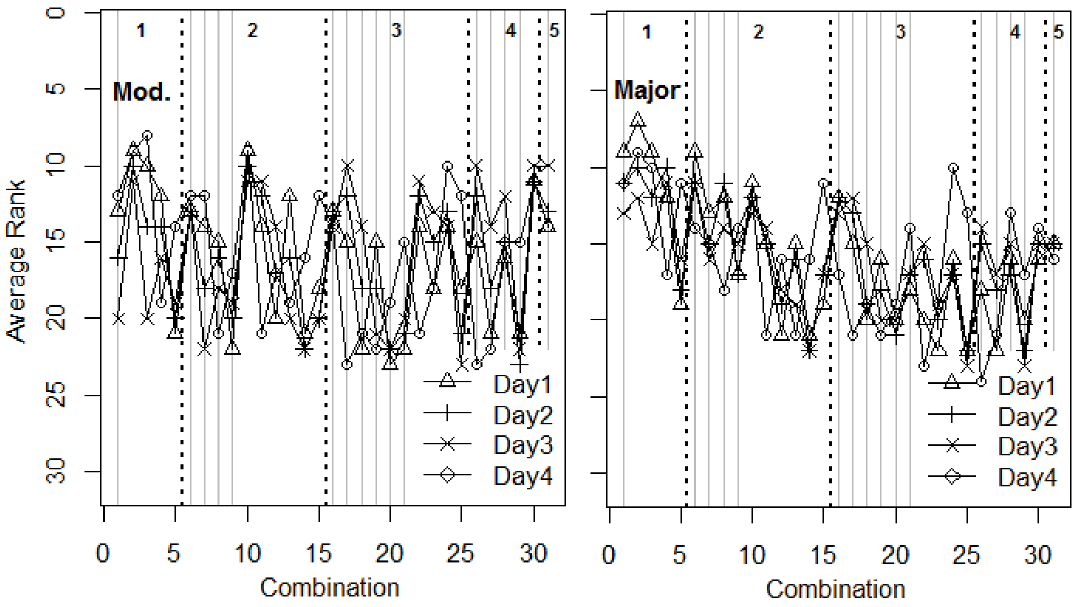


Figure 8: Average rank for each joint predictor for one to four days of lead time and the two highest flood stages. Vertical gray lines correspond to the configurations that include forecast as one of the predictors. The y-axis is reversed, so that an increasing trend indicates increasing performance.

