~~Article Submission~~Revised 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 ~~further develops the method of~~applies quantile regression (QR) to the prediction of~~predict~~ flood stage exceedance probabilities ~~of flood stages by post-processing forecasts~~based on post-processing single-value flood stage forecasts. A computationally cheap technique to predict forecast errors is valuable, because many national flood forecasting services, such as the National Weather Service (NWS), only publish deterministic single-value forecasts. ~~Using data~~The study uses data from the 82 river gages, for which the ~~National Weather Service's~~NWS' North Central River Forecast Center issues forecasts daily. ~~, this is the first QR application to U.S. American river gages.~~ Archived forecasts for lead times up to six days from 2001-20~~0~~13 were analyzed. ~~Earlier implementations of QR used the forecast itself as the only independent variable .~~ Besides the forecast itself, t~~T~~his study ~~adds~~ uses the ~~rise rate~~rate of rise of the river stage in the last 24 and 48 hours and the forecast error 24 and 48 hours ago ~~to~~ as predictors in ~~the~~ QR ~~model~~configurations. ~~Including those~~When compared to just using the forecast as independent variable, adding the latter four ~~variables~~ predictors significantly improved the forecasts, as measured by the Brier Skill Score (BSS). Mainly, the resolution increases, as the forecast-only ~~original~~ QR ~~implementation~~ configuration already delivered high reliability. Combining the forecast with the other four ~~variables~~ predictors results in much less favorable BSSs. 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 ~~model~~ configuration or at least calibration.

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

---

1 Introduction

River-stage forecasts ~~are inherently uncertain~~are no crystal ball; the future remains uncertain. The past has shown that unfortunate decisions have been made in ignorance of the potential forecast errors (Pielke, 1999; Morss, 2010)~~(e.g., Pielke, 1999; Morss, 2010)~~. For many users, such as emergency managers, forecasts are most important in ~~extreme~~ extreme situations, such as droughts and floods. Unfortunately, it is exactly in those situations that forecast errors are largest, due~~Due~~ to the ~~ir~~ infrequency of extreme events 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 river forecast would therefore be valuable, just as has been ~~weather forecasts has been strongly~~ recommended for weather forecasts in general (e.g., National Research Council, 2006)~~(e.g., National Research Council, 2006)~~.

There are two types of approaches to ~~quantify~~ estimate forecast uncertainty (e.g., Leahy, 2007; Demargne et al., 2013; Regonda et al., 2013)~~(e.g., Leahy, 2007; Demargne et al., 2013; Regonda et al., 2013)~~: Those addressing ~~certain~~ major sources of uncertainty individually 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

~~model~~configurations. On the downside, ~~it~~ the approach 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)~~(Regonda et al., 2013)~~.

The National Weather Service has chosen ~~for ensemble forecasting to quantify the uncertainty from major sources~~to quantify the most significant sources of uncertainty using ensemble techniques (Demargne et al., 2013)~~(Demargne et al., 2013)~~. ~~As of today~~Currently, the National Weather Service does not routinely publish uncertainty information along with their short-term river-stage forecast ~~(~~(Figure 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 hours and zero precipitation for the forecasts beyond 12 hours.[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).~~

~~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 and . 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~~

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

~~few days are much used to prepare for flood events, they have remained deterministic, as shown in .~~[2]

Figure 1~~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.

The ~~**Figure 12: 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/hydrograph.php?wfo=tsa&gage=como2**~~

~~**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/hydrograph.php?wfo=tsa&gage=como2**~~

NWS has developed the Hydrologic Ensemble Forecast Service (HEFS) ~~in~~ to be able to provide ~~also~~ short-term and medium-term probabilistic forecasts~~. Its implementation at all 13 river forecasts center is planned to be completed in 2014~~ (Demargne et al., 2013)~~(Demargne et al., 2013)~~. HEFS includes two types of post-processors. The Hydrologic Model Output Statistics (HMOS) Streamflow Ensemble Processor – which is also a module in NWS' main forecast tool, the Community Hydrologic Prediction System (CHPS) – corrects bias and evaluates the uncertainty of each ensemble, while Hydrologic Ensemble Post-Processing (EnsPost) corrects bias and lumps the set of ensembles into one uncertainty estimate (Demargne et al., 2013; Seo, 2008). HMOS performs a similar task as the QR approach presented here, but with two major differences. First, it relies on linear regression based on streamflows at various times as predictor, instead of using QR with several types of independent variables. Second, it does not

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

compute distributions of water levels from which confidence intervals or exceedance probabilities of flood stages can be derived, but generates ensembles (Regonda et al., 2013).

In contrast to ~~the~~ an ensemble approach ~~chosen by the NWS~~such as HEFS, the statistical 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 . To further strengthen the main advantage of this method, i.e., requiring relatively little resources,~~ To make this approach useful for actors with limited resources, we exclusively use publicly available data to ~~build our models~~define our configurations.

Most previously developed post-processors to generate probabilistic forecasts share the overall set-up but differ in their implementation. ~~Explanatory~~ 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 ~~metrics~~ 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)~~(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,~~ T~~t~~hese ~~method~~techniques differ ~~in their mathematical methods~~in a number of ways, including their sub-setting of data, and the output~~ metri~~.~~c.~~ Please see Regonda et al. (2013)~~(2013)~~ and Solomatine & Shrestha (2009)~~(2009)~~ for a summary of each ~~method~~technique. In a meta-analysis of four different post-processing ~~method~~techniques to generate confidence intervals, the quantile regression

~~method~~technique was one of the two most reliable ~~method~~techniques (Solomatine and Shrestha, 2009)~~(Solomatine and Shrestha, 2009)~~, while being the mathematically least complicated ~~method~~ and requiring few assumptions.

This paper further develops one of the ~~method~~techniques mentioned above: the Quantile Regression ~~method~~ approach to post-process river forecasts first introduced by Wood et al. (2009) and further elaborated by Weerts et al. (2011)~~(2011)~~ and López López et al. (2014). ~~.~~The Weerts~~at~~ 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. This paper combines elements of the studies mentioned above. ~~–~~In some aspects, our approach differs from ~~the original approach by Weerts et al. and López López et al. .~~those three studies. We predict the ~~probabilities that flood stages are exceeded~~exceedance probabilities of flood stages 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. found in a survey of the general U.S. public that most people are able to base decisions on those forecasts.~~ 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~~ available~~. The study does not add to the mathematical technique of quantile regression itself.

In this paper, the QR ~~method~~technique 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]~~

Identifying the best-performing set of independent variables is central to this paper. ~~To our knowledge, this paper is the first application of the QR method to the U.S. American context.~~ All possible combinations of the following predictors have been studied: forecast, the

~~The method is further developed by demonstrating the benefit – measured by an increase in Brier Skill Score (BSS) – of including the rise rate~~rate of rises of water levels in past hours, and the past forecast errors ~~as independent variables into the quantile regression~~. The performance of these joint predictors has been measured and compared using the Brier Skill Score (BSS).~~.~~ ~~For extremely high water levels the variable combination has to be 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.~~ This exercise has been repeated for various water levels and lead times. 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 Method section ~~summarizes the additions that this paper makes to the quantile regression method introduced by Weerts et al. . It~~ reviews ~~the~~

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

~~method~~quantile regression, ~~explains the additions,~~ introduces the performance ~~metric~~measure, and discusses the ~~computations~~ performed analyses 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 .~~ It then describes the results of identifying the best-performing set of independent variables. Finally, it discusses the robustness of the ~~proposed method~~studied QR configurations. The fourth and last section presents the conclusions and proposes further research ideas.

2 Method

The use of quantile regression to ~~quantify~~ estimate the error distribution of river-stage forecasts has first been ~~presented~~ introduced by Woods et al. (2009) for the Lewis River in Washington State. Later, ~~by~~ Weerts et al. (2011)~~(2011)~~ applied it to ~~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;~~elements of both studies are combined. However, our predictand is the probability of exceeding flood stages rather than confidence bounds. Additionally, this study tests ~~b) by testing~~ the robustness of the ~~method~~technique across locations, lead times, event thresholds, forecast years, and the size of training dataset is tested. ~~; c) by estimating the more decision-relevant probability of exceeding flood stages rather than confidence bounds.~~ To develop the different QR configurations ~~of quantile regression~~ and to compare their performance, the Brier Skill Score (BSS) is used.

In the following, ~~the~~ quantile regression itself and ~~,~~ the ~~proposed addition to the method~~analysis to identify the best-performing set of independent variables ~~, 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)~~(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)~~(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)~~(Koenker and Machado, 1999, p.1305)~~ and Alexander et al. (2011)~~(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)~~(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~~ When applying QR to river forecasts, Weerts et al. (2011)~~(2011)~~ transformed~~,~~ the forecast values and the corresponding forecast errors ~~are transformed~~ into the Gaussian domain using Normal Quantile Transformation (NQT) to account for heteroscedasticity. Detailed instructions to perform NQT can be found in~~, as instructed by~~ Bogner et al. (2012)~~(2012)~~. ~~to account for heteroscedasticity.~~ Building on this study, López

López et al. (2014)~~(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~~ measures (López López et al., 2014)~~(López López et al., 2014)~~. To be able to combine predictors ~~variables~~ of different nature, we ~~build a model~~ based our QR configuration on untransformed ~~variables~~predictors. The reason to do so will be discussed and illustrated later (see Figure 11 and Figure 12).

~~Using the transformed data,~~ A~~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)~~(2011)~~ give a detailed mathematical description for applying QR to river forecasts. Mathematically, the approach is formulated as follows (with and without NQT):

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

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

Field Cod

Equation 2: QR ~~implementation~~ 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

[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.~~

$fcst(t)$	– 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	– ~~model~~ configuration coefficients

The second part of the equations stands for the error estimate based on the quantile regression ~~model~~ configuration for each percentile τ and lead time. In Equation 1, that was used ~~in the original QR method proposed~~ by Weerts et al. (2011)~~(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).[5]

2.2 Brier Skill Score

The ~~original~~ QR ~~implementation~~ configuration by Weerts et al. (2011)~~(2011)~~ was evaluated by determining the fraction of observations that fell into the confidence intervals predicted by the QR ~~model~~configuration; i.e., ideally, ~~90~~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)~~(2014)~~ used a number of ~~metrics~~ measures to assess ~~model~~ configuration 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 – first introduced by Brier (1950) – to ~~compare~~ assess ~~the different versions of the~~ QR ~~model~~configurations ~~proposed in this paper. We chose to optimize~~

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

~~our QR models based on the BSS, first introduced by Brier~~ for ~~two~~ two reasons. First, to be able to optimize model performance it is best to choose a single measure. ~~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~~Second, 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, which is a hydrological characteristic inherent to the river gage. This uncertainty is different than the forecast uncertainty that the technique studied in this paper estimates. Besides the uncertainty that can be mathematically explained, it also includes natural variability. ~~Thus~~In sum, the BS' uncertainty term~~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; Wikipedia, 2014; WWRP/WGNE, 2009)~~(e.g., Jolliffe and Stephenson, 2012; Anon, 2014; WWRP/WGNE, 2009)~~.[6]

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

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

with BS – Brier Score

~~[6] Bröcker (2012)(2012) showed that the conventional decomposition of the Brier Score is biased for finite sample sizes. It systematically overestimates reliability, under- or overestimates resolution, and underestimates uncertainty. Several authors proposed less biased decompositions (e.g., Bröcker, 2012; Ferro and Fricker, 2012)(e.g., Bröcker, 2012; Ferro and Fricker, 2012). Additionally, Stephenson et al. (2008)(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)(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)(2014). After all, the Score is mainly used to compare model configurations, rather than establishing the absolute performance of each modelconfiguration.~~

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. ~~A forecast with~~ The reliability curve for the forecast representing perfect reliability would follow the diagonal in Figure 2,~~,~~ i.e., the area in Figure 2a representing reliability would equal zero (Jolliffe and Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009)~~(e.g., Jolliffe and Stephenson, 2012; Anon, 2014; WWRP/WGNE, 2009)~~. The configuration by López López et al. (2014)~~(2014)~~ performs well in terms of reliability. When estimating confidence intervals, Weerts et al. (2011)~~(2011)~~ achieved good results especially for the more extreme percentiles (i.e., 10th and 90th).

Figure 2: 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 4: 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 random guessing, 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 ~~pertains to how much better the forecast performs than taking the historical frequency (climatology) as a forecast.~~measures the difference between the predicted probability of an event on a given day and the observed average probability. When calculated for a time period longer than a day, the forecast performs better if the resolution term is higher. -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. The accumulated difference between the predicted frequency and the historical average across a time period of several days would then be zero (e.g., Jolliffe and Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009)~~(e.g., Jolliffe and Stephenson, 2012; Anon, 2014; WWRP/WGNE, 2009)~~. In Figure 2, the curve for a ~~a~~ forecast with good resolution would be steeper than the dashed line that represents climatology, i.e., the area in ~~a~~Figure 2a representing resolution would be maximized. In absolute terms, the resolution can never exceed the third term in Equation 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 detail, see Ikeda et al., 2002)~~(for more detail, see Ikeda et al., 2002)~~. The latter likewise quantifies how much better ~~a forecast is~~ than ~~random guessing~~the reference forecast (i.e., climatology) a forecast is -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; Wikipedia, 2014; WWRP/WGNE, 2009)~~(e.g., Jolliffe and Stephenson, 2012; Anon, 2014; WWRP/WGNE, 2009)~~.

A forecast possesses skill, i.e., performs better than ~~random guessing or climatology~~the reference forecast (in this case climatology), if it is inside the shaded area in ~~b~~Figure 2b. 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. Equation 4 defines the BSS' decomposition into the resolution and reliability components described above (Brown and Seo, 2013). ~~[7]~~ 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)~~(e.g., Jolliffe and Stephenson, 2012; Anon, 2014; WWRP/WGNE, 2009)~~. All measures of forecast quality were computed using the R-package "verification" (NCAR, 2014).

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
$\bar{o}$ – Frequency of binary event occurring
$\bar{o}(1-\bar{o})$ – Climatological variance

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

2.3 ~~Proposed addition: More than one independent variable~~Identifying the best-performing sets of independent variables

~~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 challenge is to identify a well-performing 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, t~~T~~he most obvious ~~variables~~ predictors to include ~~besides the forecast itself~~ are the observed water level 24 and 48 hours ago, ~~the observed rise in water level in the last 24 and 48 hours (called rise rate hereafter),~~ the forecast error 24 and 48 hours ago (i.e., the difference between the current water level at issue time of the forecast and the forecast that was produced 24/48 hours ago), or the time of the year, e.g., using month or season as categorical predictors. ~~Other~~ Additional potential ~~variables~~ independent variables are the water levels observed up- and downstream at various times, the precipitation upstream of the catchment area, and the precipitation forecast. However, r~~R~~equesting the corresponding precipitation and precipitation forecast requires an extensive effort or direct access to the ~~database.~~~~these latter variables are much more difficult to gather because of the way data is archived~~database at the National Climatic Data Center (NCDC).[8]

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

~~**Table : Variable Combinations**~~

In preliminary trials on two case studies (gages HARI2 and HYNI2), it was found that the rates of rise 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 rates of rise and forecast errors, so that these latter variables include the information of the former variable. It was also found that season and months are not significant in quantile regression configurations to predict the quantiles of the forecast error. Probably, the time of the year is already reflected in the observed water levels and forecast errors in the previous days. ~~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 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 on average and most often leads to the best out-of-sample results 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.

~~**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 hours (rr24, rr48), and the forecast error 24 and 48 hours ago (err24, err48) were tested for 82 gages that the NCRFC issues forecasts for every morning (). Based on the Bier Skill Score, a metric of forecast quality explained below, it was determined which variable combination on average and most often leads to the best out-of-sample results for various lead times and water levels.**~~ **Table 1: Joint predictors.**

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 at location X is 60%." This is done in two steps. First, a training dataset (first half of the data) is used to ~~build~~ define one quantile regression ~~model~~configuration for ~~each~~ each ~~of the following~~ percentile~~s:~~ π = [0.05, 0.1, 0.15, … , 0.85, 0.90, 0.95] and each lead time.~~-~~ The dependent variable is the water level. As described ~~above~~in Equation 5, the forecast itself, the ~~rise rates~~rates of rise and forecast errors serve as independent variables.

In the second step, these QR ~~model~~configurations 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 predicted ~~QR model~~percentile π 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~~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.~~[9]~~

To be able to compare various ~~model~~ configurations, the Brier Skill Score is determined ~~across all the days in~~based on forecast exceedance probability for all 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.

To study whether the various combinations of ~~variables~~ predictors perform equally well for high and low thresholds, these last computational steps (i.e., interpolating to determine the

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

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 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 ~~optimal~~best-performing set of independent variables, the entire procedure is repeated for each of the 31 ~~variable combination~~joint predictors 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~~ eight 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 ~4,000 river gages every day. Such~~'s daily ~~published~~ short-term river forecasts predict the stage height in six-hour intervals for up to five days ahead (20 6-hour 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)~~(2007)~~ provides a detailed description of the forecasting process.

~~[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.~~

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).~~[11]~~ In total, the NCRFC produces forecasts for 525 gages. ~~().~~ 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 (Figure 3). 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 4 shows an empirical cumulative density function of drainage areas sizes.

Figure 3: 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.

Figure 4: 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. Hardin, IL is just upstream of the confluence of the Illinois River and the Mississippi River (Figure 3). 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

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

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 (Illinois Department of Natural Resources, 2011),~~[12]~~ draining an area of ~13,500 square miles ~~(~35,000 km^2)~~ at Henry (USGS, 2015a)~~[13]~~ and ~28,700 square miles ~~(~72,000 km^2)~~ at Hardin (USGS, 2015b).~~[14]~~

~~**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/**~~

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. 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 (~~a, a~~Figure 5a, Figure 6). Extremely low water levels, defined as below the 10th percentile of observed water levels, are also well calibrated (Figure 5b, Figure 6). ~~(b, b).~~ 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 (Figure 5c, Figure 6). ~~(c, c).~~ When only looking at

~~[12] Illinois Environmental Protection Agency: "Illinois River and Lakes Fact Sheets", URL [accessed 04/24/2014]: http://dnr.state.il.us/education/aquatic/aquaticillinoisrivlakefactshts.pdf~~
~~[13] Source: http://waterdata.usgs.gov/nwis/nwisman/?site_no=05558300&agency_cd=USGS~~
~~[14] Source: http://waterdata.usgs.gov/nwis/nwisman/?site_no=05587060&agency_cd=USGS~~

~~[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 calculations on.~~

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 (Figure 5d, Figure 6). ~~(Figure 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 12 feet for water levels higher than minor flood stage. 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 calculations on.

Figure ~~6~~5: 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 6: 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.

~~**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.**~~

3.2 ~~Including more variables~~ Identifying the best-performing sets of independent variables

In total, the Brier Skill Score (BSS) for 31 ~~variable combination~~joint predictors (Table 1) across various lead times and event threshold have been compared. Across 82 river gages, it has been

~~[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.~~

analyzed (a) which combinations perform best and worst most often, and (b) which ~~sets of variables~~joint predictor delivers the best BSSs on average.

