

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

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

4 **Abstract**

5 This study applies quantile regression (QR) to the prediction of flood stage exceedance
6 probabilities based on post-processing single-value flood stage forecasts. A computationally
7 cheap technique to predict forecast errors is valuable, because many national flood forecasting
8 services, such as the National Weather Service (NWS), only publish deterministic single-value
9 forecasts. The study uses data from the 82 river gages, for which the NWS' North Central River
10 Forecast Center issues forecasts daily. Archived forecasts for lead times up to six days from
11 2001-2013 were analyzed. Besides the forecast itself, this study uses the rate of rise of the river
12 stage in the last 24 and 48 hours and the forecast error 24 and 48 hours ago as predictors in QR
13 configurations. When compared to just using the forecast as independent variable, adding the
14 latter four predictors significantly improved the forecasts, as measured by the Brier Skill Score
15 (BSS). Mainly, the resolution increases, as the forecast-only QR configuration already delivered
16 high reliability. Combining the forecast with the other four predictors results in much less
17 favorable BSSs. Lastly, the forecast performance does not strongly depend on the size of the
18 training dataset, but on the year, the river gage, lead time and event threshold that are being
19 forecast. We find that each event threshold requires a separate configuration or at least
20 calibration.

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

22

23 **1 Introduction**

24 River-stage forecasts are no crystal ball; the future remains uncertain. The past has shown that
25 unfortunate decisions have been made in ignorance of the potential forecast errors (Pielke, 1999;
26 Morss, 2010). For many users, such as emergency managers, forecasts are most important in
27 extreme situations, such as droughts and floods. Unfortunately, it is exactly in those situations
28 that forecast errors are largest, due to the infrequency of extreme events and the subsequent
29 scarcity of data. Additionally, users might only experience such an event once or twice in their
30 lifetime, so that they have no experience to what extent they can rely on forecasts in such
31 situations. Given the many sources and complexity of uncertainty and the lacking user
32 experience, it is easy to see how forecast users find it difficult to estimate the forecast error.
33 Including uncertainty in river forecast would therefore be valuable, just as has been
34 recommended for weather forecasts in general (e.g., National Research Council, 2006).

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

45 The National Weather Service has chosen to quantify the most significant sources of
46 uncertainty using ensemble techniques (Demargne et al., 2013). Currently, the National Weather
47 Service does not routinely publish uncertainty information along with their short-term river-stage
48 forecast (Figure 1).

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

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

52 The NWS has developed the Hydrologic Ensemble Forecast Service (HEFS) to be able to
53 provide short-term and medium-term probabilistic forecasts (Demargne et al., 2013). HEFS
54 includes two types of post-processors. The Hydrologic Model Output Statistics (HMOS)
55 Streamflow Ensemble Processor – which is also a module in NWS’ main forecast tool, the
56 Community Hydrologic Prediction System (CHPS) – corrects bias and evaluates the uncertainty
57 of each ensemble, while Hydrologic Ensemble Post-Processing (EnsPost) corrects bias and
58 lumps the set of ensembles into one uncertainty estimate (Demargne et al., 2013; Seo, 2008).
59 HMOS performs a similar task as the QR approach presented here, but with two major
60 differences. First, it relies on linear regression based on streamflows at various times as
61 predictor, instead of using QR with several types of independent variables. Second, it does not
62 compute distributions of water levels from which confidence intervals or exceedance
63 probabilities of flood stages can be derived, but generates ensembles (Regonda et al., 2013).

64 In contrast to an ensemble approach such as HEFS, the statistical post-processing in this
65 paper does not distinguish between sources of uncertainty, but studies the overall uncertainty in a
66 lumped fashion. To make this approach useful for actors with limited resources, we exclusively
67 use publicly available data to define our configurations.

68 Most previously developed post-processors to generate probabilistic forecasts share the
69 overall set-up but differ in their implementation. Independent variables such as the forecasted
70 and observed river stage, river flow or precipitation, and previous forecast errors are used to
71 predict the forecast error, conditional probability distribution of the forecast error or other
72 measures of uncertainty for various lead times (e.g., Kelly and Krzysztofowicz, 1997; Montanari
73 and Brath, 2004; Montanari and Grossi, 2008; Regonda et al., 2013; Seo et al., 2006; Solomatine
74 and Shrestha, 2009; Weerts et al., 2011). These techniques differ in a number of ways, including
75 their sub-setting of data, and the output. Please see Regonda et al. (2013) and Solomatine &
76 Shrestha (2009) for a summary of each technique. In a meta-analysis of four different post-
77 processing techniques to generate confidence intervals, the quantile regression technique was one
78 of the two most reliable techniques (Solomatine and Shrestha, 2009), while being the
79 mathematically least complicated and requiring few assumptions.

80 This paper further develops one of the techniques mentioned above: the Quantile
81 Regression approach to post-process river forecasts first introduced by Wood et al. (2009) and
82 further elaborated by Weerts et al. (2011) and López López et al. (2014). The Weerts study
83 achieved impressive results in estimating the 50% and 90% confidence interval of river-stage
84 forecasts for three case studies in England and Wales using QR with calibration and validation
85 datasets spanning two years each. This paper combines elements of the studies mentioned above.
86 In some aspects, our approach differs from those three studies. We predict the exceedance
87 probabilities of flood stages rather than uncertainty bounds. Additionally, we are fortunate to
88 have a much larger dataset than the three earlier studies consisting of archived forecasts for 82
89 river gages covering 11 years. The study does not add to the mathematical technique of quantile
90 regression itself.

91 In this paper, the QR technique is applied to the 82 river gages of the North Central River
92 Forecast Center (NCRFC) encompassing (parts of) Illinois, Michigan, Wisconsin, Minnesota,
93 Indiana, North Dakota, Iowa, and Missouri.

94 Identifying the best-performing set of independent variables is central to this paper. All
95 possible combinations of the following predictors have been studied: forecast, the rate of rise of
96 water levels in past hours, and the past forecast errors . The performance of these joint predictors
97 has been measured and compared using the Brier Skill Score (BSS). This exercise has been
98 repeated for various water levels and lead times. Additionally, the robustness of the resulting QR
99 configurations across different sizes of training datasets, locations, lead times, water levels, and
100 forecast year has been assessed.

101 The paper is structured as follows. The Method section reviews quantile regression,
102 introduces the performance measure, and discusses the performed analyses and data. The Results
103 section first reviews the overall forecast error for the dataset. It then describes the results of
104 identifying the best-performing set of independent variables. Finally, it discusses the robustness
105 of the studied QR configurations. The fourth and last section presents the conclusions and
106 proposes further research ideas.

107 **2 Method**

108 The use of quantile regression to estimate the error distribution of river-stage forecasts has first
109 been introduced by Woods et al. (2009) for the Lewis River in Washington State. Later, Weerts
110 et al. (2011) applied it to river catchments in England and Wales. In this paper, elements of both
111 studies are combined. However, our predictand is the probability of exceeding flood stages rather
112 than confidence bounds. Additionally, this study tests the robustness of the technique across

113 locations, lead times, event thresholds, forecast years, and the size of training dataset is tested.
114 To develop the different QR configurations and to compare their performance, the Brier Skill
115 Score (BSS) is used.

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

118 **2.1 Quantile Regression**

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

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

127 Quantile Regression was first introduced by Koenker (2005; 1978). It is different from
128 ordinary least square regression in that it predicts percentiles rather than the mean of a dataset.
129 Koenker and Machado (Koenker and Machado, 1999, p.1305) and Alexander et al. (2011)
130 demonstrate that studying the coefficients and their uncertainty for different percentiles generates
131 new insights, especially for non-normally distributed data. For example, using quantile
132 regression to analyze the drivers of international economic growths, Koenker and Machado
133 (1999) find that benefits of improving the terms of trade show a monotonously increasing trend
134 across percentiles, thus benefitting faster-growing countries proportionally more.

135 When applying QR to river forecasts, Weerts et al. (2011) transformed the forecast values
 136 and the corresponding forecast errors into the Gaussian domain using Normal Quantile
 137 Transformation (NQT) to account for heteroscedasticity. Detailed instructions to perform NQT
 138 can be found in Bogner et al. (2012). Building on this study, López López et al. (2014) compare
 139 different configurations of QR with the forecast as the only independent variable, including
 140 configurations omitting NQT. They find that no configuration was consistently superior for a
 141 range of forecast quality measures (López López et al., 2014). To be able to combine predictors
 142 of different nature, we based our QR configuration on untransformed predictors. The reason to
 143 do so will be discussed and illustrated later (see Figure 11 and Figure 12).

144 A quantile regression is run for each lead time and desired percentile with the forecast error
 145 as the dependent variable and the forecast and other variables as independent variables. To
 146 prevent the quantile regression lines from crossing each other, a fixed effects model is
 147 implemented below a certain forecast value. Weerts et al. (2011) give a detailed mathematical
 148 description for applying QR to river forecasts. Mathematically, the approach is formulated as
 149 follows (with and without NQT):

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

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

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

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

154 with $F_{\tau}(t)$ – estimated forecast associated with percentile τ and time t
 155 $fcst(t)$ – original forecast at time t
 156 $V_i(t)$ – the independent variable i (e.g., the original forecast) at time t
 157 $V_{i:NQT}(t)$ – the independent variable I transformed by NQT at time t
 158 $a_{i,\tau}, b_{\tau}$ – configuration coefficients
 159

160 The second part of the equations stands for the error estimate based on the quantile regression
 161 configuration for each percentile τ and lead time. In Equation 1, that was used by Weerts et al.
 162 (2011), this estimation was executed in the Gaussian domain using only the forecast as
 163 independent variable. Our study mainly uses Equation 2, i.e., it does not transform the predictors
 164 and the predictand. All quantile regressions were done using the command *rq()* in the R-package
 165 “*quantreg*” (Koenker, 2013).

166 2.2 Brier Skill Score

167 The QR configuration by Weerts et al. (2011) was evaluated by determining the fraction of
 168 observations that fell into the confidence intervals predicted by the QR configuration; i.e.,
 169 ideally, 80% of the observations should be larger than the predicted 10th percentile for that day,
 170 and smaller than the predicted 90th percentile. López López et al. (2014) used a number of
 171 measures to assess configuration performance, e.g., the Brier Skill Score (BSS), the mean
 172 continuous ranked probability (skill) score (RPSS), the relative operating characteristic (ROC),
 173 and reliability diagrams to compare QR configurations.

174 We use the Brier Skill Score – first introduced by Brier (1950) – to assess QR
 175 configurations for two reasons. First, to be able to optimize model performance it is best to
 176 choose a single measure. Second, out of the available measures the Brier Score is attractive,
 177 because it can be decomposed into two different measures of forecast quality (see Equation 3):

178 Reliability and resolution. The third component is uncertainty, which is a hydrological
 179 characteristic inherent to the river gage. This uncertainty is different than the forecast uncertainty
 180 that the technique studied in this paper estimates. Besides the uncertainty that can be
 181 mathematically explained, it also includes natural variability. In sum, the BS' uncertainty term is
 182 not subject to the forecast quality. Equation 3 gives the definition of the (de-composed) Brier
 183 Score (e.g., Jolliffe and Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009).

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

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

185 with BS – Brier Score
 186 N – number of forecasts
 187 K – the number of bins for forecast probability of binary event occurring on each
 188 day
 189 n_k – the number of forecasts falling into each bin
 190 \bar{o}_k – the frequency of binary event occurring on days in which forecast falls into bin
 191 k
 192 f_k – forecast probability
 193 \bar{o} – frequency of binary event occurring
 194 f_t – forecast probability at time t
 195 o_t – observed event at time t (binary: 0 – event did not happen, 1 – event happened)

196 The Brier Score pertains to binary events, e.g., the exceedance of a certain river stage or
 197 flood stage. Reliability compares the estimated probability of such an event with its actual
 198 frequency. For example, perfect reliability means that on 60% of all days for which it was
 199 predicted that the water level would exceed flood stage with a 60% probability, it actually does
 200 so. The reliability curve for the forecast representing perfect reliability would follow the diagonal
 201 in Figure 2, i.e., the area in Figure 2a representing reliability would equal zero (Jolliffe and

202 Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009). The configuration by López López
203 et al. (2014) performs well in terms of reliability. When estimating confidence intervals, Weerts
204 et al. (2011) achieved good results especially for the more extreme percentiles (i.e., 10th and
205 90th).

206 **Figure 2: Theory behind Brier Skill Score illustrated for an imaginary forecast (red line): (a)**
207 **reliability and resolution; (b) skill. In figure a, the area representing reliability should be as small,**
208 **and for resolution as large as possible. The forecast has skill (BSS > 0), i.e., performs better than the**
209 **reference forecast, if it is inside the shaded area in the figure b. Ideally, the forecast would follow**
210 **the diagonal (BSS=1). (Adapted from Hsu and Murphy, 1986; Wilson, n.d.).**

211 Resolution measures the difference between the predicted probability of an event on a
212 given day and the observed average probability. When calculated for a time period longer than a
213 day, the forecast performs better if the resolution term is higher. For example, for a gage where
214 flood stage is exceeded on 5% of the days in a year, simply using the historical frequency as the
215 forecast would mean forecasting that the probability of the water level exceeding flood stage is
216 5% on any given day. The accumulated difference between the predicted frequency and the
217 historical average across a time period of several days would then be zero (e.g., Jolliffe and
218 Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009). In Figure 2, the curve for a forecast
219 with good resolution would be steeper than the dashed line that represents climatology, i.e., the
220 area in Figure 2a representing resolution would be maximized. In absolute terms, the resolution
221 can never exceed the third term in Equation 3 representing the uncertainty inherent to the river
222 gage. Through the resolution component, the Brier Score is related to the area under the relative
223 operating characteristic (ROC) curve (for more detail, see Ikeda et al., 2002). The latter likewise
224 quantifies how much better than the reference forecast (i.e., climatology) a forecast is in

225 detecting a binary event; though unlike the Brier Score it focuses on the ratios of false and
226 missed alarms (e.g., Jolliffe and Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009).

227 A forecast possesses skill, i.e., performs better than the reference forecast (in this case
228 climatology), if it is inside the shaded area in Figure 2b. The Brier *Skill* Score (BSS) equals the
229 Brier Score normalized by climatology to make the score comparable across gages with different
230 frequencies of a binary event. Equation 4 defines the BSS' decomposition into the resolution and
231 reliability components described above (Brown and Seo, 2013). The BSS can range from minus
232 infinity to one. A BSS below zero indicates no skill; the perfect score is one (e.g., Jolliffe and
233 Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009). All measures of forecast quality
234 were computed using the R-package "verification" (NCAR, 2014).

235 **Equation 4: Decomposition of Brier Skill Score**

$$236 \quad BSS = 1 - \frac{BS}{\bar{o}(1-\bar{o})} = \frac{RES}{\bar{o}(1-\bar{o})} - \frac{REL}{\bar{o}(1-\bar{o})}$$

237 with BSS – Brier Skill Score
238 BS – Brier Score
239 RES – Resolution
240 REL – Reliability
241 \bar{o} – Frequency of binary event occurring
242 $\bar{o}(1 - \bar{o})$ – Climatological variance

243 **2.3 Identifying the best-performing sets of independent variables**

244 The challenge is to identify a well-performing set of predictors that is both parsimonious and
245 comprehensive. Wood et al. (2009) found rate of rise and lead time to be informative
246 independent variables. Weerts et al. (2011) achieved good results using only the forecast itself as
247 predictor. Besides these variables, the most obvious predictors to include are the observed water
248 level 24 and 48 hours ago, the forecast error 24 and 48 hours ago (i.e., the difference between the

249 current water level at issue time of the forecast and the forecast that was produced 24/48 hours
250 ago), or the time of the year, e.g., using month or season as categorical predictors. Additional
251 potential independent variables are the water levels observed up- and downstream at various
252 times, the precipitation upstream of the catchment area, and the precipitation forecast. However,
253 requesting the corresponding precipitation and precipitation forecast requires an extensive effort
254 or direct access to the database at the National Climatic Data Center (NCDC).

255 In preliminary trials on two case studies (gages HARI2 and HYNI2), it was found that the
256 rates of rise and the forecast errors are better predictors than the water levels observed in
257 previous days. After all, the observed water levels are used to compute the rates of rise and
258 forecast errors, so that these latter variables include the information of the former variable. It was
259 also found that season and months are not significant in quantile regression configurations to
260 predict the quantiles of the forecast error. Probably, the time of the year is already reflected in
261 the observed water levels and forecast errors in the previous days.

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

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

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

271 with $F_{\tau}(t)$ – estimated forecast associated with percentile τ and time t
 272 $fcst(t)$ – original forecast at time t
 273 $rr24(t), rr48(t)$ – rates of rise in the last 24 and 48 hours at time t
 274 $err24(t), err48(t)$ – forecast errors 24 and 48 hours ago (e.g., the original forecast) at
 275 time t
 276 $a_{xx,\tau}, b_{\tau}$ – configuration coefficients; forced to be zero if the predictor is
 277 excluded from the joint predictor that is being studied.

278 **Table 1: Joint predictors.**

279 **2.4 Computations**

280 The output of our QR application to river forecasts is the probability that a certain water level in
 281 the river or flood stage is exceeded on a given day, e.g., “On the day after tomorrow, the
 282 probability that the river exceeds 15 feet at location X is 60%.” This is done in two steps. First, a
 283 training dataset (first half of the data) is used to define one quantile regression configuration for
 284 each percentile $\pi = [0.05, 0.1, 0.15, \dots, 0.85, 0.90, 0.95]$ and each lead time. The dependent
 285 variable is the water level. As described in Equation 5, the forecast itself, the rates of rise and
 286 forecast errors serve as independent variables.