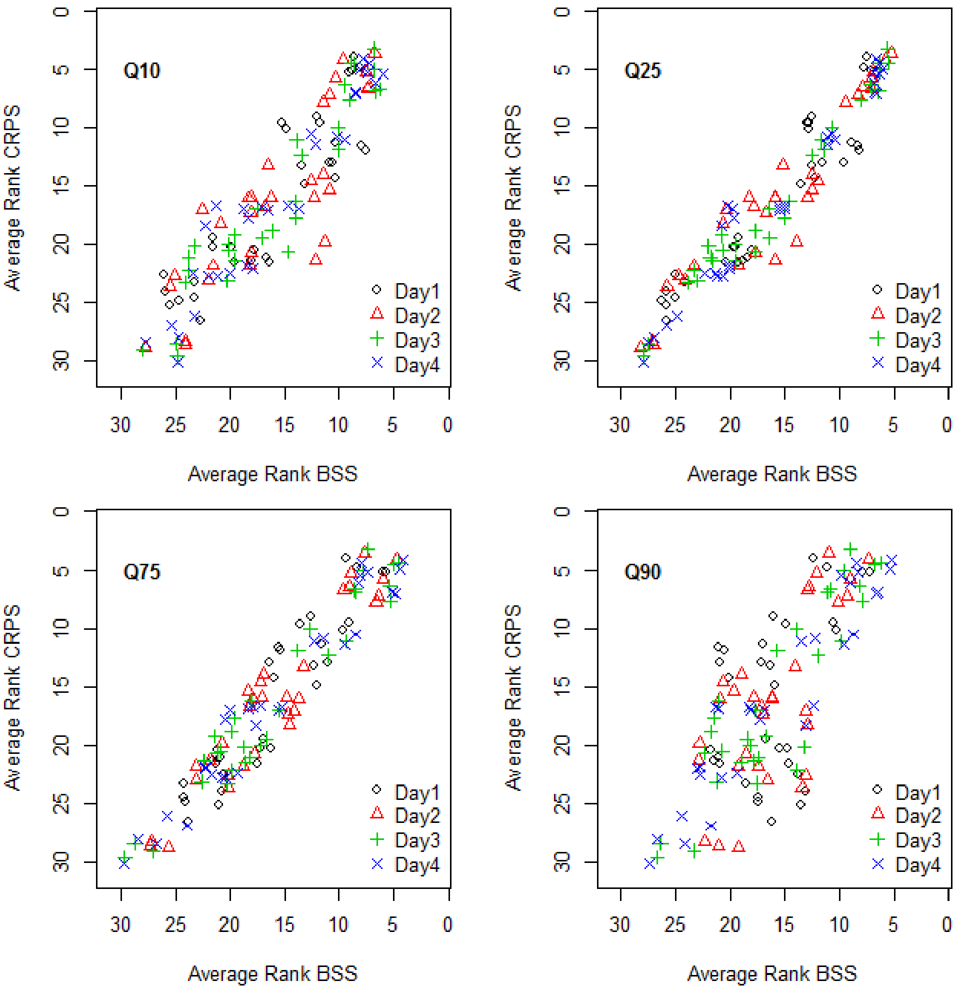


Figure 9: Comparing average rank across 82 gages based on Brier Skill Score and CRPS.

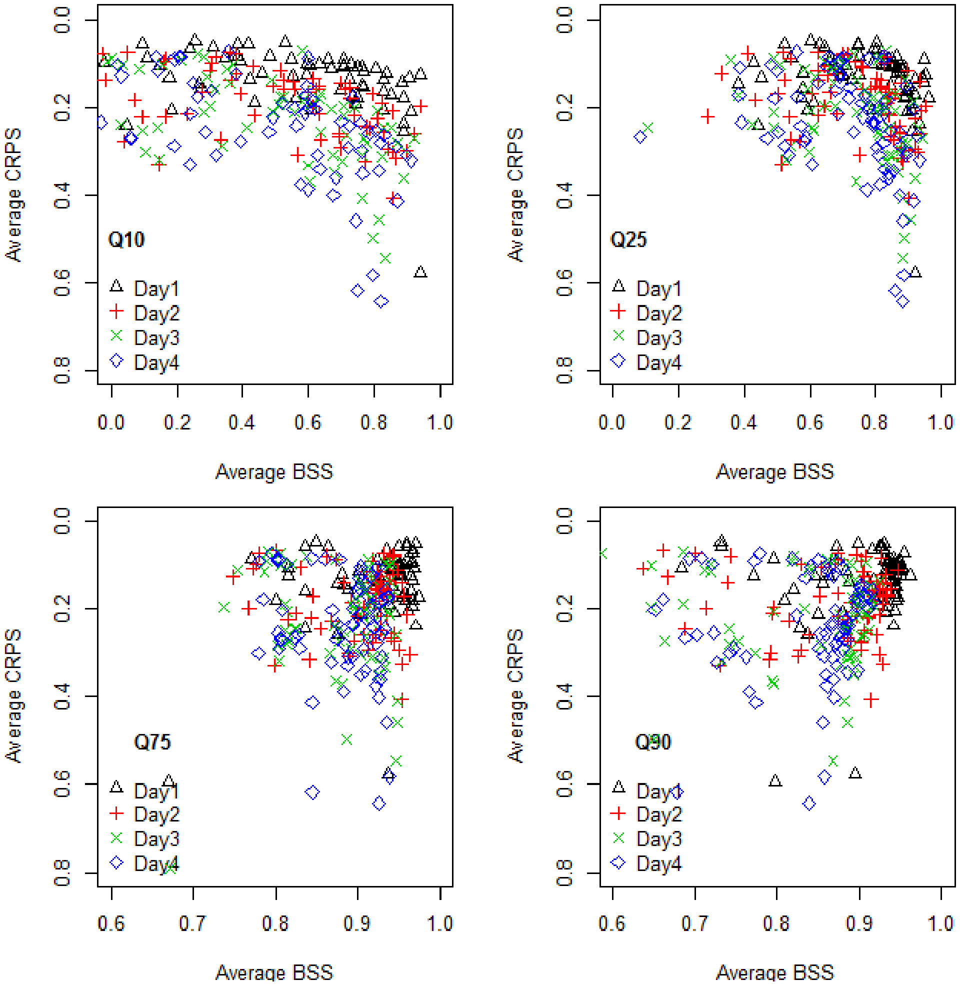


Figure 10: Comparing the performance of combination 30 [err24, err48, rr24, rr48] as measured Brier Skill Score and as measured by the Continuous Ranked Probability Score. Each data point corresponds with a gage at a certain lead time. Since the CRPS' perfect score equals zero, the y-axis has been reversed.