3.2.1 Frequency Analysis

For ~~each~~ the four lead time (i.e., one to four days) and ~~various~~ the eight event thresholds (i.e., 10th, 25th, 75th, 90th percentiles as well as the four flood stages), we counted ~~how often~~at how many river gages each ~~variable combination~~joint predictor resulted in the highest and the lowest BSS ~~across the 82 river gages~~. Figure 7 shows that for water levels below the 50th percentile ~~variable combination~~joint predictors with four or more independent variables return the best BSSs most often, while those with one and two ~~variables~~ predictors perform worst most often. For thresholds higher than the 50th percentile the distributions gradually become ~~more flat~~flatter. For the 90th percentile, a clear trend is no longer detectable. Given that the frequency distributions for the extreme events in Figure 7 are relatively uniform, it seems as if extreme events are characterized by different processes at different gages. The same set of histograms for the four flood stages (i.e., action, minor, moderate, and major) confirms this (Figure 8). Across lead times, there is a slight trend noticeable that single ~~variables~~ predictors tend to be the worst combination more often for longer lead times. This suggests that~~us,~~ the further out one is forecasting, the more important it becomes to include more data in the ~~model~~configuration.

Figure 7: Histograms of ~~variable combination~~joint predictors 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 combination~~joint predictors with different numbers of independent variables.

Figure 8: Histograms of ~~variable combination~~joint predictors 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 combination~~joint predictors with different numbers of independent variables.

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 independent variables are included in a ~~set~~joint predictor, the higher that set of ~~variables~~ predictors will rank on average (Figure 9). However, for extremely high water levels, this trend gradually reverses (Figure 10). For action stage~~[17]~~ and minor flood stage,~~[18]~~ a slightly increasing trend is still visible. For moderate ~~[19]~~ and major flood stage,~~[20]~~ combinations with fewer independent variables rank higher on average. The most likely explanation is that extreme events like major and moderate flood stage are infrequent. After all, major flood stage equals 90th to 100th percentiles at the various gages. This data scarcity can lead to overfitting when using more predictors.

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

~~[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.~~

~~[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.~~

The combinations including the forecast (indicated by gray vertical lines in Figure 9 and Figure 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 (Figure 11, Figure 12). Without a transformation into the normal domain, the ~~forecast does not provide a lot of information for the QR model~~ scatterplot of forecast and forecast error does not show a trend. After NQT, the percentiles show trends laid out like a fan. ~~.~~In contrast, ~~the other four variables do not lend themselves for linear quantile regression after performing NQT~~the other four predictors become uniform distributions after NQT transformation. There is no trend detectable anymore. Further research is necessary to reconcile these two types of ~~variables~~predictors. A possible solution could be to ~~build~~ define QR ~~model~~configurations for subsets of the transformed dependent and independent variable.

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

Figure 10: Average rank for each ~~variable combination~~joint predictor for one to four days of lead time and four flood stages. Vertical gray lines indicate ~~variable combination~~joint predictors 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~~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: ~~rise rates~~rates of rise.

3.2.3 Brier Skill Score

Figure 13 illustrates the BSS when using the forecast as the only predictor as studied by Weerts et al. (2011). Confirming Wood et al.'s findings (2009), additionally i~~I~~ncluding the ~~rise rate~~rate of rise and forecasts errors as independent variables into the QR ~~model~~configuration improves the Brier Skill Score (BSS) significantly. ~~. illustrates the BSS when using the model as originally introduced by Weerts et al. .~~Using the best performing ~~variable combination~~joint predictors ~~instead,~~ gives an upper bound of the BSSs that can be achieved at best. This configuration increases the mean and decreases the standard deviation ~~(,).~~(Figure 14, Figure 16). The performance improves most where all ~~model~~ configurations perform worst: at the 10th percentile. Possibly, the configurations 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 (Figure 16).~~[21]~~ The decrease of the BSSs with lead time also becomes considerably less with this configuration. Additionally, a~~n~~ one-size-fits-all approach was tested to investigate, whether customizing the QR ~~model~~configuration to each river gage would be worth it. In this configuration, the ~~rise rates~~rates of rise in the past 24 and 48 hours and the forecast errors 24 and 48 hours ago serve as the independent variables (combination 30). It was found that this approach returns only slightly worse results than working with the best

~~[21]~~ ~~**Possibly, the model**~~**configurations** ~~**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 (**~~**Figure 6: 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.**

~~Table 2).~~

performing configuration for each river gage deviation (Figure 15, Figure 16). ~~(;).~~ Accordingly, the same ~~variable combination~~joint predictor can be used for all river gages.

As ~~shown in ,~~already discussed earlier, this last conclusion is not true for extremely high water levels. Including more independent variables does improve the BSSs considerably deviation (Figure 17,18, and 19). ~~(and ;).~~ However, for each river gage the best ~~combination of variables~~joint predictor needs to be identified separately. Because data to ~~build models~~define configurations is scarce for extreme levels, the QR ~~model~~configurations all perform less well for each increase in flood stage.

~~Table 3: Mean and standard deviation 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 hours and the forecast errors 24 and 48 hours ago as independent variable (one-size-fits-all solution).~~

Figure 13: Brier Skill Scores of the ~~original~~ forecast-only QR ~~model~~configuration (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~~joint predictor for each river gage as independent variables in the QR ~~model~~configuration.

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~~configuration.

Figure 16: Empirical cumulative density functions of three QR configurations predicting exceedance probabilities of the 10^{th}, 25^{th}, 75^{th}, and 90^{th} percentile: 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]; rates of rise in the past 24 and 48

hours and the forecast errors 24 and 48 hours ago as independent variable (one-size-fits-all solution) [rr+err24/48].

Figure 17: Brier Skill Scores of the ~~original~~ forecast-only QR ~~model~~configuration (i.e., using the transformed forecast as the only independent variable) for four lead times and flood stages.

Figure 18: Brier Skill Scores for four lead times and flood stages of observed water levels using the best ~~variable combination~~joint predictor for each river gage as independent variables in the QR ~~model~~configuration.

Figure 19: 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]

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~~predictors than just the forecast. ~~Figure 18~~Figure 20 shows that the ~~original~~ forecast-only QR ~~model~~ configuration as studied by Weerts et al. (2011)~~(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~~rates of rise and forecast errors as independent variables, because the resolution increases. This confirms the finding by Wood et al. (2009) that QR error models should be based on rate of rise (as well as lead time). The forecast quality improves along other ~~dimensions~~ metrics 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)~~(e.g., WWRP/WGNE, 2009)~~.

Figure ~~18~~20: Comparison of the ~~original~~ forecast-only QR ~~model~~configuration (i.e., only transformed forecast as independent variables) and the one-size-fits-all approach (i.e., ~~rise rates~~rates of rise 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.

3.3 Robustness

The impact of the length of the training dataset on the ~~model~~configuration's performance measured by the Brier Skill Score (BSS) was assessed for the one-size-fits-all QR ~~model~~configuration (i.e., ~~rise rates~~rates of rise and forecast errors as independent variables for all gages) for Hardin and Henry on the Illinois River. We were particularly interested in testing how many years of training data are necessary to achieve satisfactory forecasting results. Each year between 2003 and 2013 was forecast by QR ~~model~~configurations trained on ~~on one year up to~~ however many years of archived forecasts were ~~available~~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 21 and Figure 22 show that training datasets shorter than three years result in very low BSSs for low event thresholds (Q10) at Henry and Hardin.~~show that for those gages,~~ For the other event thresholds, it ~~does not~~barely matters ~~-~~for the BSS how many years are included in the training dataset. That is good new~~news~~s, if stationarity cannot be assumed (Milly et al., 2008)~~(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. Only needing short time series to define a skillful QR configuration implies that the configuration parameters can be updated regularly. This way, changing relationships between predictors etc. can be taken into account.

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

To generalize the result, the same analysis as just described for Hardin and Henry was repeated 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 (Table 2). The forecast performance was found to vary statistically significantly across all those dimensions except the number of training years. This results in a very wide range of Brier Skill Scores (Figure 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 QR configuration based on the forecast only (not shown).

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

A closer look at the regression coefficients (Table 2) provides interesting insights. For low event thresholds, the BSSs are much worse than for high thresholds. The QR configurations 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. The illustrative cases of Henry and Hardin, described above, indicate that using longer time series to predict exceedance probabilities of low event thresholds improves forecast performance.

As expected, the BSSs slightly decrease with lead time. 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 2).

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 2 for the sake of brevity. Instead, Figure 23 maps the geographic position of the river gages with the color code indicating each gage's regression coefficient. The coefficients are lower, and therefore the Brier Skill Scores are lower, for gages far upstream a river and those close to confluences. 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.

Figure 21~~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 -22~~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~~23: Geographical position of rivers. Colors indicate the regression coefficient of each station with the Brier Skill Score as dependent variable.

Figure ~~22~~24: Minimum (black) and maximum (red) Brier Skill Scores for various lead times and event thresholds across locations, size of training dataset and forecast years.

4 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., 10^{th}, 25^{th}, 75^{th}, 90^{th} 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 U.S. American context. Additionally, it~~It further develops the ~~method~~ application of QR to estimating river forecast uncertainty ~~by~~ (a) ~~including more~~comparing different sets of independent variables, (b) and testing the ~~method~~technique's robustness across locations, lead times, event thresholds, forecast years and sizes of training dataset.

~~Most importantly~~When compared to the configuration using only the forecast, it was found that including ~~rise rates~~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 ~~model~~configuration, as measured by the Brier Skill Score. This confirms Wood et al.'s (2009) finding that QR error models should be a function of rate of rise and lead time. ~~Since the reliability was already high with the original QR method as proposed by Weerts et al. ,~~ 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~~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 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 ~~variables~~ 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. ~~a one-size-fits-all approach (i.e. the rise rates and forecasts errors as independent variables) performs satisfactorily for those cases.~~

When forming a joint predictor, the independent variables rates of rise and forecast errors do not combine well with the forecast itself. To account for heteroscedasticity, the forecast was transformed into the Gaussian domain. However, no trend is detectable anymore between forecast error and the rates of rise or the previous forecast errors after applying NQT to those variables. Therefore, it is difficult to combine these two predictors. A possible solution could be to define QR configurations for subsets of the transformed data. However, such an approach drastically decreases the amount of data available for each configuration.

~~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. and López López et al. . 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~~ studied QR ~~method~~ configurations are~~is~~ relatively robust to the size of training dataset, which is convenient if stationarity cannot be assumed (Milly et al., 2008)~~(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~~technique ~~does depend~~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.

The ~~proposed~~ studied QR ~~approach~~ configurations perform~~s~~ 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 QR ~~model~~configurations 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~~configuration's performance.

Future Work

This~~e~~ ~~method~~technique~~s~~ 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~~technique, classification trees, showed that the observed precipitation, the precipitation forecast (i.e., POP – probability of precipitation) and the upstream water levels significantly improve ~~models~~forecasting performance. Presumably, this is the case, because the ~~QPF~~ forecast used in this study includes the precipitation forecast ~~only~~ 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.~~[23]~~ For a period of 12 years, requesting such data for several weather stations~~[24]~~ is obviously time-consuming; n~~.~~ot 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 ~~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~~performance also warrant consideration. Particularly, clustering the data by variability seems promising. However, early trials indicated that this ~~method~~technique is very sensitive to the training dataset.

~~[23] URL [accessed July 2014]: http://cdo.ncdc.noaa.gov/pls/plhas/HAS.FileAppSelect?datasetname=9957ANX~~

~~[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~~ approach 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~~Additionally, the ~~proposed~~ studied ~~method~~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 ~~have~~ has had some of the most unfortunate impacts on decision-making in Grand Forks, ND and Fargo, ND (Pielke, 1999; Morss, 2010)~~(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 two 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).~~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, 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.

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.

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.

~~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.~~

~~Anon: Brier score, Wikipedia Free Encycl. [online] Available from: http://en.wikipedia.org/w/index.php?title=Brier_score&oldid=619686224 (Accessed 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(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.~~

~~Bröcker, J.: Estimating reliability and resolution of probability forecasts through decomposition of the empirical score, Clim. Dyn., 39(3-4), 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, Bull. Am. Meteorol. Soc., 95(1), 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. R. Meteorol. Soc., 138(668), 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. 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.~~

~~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.~~

~~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(2), 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. [online] Available from: http://www.nap.edu/catalog.php?record_id=11699 (Accessed 18 September 2014), 2006.~~

~~NCAR-Research Applications Laboratory, N.-R. A.: verification: Weather Forecast Verification Utilities. [online] Available from: http://cran.r-project.org/web/packages/verification/index.html (Accessed 27 August 2014), 2014.~~

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

~~R Core Team: R: A language and environment for statistical computing., [online] Available from: http://ww.R-project.org/., 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(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.~~

~~Stephenson, D. B., Coelho, C. A. S. and Jolliffe, I. T.: Two Extra Components in the Brier Score Decomposition, Weather Forecast., 23(4), 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(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.~~

~~Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press, San Diego., 1995.~~

~~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.~~

~~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: ~~Variable Combination~~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~~rate of rise in the past 24 and 48 hours;
err24, err 48 = forecast error 24 and 48 hours ago
The forecast error equals the difference between the current (i.e., at issue time of the forecast) water level and the forecast that was produced 24/48 hours 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.**~~

Average errors			Lead Time			
of 82 gages	**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: Mean and standard deviation 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 hours and the forecast errors 24 and 48 hours ago as independent variable (one-size-fits-all solution).**~~

	~~**Q10**~~	~~**Q25**~~	~~**Q75**~~	~~**Q90**~~	~~**Q10**~~	~~**Q25**~~	
			~~**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)~~	~~0.8~~
~~**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)~~	~~0.9~~
~~**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)~~	~~0.9~~
			~~**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)~~	~~0.7~~
~~**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)~~	~~0.8~~
~~**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)~~	~~0.8~~
	~~**Action**~~	~~**Minor**~~	~~**Moderate**~~	~~**Major**~~	~~**Action**~~	~~**Minor**~~	~~**M**~~
			~~**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.~~
~~**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.~~
			~~**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.~~
~~**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.~~

~~* Combination 30~~

Table 2: Regression results

	Coef.	**St.Dev.**	
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			

Figures

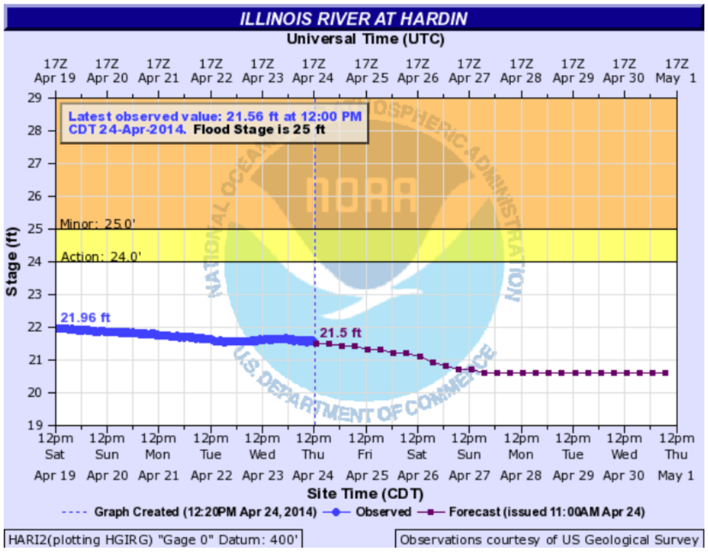
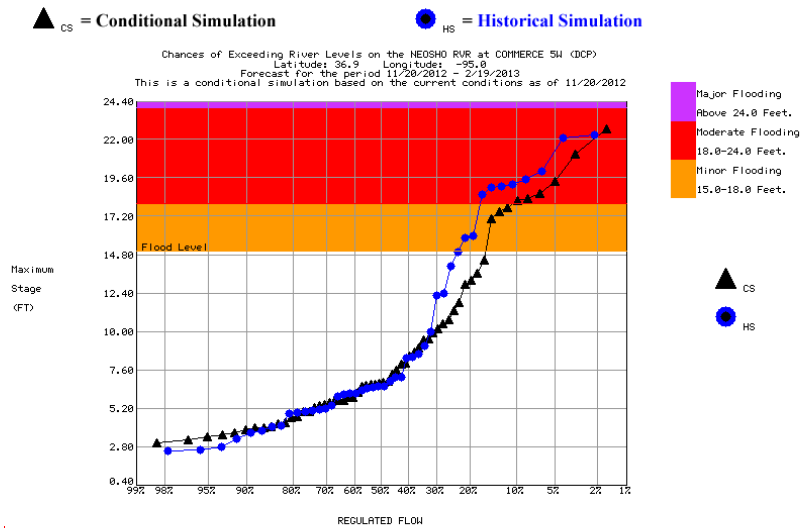
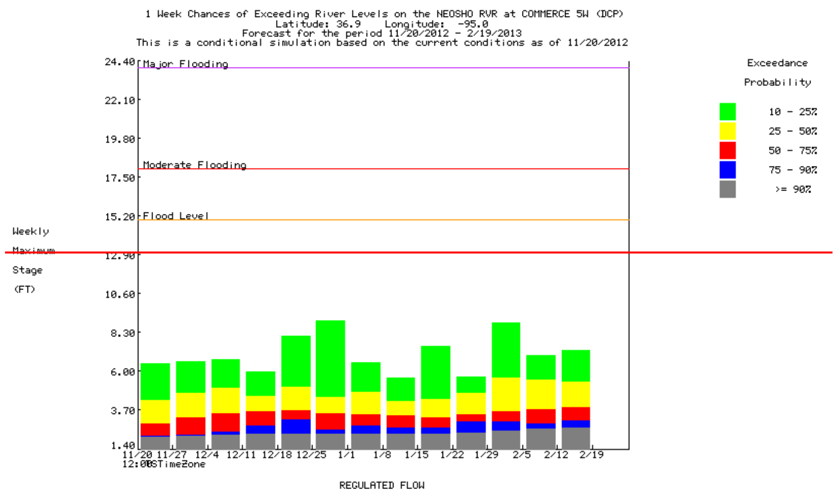


Figure 1: Deterministic short-term weather forecast in six hour intervals as published by the NWS for Hardin, IL on 24 April 2014.

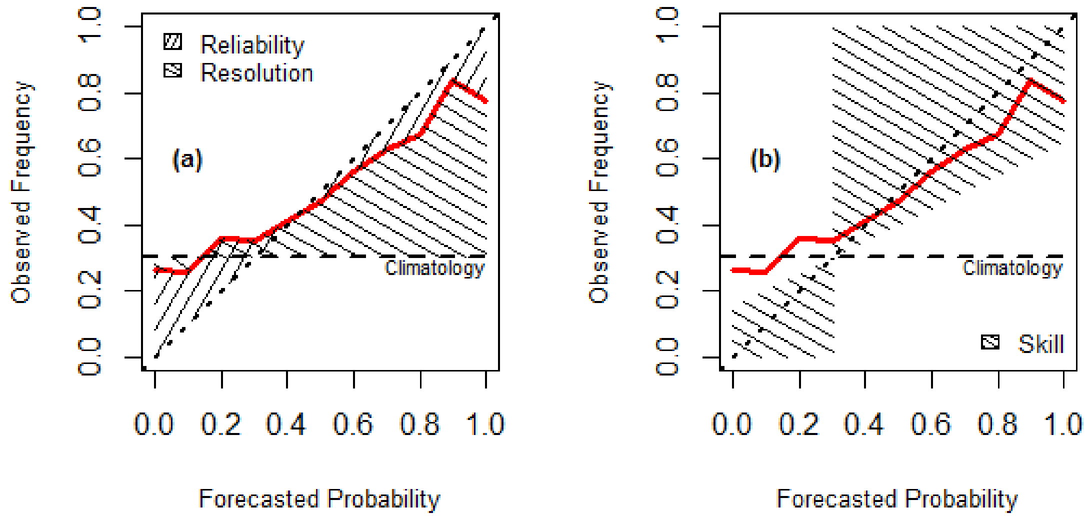
Source:http://water.weather.gov/ahps2/hydrograph.php?wfo=lsx&gage=hari2.