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

291 Then, we calculate the probability with which various water levels (called event
292 thresholds hereafter) will be exceeded. The probability of exceeding each water level is
293 computed by linearly interpolating between the points of the discrete probability distribution that
294 was computed in the previous step.

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

299 To study whether the various combinations of predictors perform equally well for high
300 and low thresholds, these last computational steps (i.e., interpolating to determine the exceedance
301 probability for a certain water level and calculating the BSS) were done for the 10th, 25th, 75th,
302 and 90th percentile of observed water levels and the four decision-relevant flood stages (action
303 stage, and minor, moderate, and major flood stage) of each gage. Flood stages indicated when
304 material damage or substantial hinder is caused by high water levels. Therefore, the flood stages
305 correspond with different percentiles at different river gages.

306 To determine the best-performing set of independent variables, the entire procedure is
307 repeated for each of the 31 joint predictors in Table 1, thus using a different set of independent
308 variables each time. To test the robustness of this approach, the procedure was also repeated for
309 each river gage and for several lead times. The result is 31 BSSs for 82 river gages for four
310 different lead times (one to four days) and for eight event thresholds (i.e., flood stages or
311 percentiles of the observed water level).

312 **2.5 Data**

313 The National Weather Service (NWS)'s daily short-term river forecasts predict the stage height
314 in six-hour intervals for up to five days ahead (20 6-hour intervals). When floods occur and
315 increased information is needed, the local river forecast center (RFC) can decide to publish river-
316 stage forecasts more frequently and for more locations. Welles et al. (2007) provides a detailed
317 description of the forecasting process.

318 For this paper, all forecasts published by the North Central River Forecast Center
319 (NCRFC) between 1 May 2001 and 31 December 2013 were requested from the NCDC's HDSS
320 Access System (National Climatic Data Center, 2014; Station ID: KMSR, Bulletin ID: FGUS5).
321 In total, the NCRFC produces forecasts for 525 gages. For 82 of those gages, forecasts have been
322 published daily for a sufficient number of years, and are not inflow forecasts. The latter have
323 been excluded from the forecast error analysis because they forecast discharge rather than water
324 level. About half of the analyzed gages are along the Mississippi River (Figure 3). The Illinois
325 River and the Des Moines River are two other prominent rivers in the region. The drainage areas
326 of the 82 river gages average 61,500 square miles (minimum 200 sq.miles; maximum 708,600
327 sq.miles). Figure 4 shows an empirical cumulative density function of drainage areas sizes.

328 **Figure 3: River gages for which the North Central River Forecast Centers publishes forecasts daily.**
329 **Henry (HYN12) and Hardin (HARI2) are indicated by the upper and lower red arrow respectively.**
330 **For gages indicated by black dots the basin size is missing.**

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

333 Two river gages serve as an illustration for the points made throughout this paper.
334 Hardin, IL is just upstream of the confluence of the Illinois River and the Mississippi River

335 (Figure 3). Therefore, it probably experiences high water levels through backwatering, when the
336 high water levels in the Mississippi River prevent the Illinois River from draining. Henry, IL is
337 located ~200 miles upstream of Hardin, having a difference in elevation of ~25 feet. The Illinois
338 River is ~330 miles long (Illinois Department of Natural Resources, 2011), draining an area of
339 ~13,500 square miles at Henry (USGS, 2015a) and ~28,700 square miles at Hardin (USGS,
340 2015b).

341 **3 Results**

342 **3.1 Forecast error at NCRFC's gages**

343 In general, the NCRFC's forecasts are well calibrated across the entire dataset. The average
344 error, defined as observation minus the forecast, is zero for most gages. For lead times longer
345 than three days, a slight underestimation by the forecast is noticeable. By a lead time of 6 days
346 this underestimation averages 0.41 feet only (Figure 5a, Figure 6). Extremely low water levels,
347 defined as below the 10th percentile of observed water levels, are also well calibrated (Figure 5b,
348 Figure 6). However, when considering higher water levels the picture changes. The
349 underestimation becomes more pronounced, averaging 0.29 feet for three days of lead time and
350 1.14 feet for six days of lead time, when only observations exceeding the 90th percentile of all
351 observations are considered (Figure 5c, Figure 6). When only looking at observations that
352 exceeded the minor flood stages corresponding to each gage, the underestimation averages 0.45
353 feet for three days of lead time and 1.51 feet for 6 days of lead time (Figure 5d, Figure 6).
354 However, some gages, such as Morris (MORI2), Marseilles Lock/Dam (MMOI2) – both on the
355 Illinois River – and Marshall Town on the Iowa River (MIWI4) experience *average* errors of 5 to
356 12 feet for water levels higher than minor flood stage. The gages MORI2 and MMOI2 are

357 upstream of a dam. It is likely that the forecasts performed so poorly there, because the dam
358 operators deviated from the schedules that they provide the river forecast centers to base their
359 calculations on.

360 **Figure 5: Forecast error for 82 river gages that the NCRFC publishes daily forecasts for. In anti-**
361 **clockwise direction starting at the top left: (a) Average error; (b) error on days that the water level**
362 **did not exceed the 10th percentile of observations; (c) error on days that the water level exceeded the**
363 **90th percentile of observations; (d) error on days that the water level exceeded minor flood stage.**

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

366 3.2 Identifying the best-performing sets of independent variables

367 In total, the Brier Skill Score (BSS) for 31 joint predictors (Table 1) across various lead times
368 and event threshold have been compared. Across 82 river gages, it has been analyzed (a) which
369 combinations perform best and worst most often, and (b) which joint predictor delivers the best
370 BSSs on average.

371 3.2.1 Frequency Analysis

372 For the four lead time (i.e., one to four days) and the eight event thresholds (i.e., 10th, 25th, 75th,
373 90th percentiles as well as the four flood stages), we counted at how many river gages each joint
374 predictor resulted in the highest and the lowest BSS. Figure 7 shows that for water levels below
375 the 50th percentile joint predictors with four or more independent variables return the best BSSs
376 most often, while those with one and two predictors perform worst most often. For thresholds
377 higher than the 50th percentile the distributions gradually become flatter. For the 90th percentile, a
378 clear trend is no longer detectable. Given that the frequency distributions for the extreme events
379 in Figure 7 are relatively uniform, it seems as if extreme events are characterized by different

380 processes at different gages. The same set of histograms for the four flood stages (i.e., action,
381 minor, moderate, and major) confirms this (Figure 8). Across lead times, there is a slight trend
382 noticeable that single predictors tend to be the worst combination more often for longer lead
383 times. This suggests that the further out one is forecasting, the more important it becomes to
384 include more data in the configuration.

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

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

393 **3.2.2 Best performing combinations on average**

394 For each river gage, the combinations have been ranked by BSSs. It was found that the more
395 independent variables are included in a joint predictor, the higher that set of predictors will rank
396 on average (Figure 9). However, for extremely high water levels, this trend gradually reverses
397 (Figure 10). For action stage and minor flood stage, a slightly increasing trend is still visible. For
398 moderate and major flood stage, combinations with fewer independent variables rank higher on
399 average. The most likely explanation is that extreme events like major and moderate flood stage
400 are infrequent. After all, major flood stage equals 90th to 100th percentiles at the various gages.
401 This data scarcity can lead to overfitting when using more predictors.

402 Considering these findings and those of the frequency analysis earlier, the configurations
403 for the various river gages can generally be based on the same joint predictor of four or more

404 independent variables. But for extremely high water levels, a configuration specific to each river
405 gage has to be built in order to achieve high BSSs.

406 The combinations including the forecast (indicated by gray vertical lines in Figure 9 and
407 Figure 10) perform less well than those that exclude it. Plotting the independent variables against
408 the forecast error as the dependent variable makes the reason visible (Figure 11, Figure 12).
409 Without a transformation into the normal domain, the scatterplot of forecast and forecast error
410 does not show a trend. After NQT, the percentiles show trends laid out like a fan. In contrast, the
411 other four predictors become uniform distributions after NQT transformation. There is no trend
412 detectable anymore. Further research is necessary to reconcile these two types of predictors. A
413 possible solution could be to define QR configurations for subsets of the transformed dependent
414 and independent variable.

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

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

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

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

425 **3.2.3 Brier Skill Score**

426 Figure 13 illustrates the BSS when using the forecast as the only predictor as studied by Weerts
427 et al. (2011). Confirming Wood et al.'s findings (2009), additionally including the rate of rise

428 and forecasts errors as independent variables into the QR configuration improves the Brier Skill
429 Score (BSS) significantly . Using the best performing joint predictors gives an upper bound of
430 the BSSs that can be achieved at best. This configuration increases the mean and decreases the
431 standard deviation (Figure 14, Figure 16). The performance improves most where all
432 configurations perform worst: at the 10th percentile. Possibly, the configurations do not perform
433 well for low percentiles, because the dependent variable – the forecast error – exhibits very little
434 variance at those water levels, i.e., the average error is very small (Figure 16). The decrease of
435 the BSSs with lead time also becomes considerably less with this configuration. Additionally, a
436 one-size-fits-all approach was tested to investigate, whether customizing the QR configuration to
437 each river gage would be worth it. In this configuration, the rates of rise in the past 24 and 48
438 hours and the forecast errors 24 and 48 hours ago serve as the independent variables
439 (combination 30). It was found that this approach returns only slightly worse results than
440 working with the best performing configuration for each river gage deviation (Figure 15, Figure
441 16). Accordingly, the same joint predictor can be used for all river gages.

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

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

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

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

455 **Figure 16: Empirical cumulative density functions of three QR configurations predicting**
456 **exceedance probabilities of the 10th, 25th, 75th, and 90th percentile: the configuration using the**
457 **transformed forecast as the only independent variable [NQT fcst]; the best performing combination**
458 **for each river gage (upper performance limit) [Best combis]; rates of rise in the past 24 and 48**
459 **hours and the forecast errors 24 and 48 hours ago as independent variable (one-size-fits-all**
460 **solution) [rr+err24/48].**

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

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

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

469 The fact that the Brier Score can be de-composed into reliability, resolution and
470 uncertainty allows a closer look at which improvements are being achieved by including more
471 predictors than just the forecast. Figure 20 shows that the forecast-only QR configuration as
472 studied by Weerts et al. (2011) has high reliability (i.e., the reliability is close to zero). The Brier
473 Score and the Brier Skill Score mainly improve when using rates of rise and forecast errors as
474 independent variables, because the resolution increases. This confirms the finding by Wood et al.
475 (2009) that QR error models should be based on rate of rise (as well as lead time). The forecast

476 quality improves along other metrics as well, i.e., the areas under the ROC curves and the ranked
477 probability skill score (RPSS) increase. The first weighs missed alarms against false alarms and
478 has a perfect score equal to one. The latter is a version of the Brier Skill Score. While the Brier
479 Skill Score pertains to a binary event, the RPSS can take into account various event categories.
480 Its perfect score equals one (e.g., WWRP/WGNE, 2009).

481 **Figure 20: Comparison of the forecast-only QR configuration (i.e., only transformed forecast as**
482 **independent variables) and the one-size-fits-all approach (i.e., rates of rise and forecast errors as**
483 **independent variables) using various measures of forecast quality: Brier Score (BS), Brier Skill**
484 **Score (BSS), Reliability (Rel), Resolution (Res), Uncertainty (Unc), Area under the ROC curve**
485 **(ROCA), ranked probability score (RPS), ranked probability skill score (RPSS). Lead time: 3 days;**
486 **75th percentile of observation levels as threshold. The left figure zooms in on the right figure to**
487 **make changes in reliability and resolution better visible.**

488 3.3 Robustness

489 The impact of the length of the training dataset on the configuration's performance measured by
490 the Brier Skill Score (BSS) was assessed for the one-size-fits-all QR configuration (i.e., rates of
491 rise and forecast errors as independent variables for all gages) for Hardin and Henry on the
492 Illinois River. We were particularly interested in testing how many years of training data are
493 necessary to achieve satisfactory forecasting results. Each year between 2003 and 2013 was
494 forecast by QR configurations trained on however many years of archived forecasts were
495 available in that year, i.e., the forecasts for 2005 is produced by a model trained on less data than
496 those for 2013. Then, the BSS for that year (e.g., 2005 or 2013) was computed.

497 Figure 21 and Figure 22 show that training datasets shorter than three years result in very
498 low BSSs for low event thresholds (Q10) at Henry and Hardin. For the other event thresholds, it
499 barely matters for the BSS how many years are included in the training dataset. That is good

500 news, if stationarity cannot be assumed (Milly et al., 2008), a step-change in river regime has
501 occurred, or forecast data have not been archived in the past. In those cases, only short training
502 datasets are available. Only needing short time series to define a skillful QR configuration
503 implies that the configuration parameters can be updated regularly. This way, changing
504 relationships between predictors etc. can be taken into account.

505 **However, the BSS varies considerably for what year is being forecast.** The forecast
506 performance varies greatly, especially for the 10th and 25th percentile of observed water levels. It
507 is likely, that a very large dataset, including more infrequent events, would improve these results.
508 However, most river forecast centers only recently started archiving forecasts in a text-format, so
509 that even having ten years' worth of data is an exception. To illustrate that point, the National
510 Climatic Data Center has archived data from 2001 onwards available in their HDSS Access
511 System.

512 To generalize the result, the same analysis as just described for Hardin and Henry was
513 repeated for all 82 gages. Following that, a regression analysis was executed with the BSS score
514 as the dependent variable and the river gages and forecast years as factorial independent
515 variables and the lead time, event thresholds, and number of training years as numerical
516 independent variables (Table 2). The forecast performance was found to vary statistically
517 significantly across all those dimensions **except the number of training years.** This results in a
518 very wide range of Brier Skill Scores (Figure 22). Accordingly, for the user, it is particularly
519 difficult to know how much to trust a forecast, if the performance depends so much on context.
520 Likewise, this is case for the QR configuration based on the forecast only (not shown).

521 A closer look at the regression coefficients (Table 2) provides interesting insights. For
522 low event thresholds, the BSSs are much worse than for high thresholds. The QR configurations

523 might be performing less well for low event thresholds, because the variance in the dependent
524 variable – the forecast error – is smaller. After all, river forecasts have much smaller errors for
525 lower water levels. The illustrative cases of Henry and Hardin, described above, indicate that
526 using longer time series to predict exceedance probabilities of low event thresholds improves
527 forecast performance.

528 As expected, the BSSs slightly decrease with lead time. Regarding the forecast quality for
529 each forecast year, the regression is slightly biased. The earlier years are included less often in
530 the dataset with on average less years' worth of data in their training dataset, because, for
531 example, unlike for the year 2013, ten years of training data were not available for the year 2006.
532 Nonetheless, the regression indicates that 2008 was particularly difficult to forecast and 2012
533 relatively easy, i.e., they are associated with relatively low and high coefficients respectively
534 (Table 2).

535 The performance of the forecast additionally depends on the river gage. The coefficients
536 of the river gages, included as factors in the regression, have been excluded from Table 2 for the
537 sake of brevity. Instead, Figure 23 maps the geographic position of the river gages with the color
538 code indicating each gage's regression coefficient. The coefficients are lower, and therefore the
539 Brier Skill Scores are lower, for gages far upstream a river and those close to confluences. At
540 least for the gages at confluences, the QR model could **probably** be improved by including the
541 rise rates at the river gages on the other joining river into the regression.

542 **Figure21: Brier Skill Score for various forecast years and various sizes of training dataset across**
543 **different lead times (colors) and event thresholds (plots) for Hardin, IL (HARI2). The filled-in end**
544 **point of each line indicates the BSS for the forecast year on the x-axis with one year in the training**
545 **dataset. Each point further to the left stands for one additional training year for that same forecast**
546 **year.**

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

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

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

556 **4 Conclusion**

557 In this study, quantile regression (QR) has been applied to estimate the probability of the river
558 water level exceeding various event thresholds (i.e., 10th, 25th, 75th, 90th percentiles of observed
559 water levels as well as the four flood stages of each river gage). It further develops the
560 application of QR to estimating river forecast uncertainty (a) comparing different sets of
561 independent variables, (b) and testing the technique's robustness across locations, lead times,
562 event thresholds, forecast years and sizes of training dataset.

563 When compared to the configuration using only the forecast, it was found that including
564 rates of rise in the past 24 and 48 hours and the forecast errors of 24 and 48 hours ago as
565 independent variables improves the performance of the QR configuration, as measured by the
566 Brier Skill Score. This confirms Wood et al.'s (2009) finding that QR error models should be a
567 function of rate of rise **and lead time**. The configuration with the forecast as the only independent
568 variable, as studied by Weerts et al. (2011), produced estimates with high reliability. Including
569 the other four predictors mentioned above mainly increases the resolution.

570 For extremely high water levels, the combinations of independent variables that perform best
571 vary across stations. On those days, combinations of fewer independent variables perform better
572 than those that include more. The most likely explanation is that QR configurations based on
573 large joint predictors result in overfitting the data. In contrast to these extremely high event
574 thresholds, larger sets of predictors work better than smaller ones for non-extreme and low event
575 thresholds. Additionally, customizing the set of predictors to the event thresholds does not
576 improve the BSS much.