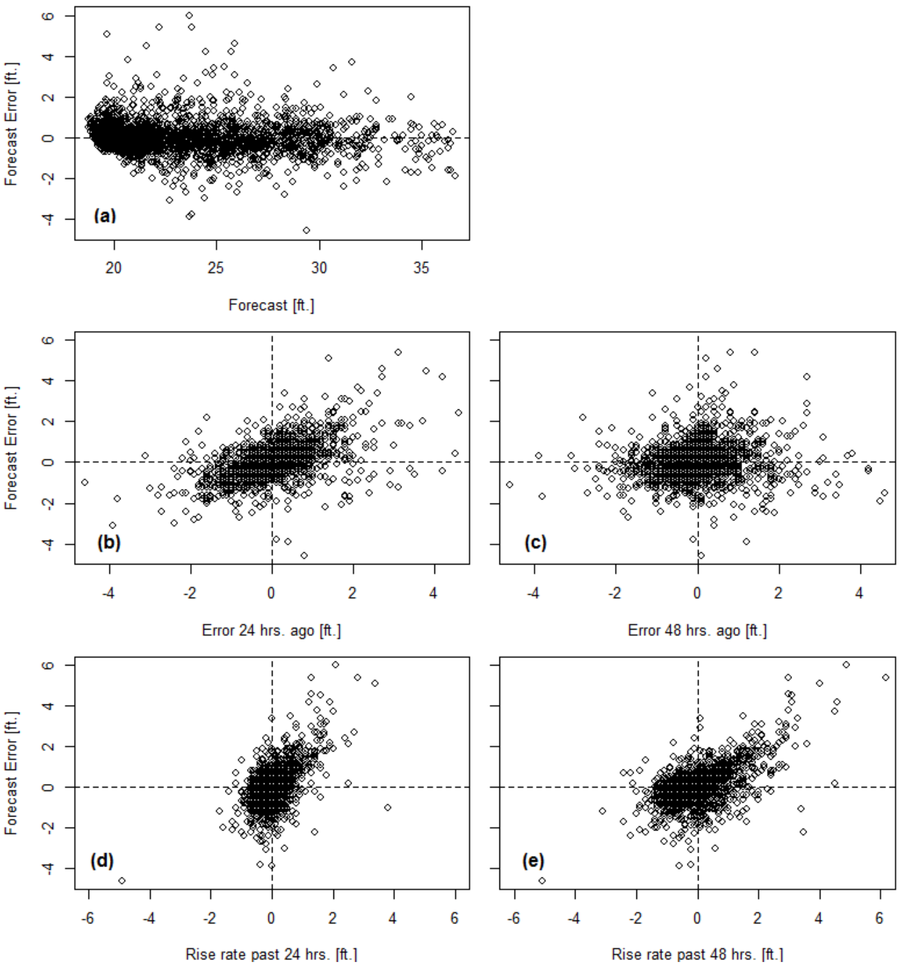


Figure 11: Independent variables plotted against the forecast error for Hardin IL with 3 days of lead time. First row: Forecast; second row: past forecast errors; third row: rates of rise.