~~**Figure 2: Probabilistic long-term forecast as published by the NWS for Commerce, OK on December 14th, 2012: Exceedance curve for three months period. (Not available for Hardin, IL). Source: http://water.weather.gov/ahps2/hydrograph.php?wfo=tsa&gage=como2**~~



~~**Figure 3: Probabilistic long-term forecast as published by the NWS for Commerce, OK on December 14th, 2012: Bar plot for each week of a three months period. (Not available for Hardin, IL). Source: http://water.weather.gov/ahps2/hydrograph.php?wfo=tsa&gage=como2**~~



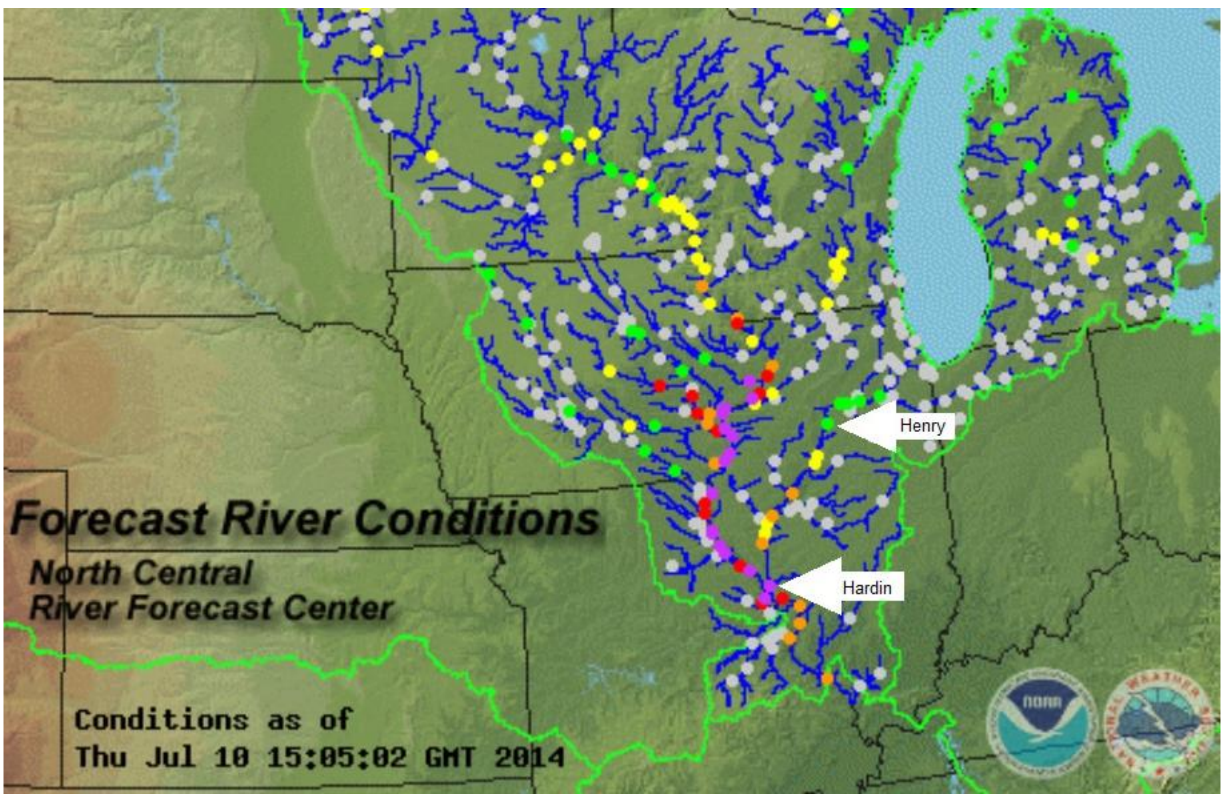


Figure 24: Theory behind Brier Skill Score illustrated for an imaginary forecast (red line): (a) reliability and resolution; (b) skill. In figure a, the area representing reliability should be as small, and for resolution as large as possible. The forecast has skill (BSS > 0), i.e., performs better than ~~random guessing~~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.)__~~(Adapted from Hsu and Murphy, 1986; Wilson, n.d.)~~.

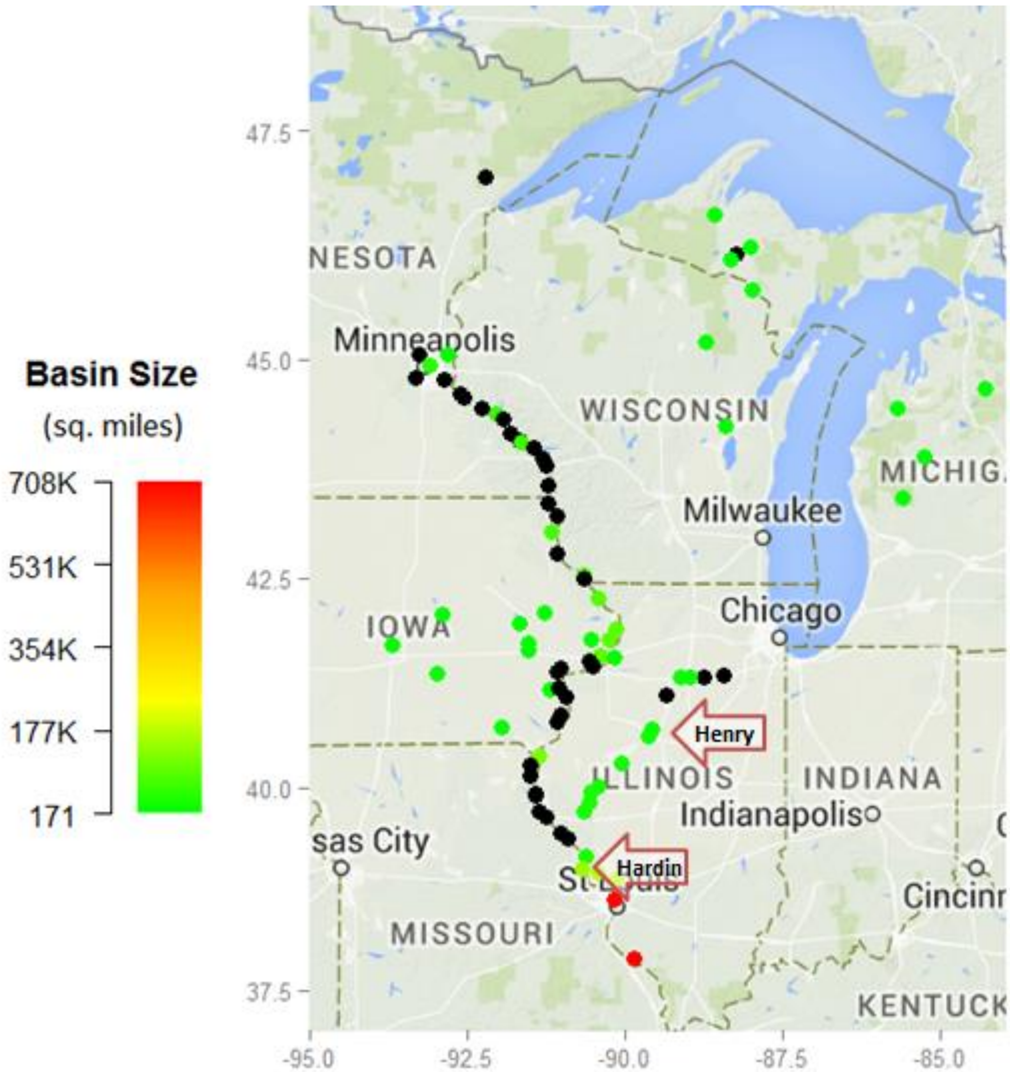


Figure 3~~5~~: ~~Portion~~ River gages for which the ~~of the~~ North Central River Forecast Centers ~~river gages with~~publishes forecasts daily. Henry (HYNI2) and Hardin (HARI2) are indicated by the upper and lower red arrow respectively. For gages indicated by black dots the basin size is missing.~~Source:~~

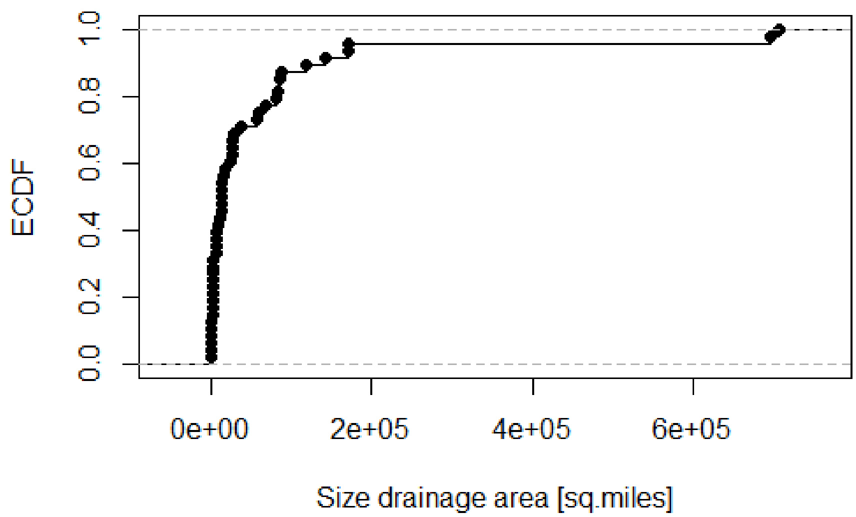
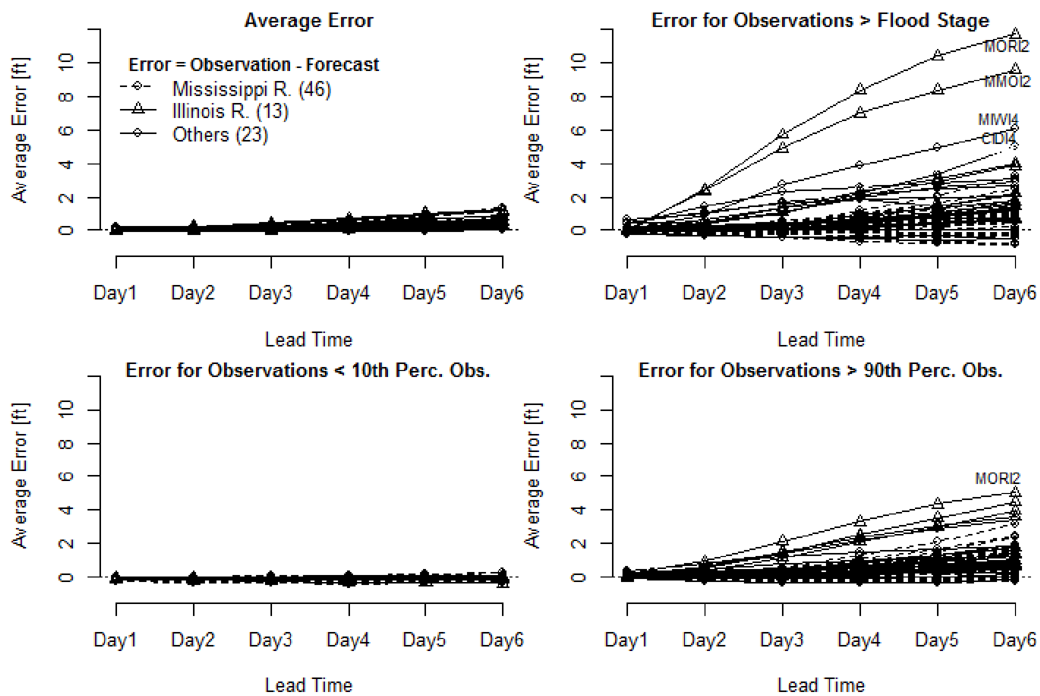


Figure 4: Empirical cumulative density function (ecdf) of sizes of drainage area for the river gages that are being forecasted daily by the NCRFC.


Average Error
Error = Observation - Forecast
Mississippi R. (46)
Illinois R. (13)
Others (23)
Average Error [ft]
Lead Time
Day1
Day2
Day3
Day4
Day5
Day6
Error for Observations > Flood Stage
MORI2
MMOI2
MIWI4
CIDI4
Error for Observations < 10th Perc. Obs.
Error for Observations > 90th Perc. Obs.
MORI2

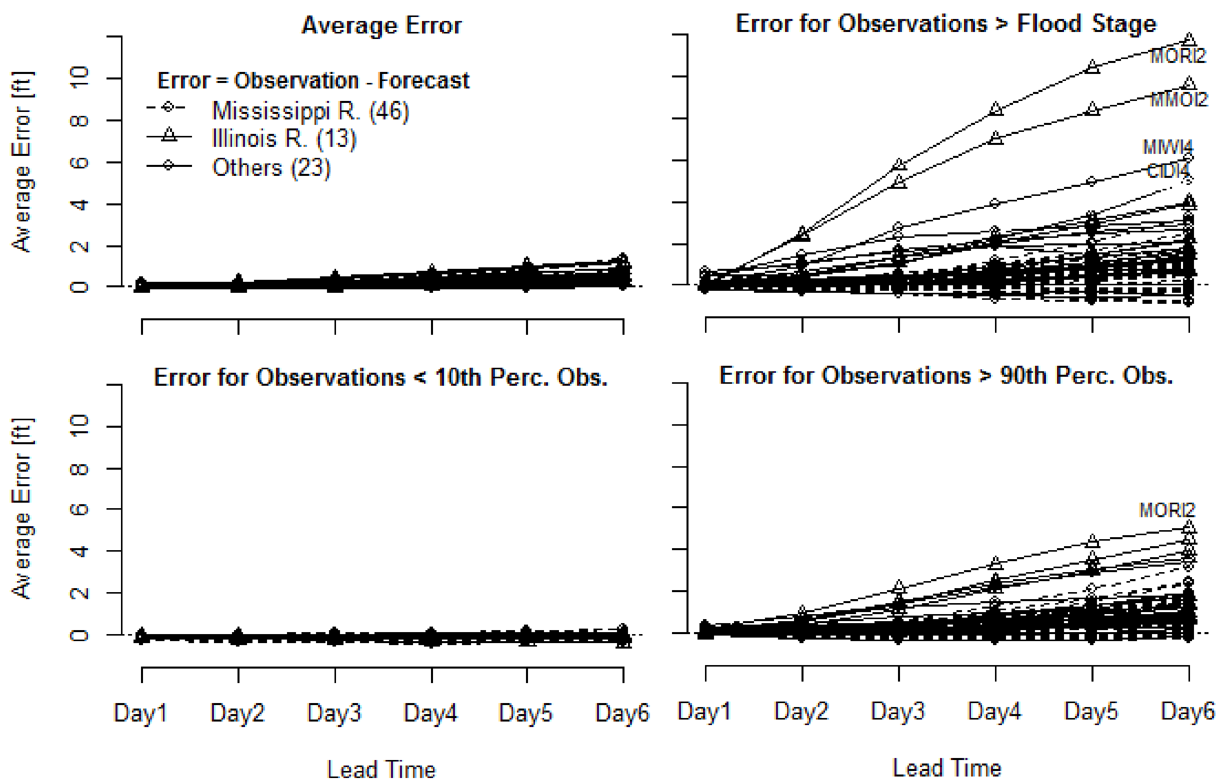

Average Error
Error = Observation - Forecast
Mississippi R. (46)
Illinois R. (13)
Others (23)
Average Error [ft]
Error for Observations > Flood Stage
MORI2
MMOI2
MIWI4
CIDI4
Error for Observations < 10th Perc. Obs.
Error for Observations > 90th Perc. Obs.
MORI2
Day1
Day2
Day3
Day4
Day5
Day6
Lead Time

Figure ~~5~~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.