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

584 The studied QR configurations are relatively robust to the size of training dataset, which is
585 convenient if stationarity cannot be assumed (Milly et al., 2008), a step-change in the river
586 regime has occurred, or – as is the case for most river forecast centers – only recent forecast data
587 have been archived. However, the performance of the technique depends heavily on the river
588 gage, the lead time, event threshold and year that are being forecast. This results in a very wide
589 range of Brier Skill Scores. This means that the danger remains that forecast users make good
590 experiences with a forecast one year or at one location and assume it is equally reliable in other
591 locations and every year. As is the case with most other forecasts, an indication of forecast

592 uncertainty needs to be communicated alongside the exceedance probabilities generated by our
593 approach.

594 The studied QR configurations perform less well for longer lead times, for gages far
595 upstream a river or close to confluences, for low event thresholds and extremely high ones. The
596 QR configurations might be performing less well for low event thresholds, because the variance
597 in the dependent variable – the forecast error – is smaller. After all, river forecasts have much
598 smaller errors for lower water levels. In turn, for extremely high water levels, the scarcity of data
599 decreases the configuration’s performance.

600 *Future Work*

601 This technique can be further developed in several ways to achieve higher Brier Skill Scores and
602 more robustness. First, more independent variables can be added. Trials with a different
603 technique, classification trees, showed that the observed precipitation, the precipitation forecast
604 (i.e., POP – probability of precipitation) and the upstream water levels significantly improve
605 forecasting performance. Presumably, this is the case, because the forecast used in this study
606 includes the precipitation forecast for only the next 12 hours. However, currently, the
607 precipitation data and forecasts can only be requested in chunks of a month, three chunks per
608 day, from the NCDC’s HDSS Access System. For a period of 12 years, requesting such data for
609 several weather stations is obviously time-consuming; not least, because the geographical units
610 of the weather forecasts bulletins do not correspond with those of the river forecast bulletins.
611 Upstream water levels can easily be included after manually determining the upstream gage(s)
612 for each of the 82 NCRFC gages. To improve performance at gages close to river confluences,
613 the upstream water level of the gages on the joining river should be included as well.

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

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

621 Additionally, the studied technique would need to be verified for gages for which the
622 NCRFC does not publish daily forecasts. Ignorance of the uncertainty inherent in river forecasts
623 has had some of the most unfortunate impacts on decision-making in Grand Forks, ND and
624 Fargo, ND (Pielke, 1999; Morss, 2010). Both of those stages are discontinuously forecast
625 NCRFC gages.

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

630 *Acknowledgements:*

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

References

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

Bogner, K., Pappenberger, F. and Cloke, H. L.: Technical Note: The normal quantile transformation and its application in a flood forecasting system, *Hydrol. Earth Syst. Sci.*, 16(4), 1085–1094, doi:10.5194/hess-16-1085-2012, 2012.

Brier, G. W.: Verification of Forecasts Expressed in Terms of Probability, *Mon. Weather Rev.*, 78(1), 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2, 1950.

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

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H. D., Fresch, M., Schaake, J. and Zhu, Y.: The Science of NOAA's Operational Hydrologic Ensemble Forecast Service, *Bull. Am. Meteorol. Soc.*, 95(1), 79–98, doi:10.1175/BAMS-D-12-00081.1, 2013.

Hsu, W. and Murphy, A. H.: The attributes diagram A geometrical framework for assessing the quality of probability forecasts, *Int. J. Forecast.*, 2(3), 285–293, doi:10.1016/0169-2070(86)90048-8, 1986.

Ikeda, M., Ishigaki, T. and Yamauchi, K.: Relationship between Brier score and area under the binormal ROC curve, *Comput. Methods Programs Biomed.*, 67(3), 187–194, doi:10.1016/S0169-2607(01)00157-2, 2002.

Illinois Department of Natural Resources: Aquatic Illinois - Illinois Rivers and Lakes Fact Sheets, [online] Available from:

<http://dnr.state.il.us/education/aquatic/aquaticillinoisrivlakefactshts.pdf> (Accessed 3 February 2015), 2011.

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

Kelly, K. S. and Krzysztofowicz, R.: A bivariate meta-Gaussian density for use in hydrology, *Stoch. Hydrol. Hydraul.*, 11(1), 17–31, doi:10.1007/BF02428423, 1997.

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

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

Koenker, R. and Bassett, G.: Regression Quantiles, *Econometrica*, 46(1), 33, doi:10.2307/1913643, 1978.

Koenker, R. and Machado, J. A. F.: Goodness of Fit and Related Inference Processes for Quantile Regression, *J. Am. Stat. Assoc.*, 94(448), 1296–1310, doi:10.1080/01621459.1999.10473882, 1999.

Leahy, C. P.: Objective Assessment and Communication of Uncertainty in Flood Warnings., 2007.

López López, P., Verkade, J. S., Weerts, A. H. and Solomatine, D. P.: Alternative configurations of Quantile Regression for estimating predictive uncertainty in water level forecasts for the Upper Severn River: a comparison, *Hydrol. Earth Syst. Sci. Discuss.*, 11(4), 3811–3855, 2014.

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P. and Stouffer, R. J.: Stationarity Is Dead: Whither Water Management?, *Science*, 319(5863), 573–574, doi:10.1126/science.1151915, 2008.

Montanari, A. and Brath, A.: A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, *Water Resour. Res.*, 40(1), W01106, doi:10.1029/2003WR002540, 2004.

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

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

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

National Research Council: Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts, National Academies Press, Washington, 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.

Tables

Table 1: Joint predictors

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

fcst = forecast; rr24, rr48 = 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: 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			***
<i>For the sake of brevity, the 82 river gages included in the regression as factors are omitted here.</i>			
R²		0.26	
Adjusted R²		0.25	
P-Values: *** – <0.001; ** – 0.01; * – 0.05; . – 0.1			

Figures

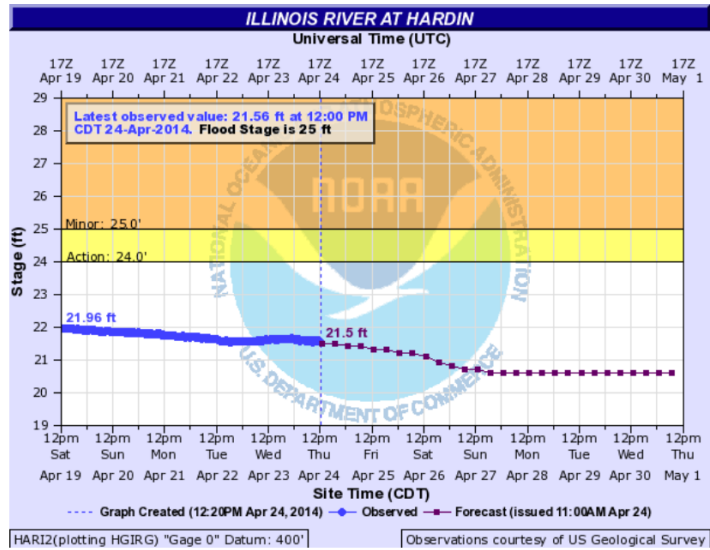


Figure 1: Deterministic short-term weather forecast in six hour intervals as published by the NWS for Hardin, IL on 24 April 2014.

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

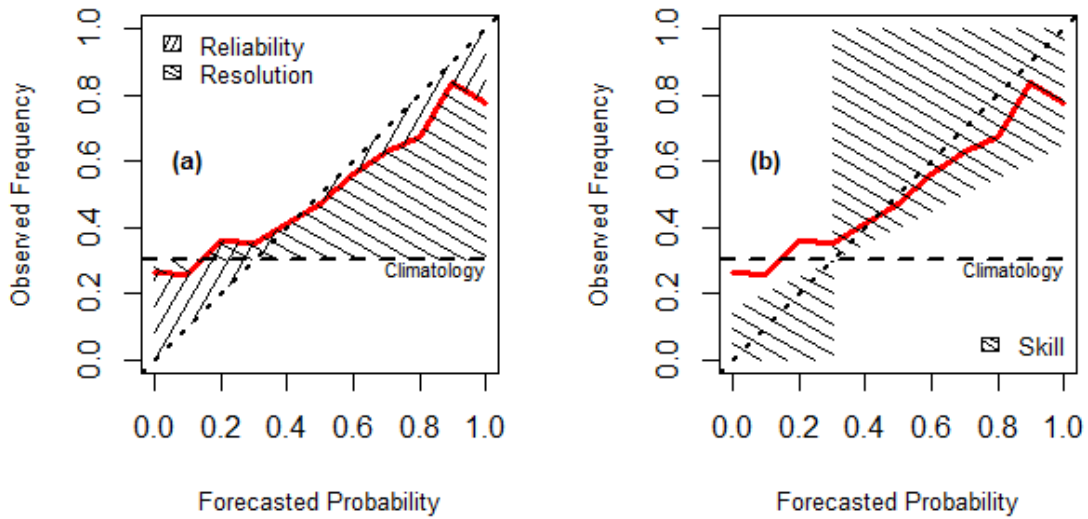


Figure 2: Theory behind Brier Skill Score illustrated for an imaginary forecast (red line): (a) reliability and resolution; (b) skill. In figure a, the area representing reliability should be as small, and for resolution as large as possible. The forecast has skill ($BSS > 0$), i.e., performs better 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.).

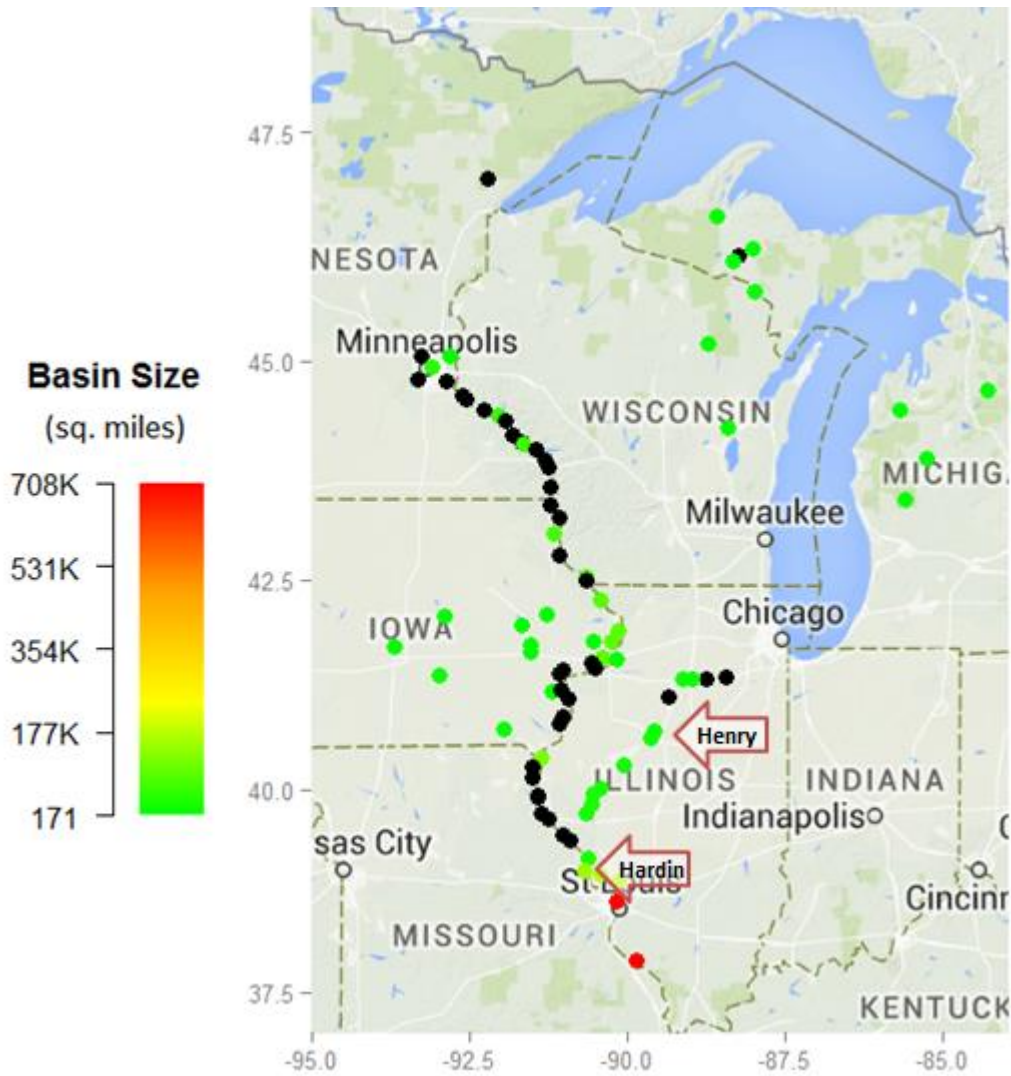


Figure 3: River gages for which the North Central River Forecast Centers publishes forecasts daily. Henry (HYNI2) and Hardin (HARI2) are indicated by the upper and lower red arrow respectively. For gages indicated by black dots the basin size is missing.

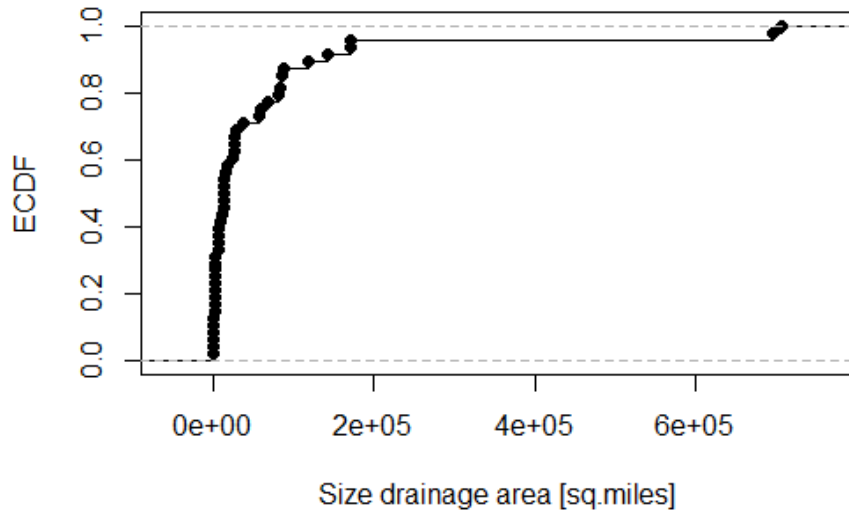


Figure 4: Empirical cumulative density function (ecdf) of sizes of drainage area for the river gages that are being forecasted daily by the NCRFC.

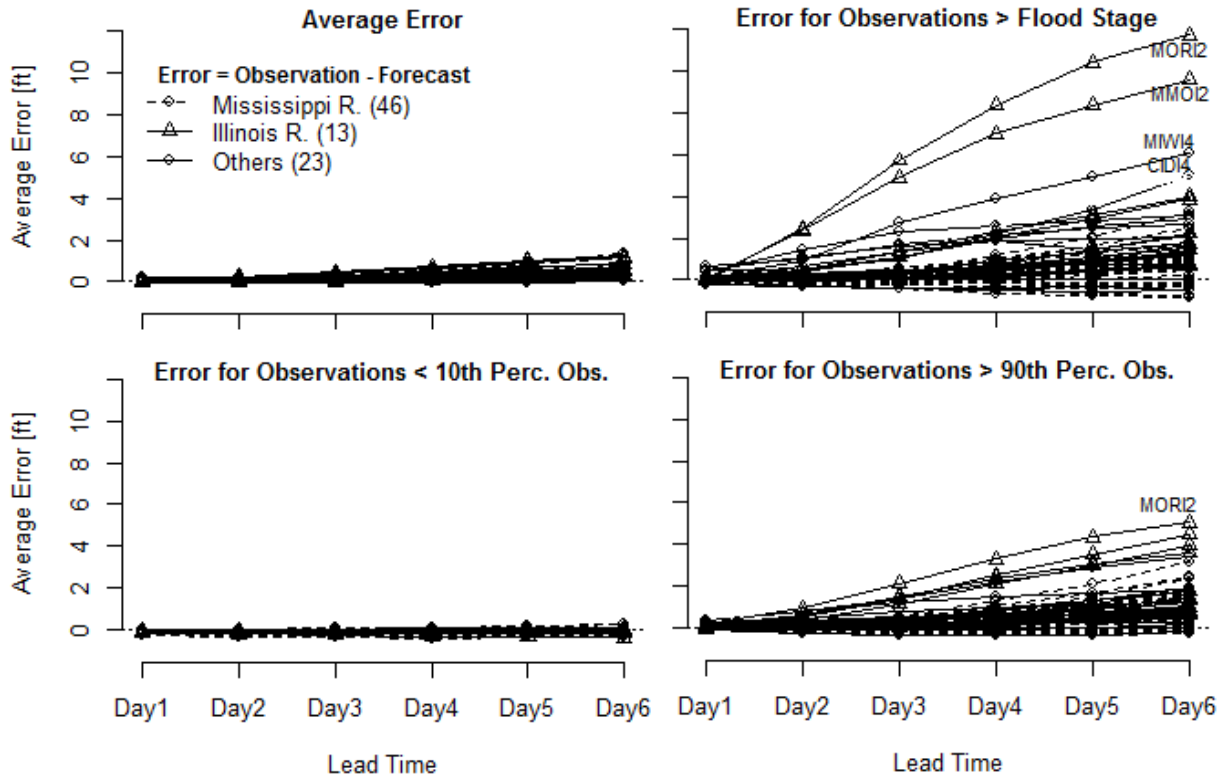


Figure 5: Forecast error for 82 river gages that the NCRFC publishes daily forecasts for. In anti-clockwise direction starting at the top left: (a) Average error; (b) error on days that the water level did not exceed the 10th percentile of observations; (c) error on days that the water level exceeded the 90th percentile of observations; (d) error on days that the water level exceeded minor flood stage.