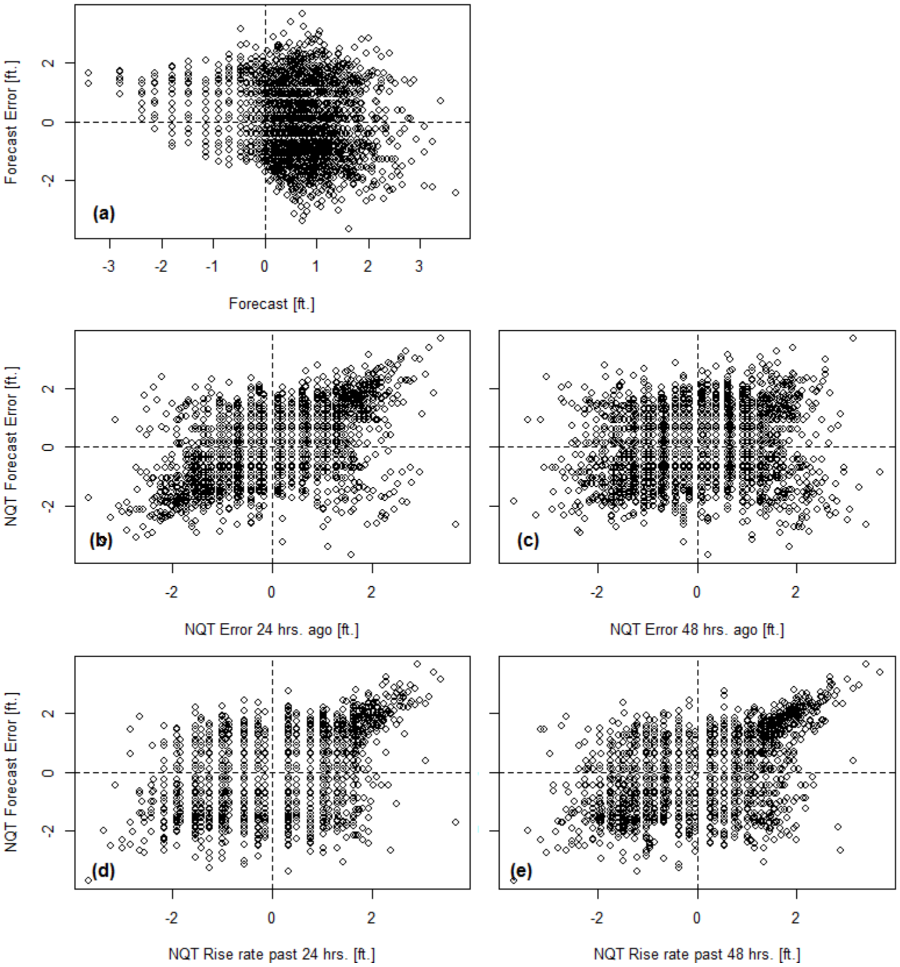


Figure 12: Independent variables after transforming into the Gaussian domain plotted against the forecast error for Hardin IL with 3 days of lead time. First row: Forecast; second row: past forecast errors; third row: rates of rise.

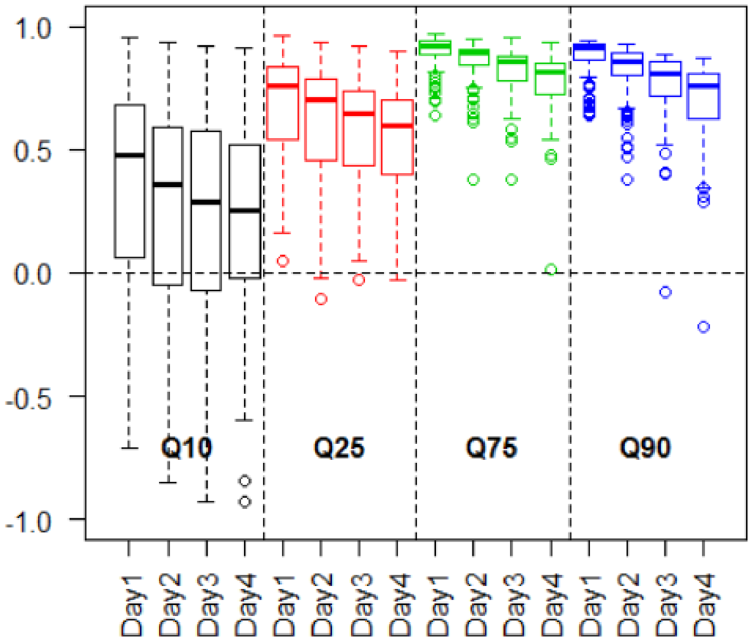


Figure 13: Brier Skill Scores (BSS) for forecast-only configuration for different lead times and event thresholds. The BSS' perfect score equals one. A BSS of zero indicates a forecast without skill.