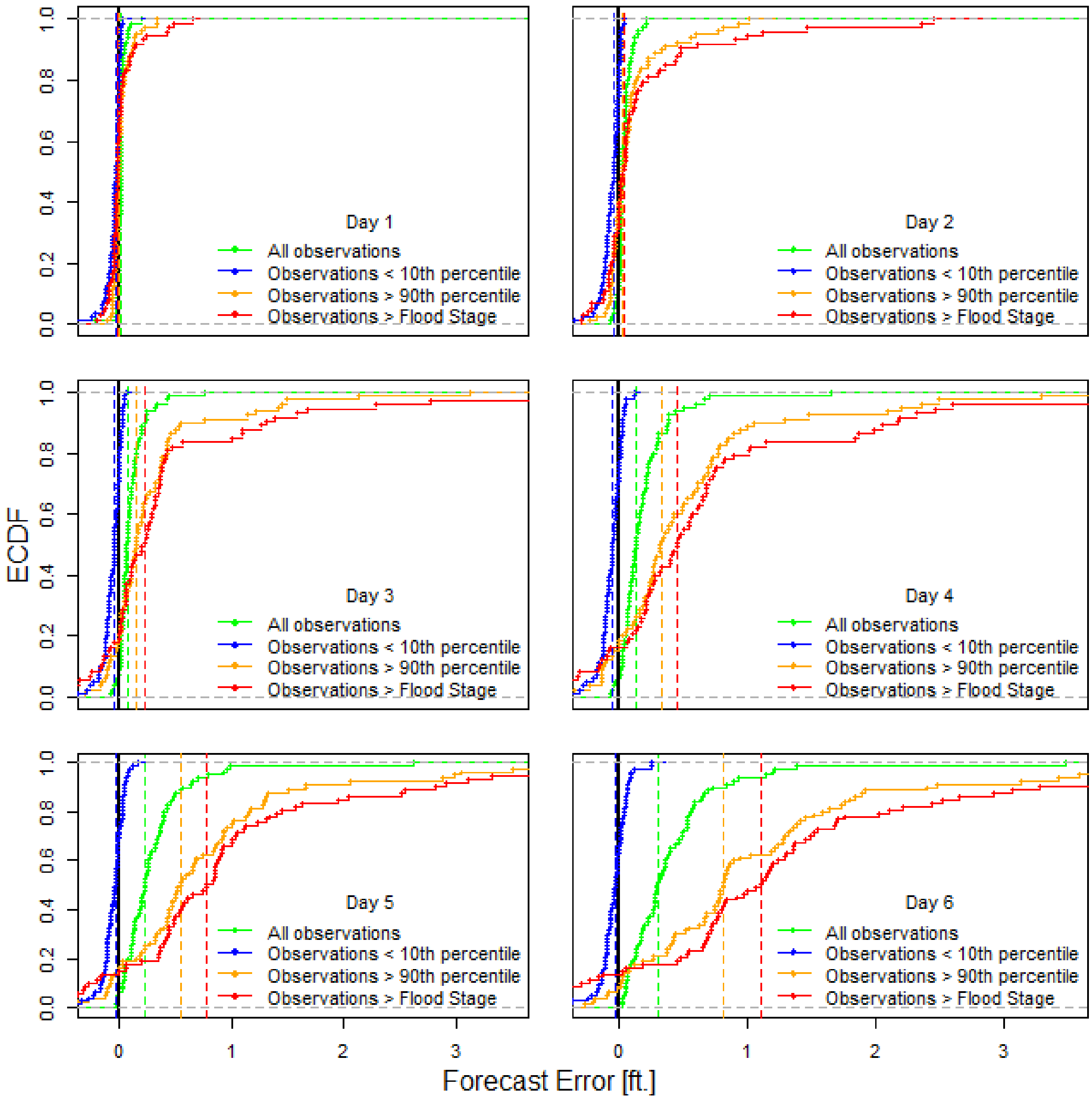
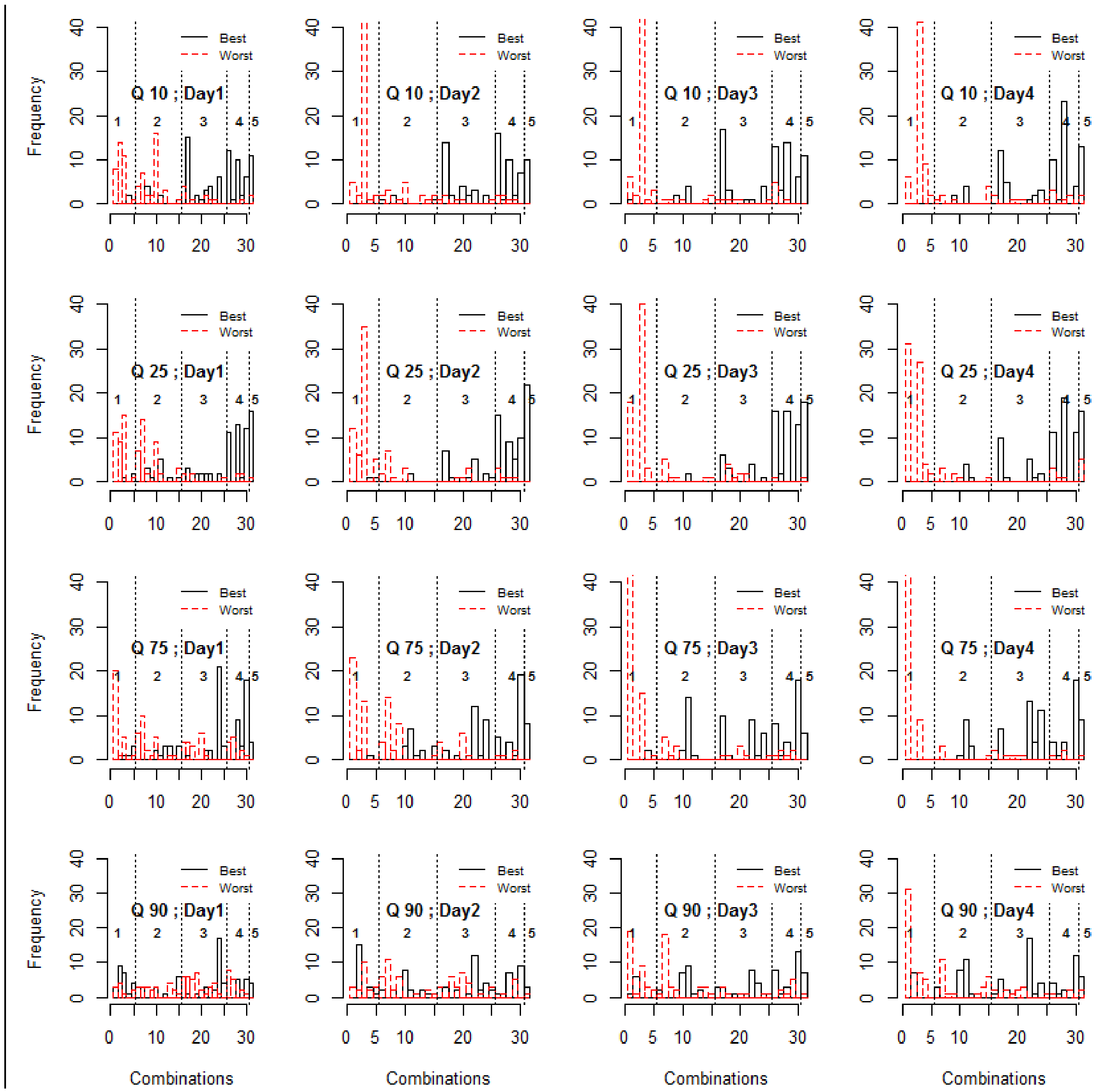


Figure 6: Empirical cumulative distribution function (ecdf) of forecast error at 82 river gages for six lead times. Vertical lines show the median forecast error of the corresponding subset.


Q 10 ; Day1
Q 10 ; Day2
Q 10 ; Day3
Q 10 ; Day4
Q 25 ; Day1
Q 25 ; Day2
Q 25 ; Day3
Q 25 ; Day4
Q 75 ; Day1
Q 75 ; Day2
Q 75 ; Day3
Q 75 ; Day4
Q 90 ; Day1
Q 90 ; Day2
Q 90 ; Day3
Q 90 ; Day4
Best
Worst
Frequency
Combinations

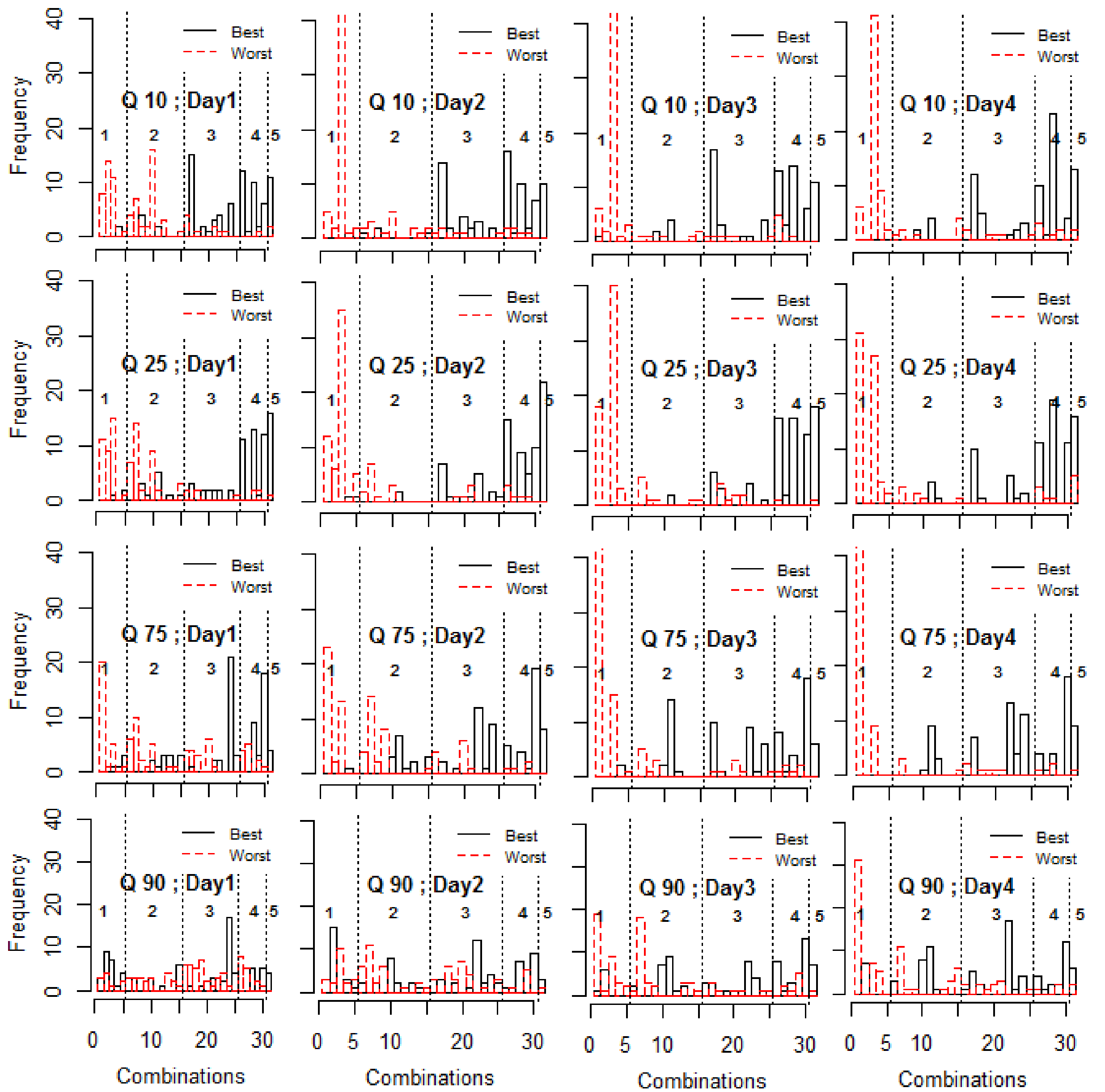
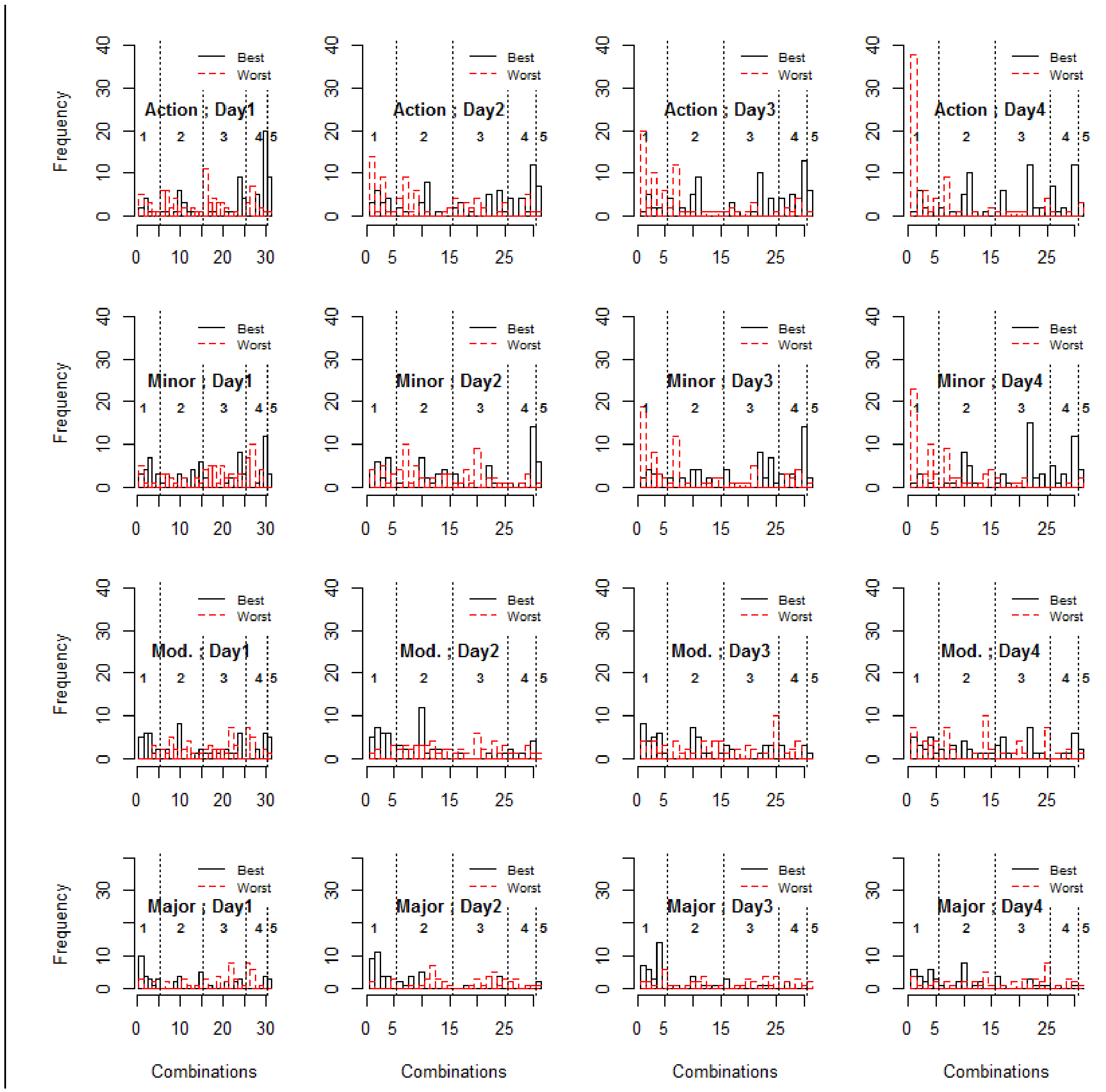


Figure 7: Histograms of ~~variable combination~~joint predictors returning the best and worst Brier Skill Scores across 82 river gages. Each row of histograms refers to an event threshold defined as a percentile of the observed water levels, and each column to a lead time. The dotted vertical lines in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


Action ; Day1
Action ; Day2
Action ; Day3
Action ; Day4
Minor ; Day1
Minor ; Day2
Minor ; Day3
Minor ; Day4
Mod. ; Day1
Mod. ; Day2
Mod. ; Day3
Mod. ; Day4
Major ; Day1
Major ; Day2
Major ; Day3
Major ; Day4
Best
Worst
Frequency
Combinations

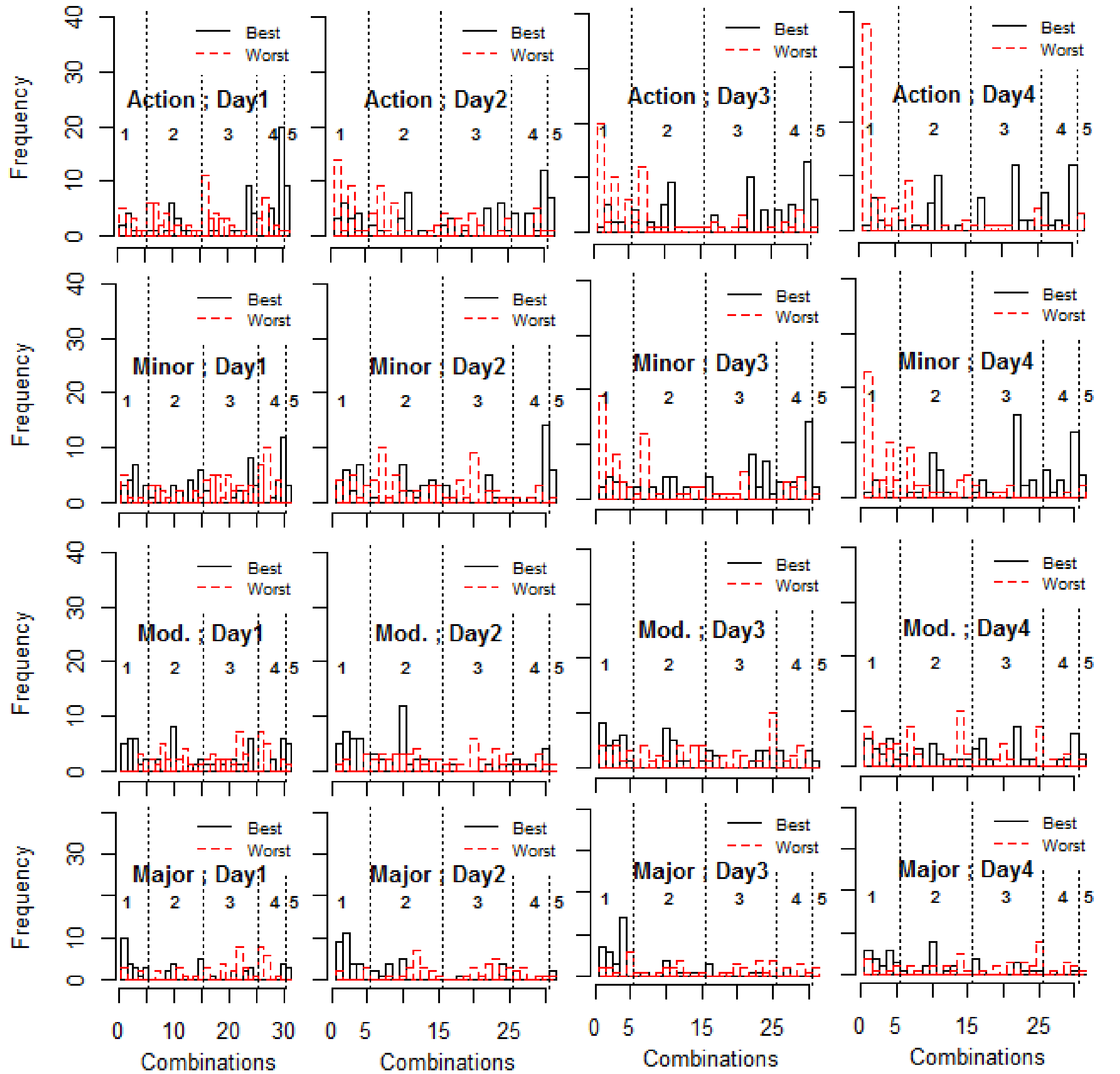
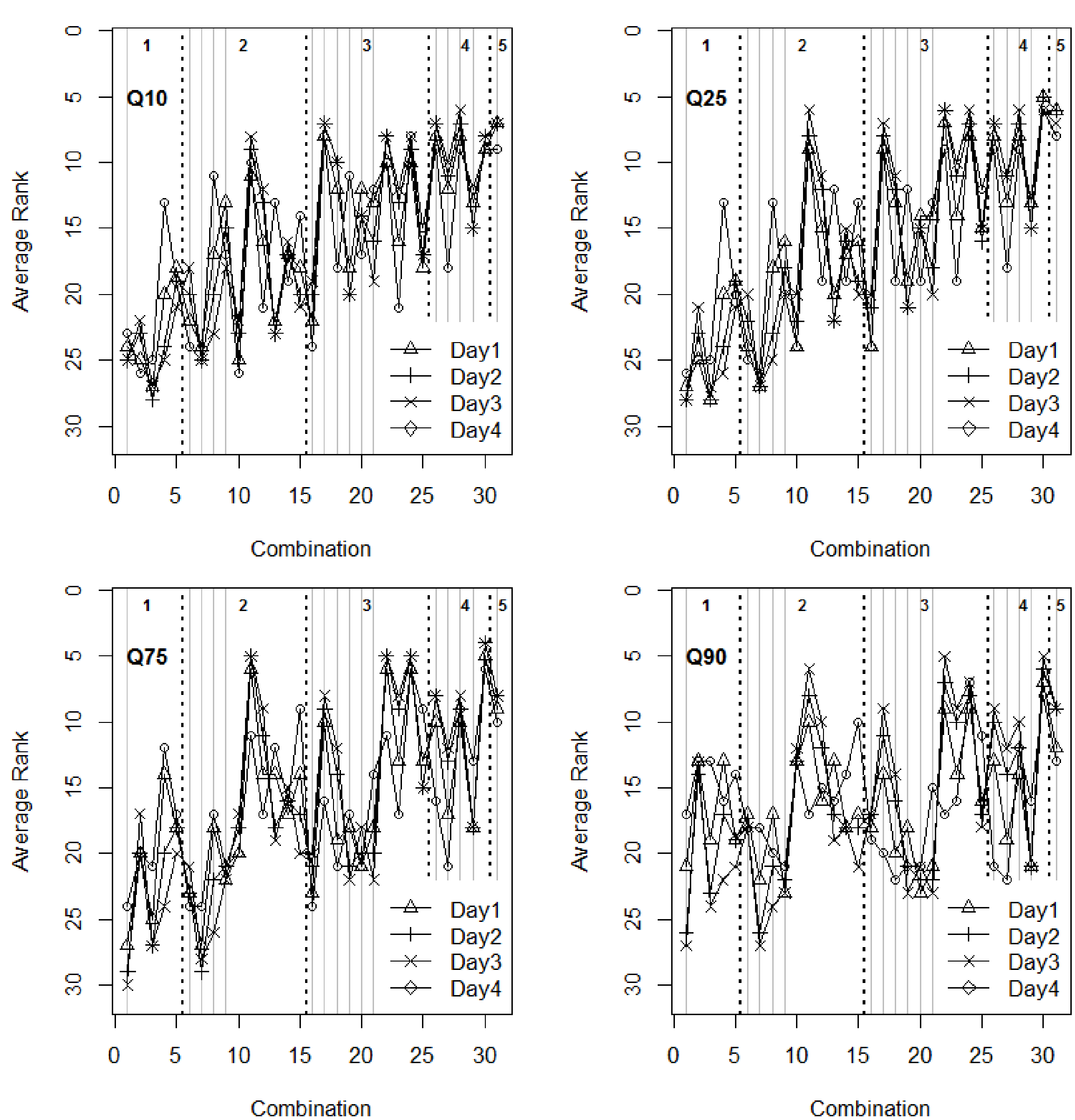