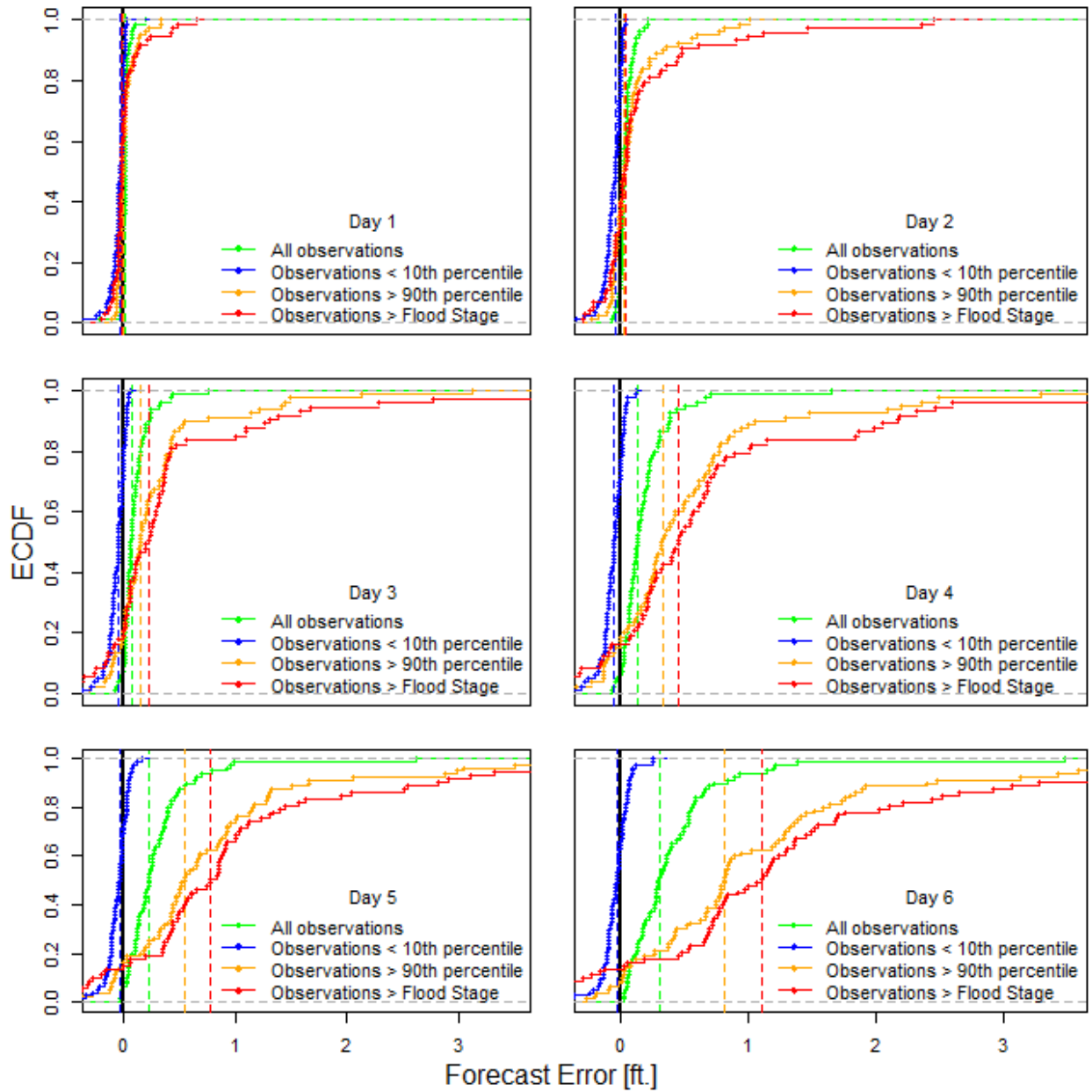


Figure 6: Empirical cumulative distribution function (ecdf) of forecast error at 82 river gages for six lead times. Vertical lines show the median forecast error of the corresponding subset.

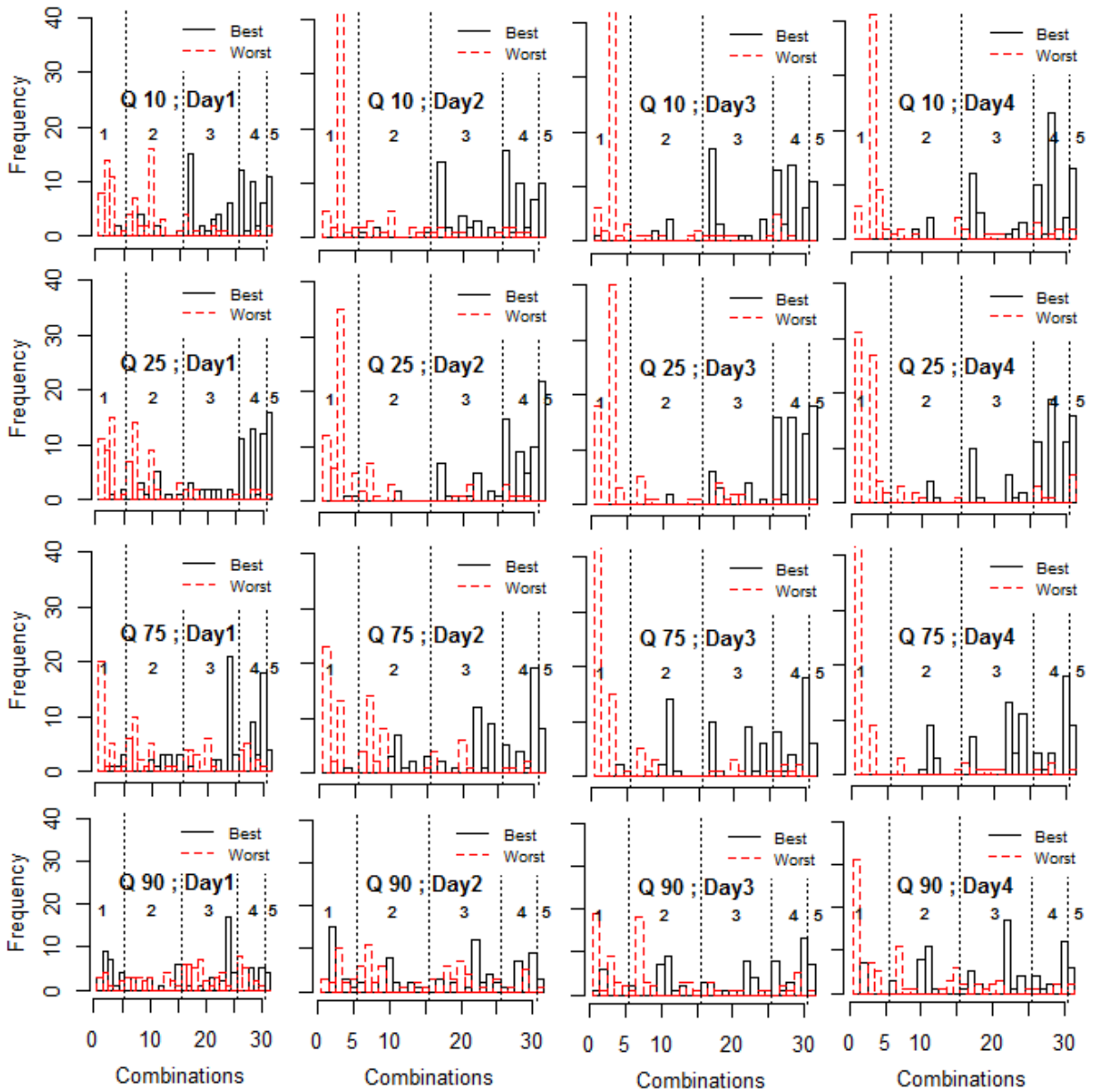


Figure 7: Histograms of joint predictors returning the best and worst Brier Skill Scores across 82 river gages. Each row of histograms refers to an event threshold defined as a percentile of the observed water levels, and each column to a lead time. The dotted vertical lines in the histograms distinguish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

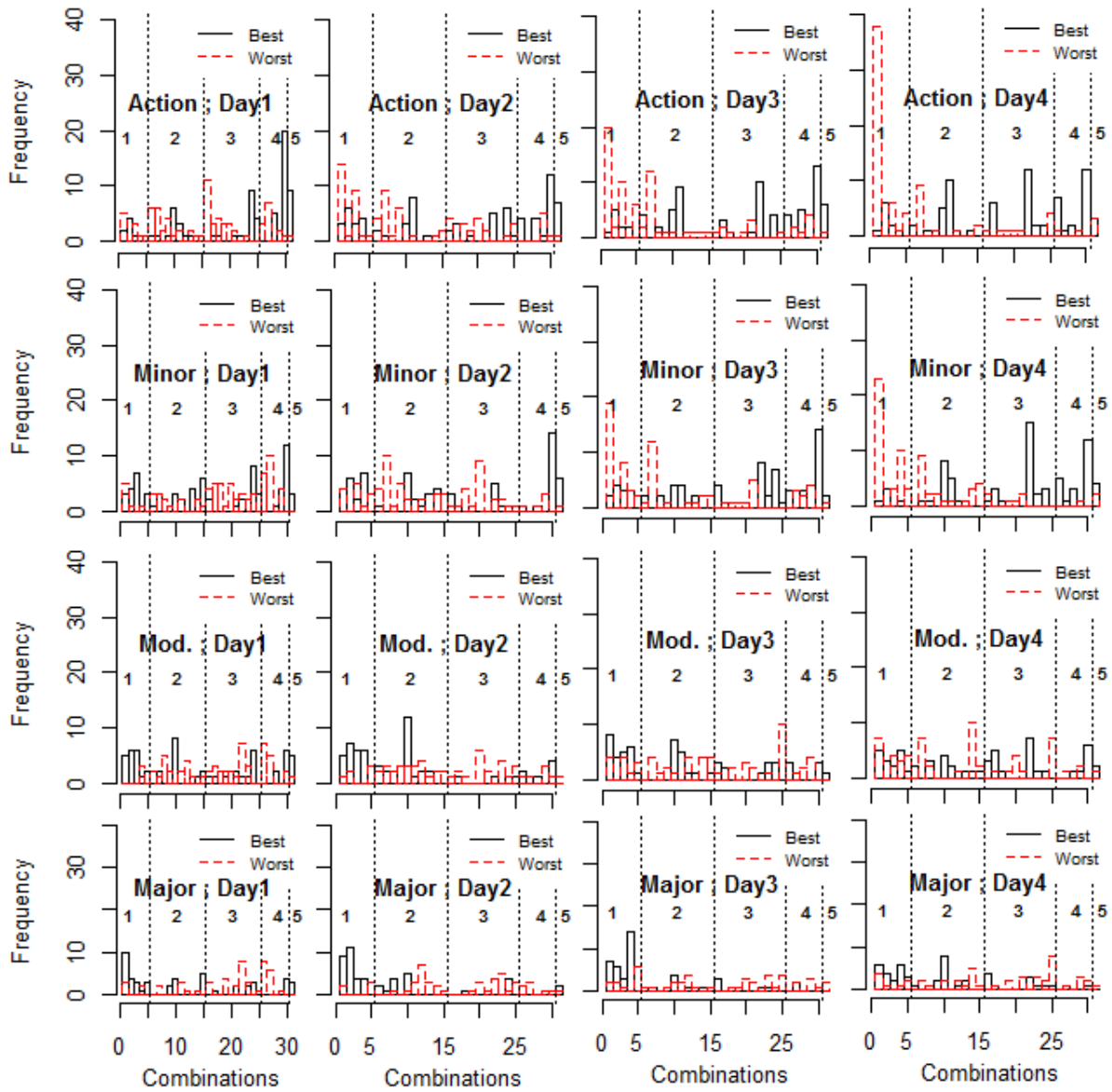


Figure 8: Histograms of joint predictors returning the best and worst Brier Skill Scores across 82 river gages. Each row of histograms refers to a flood stage, and each column to a lead time. The dotted vertical lines in the histograms distinguish joint predictors with different numbers of independent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

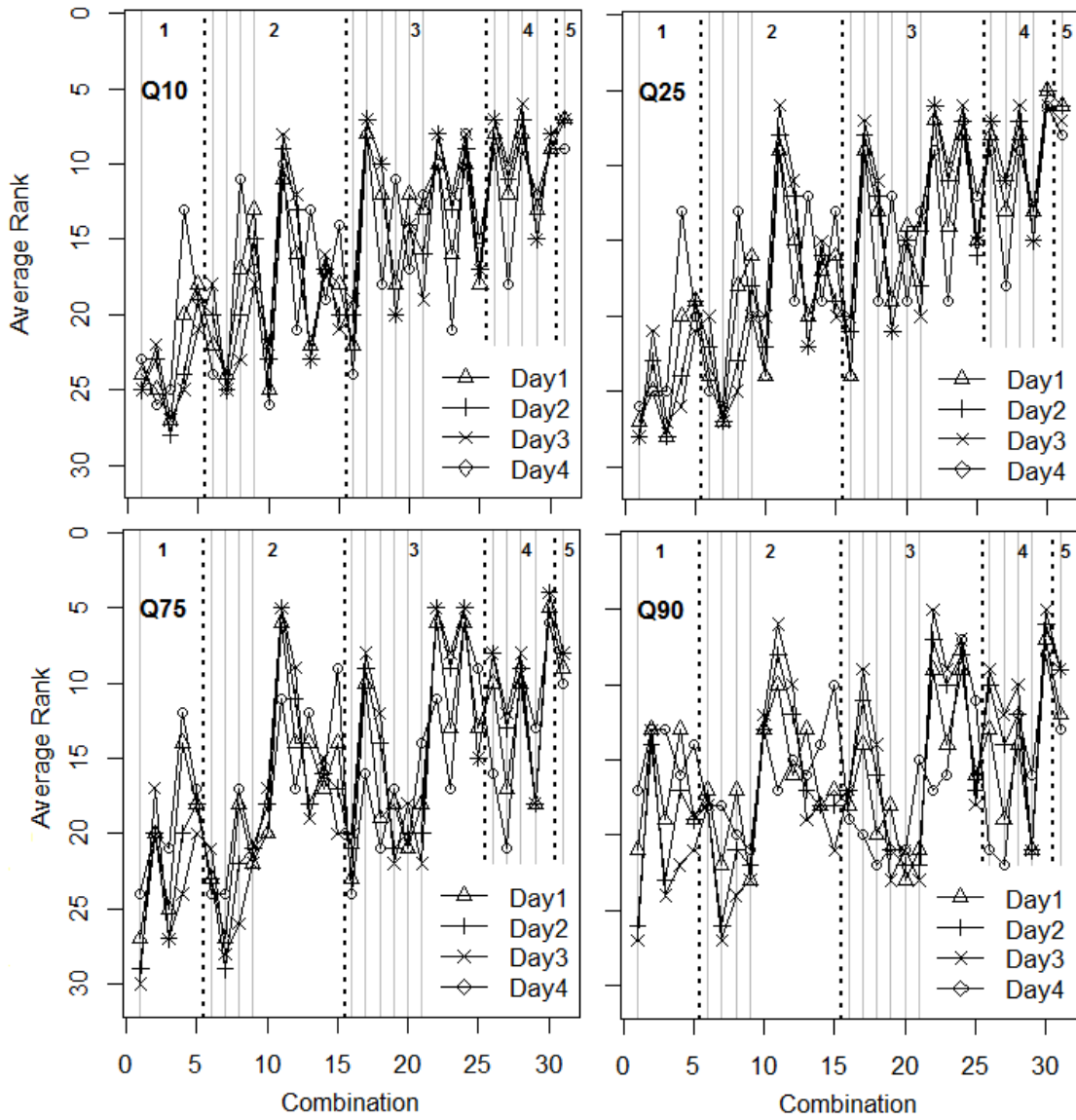


Figure 9: Average rank for each joint predictor for one to four days of lead time and four percentiles of observed water levels. Vertical gray lines indicate joint predictors including the forecast.