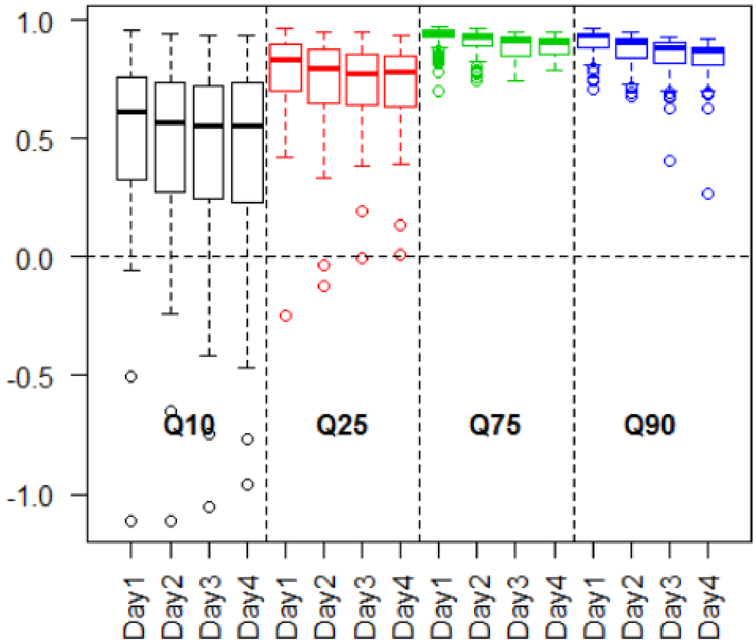


Figure 14: Brier Skill Scores (BSS) for best performing the joint predictor at each gage for different lead times and event thresholds. The BSS' perfect score equals one. A BSS of zero indicates a forecast without skill.