Figure 8: Histograms of ~~variable combination~~joint predictors returning the best and worst Brier Skill Scores across 82 river gages. Each row of histograms refers to a flood stage, and each column to a lead time. The dotted vertical lines in the histograms distinguish ~~variable combination~~joint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


Q10
Q25
Q75
Q90
1
2
3
4
5
Average Rank
Combination
Day1
Day2
Day3
Day4

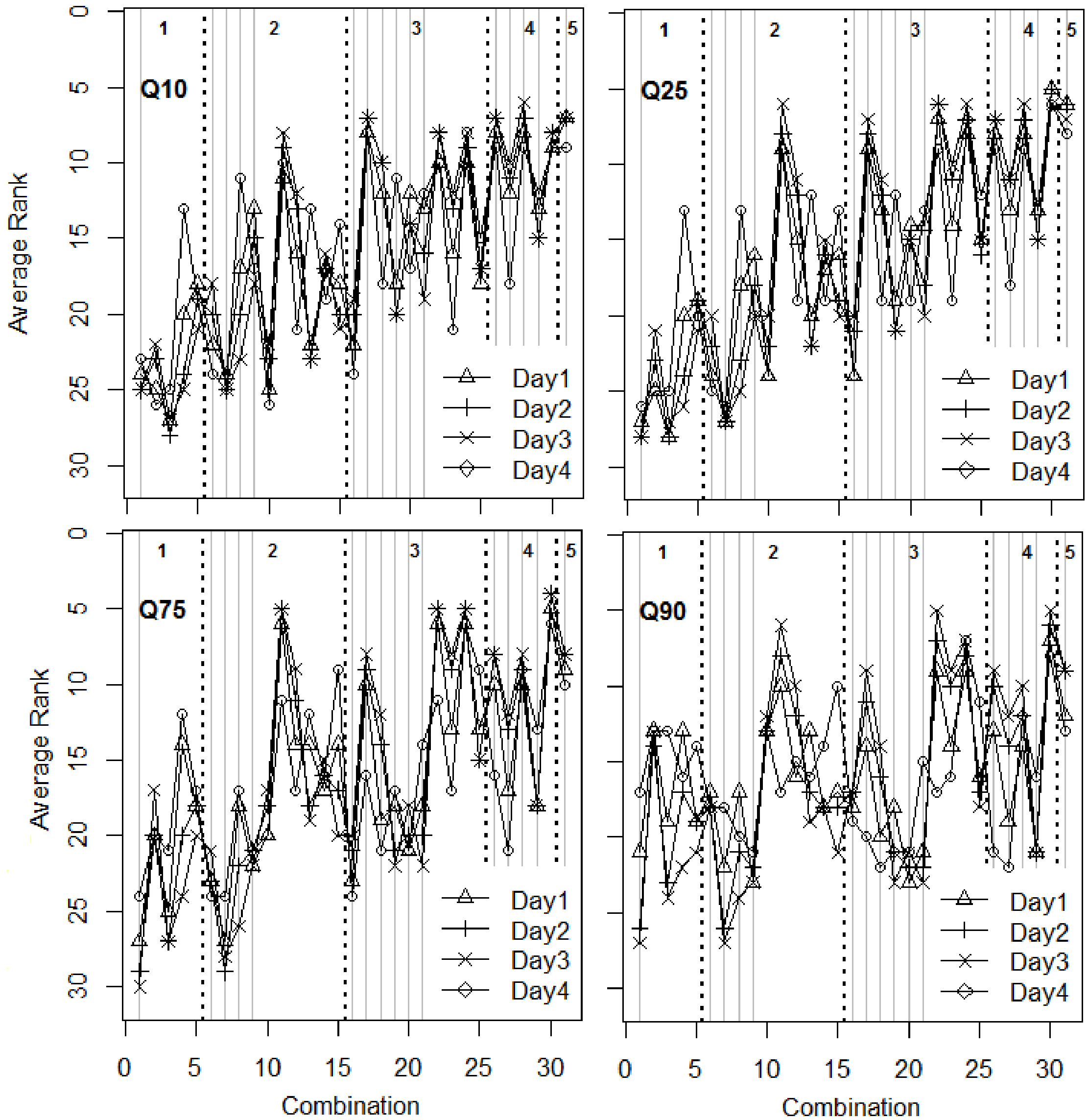
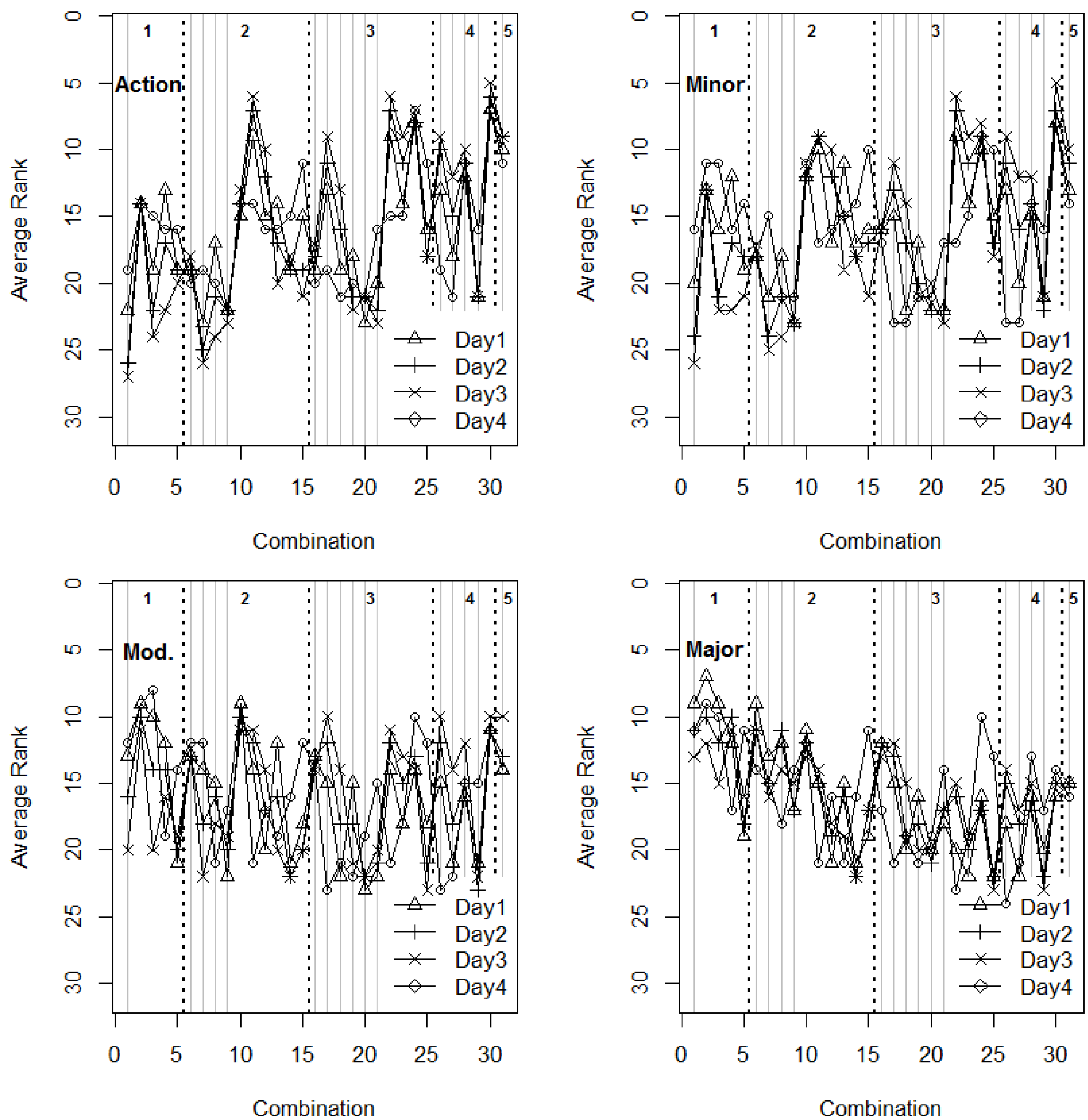


Figure 9: Average rank for each ~~variable combination~~joint predictor for one to four days of lead time and four percentiles of observed water levels. Vertical gray lines indicate ~~variable combination~~joint predictors including the forecast.


Action
Minor
Mod.
Major
Average Rank
Combination
1
2
3
4
5
Day1
Day2
Day3
Day4

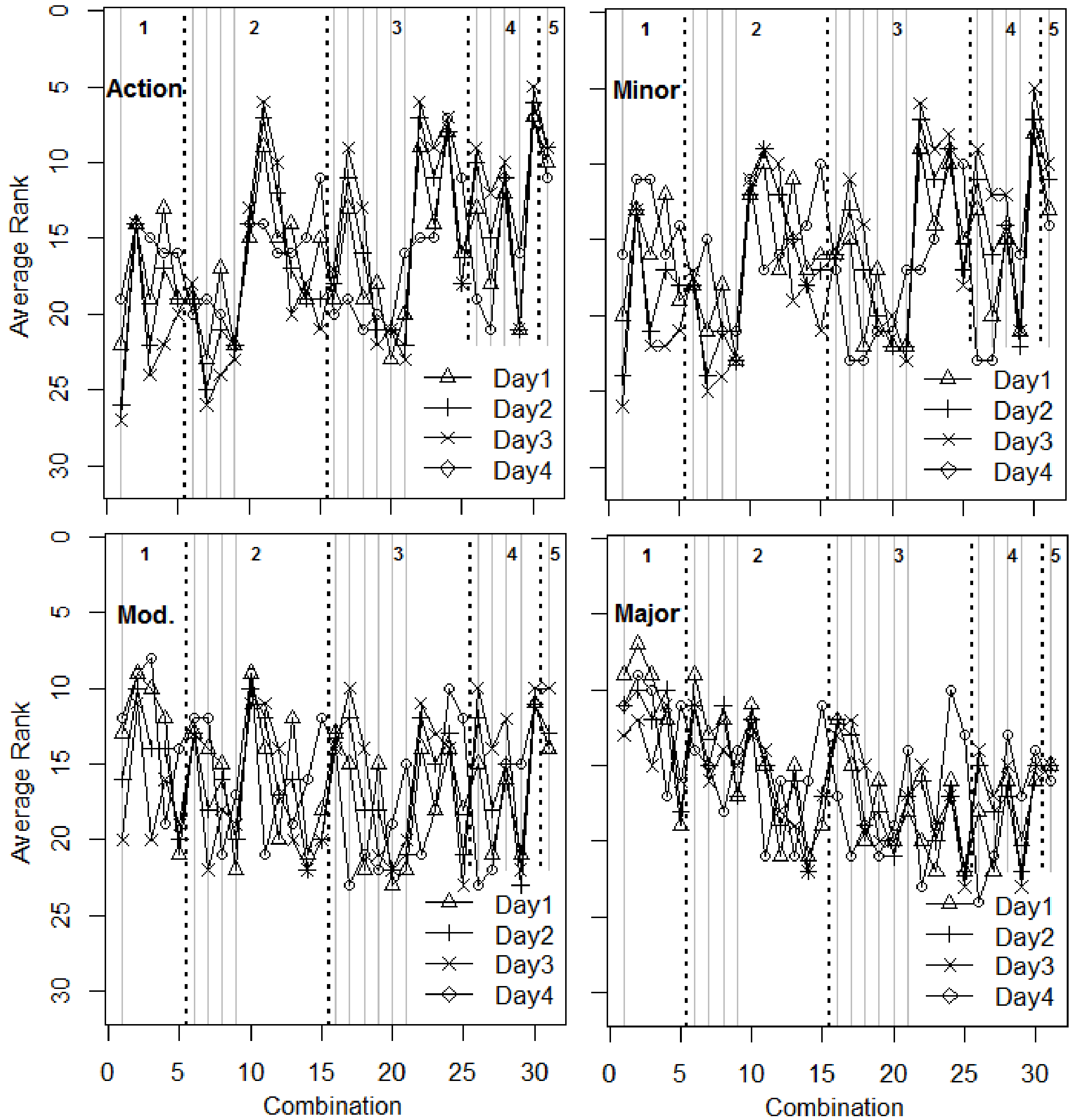
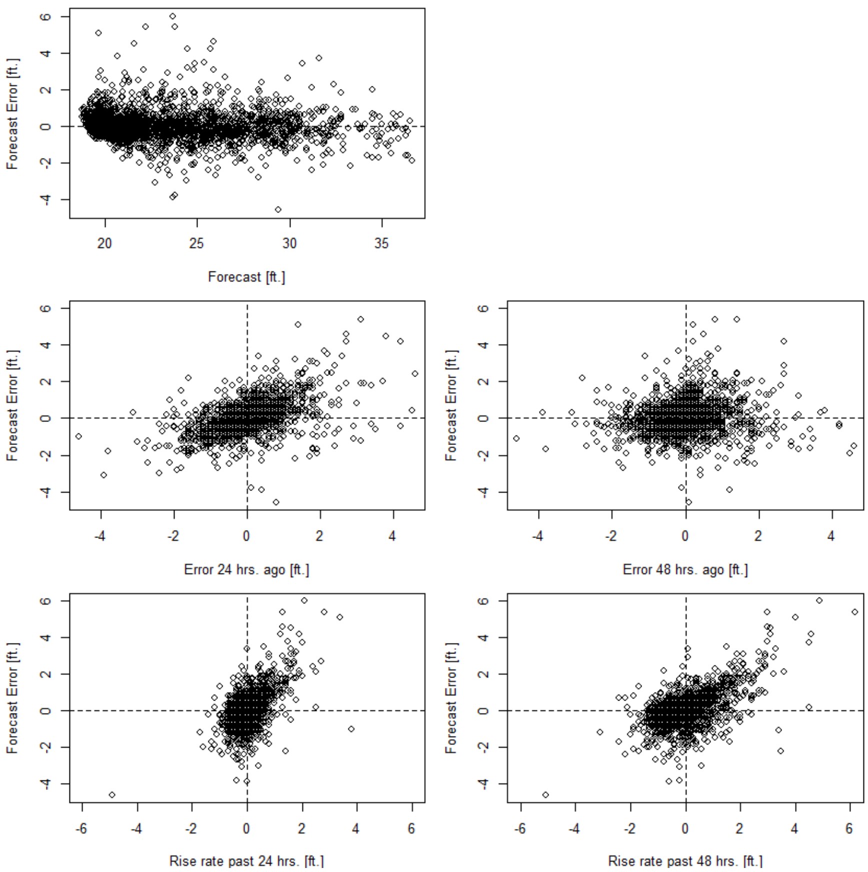


Figure 10: Average rank for each ~~variable combination~~joint predictor for one to four days of lead time and four flood stages. Vertical gray lines indicate ~~variable combination~~joint predictors including the forecast.


Forecast Error [ft.]
Forecast [ft.]
Error 24 hrs. ago [ft.]
Error 48 hrs. ago [ft.]
Rise rate past 24 hrs. [ft.]
Rise rate past 48 hrs. [ft.]