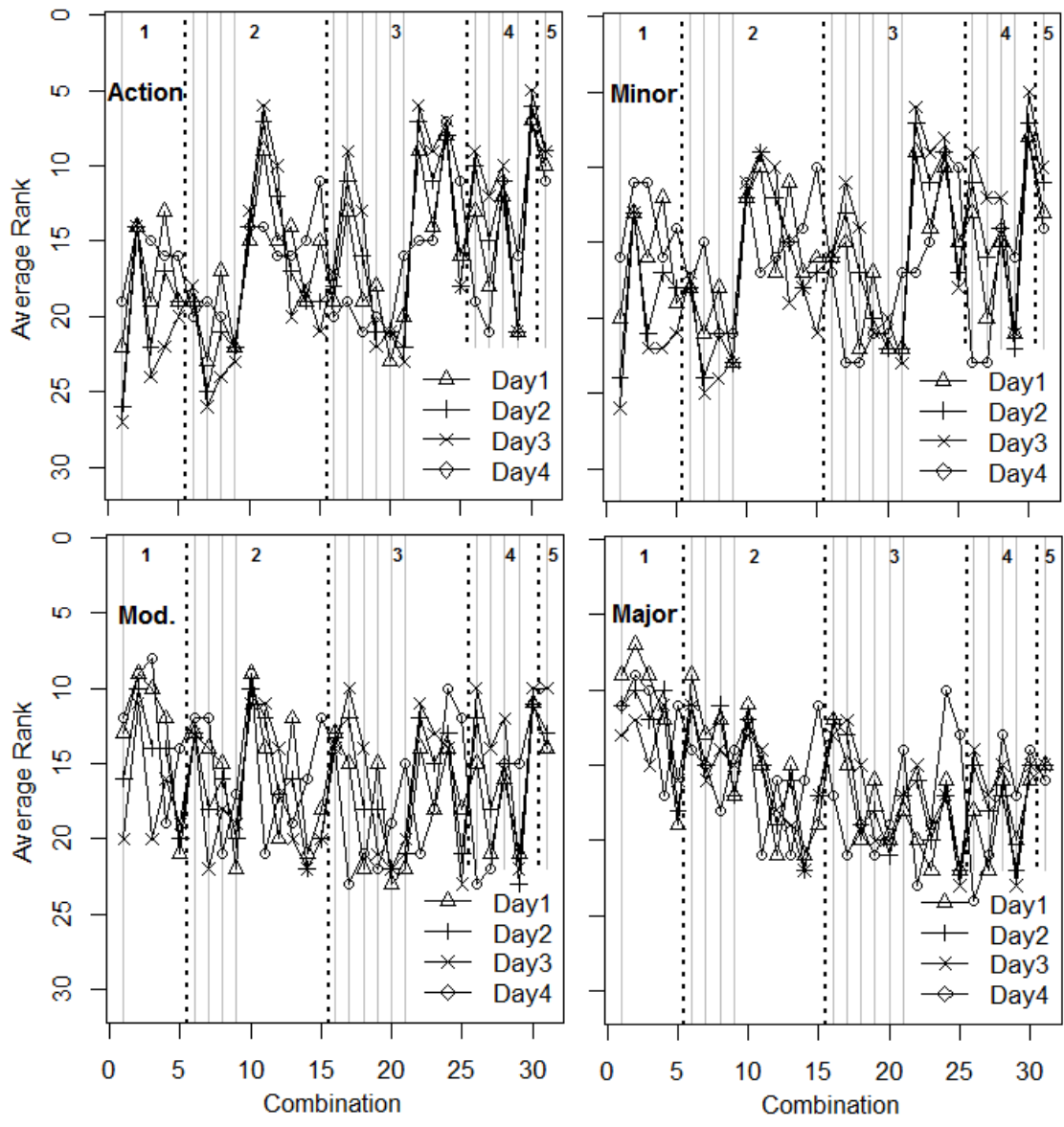


Figure 10: Average rank for each joint predictor for one to four days of lead time and four flood stages. Vertical gray lines indicate joint predictors including the forecast.

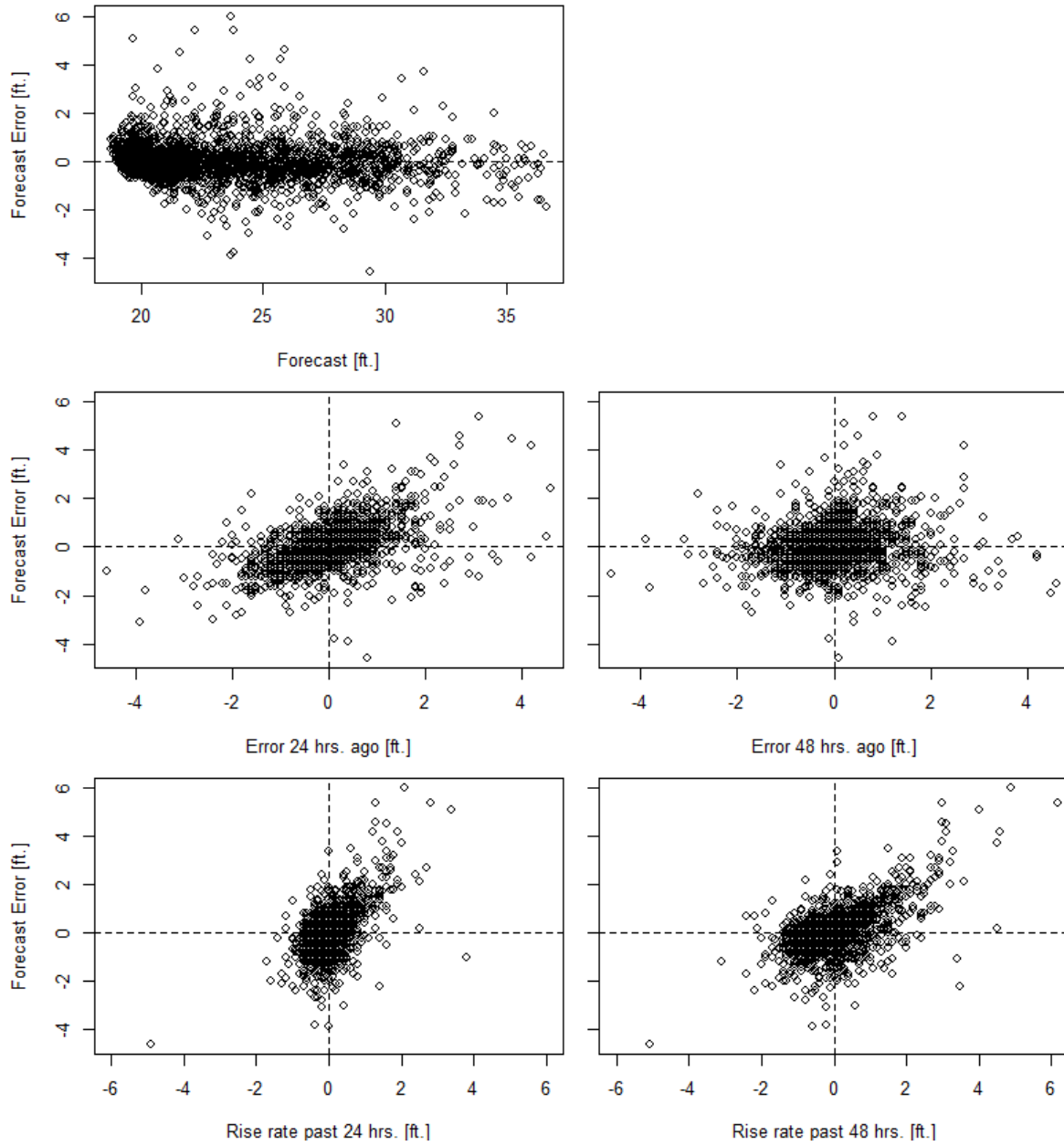


Figure 11: Independent variables plotted against the forecast error for Hardin IL with 3 days of lead time. First row: Forecast; second row: past forecast errors; third row: rates of rise.

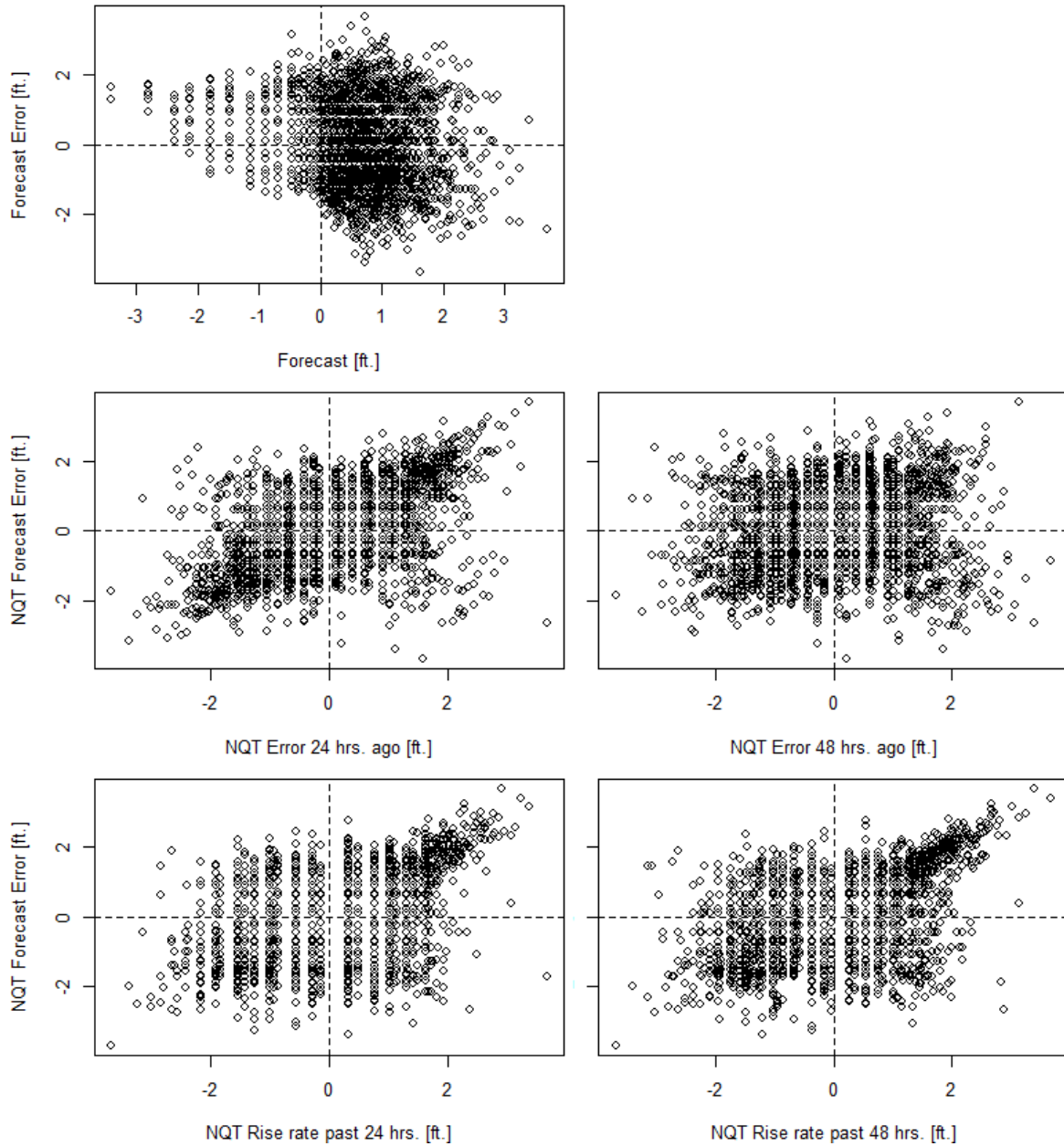


Figure 12: Independent variables after transforming into the Gaussian domain plotted against the forecast error for Hardin IL with 3 days of lead time. First row: Forecast; second row: past forecast errors; third row: rates of rise.

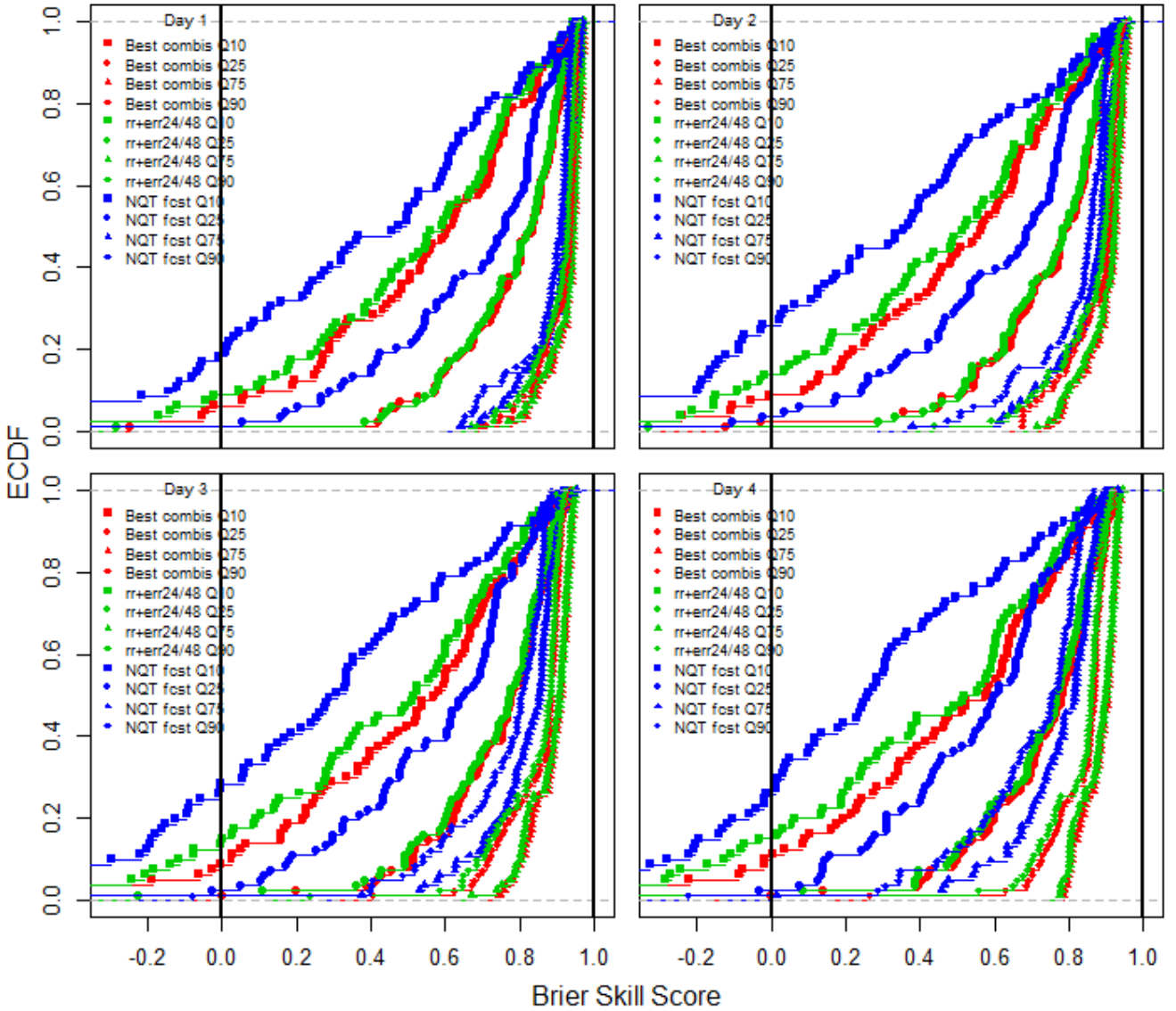


Figure 16: Empirical cumulative density functions of three QR configurations predicting exceedance probabilities of the 10th, 25th, 75th, and 90th percentile: the configuration using the transformed forecast as the only independent variable [NQT fcst]; the best performing combination for each river gage (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].

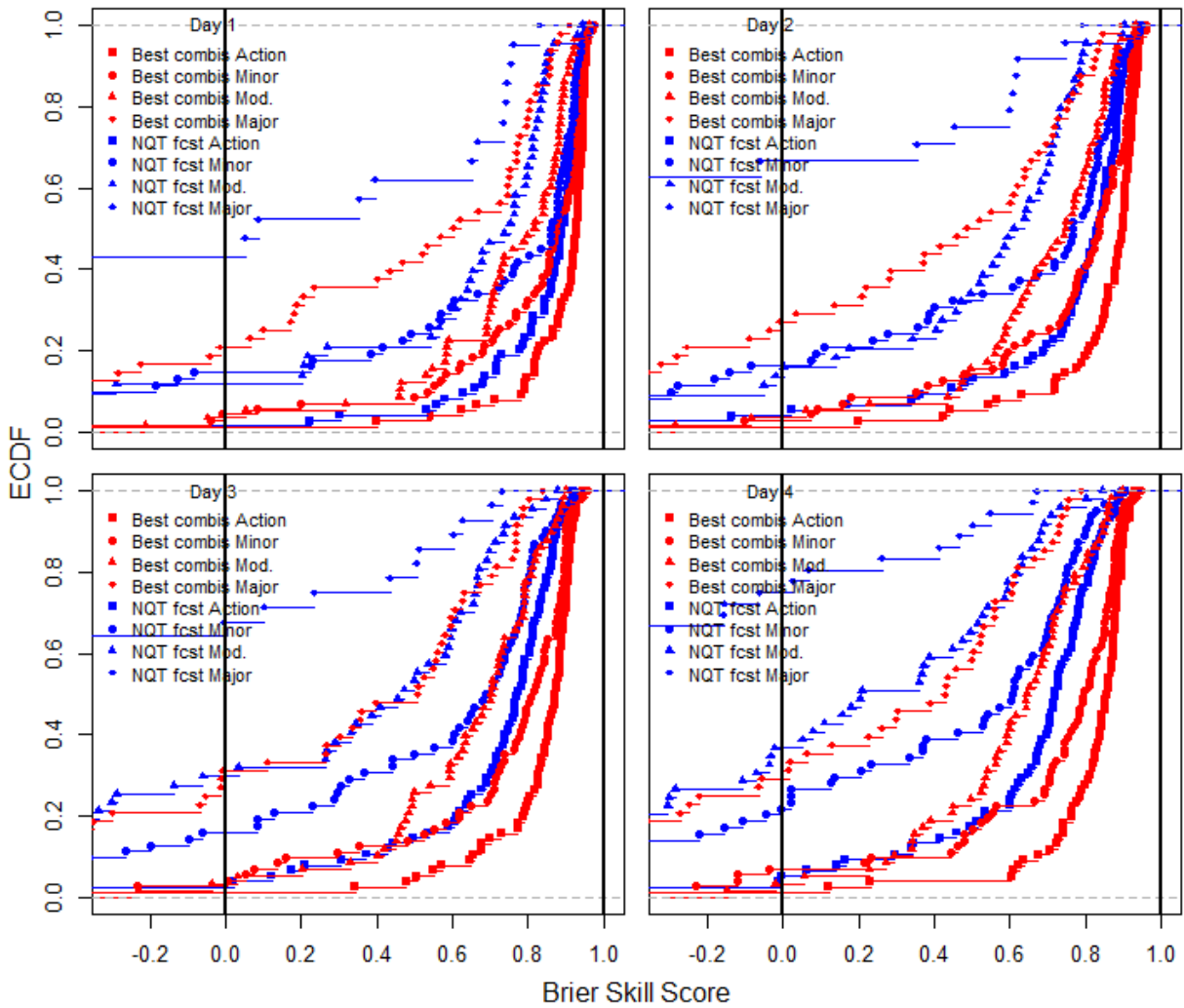


Figure 19: Empirical cumulative density functions of three QR configurations predicting exceedance probabilities of the Action, Minor, Moderate, and Major Flood Stage: the configuration using the transformed forecast as the only independent variable [NQT fcst]; the best performing combination for each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

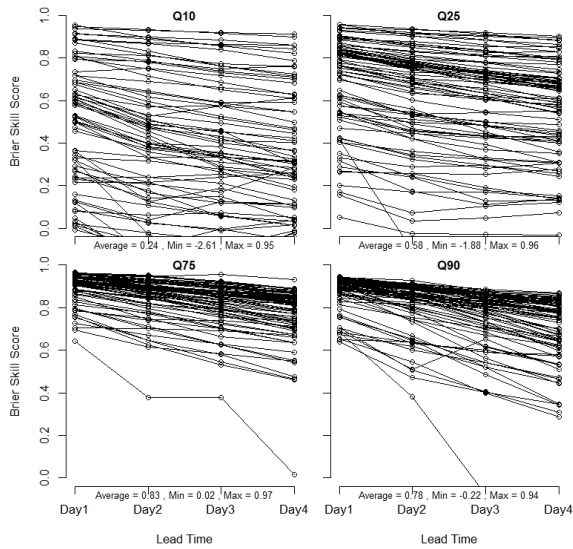


Figure 13: Brier Skill Scores of the forecast-only QR configuration (i.e., using the transformed forecast as the only independent variable) for four lead times and percentiles of observed water levels.