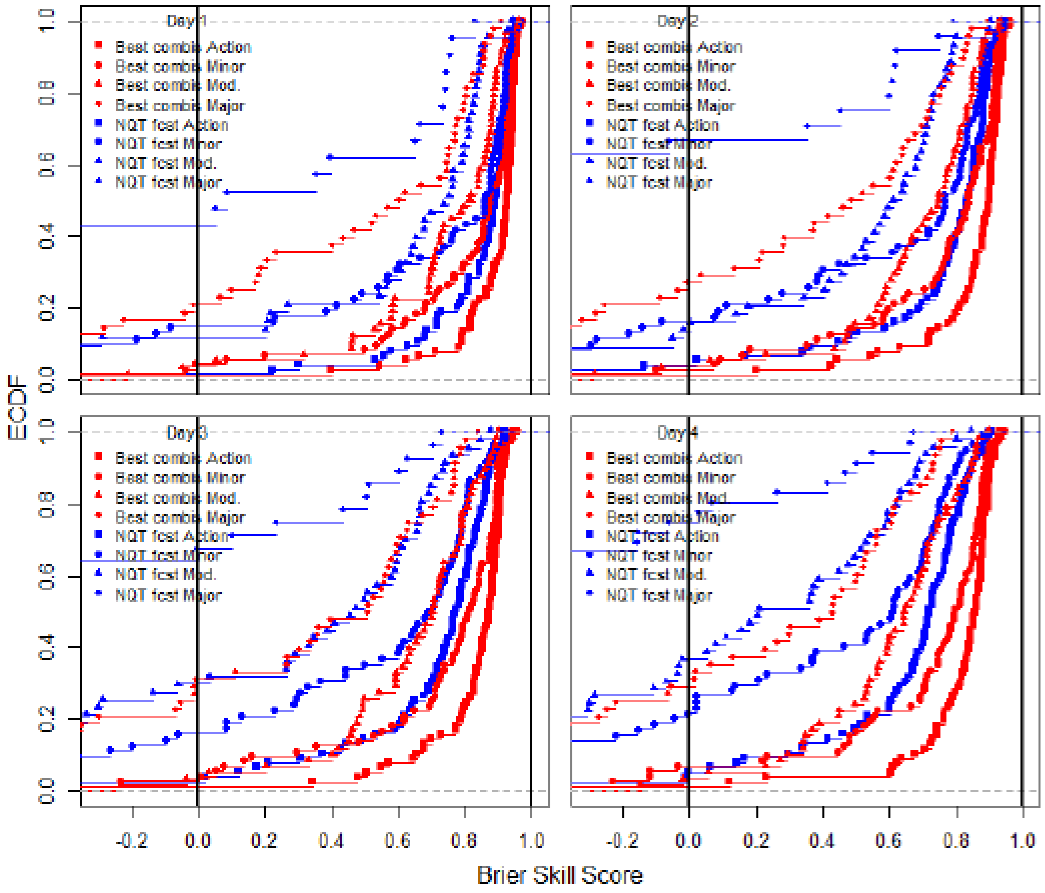


Figure 15: Empirical cumulative density functions of three QR configurations predicting exceedance probabilities of the Action, Minor, Moderate, and Major Flood Stage: the configuration using the transformed forecast as the only independent variable [NQT fcst]; the best performing combination for each river gage (upper performance limit) [Best combis]

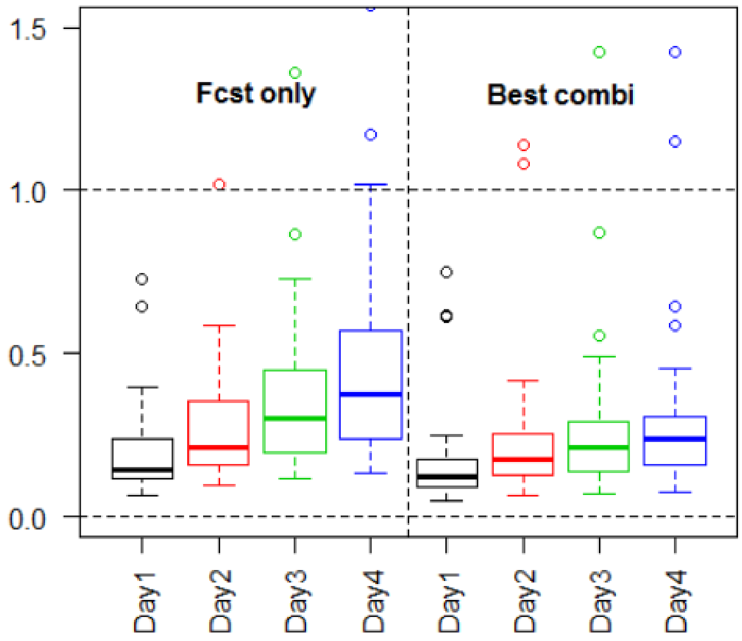


Figure 16: Continuous Ranked Probability Score (CRPS) for the forecast-only configuration and for the best performing the joint predictor at each gage for different lead times and event thresholds. The CRPS' perfect score equals zero.

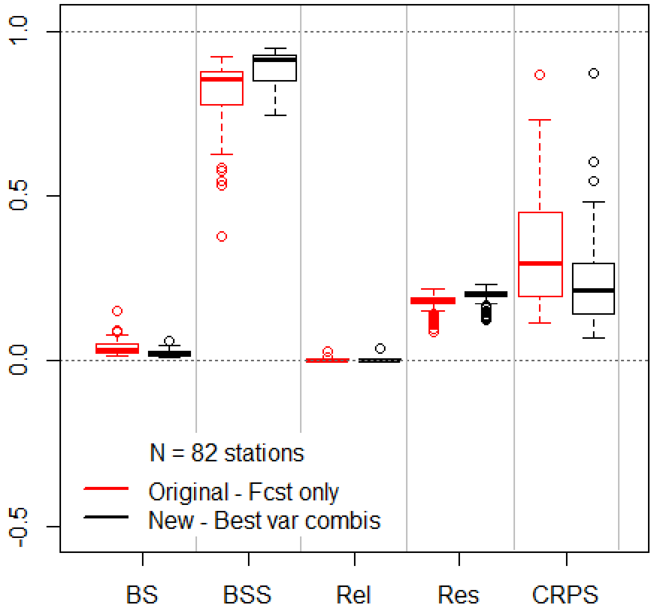


Figure 17: Comparison of the forecast-only QR configuration (i.e., only transformed forecast as independent variables) and using the best-performing joint predictor at each gage along various measures of forecast quality: Brier Score (BS), Brier Skill Score (BSS), Reliability (Rel), Resolution (Res), and continuous ranked probability score (CRPS). Lead time: 3 days; 75th percentile of observation levels as threshold.