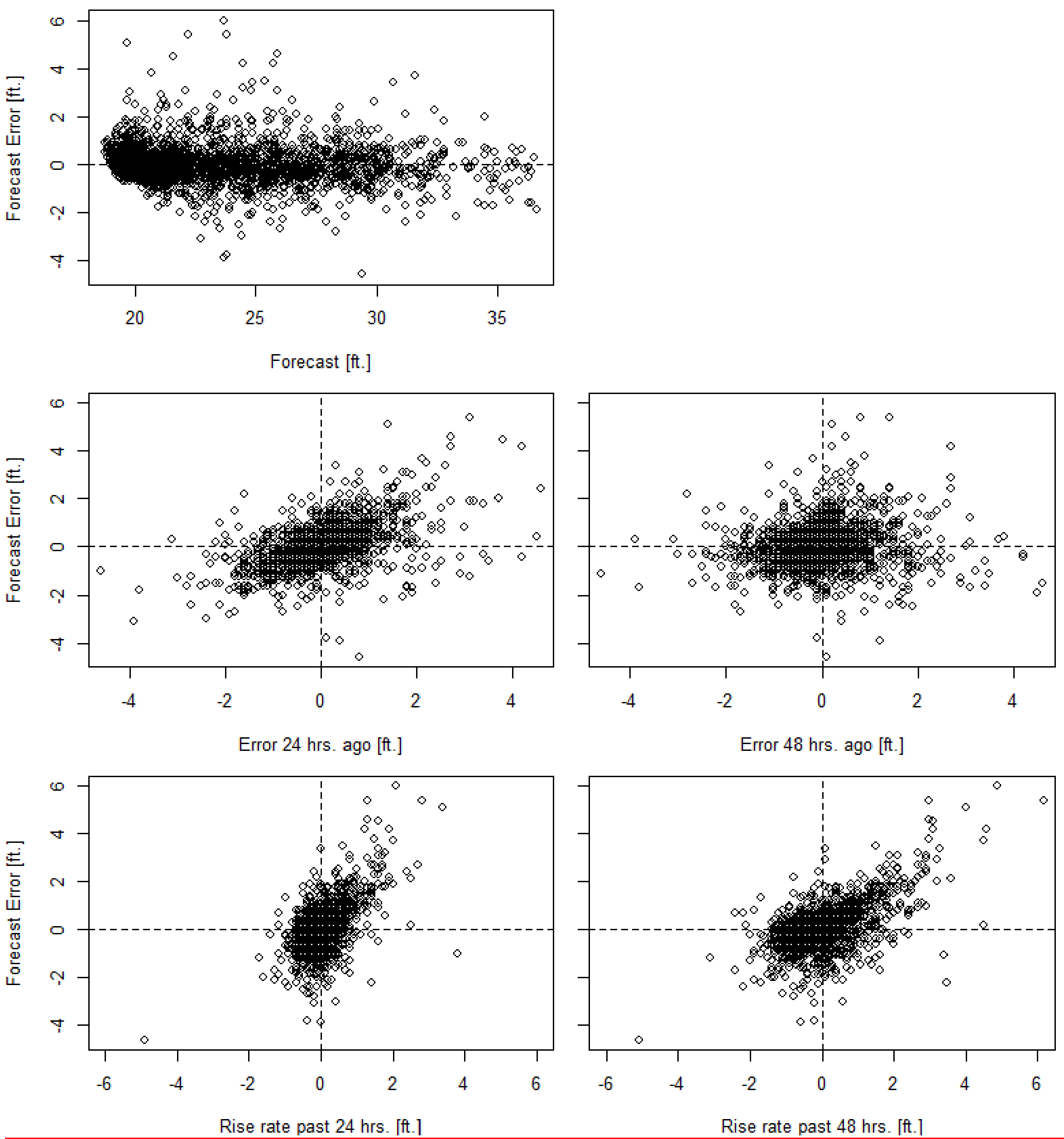
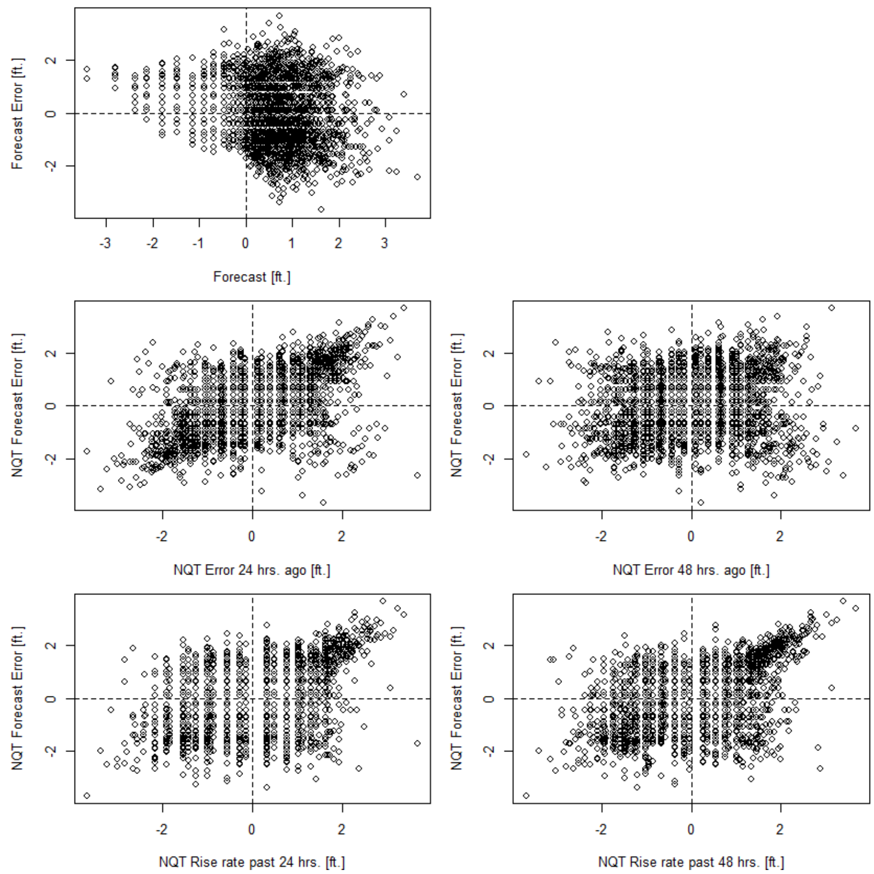


Figure 11: Independent variables plotted against the forecast error for Hardin IL with 3 days of lead time. First row: Forecast; second row: past forecast errors; third row: ~~rise rates~~rates of rise.


Forecast Error [ft.]
Forecast [ft.]
NQT Forecast Error [ft.]
NQT Error 24 hrs. ago [ft.]
NQT Error 48 hrs. ago [ft.]
NQT Rise rate past 24 hrs. [ft.]
NQT Rise rate past 48 hrs. [ft.]

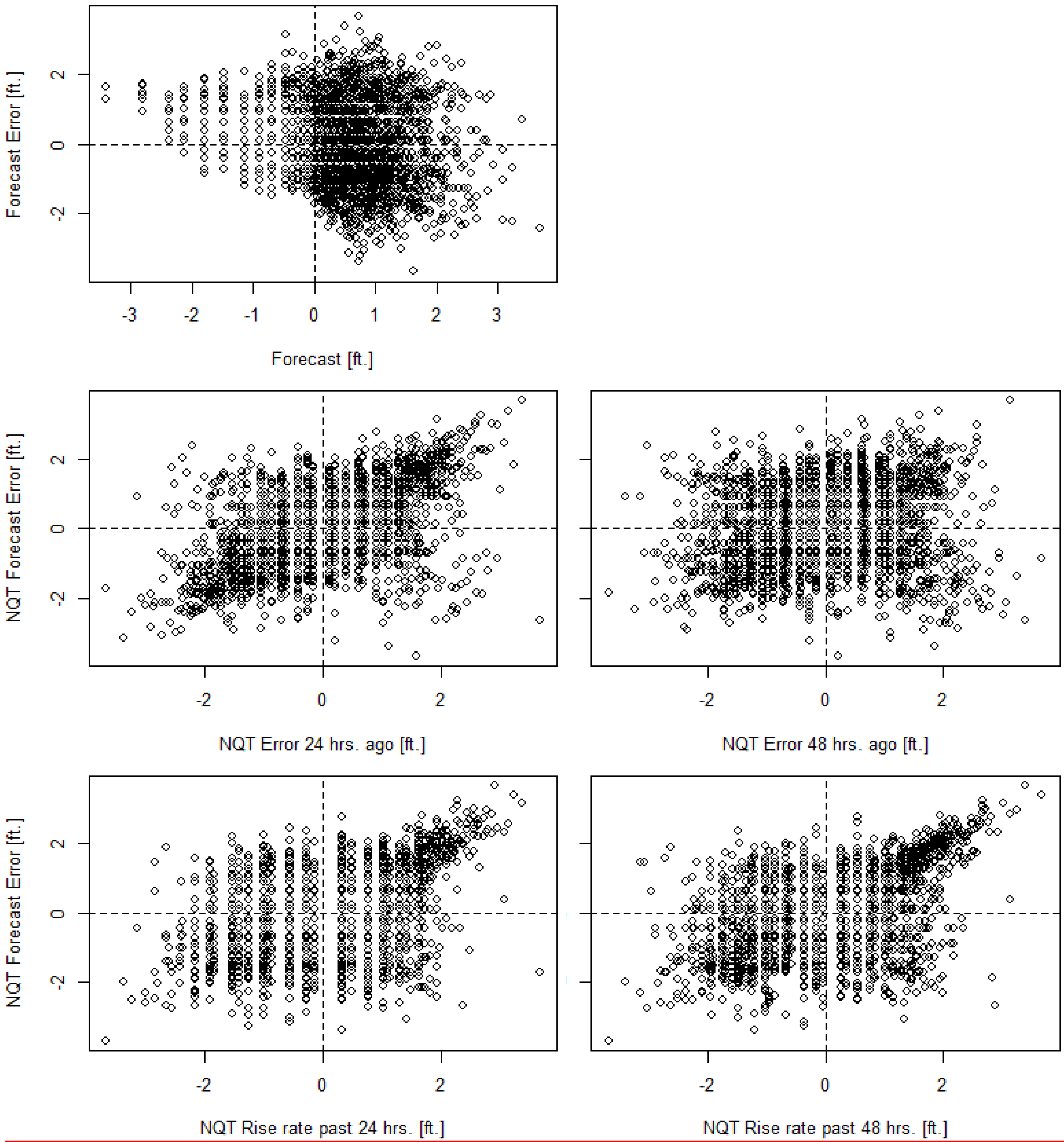
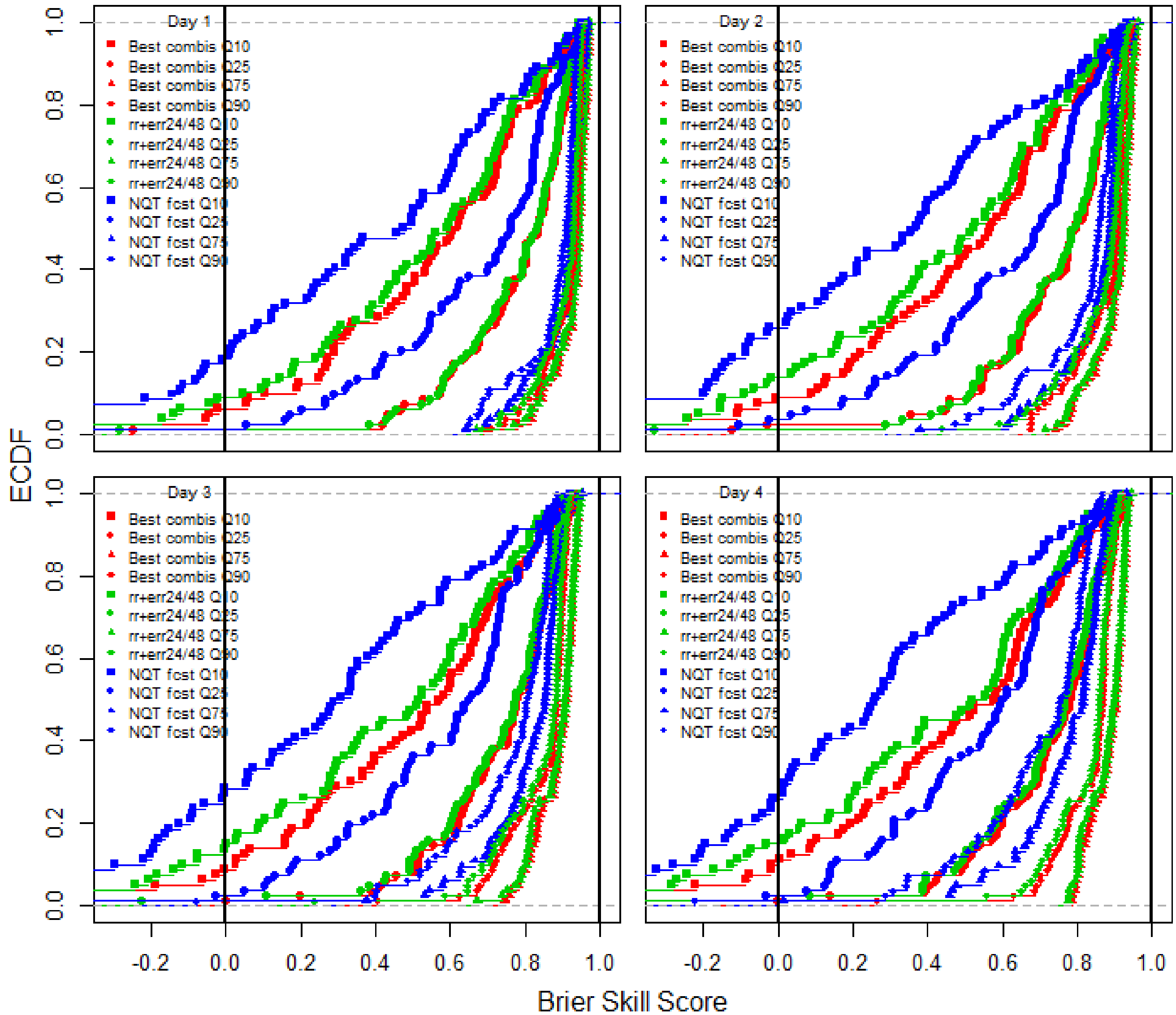


Figure 12: Independent variables after transforming into the Gaussian domain plotted against the forecast error for Hardin IL with 3 days of lead time. First row: Forecast; second row: past forecast errors; third row: ~~rise rates~~rates of rise.



<u>**Figure 16: Empirical cumulative density functions of three QR configurations predicting exceedance probabilities of the 10th, 25th, 75th, and 90th percentile: 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]; rates of rise in the past 24 and 48 hours and the forecast errors 24 and 48 hours ago as independent variable (one-size-fits-all solution) [rr+err24/48].**</u>

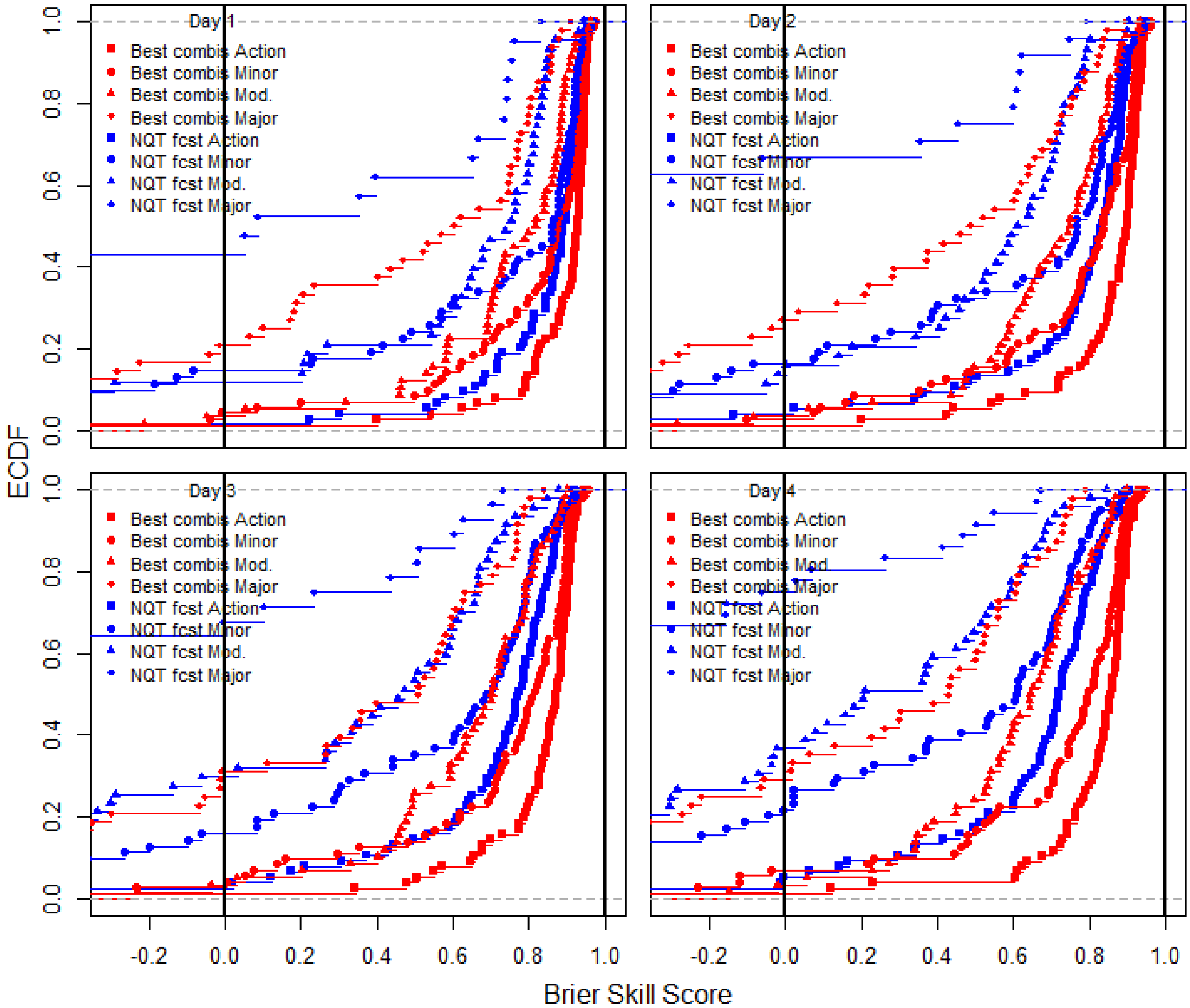
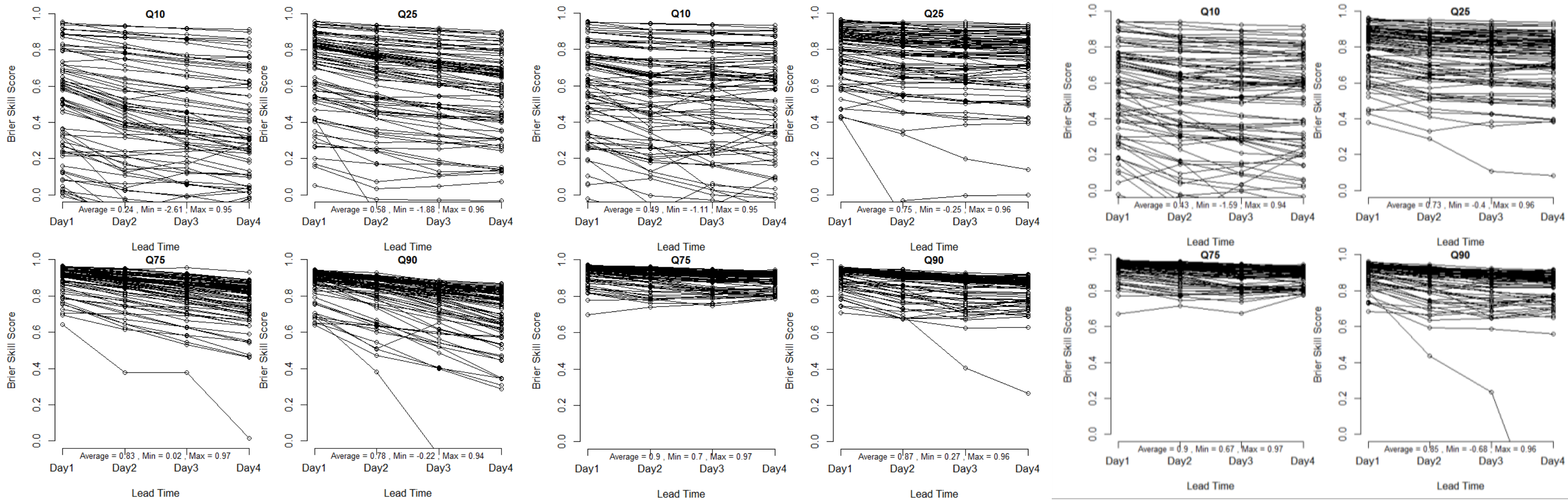


Figure 19: Empirical cumulative density functions of three QR configurations predicting exceedance probabilities of the Action, Minor, Moderate, and Major Flood Stage: the configuration using the transformed forecast as the only independent variable [NQT fcst]; the best performing combination for each river gage (upper performance limit) [Best combis]


Q10
Average = 0.24 , Min = -2.61 , Max = 0.95
Q25
Average = 0.58 , Min = -1.88 , Max = 0.96
Q10
Average = 0.49 , Min = -1.11 , Max = 0.95
Q25
Average = 0.75 , Min = -0.25 , Max = 0.96
Q10
Average = 0.43 , Min = -1.59 , Max = 0.94
Q25
Average = 0.73 , Min = -0.4 , Max = 0.96
Q75
Average = 0.83 , Min = 0.02 , Max = 0.97
Q90
Average = 0.78 , Min = -0.22 , Max = 0.94
Q75
Average = 0.9 , Min = 0.7 , Max = 0.97
Q90
Average = 0.87 , Min = 0.27 , Max = 0.96
Q75
Average = 0.9 , Min = 0.67 , Max = 0.97
Q90
Average = 0.85 , Min = -0.68 , Max = 0.96
Brier Skill Score
Lead Time
Day1
Day2
Day3
Day4

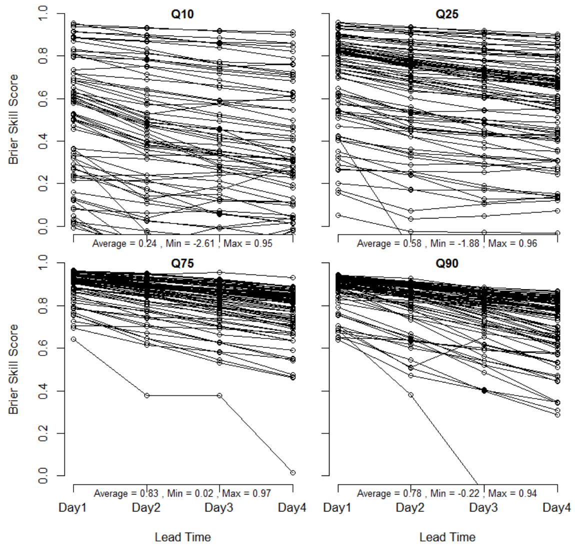


Figure 13: Brier Skill Scores of the ~~original~~ forecast-only QR ~~model~~configuration (i.e., using the transformed forecast as the only independent variable) for four lead times and percentiles of observed water levels.