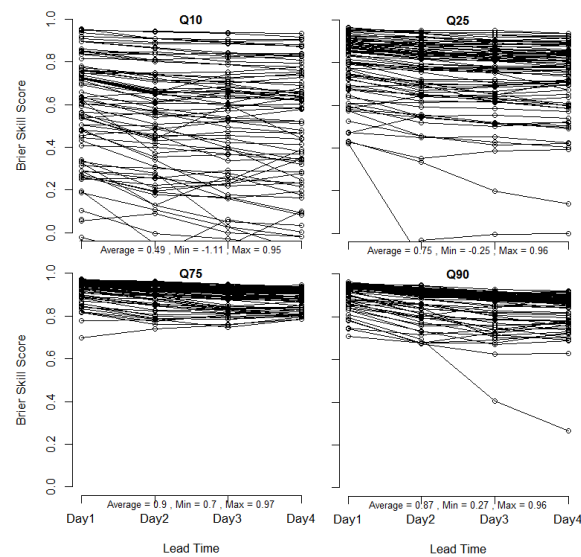


Figure 14: Brier Skill Scores for four lead times and percentiles of observed water levels using the best joint predictor for each river gage as independent variables in the QR configuration.

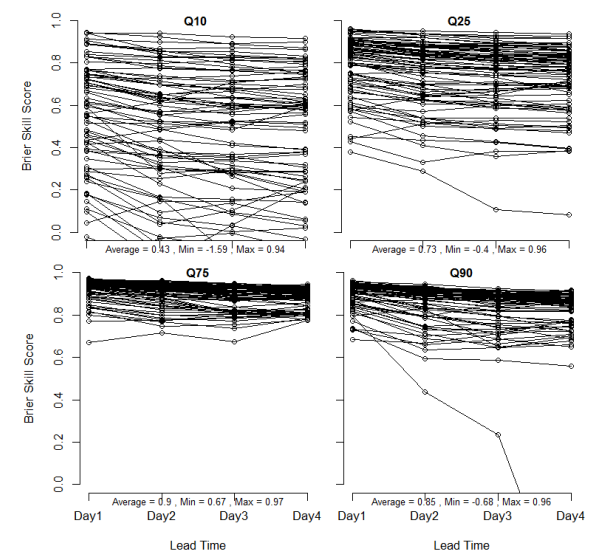


Figure 15: Brier Skill Scores for four lead times and percentiles of observed water levels using a one-size-fits-all approach (i.e., rr24, rr48, err24, err48) for the independent variables in the QR configuration.

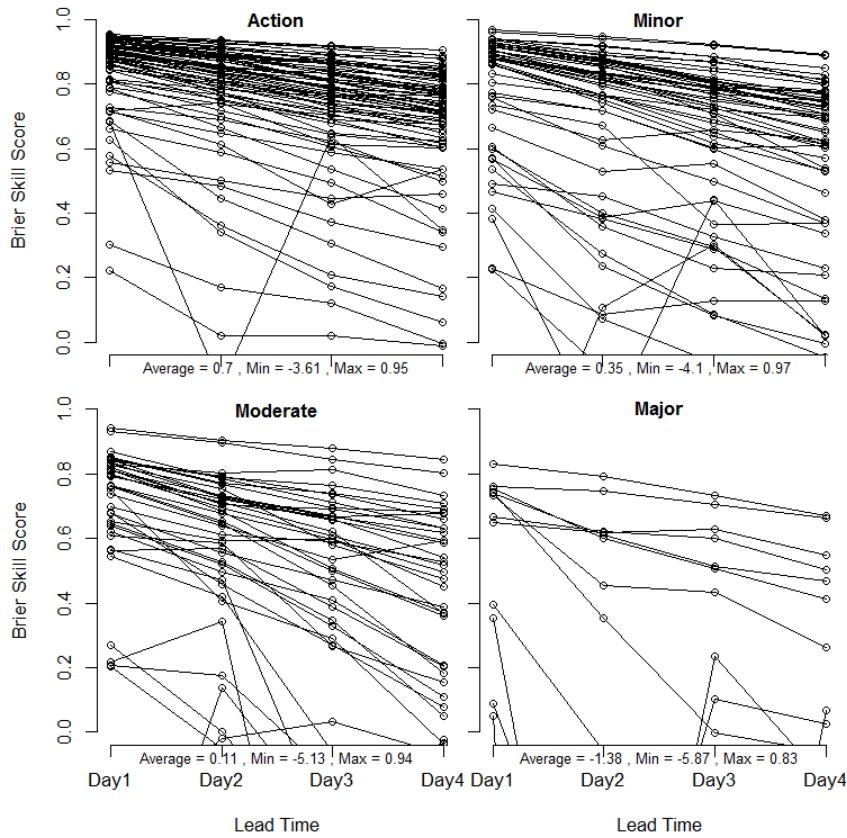


Figure 17: Brier Skill Scores of the forecast-only QR configuration (i.e., using the transformed forecast as the only independent variable) for four lead times and flood stages.

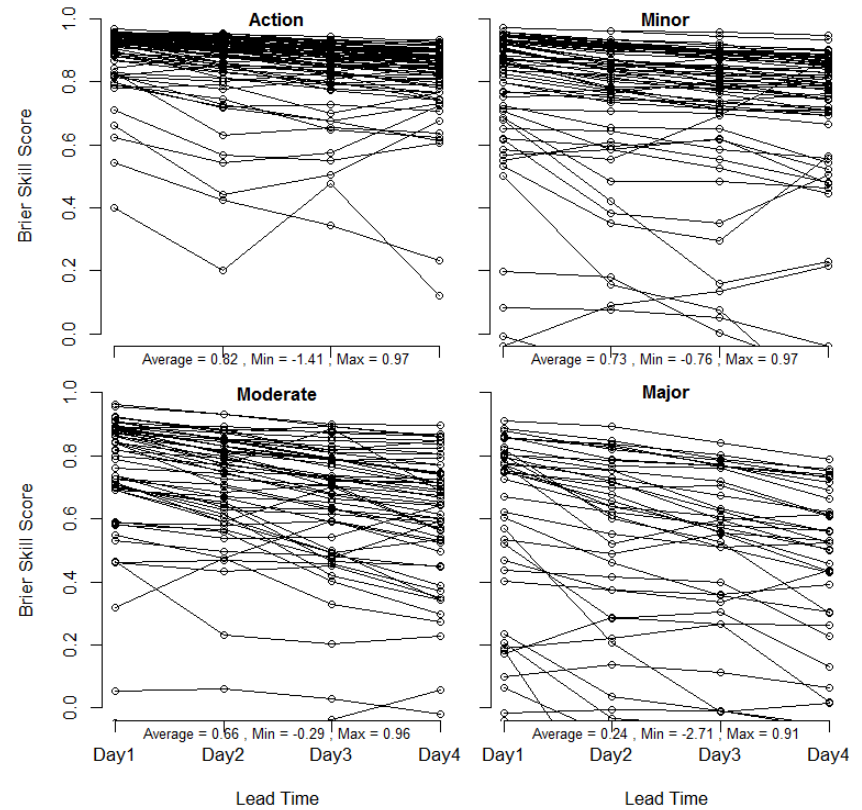


Figure 18: Brier Skill Scores for four lead times and flood stages of observed water levels using the best joint predictor for each river gage as independent variables in the QR configuration.

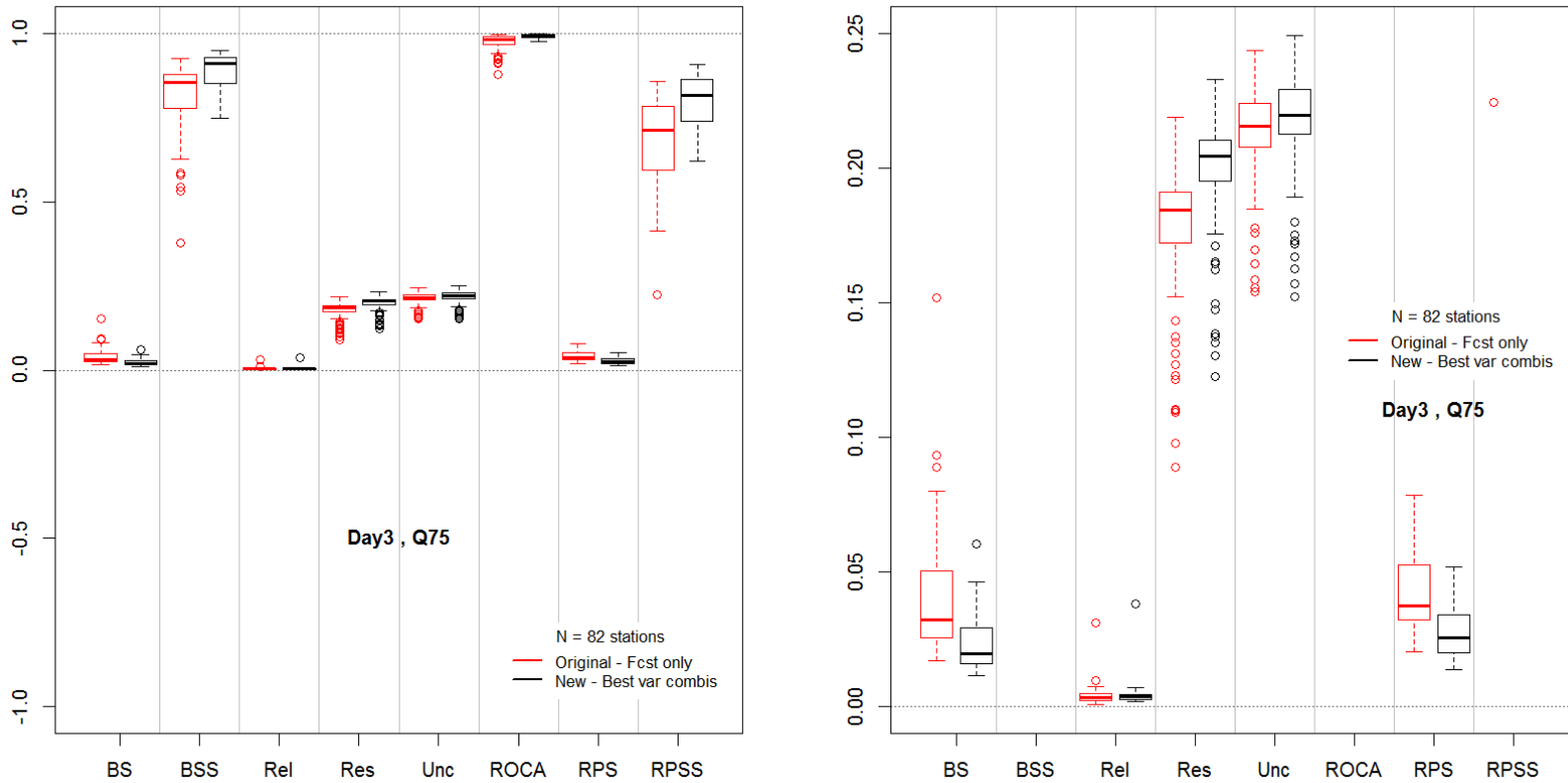


Figure 20: Comparison of the forecast-only QR configuration (i.e., only transformed forecast as independent variables) and the one-size-fits-all approach (i.e., rates of rise and forecast errors as independent variables) using various measures of forecast quality: Brier Score (BS), Brier Skill Score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

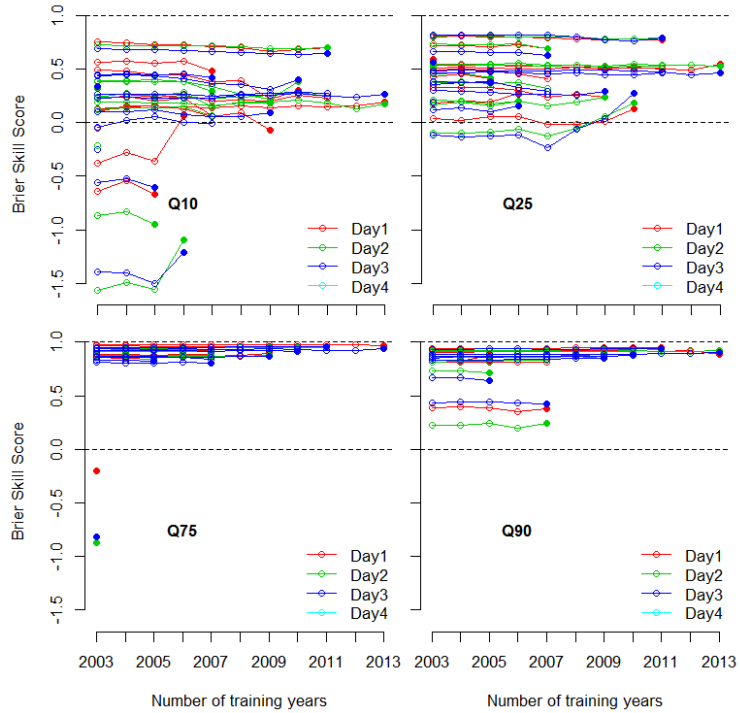


Figure 21: Brier Skill Score for various forecast years and various sizes of training dataset across different lead times (colors) and event thresholds (plots) for Hardin, IL (HARI2). The filled-in end point of each line indicates the BSS for the forecast year on the x-axis with one year in the training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

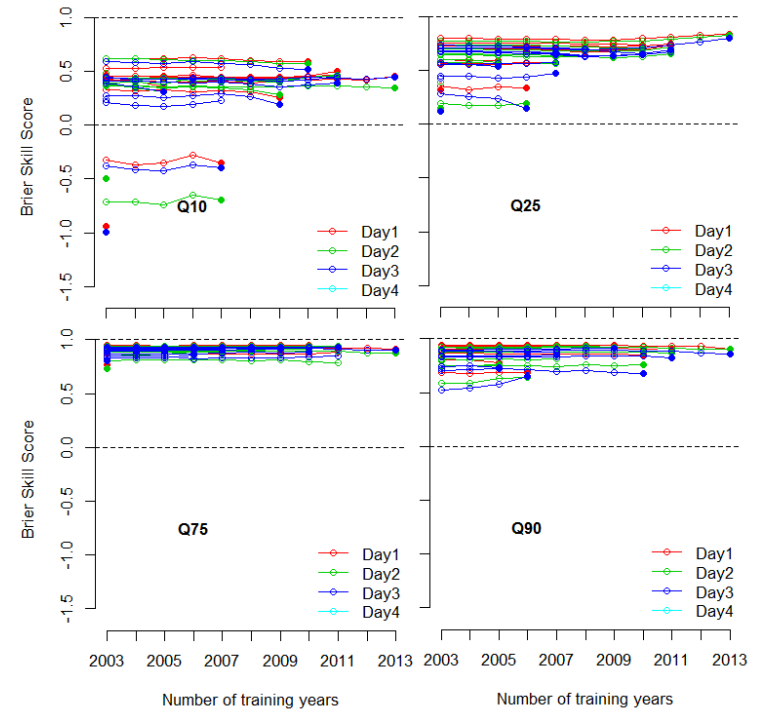


Figure 22: Brier Skill Score for various forecast years and various sizes of training dataset across different lead times (colors) and event thresholds (plots) for Henry, IL (HNYI2). The filled-in end point of each line indicates the BSS for the forecast year on the x-axis with one year in the training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