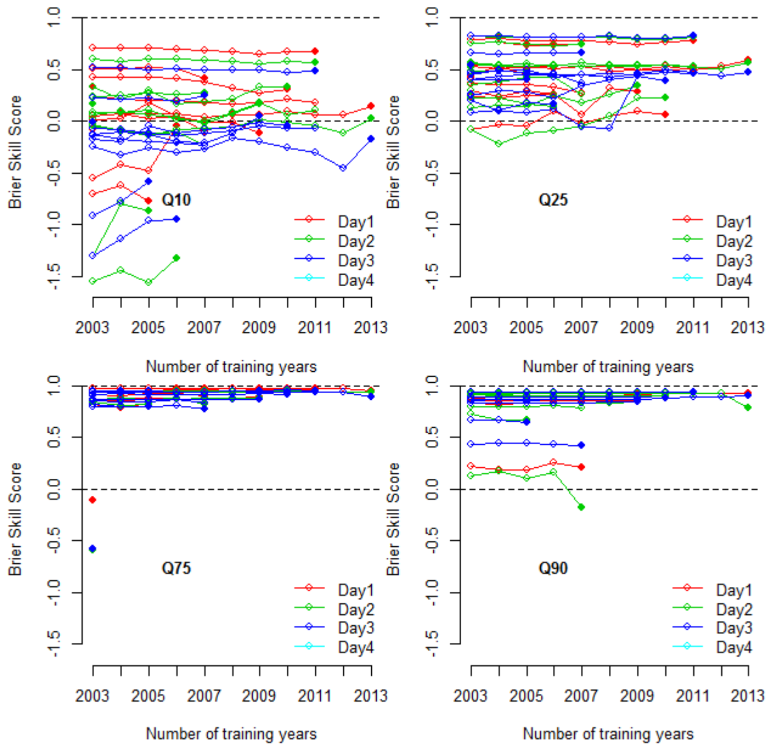


Figure 18: Brier Skill Score for various forecast years and various sizes of training dataset across different lead times (colors) and event thresholds (plots) for Hardin, IL (HARI2). The filled-in end point of each line indicates the BSS for the forecast year on the x-axis with one year in the training dataset. Each point further to the left stands for one additional training year for that same forecast year.

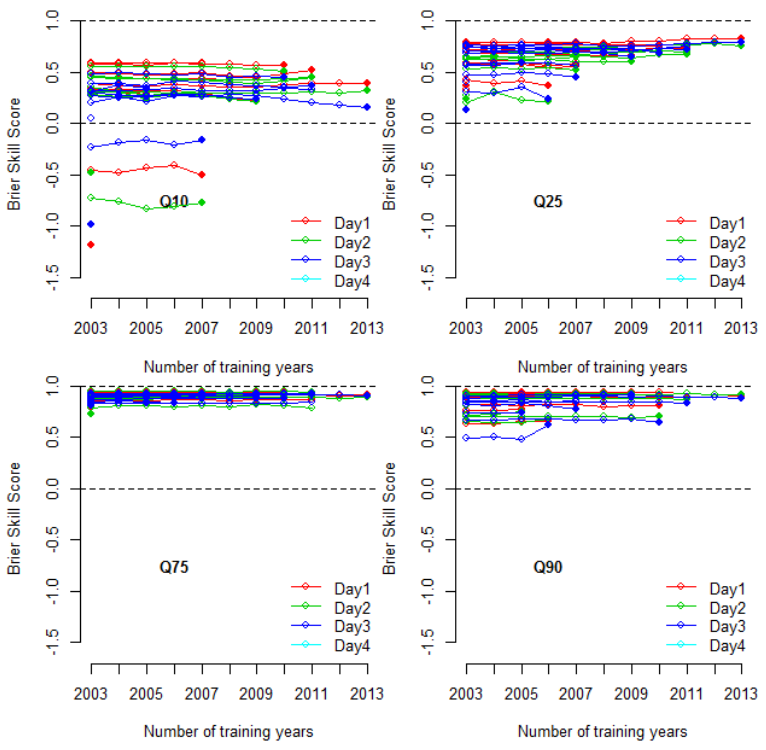


Figure 19: Brier Skill Score for various forecast years and various sizes of training dataset across different lead times (colors) and event thresholds (plots) for Henry, IL (HNYI2). The filled-in end point of each line indicates the BSS for the forecast year on the x-axis with one year in the training dataset. Each point further to the left stands for one additional training year for that same forecast year.