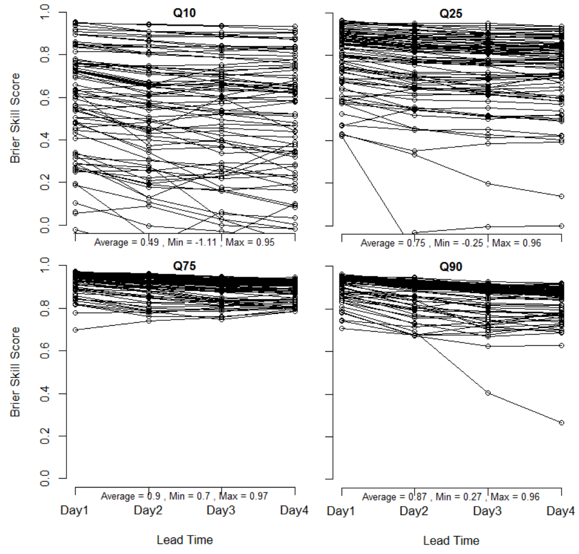


Figure 14: Brier Skill Scores for four lead times and percentiles of observed water levels using the best ~~variable combination~~joint predictor for each river gage as independent variables in the QR ~~model~~configuration.

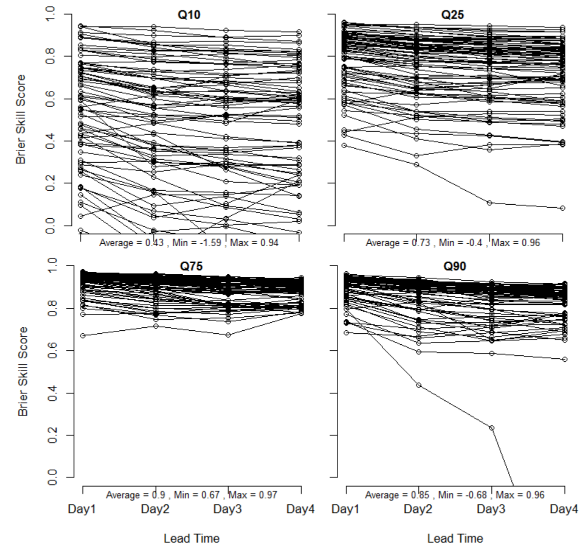
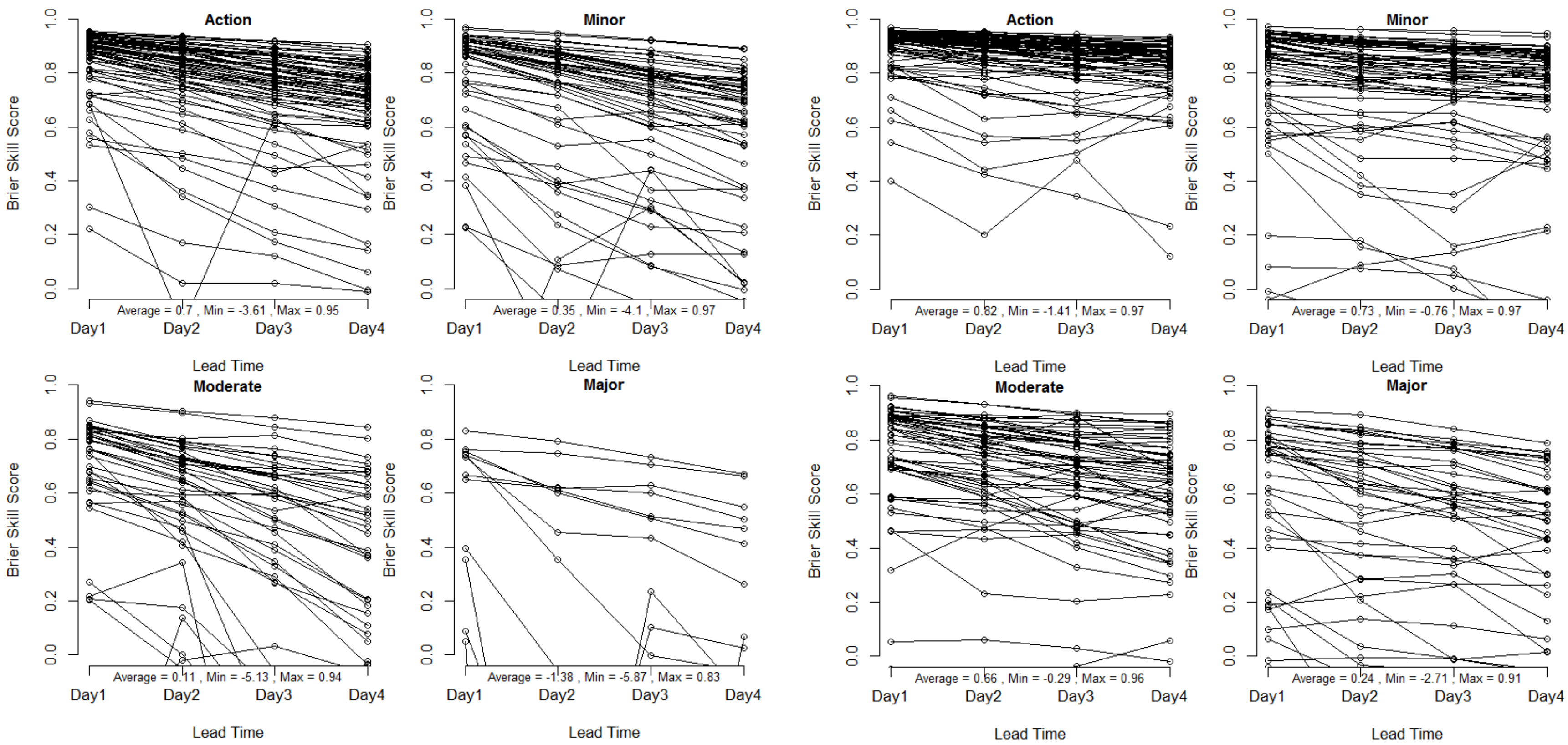


Figure 15: Brier Skill Scores for four lead times and percentiles of observed water levels using a one-size-fits-all approach (i.e., rr24, rr48, err24, err48) for the independent variables in the QR ~~model~~configuration.


Action
Brier Skill Score
0.0 0.2 0.4 0.6 0.8 1.0
Average = 0.7 , Min = -3.61 , Max = 0.95
Day1 Day2 Day3 Day4
Lead Time
Minor
Brier Skill Score
0.0 0.2 0.4 0.6 0.8 1.0
Average = 0.35 , Min = -4.1 , Max = 0.97
Day1 Day2 Day3 Day4
Lead Time
Action
Brier Skill Score
0.0 0.2 0.4 0.6 0.8 1.0
Average = 0.82 , Min = -1.41 , Max = 0.97
Day1 Day2 Day3 Day4
Lead Time
Minor
Brier Skill Score
0.0 0.2 0.4 0.6 0.8 1.0
Average = 0.73 , Min = -0.76 , Max = 0.97
Day1 Day2 Day3 Day4
Lead Time
Moderate
Brier Skill Score
0.0 0.2 0.4 0.6 0.8 1.0
Average = 0.11 , Min = -5.13 , Max = 0.94
Day1 Day2 Day3 Day4
Lead Time
Major
Brier Skill Score
0.0 0.2 0.4 0.6 0.8 1.0
Average = -1.38 , Min = -5.87 , Max = 0.83
Day1 Day2 Day3 Day4
Lead Time
Moderate
Brier Skill Score
0.0 0.2 0.4 0.6 0.8 1.0
Average = 0.66 , Min = -0.29 , Max = 0.96
Day1 Day2 Day3 Day4
Lead Time
Major
Brier Skill Score
0.0 0.2 0.4 0.6 0.8 1.0
Average = 0.24 , Min = -2.71 , Max = 0.91
Day1 Day2 Day3 Day4
Lead Time

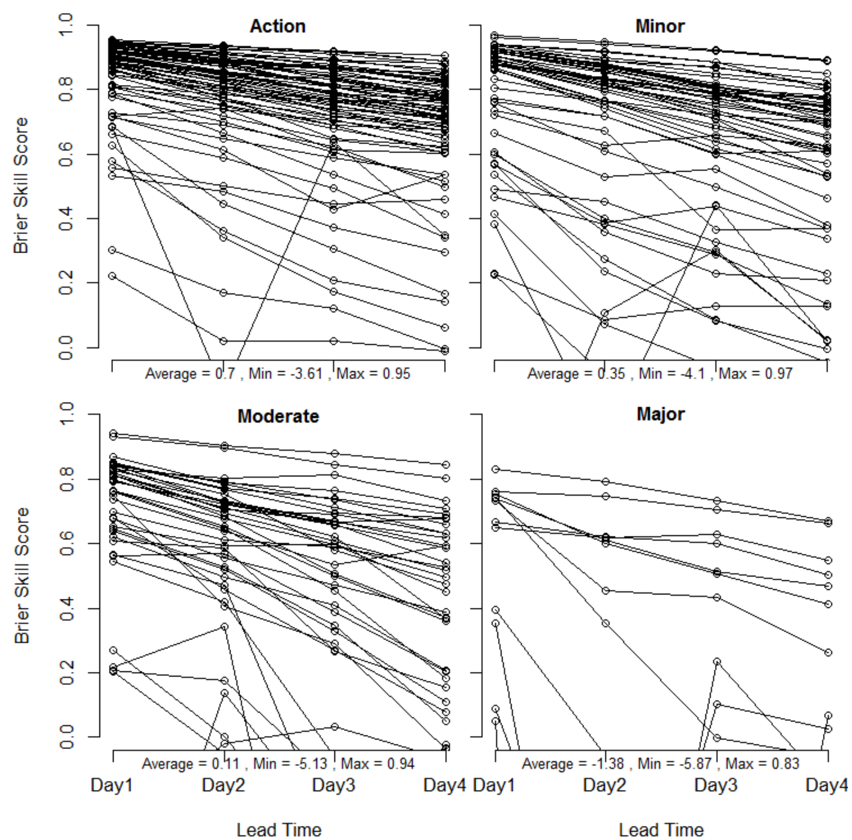


Figure 17~~6~~: Brier Skill Scores of the forecast-only ~~original~~ QR ~~model~~configuration (i.e., using the transformed forecast as the only independent variable) for four lead times and flood stages.

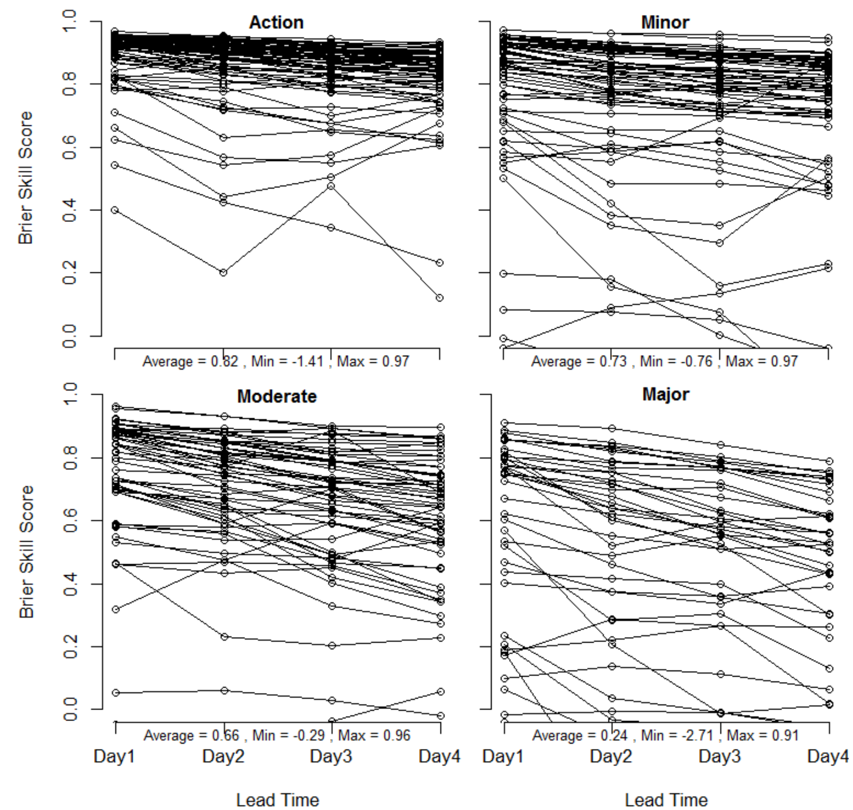


Figure 18~~7~~: Brier Skill Scores for four lead times and flood stages of observed water levels using the best ~~variable combination~~joint predictor for each river gage as independent variables in the QR ~~model~~configuration.

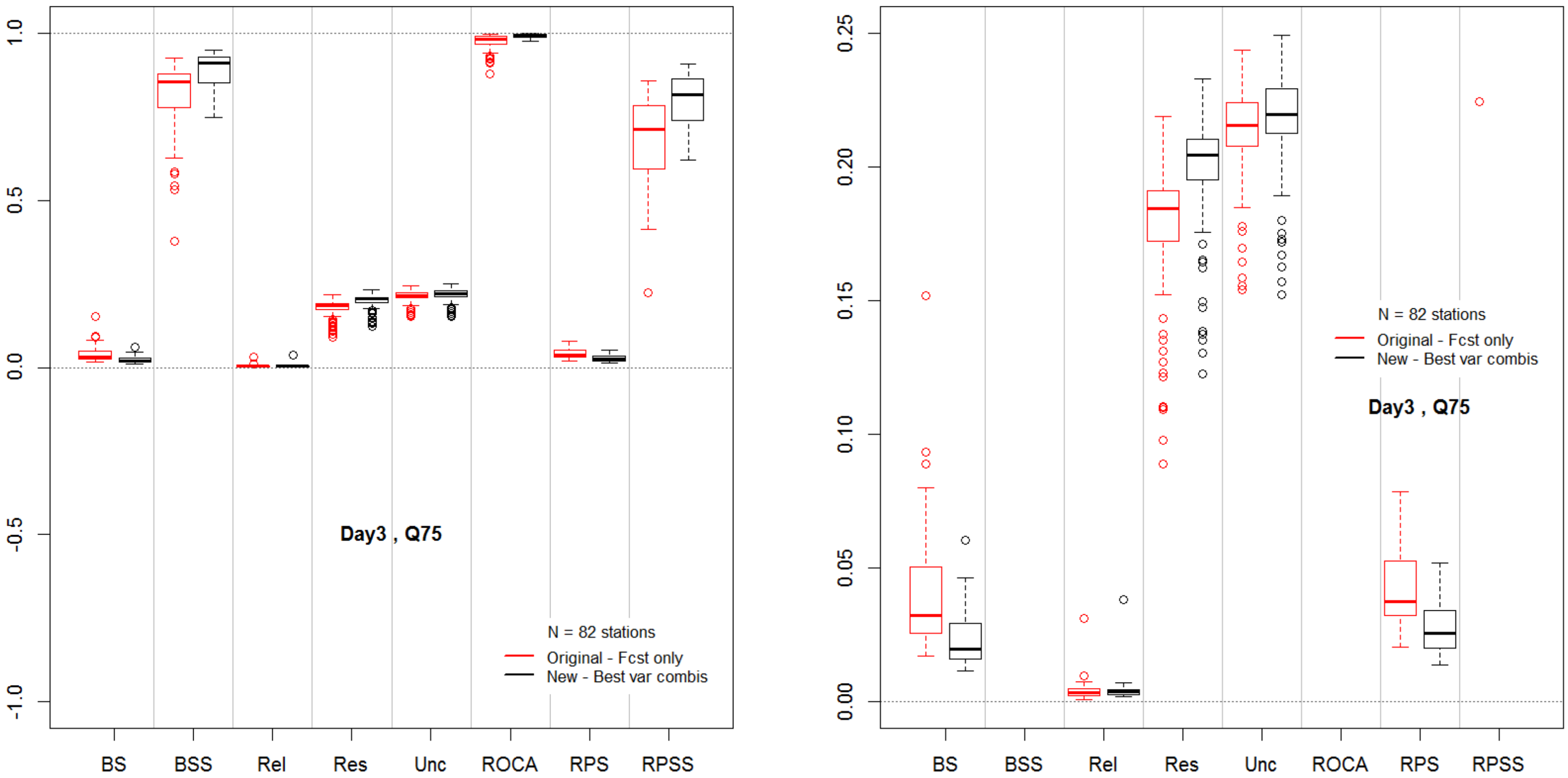
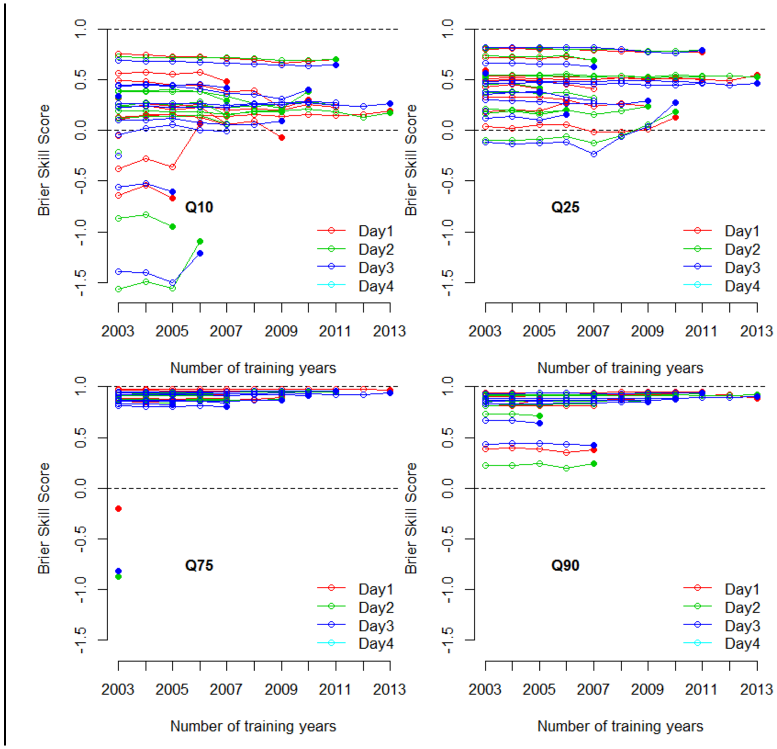
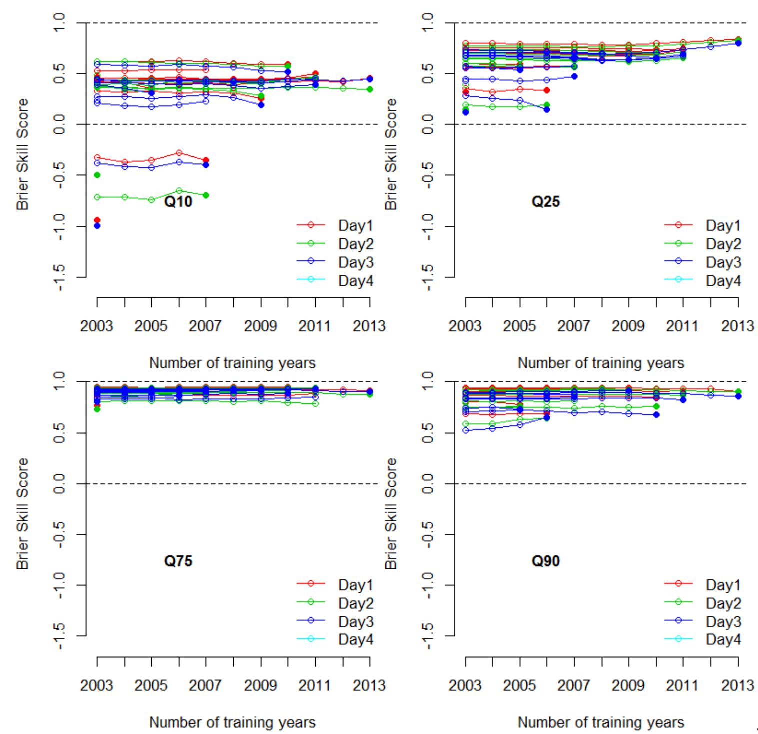