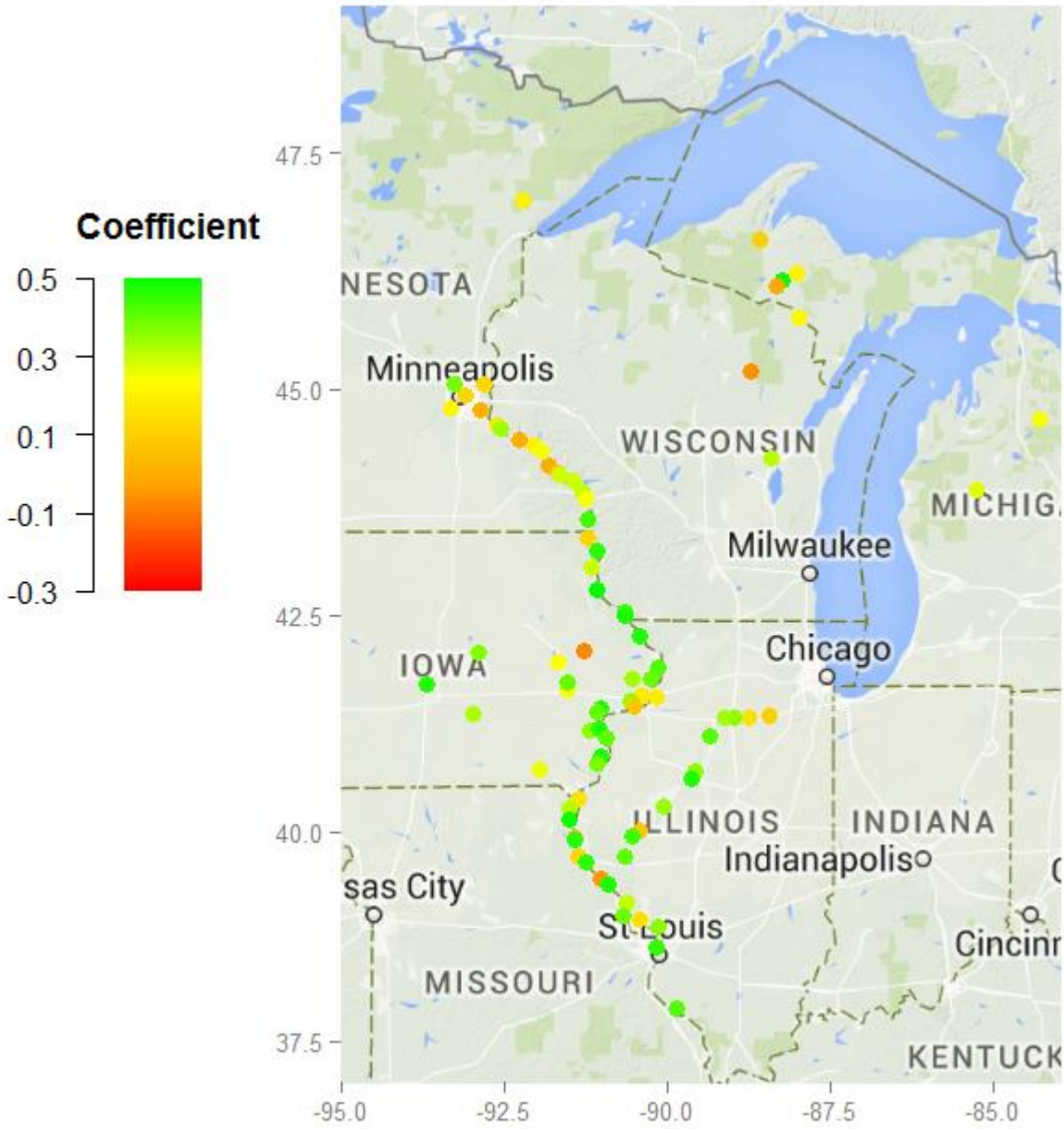


Figure 23: Geographical position of rivers. Colors indicate the regression coefficient of each station with the Brier Skill Score as dependent variable.

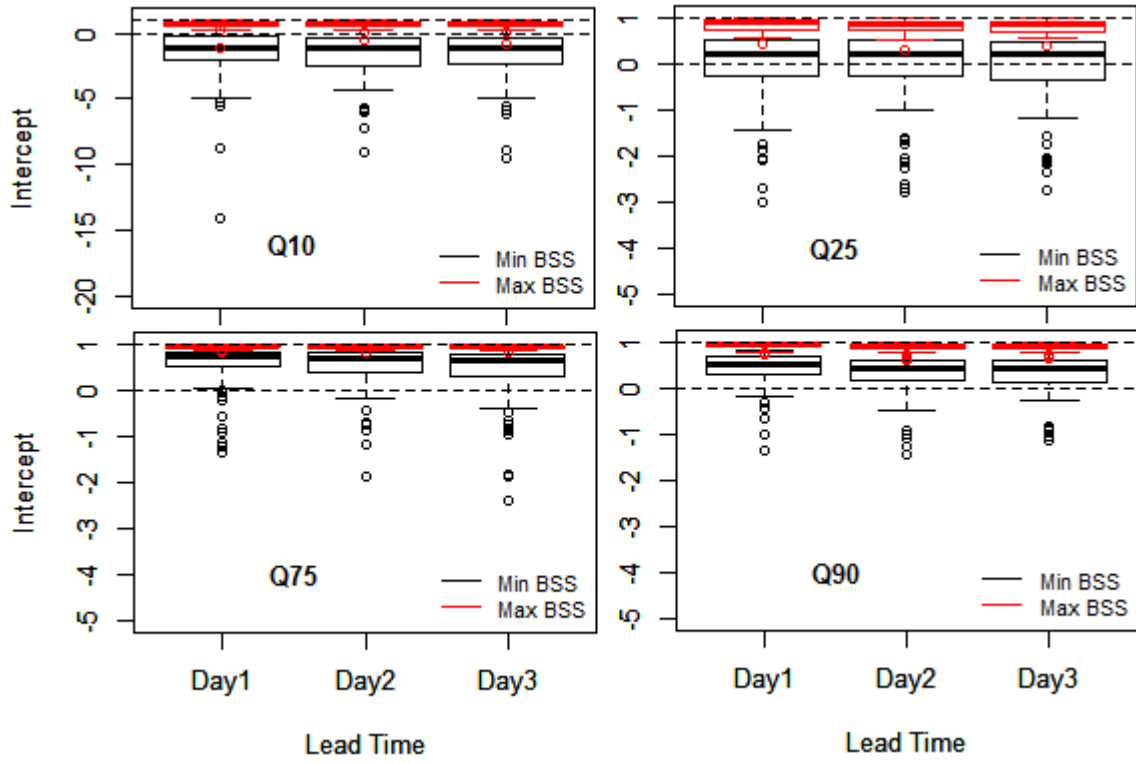


Figure 24: Minimum (black) and maximum (red) Brier Skill Scores for various lead times and event thresholds across locations, size of training dataset and forecast years.


February 4th, 2015

Revision to Journal Paper

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 Fischbeck

Dear Jan,

This letter outlines the changes we have made to our journal paper "Ten egies to Systematically Exploit All Options to Cope with Anthropogenic Climate Change".

General Comments:

1) The manuscript could benefit from a more substantial "hydrological analysis" of the forecasts made. Post-processors can be used to find statistical relations between predictors and predictands. There needs to be correlation and causality. The paper could benefit from a more in-depth analysis of the latter: what does the 'forecast error' depend on? Here, the authors choose rate of rise and past forecast error: these appear to be more or less randomly chosen, and are subsequently applied to all forecasting locations considered. However, I think that an analysis of the hydrology of the basins considered, in conjunction with the forecasting models for those basins, could reveal important information on how those models are expected to perform. How are the models calibrated? What does this mean for extreme events? Is the relation between predictors and predictand stationary across 'normal flow regimes' and 'extremes'? This likely varies with basin, and therefore one should consider varying post-processing configurations with basin also.

(1) How were the independent variables chosen:

The independent variables were not randomly chosen. It says in the paper:

"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 level and forecast error in the previous days.**"

For the sake of brevity, I did not include the results of these regressions. I rather wanted to use those pages to describe our results in depth. I added the bold part to the text excerpt above to clarify my intuition why this choice of variables makes sense. It was also explained that other in

I see no reasoning on why there could be causal relations between the predictand and the predictors you have chosen.

(2) Thoughts on the analysis:

As I have also explained in my answer to your special comment 7 below, I – like Wood et al. – see this post-processor as something that small organizations can use to make quick estimates of uncertainty.

As to extreme events vs. normal flow, I do analyze the performance of QR configurations for eight event thresholds separately. I find that a one-size-fits-all approach performs well for all gages unless extremely high events are forecasted. In the robustness section, I describe that forecast performance depends very much on river gage. **So the hydrological circumstances at each river gage do seem to make a difference.** I comment on basin-based analysis in response to your comment 295,7.

2) *There is one important assumption underlying the use of statistical post-processors: stationarity of the joint predictor, predict and distributions. The paper would benefit from a discussion thereof, particularly in relation to the results section, and the ‘robustness’ section contained therein.*

Added sentences indicated in bold in “Robustness” section:

“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. For the other event thresholds, it barely matters for the BSS how many years are included in the training dataset. That is good news, if stationarity cannot be assumed (Milly et al., 2008), a step-change in river regime has occurred, or forecast data have not been archived in the past. In those cases, only short training datasets are available. **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.**”

3) *“First US application” is irrelevant to the science and also incorrect, as Wood et al (see reference in Weerts et al, 2011) applied QR previously. This comes back a couple of times in the paper. Also, QR was originally devised by Roger Koenker; not by Weerts et al (I wish!).*

I deleted all references of this being the first application of QR to the American context throughout the paper and referenced Wood’s presentation throughout the paper. See the letter to the other reviewer ^{ok} for more detail.

In section 2.1 it already said:

“Quantile Regression was first introduced by Koenker (2005; 1978).”

4) *Different users have different needs for uncertainty information; it is not universally true that users benefit most from probabilities of exceedance or non-exceedance. Likewise, not all users are interested in extreme ^{ok} events per sé. This comes back a couple of times in the paper.*

True. I was writing another paper on emergency management, so that that group of clients was dominant in my head. I removed this claim throughout the paper.

5) *I would recommend to streamline use of terms:*

- ‘predictor’ or ‘independent variable’

- ‘predictand’ or ‘dependent variable’
- preferably omit use of ‘variable’ in context of statistical post-processors, as its interpretation can be ambiguous **ok**
- ‘configuration’ rather than ‘model’ (to avoid confusion with underlying hydrological models)
Updated this throughout the paper.

6) Please consider removing the footnotes. If the text contained therein is important, include it in the main body of the paper. If not, you may want to consider omitting it altogether.
ok
Footnotes were removed throughout the paper.

7) Practicalities of data access are not too relevant to the science and I would suggest omitting descriptions of why certain data sources could (not) be accessed and how much effort that would require. Instead, you could turn the argument around and say: “this and this is available and we’re trying to assess if there is any signal that can contribute to better probabilistic forecasts.”

The available data is not too relevant to the science and I would suggest omitting descriptions of why certain data sources could (not) be accessed and how much effort that would require. Instead, you could turn the argument around and say: “this and this is available and we’re trying to assess if there is any signal that can contribute to better probabilistic forecasts.”
as I said, I would turn the argument around. your choices are reasonable enough: explain how much signal you want to make clear, and how much effort it would take to access the data. I have not used the data because it is so difficult to access. I want readers to be aware that there is a way forward if they wish to further develop this technique.

Specific comments

Introduction:

1) Some elements can be safely omitted from the introduction:

- Discussion on QPF forecasts
- Discussion of PFC produced “outlooks”
ok
Okay, I deleted these parts.

2) Verifying by means of BSS only is somewhat limited I think, but it does fit with the authors’ wish to verify exceedance probabilities only. Why not, however, use a range of verification metrics? See, for example, some of the recent Brown and Seo papers as well as some of my own work (where the verification approach was inspired on the Brown/Seo papers).

The reason why I only use one metric, the BSS, is simple. When optimizing, I need an objective function. I don’t really see how/where you are optimising. do you mean it’s difficult to assess forecast performance with multiple metrics, I included Figure 18 (now Figure 20).

3) “Rate of rise” is more commonly used than “rise rate” I think.

Okay, I changed you need to change the axis labels in the figures also.

2.2 Brier Skill Score:

4) The ‘method’ section would benefit from a subsection on verification metrics. That section would then include the current sub-section on BSS, but also some discussion of other metrics now included in the ‘results’ section.

As described in my comment above, the Brier Skill Score plays a central role in optimizing the QR configurations. In my opinion, it needs therefore thorough discussion.

The other metrics are mentioned in the Results section in order to give the interested reader a feeling of what the BSS-based optimization achieves measured by those metric. A very short description of each metric is provided in the Appendix. It's not critical to paper acceptance, but personally I don't think it makes sense to spread the reader has to go back to the Appendix for explanations unnecessary.

5) A decomposition of Brier's probability score is included; what's missing, is a note on how these decompositions are computed in terms of skill. See one of the Brown and Seo papers for how that's done. Also, no quantified decompositions are shown in the results/analysis section? I added the equation below. Figure 18 (now Figure 20) already showed the performance in terms of quantified decompositions. Yes, better, but RES and REL remain undefined. I think you'd do well to describe these a

“Equation 4 defines the decomposition into resolution and reliability components described above (Brown and Seo, 2013).

Equation 1: 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 “

2.3 Proposed addition

6) The current title “Proposed addition: more than one independent variable” suggests that it is the *number* of predictors that's important. This is not necessarily so - it's content, not just quantity that's relevant. Please consider retitling this section.

The new title is:
 “Identifying the best-performing sets of independent variables”

7) This section could really benefit from some ‘hydrological intelligence’: what are the factors determining level of accuracy of model predictions? Are these already included in the model itself somehow? If so, how? If not, why not? To me, it is still an open question: what to include in a model, and what to include in a post-processor? Where is the boundary between statistical modeling and modeling of physical processes? This point is one that the authors should also revisit in the discussion/conclusions section.

I think, this discussion goes beyond the scope of this paper. Yes, variables as rate of rise are at least indirectly included in the “physical” model, referred to as hydrological model hereafter. However, I started researching post-processors thinking that small consultancies could offer statistical post-processors to clients, such as emergency management agencies. As long as NWS is not providing uncertainty information (which it might not do for short-term forecasts for many more years), that would be a valuable service. Coincidentally, that is exactly the application

that Wood talks about in his presentation in 2009. In short, I did not see post-processors as part of the traditional forecast process taking place at NWS.

Lastly, the post-processor discussed here has a different objective than the current hydrological models. It estimates uncertainty. It is my understanding that the hydrological models can only estimate uncertainty by producing ensembles. Since that means running the hydrological model with different input etc., the model itself does not produce an uncertainty estimate.

Having said that, I assume that variables such as rate of rise would have no explanatory power, if they had been sufficiently included in the hydrological model, and if that model had been well calibrated. As long as those variables add to the performance of the post-processor, I do not see why they should not be included. I do not have access to the NWS models, so I cannot assess, why those variables have explanatory power in the post-processor, even though they have probably at least implicitly been included in the hydrological model.

My personal preference would be to build a hydrological model for the whole watershed and to use post-processors to improve performance and reduce bias for single gages and flood stages. Similarly, I would intuitively opt for including hydrological knowledge of the basin in the statistical model and use purely mathematical/statistical methods in the post-processor to remove (local) biases, etc. At the end, I don't think that there can be or should be a strict separation. Many statistical methods are based on variables which ultimately have a physical meaning. They might add local information that cannot be account for in the larger hydrological model.

This is such a fundamental discussion that it would warrant a separate discussion paper rather than a section in the discussion section of this paper. **Let me know if you want to write one together! ;)**

3) Table 1: “forecast error 24 hours ago”. I understand this to be the difference between the current (i.e. at issue time of the forecast) water level and the forecast that was produced 24/48 hours ago - correct? Maybe good to state this.

Correct. The following sentence has been added to Table 1 and in section 2.3:

“The forecast error equals the difference between the current (i.e. at issue time of the forecast) water level **ok** and the forecast that was produced 24/48 hours ago.”

2.5 Data:

8) First sentence may be omitted, or moved to the introduction.

I merged the first two sentences of this section to be:

“The National Weather **ok** Service (NWS)’s daily short-term river forecasts predict the stage height in six-hour intervals for up to five days ahead (20 6-hour intervals).”

9) The manuscript would benefit from a custom made map showing the forecasting locations and basin delineations.