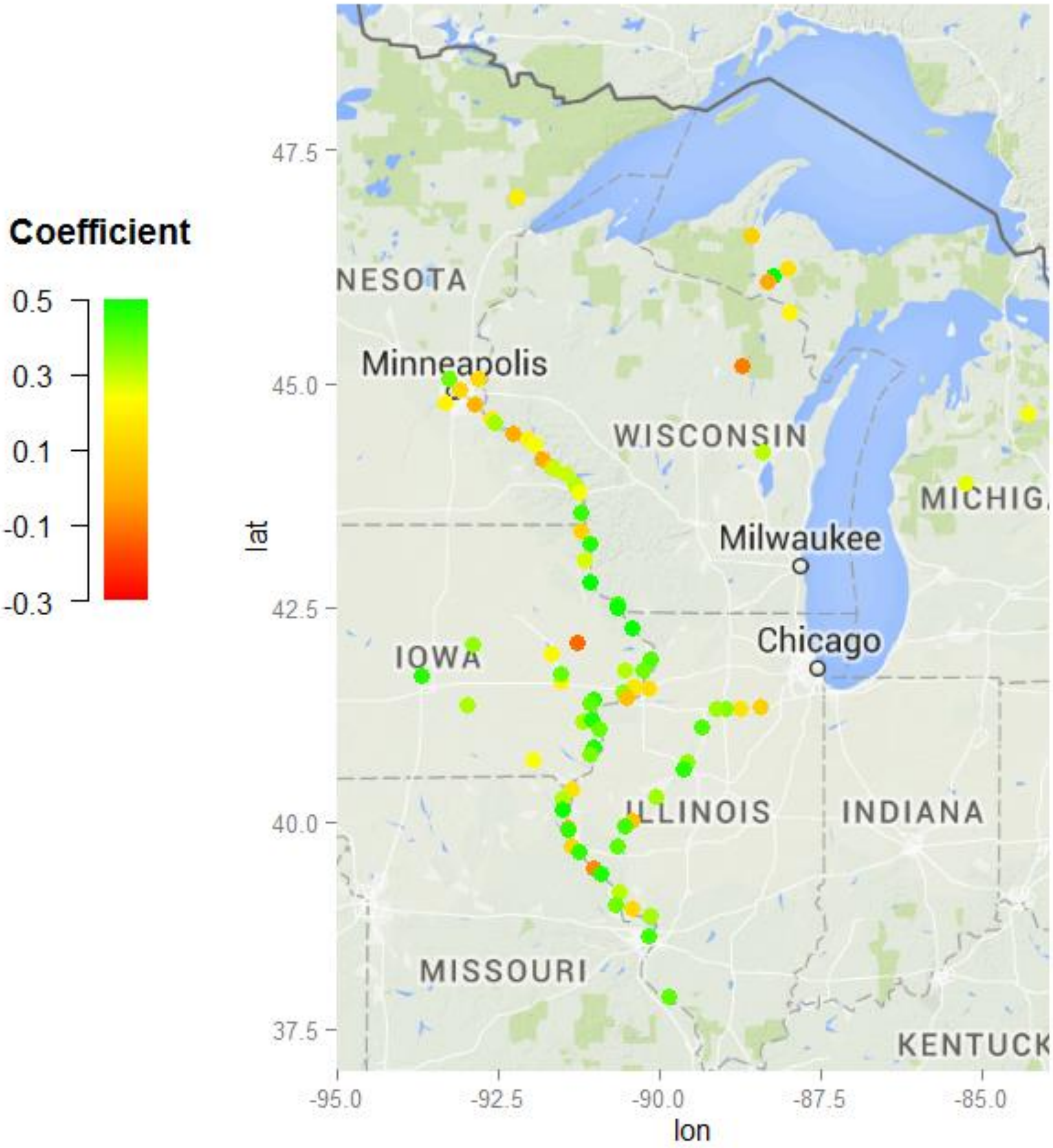


Figure 20: Geographical position of rivers. Colors indicate the regression coefficient of each station with the Brier Skill Score as dependent variable.