Figure 20~~18~~: Comparison of the forecast-only ~~original~~ QR ~~model~~configuration (i.e., only transformed forecast as independent variables) and the one-size-fits-all approach (i.e., ~~rise rates~~rates of rise and forecast errors as independent variables) using various measures of forecast quality: 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.


Brier Skill Score
Q10
Q25
Q75
Q90
Day1
Day2
Day3
Day4
1.0
0.5
0.0
-0.5
-1.0
-1.5
2003
2005
2007
2009
2011
2013
Number of training years


Brier Skill Score
Q10
Q25
Q75
Q90
Day1
Day2
Day3
Day4
1.0
0.5
0.0
-0.5
-1.0
-1.5
2003
2005
2007
2009
2011
2013
Number of training years

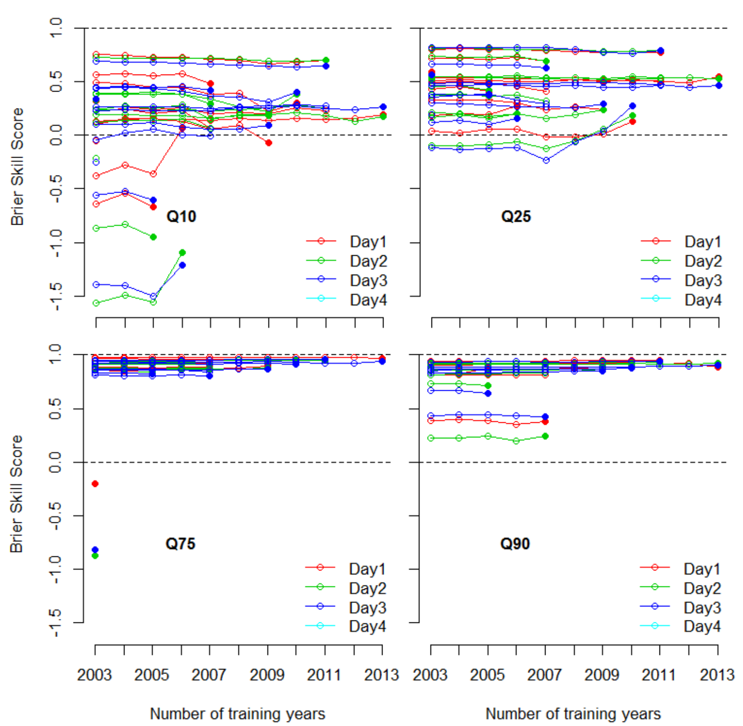


Figure 21~~19~~: Brier Skill Score for various forecast years and various sizes of training dataset across different lead times (colors) and event thresholds (plots) for Hardin, IL (HARI2). The filled-in end point of each line indicates the BSS for the forecast year on the x-axis with one year in the training dataset. Each point further to the left stands for one additional training year for that same forecast year.

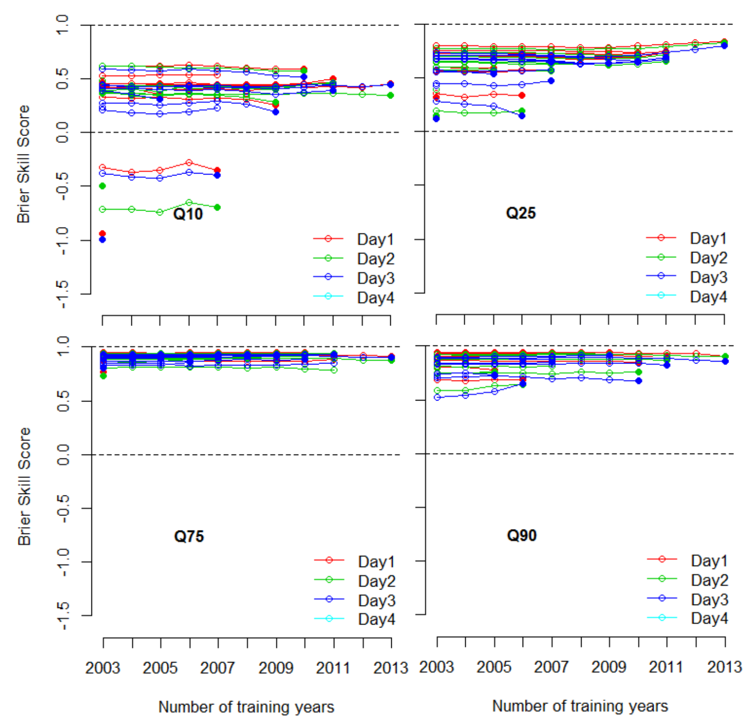
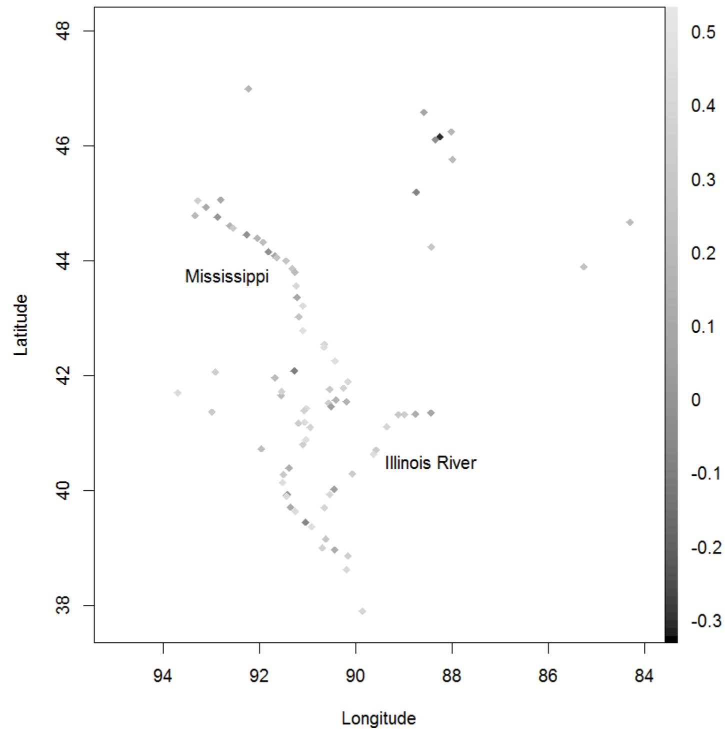
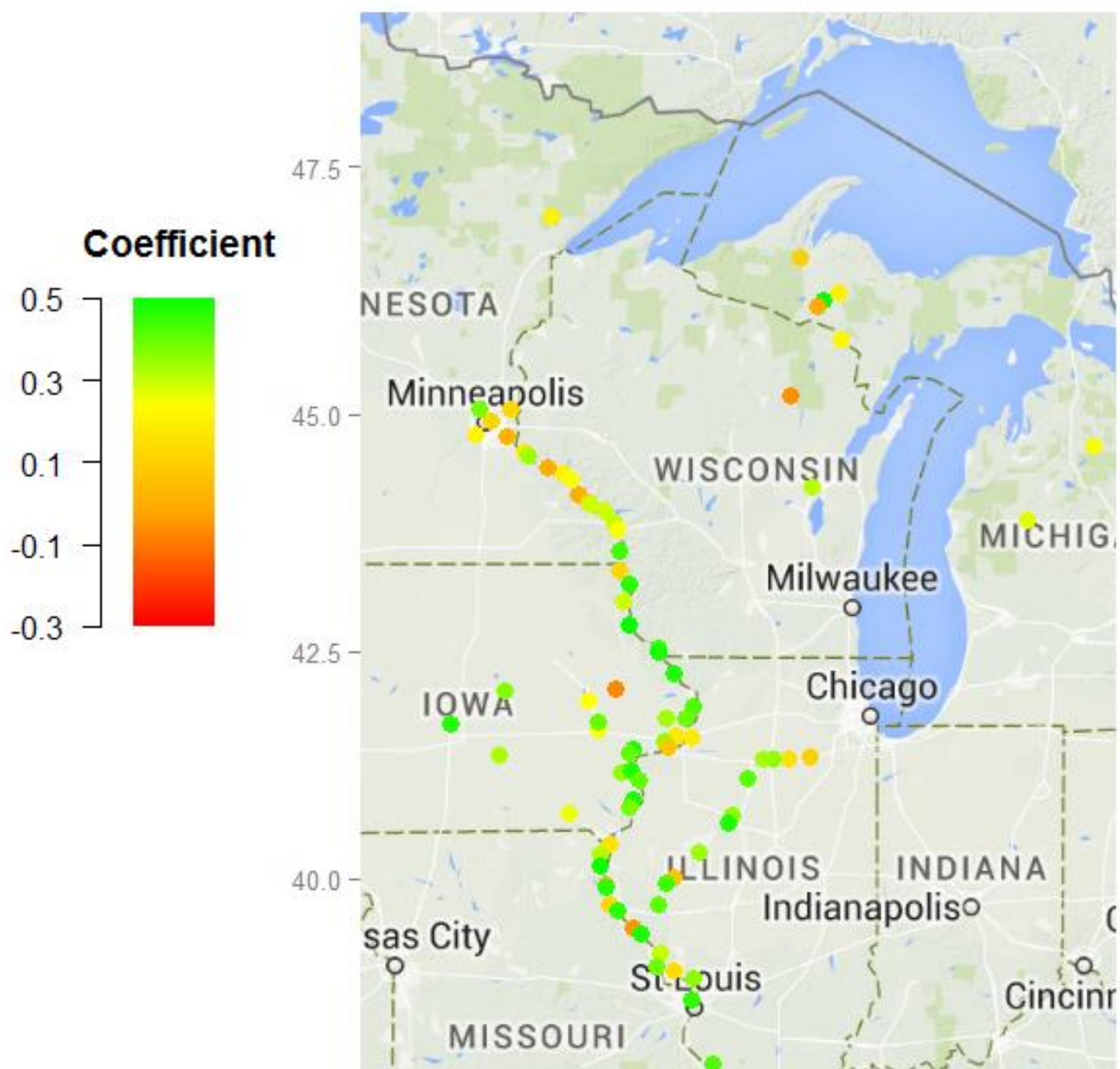


Figure 22~~0~~: Brier Skill Score for various forecast years and various sizes of training dataset across different lead times (colors) and event thresholds (plots) for Henry, IL (HNYI2). The filled-in end point of each line indicates the BSS for the forecast year on the x-axis with one year in the training dataset. Each point further to the left stands for one additional training year for that same forecast year.

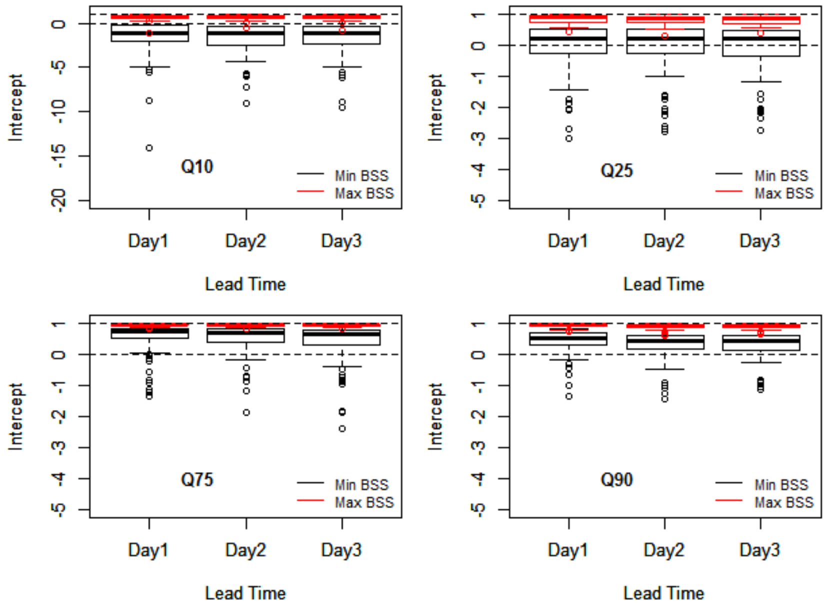


Latitude
Longitude
Mississippi
Illinois River
48
46
44
42
40
38
94
92
90
88
86
84
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3



Coefficient
0.5
0.3
0.1
-0.1
-0.3
47.5
45.0
42.5
40.0
NESOTA
Minneapolis
WISCONSIN
MICHIG
Milwaukee
Chicago
IOWA
ILLINOIS
INDIANA
Indianapolis
sas City
St Louis
Cincinn
MISSOURI

Figure 23~~1~~: Geographical position of rivers. Colors indicate the regression coefficient of each station with the Brier Skill Score as dependent variable.



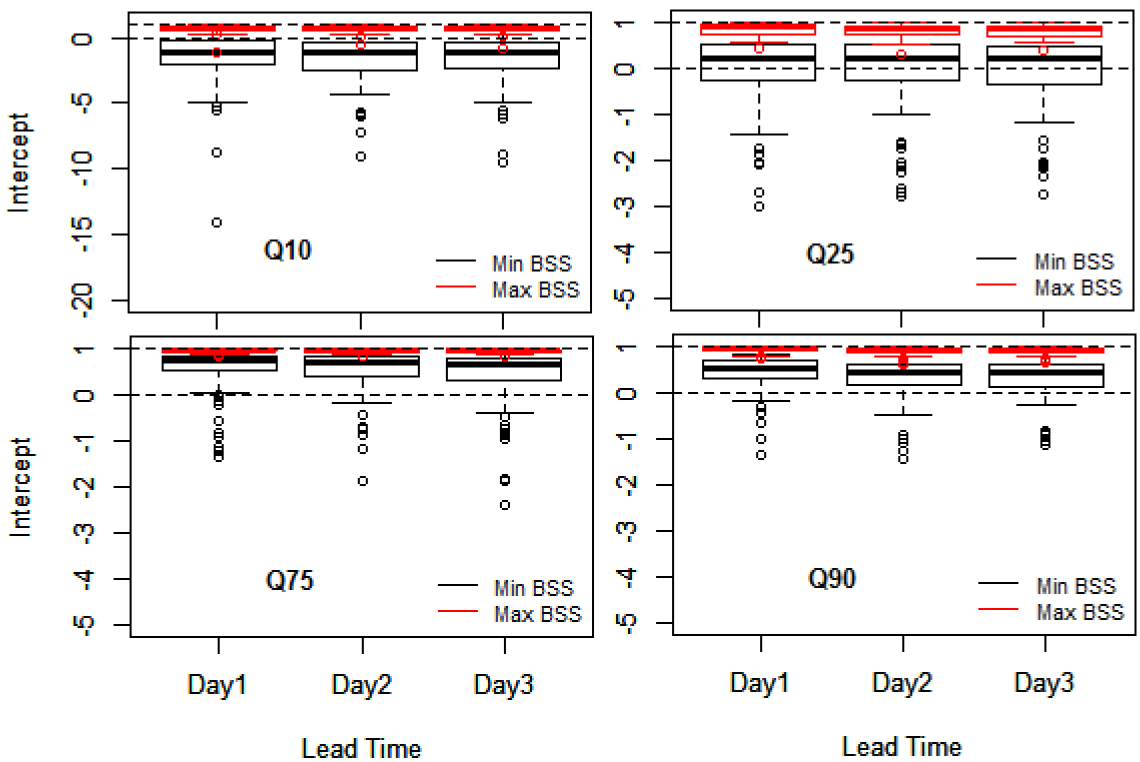


Figure 242: Minimum (black) and maximum (red) Brier Skill Scores for various lead times and event thresholds across locations, size of training dataset and forecast years.