I included the basin sizes in the figure, because those are in my opinion more relevant for this study than the delineations: **ok**

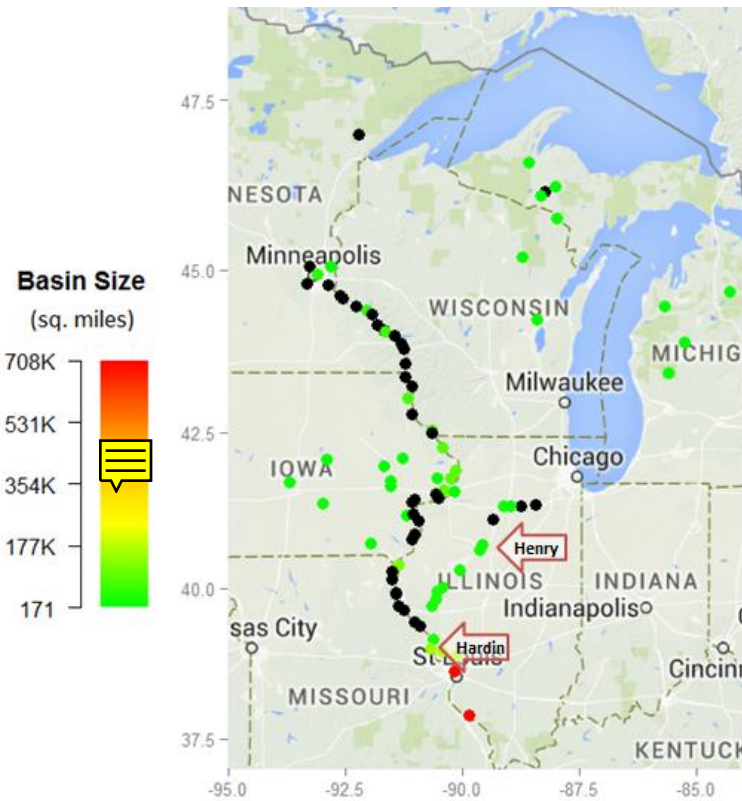


Figure 3: River gages for which the North Central River Forecast Centers publishes forecasts daily. Henry (HYN12) and Hardin (HARI2) are indicated by the upper and lower red arrow respectively. For gages indicated by black dots the basin size is missing.

3.2.2 Best performing combinations

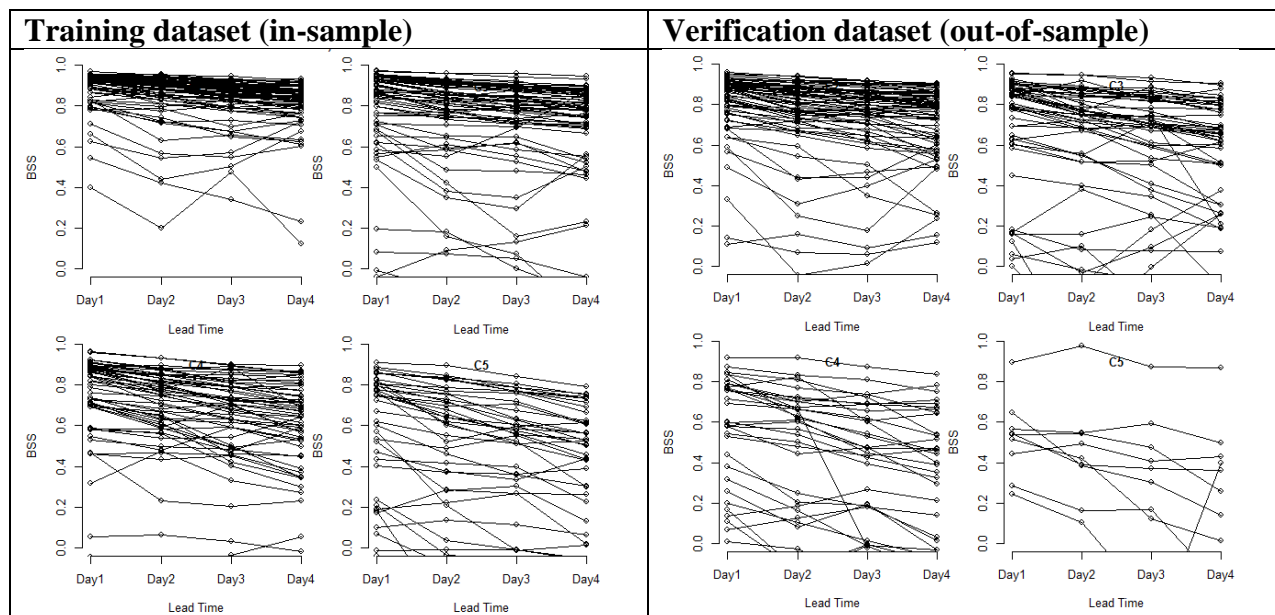
10) *The forecasts for extreme conditions perform worse when using multiple predictors. Why - overfitting? Some in-depth analysis would be good.*

Yes, that is my intuition, too. I added the following sentence:

“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.”

I re-ran some of the analysis in-sample, and indeed the model does perform much better for the training than for the verification dataset, see figures below. That is sure sign of overfitting.

re-applying quantile regression to the data it was trained on should yield perfect reliability



3.3 Robustness

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

I think the length of the training dataset is very important. In an ideal world, one would want to build reliable, skillful models, but save computation time, but alleviate climate variability and climate change should not be assumed. Urbanization and other human interventions are just too ubiquitous. I was interested to find out, how short training time series can be before the results start dropping significantly.

In sum, I prefer sticking with the current method. I added a qualifying statement though, that the small size of the training dataset leads to small BSSs for low thresholds (Q10):

“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. For the other event thresholds, it barely matters for the BSS how many years are included in the training dataset. That is good news, ...

...

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.

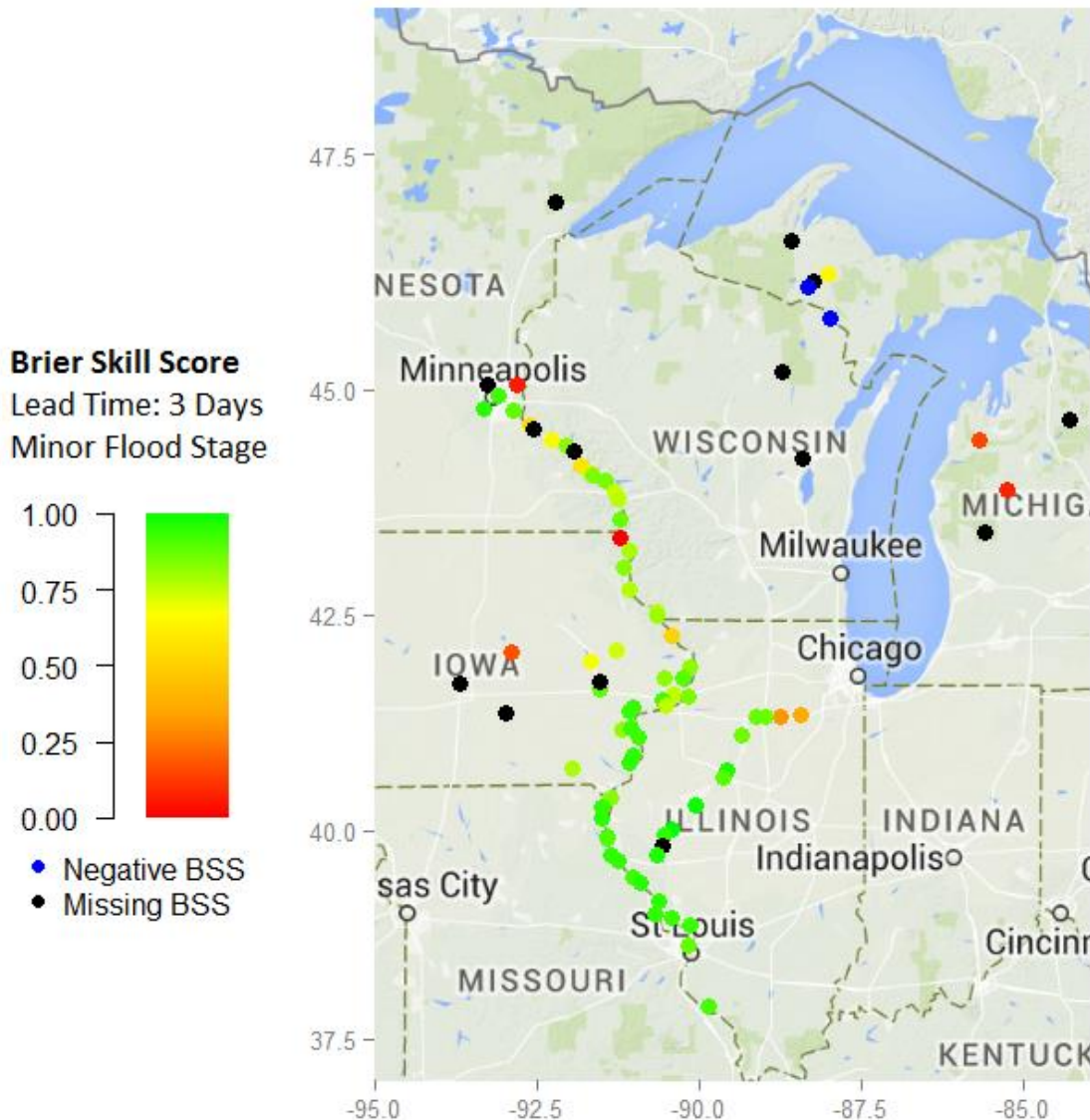
...

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.”

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

Besides watershed size and location (see comment 295,7) and the predictors mentioned in response (2) to your general comment 1, I currently don't have more data on the individual gages. A possible dataset to add in would be GAGES (http://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml), but that is for another paper. Like I have written in response to comment 7, **that level of detail does not belong into a post-processor,** in my opinion.

For your convenience, I plotted part of the upper right plot in Figure 17 onto a map, see below. It confirms what I said in response to comment 295,7.



“Future work”

13) Yes, more analysis on which predictors to use could work. Please refer to my earlier comments also on statistical modeling versus numerical modeling of physical processes, and on using knowledge of the hydrology of basins to determine meaningful predictors. Please see my answer to your specific comment 7.

Figures:

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

Additional specific comments

Additional specific comments are included in attached, annotated PDF.

You reviewed the paper very, very thoroughly. 109 comments! Thank you, this is valuable feedback!

282, 14: *These are two contradicting statements on the effect of adding four additional predictors.*

The configuration adding the other four variables to the forecast does perform better than the forecast-only configuration. But the configurations omitting the forecast, perform even better. So this is not necessarily a contradiction.

282,18: *as a philosophical side note, I am not sure if *forecasts* are uncertain. the future value of the variable of interest is, yes, but isn't the forecast certain as soon as it is issued?as a philosophical side note, I am not sure if *forecasts* are uncertain. the future value of the variable of interest is, yes, but isn't the forecast certain as soon as it is issued?*

The sentence now reads: “River-stage forecasts are no crystal ball; the future remains uncertain.”

283,1 *This statement doesn't really fit the flow of the paragraph. Would recommend to link it to river stage forecasts.*

This sentence now reads: “Including uncertainty in river forecast would therefore be valuable, just as has been recommended for weather forecasts in general (e.g., National Research Council, 2006).”

283,4: *Personally, I prefer “estimate” over “quantify”*

Changed throughout paper.

283,4: **Certain* sources of uncertainty is somewhat unfortunate. Check the Regonda paper for a useful formulation.*

The sentence now reads: “Those addressing 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.”

283,10: *Define “it”.*

The sentence now reads: “On the downside, the approach is expensive to develop, maintain and run.”

283,15: *What are these “major sources”?*

The sentence now reads: “I think the reader would benefit from some indication of what those major sources are. Future

283,15:

The sentence now reads: “Currently, the National Weather Service does not routinely publish uncertainty information along with their short-term river-stage forecast (Figure 1).”

283,18 & 283,22 & 283, 26 & 284,8:

I omitted those sections.

284,11: *What’s the relevance of this paragraph.*

I deleted the sentence on implementation in the RFCs. The paragraph provides background on post-processors used in river forecasting. The editor had explicitly asked for a more comprehensive literature review.

284,16: *Do Solomatine and Shrestha provide evidence for this statement, or do they merely state this?*

I deleted that sentence. It is not relevant for the argumentation.

284,18: *Publicly available does not equate relatively resources. Please rephrase or better even, omit altogether.*

The sentence now reads: “To make this approach useful for actors with limited resources, we exclusively use publicly available data. **ok - I do recommend that you address this more explicitly in the introduction: looking for**

284,23: **metrics* should maybe be *measure*?*

Correct. Changed **ok** throughout paper.

284,26: *I am not a fan of “method”, either. How about “technique”?*

Changed throughout paper.

284,26: *I am not a fan of “among others”.*

The sentence now reads: **ok** “These techniques differ in a number of ways, including their sub-setting of data, and the output.”

285,11: *Is that probability of exceedance the dependent variable? Or are you predicting distributions and then, from those distributions, determining the probs of exceedance?*

Technically latter, effectively both. The former is the dependent variable. The performance measure is the latter. **ok** hmmm I think you should make it very explicit what you're doing. if i remember correctly, you use a

285,14: *Can you sub...*

I removed **ok**'s claim throughout the paper.

285,24: *... there have been applications in the US context so your statement needs qualification.*

Changed throughout the paper. See **ok** my answer to general comment 3.

286,1 & 286,6:

Reacting to a comment by **ok** other reviewer, I omitted this paragraph.

286,10: *As much as I wish we had introduced QR, I think we merely applied it to hydrologic forecasting...*

The sentence now reads: “The paper is structured as follows. The Method section reviews quantile regression, introduces the **ok** performance measure, and discusses the performed analyses and data.”

286,19: *Omit “the”.*

ok
Done.

286,25: *... if you're extracting Pexc from a QR-estimated distribution then that's hardly “a way to further develop” a technique.*

Re-phrased paragraph: “. In this paper, elements of both studies are combined. However, our predictand is the probability of exceeding flood stages rather than confidence bounds. Additionally, this study tests the robustness of the technique across locations, lead times, event thresholds, forecast years, and the size of training dataset is tested. To develop the different QR configurations and to compare their performance, the Brier Skill Score (BSS) is used.”

287, 13: *QR and OLS regression differ in that assumption of how the data is distributed (non-parametrically vs. normally distributed).*

That discussion is similar to the comment in 285,11. Technically, you are right. However, I think that *effectively* QR predicts a percentile while OLS predicts a mean. In any case, if I remember correctly, it was a very easy-to-understand explanation, so I would like to leave it that way.

287,18: *rationale for probabilistic forecasting should be mentioned in the introduction, and surely there are better examples.*

This is a review of the quantile regression itself, not its application to hydrology. I think there is value to show that it has been found to be useful in hydrology. However, you start the section with "in the context of river forecasts" so the reader is

287, 23: *A 2012 paper is unlikely to instruct a 2011 paper.*

The sentence now reads: “Detailed instructions to perform NQT can be found in Bogner et al. (2012).”

288, 13: *If you are not going to use NQT, then I would omit this elaborate description thereof. What's the point?*

The point is that it later turns out that forecast cannot be combined well with the other independent variables exactly because of NQT.

288, footnote: *What's the relevance of this footnote?*

As suggested by the other reviewer, I omitted all footnotes.

289,4: *This = that of Weerts or yours?*

Ours. Changed.

289, 8:

True! Changed.

289,14: *Yes, but why not use additional verification metrics?*

As I have written in answer to one of your earlier comments, the reason why I only use one metric, the BSS, is simple. When optimizing, I need an objective function. I cannot optimize configuration performance for more than one variable. However, to give the reader some sense of how well the configurations perform in terms of other metrics, I included Figure 18 (now Figure 20).

289,21: *This uncertainty is different from the predictive uncertainty you are estimating. I would add a brief clarification to that extent.*

I added the following sentence: “. 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.”

289, footnote: *I would recommend not using footnotes.*

As already suggested by the **ok**st reviewer, I omitted all footnotes.

290, footnote: *Wilks, 1995, is unlikely to refer to the R package.*

True. But the R-package is based **is it?** Wilks' work.

291, 3: *The reliability curve for the forecast representing...*

Nice. New sentence: "The reliability curve for the forecast representing perfect reliability would follow the diagonal."

291,9: *In terms of sharpness? All of the scores and decompositions pertain to performance vs. climatology.*

Better explained: "Resolution 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."

291,14: *The curve for a forecast*

Changed according **ok**ly.

291,18: *What's the purpose of this statement pertaining to ROC?*

My adviser thought this was useful, if anybody else was going to try to apply the QR technique to different (non-hydrological) types of forecasts. In other fields of study, e.g., safety, the ROC is a very common measure of performance, especially in safety professions like emergency management.

292,1: *skill less than that of the reference forecast. Theoretically, the reference forecast could be very good. It is then unfair to say that the other forecast is devoid of skill maybe?*

The reference forecast is climatology here, i.e., predicting the average probability of an event every day. Is this formulation better?

"A forecast possesses skill, i.e., performs better than the reference forecast (in this case climatology), if it is inside the shaded area in **Error! Reference source not found.b** (now Figure 5b)."

292,4: *I disagree. The additional information may well constitute noise rather than a signal.*

Point taken. How about this: "The challenge is to identify a well-performing set of predictors that is both parsimonious and comprehensive."

292,8: *rate of rise*

Changed throughout **ok** paper.

292,9: *"additional potential independent variables"*

Change **ok**

292,15: *I think I know what you mean, but his formulation is ambiguous. Do you mean stratifying per month/season? Or using the date as another independent variable somehow? Please clarify.*

I meant the latter. The sentence explicitly lists potential predictors, there is no mentioning of stratification. I clarified: "...or the time of the year, e.g., using month or season as categorical predictors."

292,18: *True, but this still doesn't quite explain why rate of rise is a better predictor than water level observation.*

See my answer to your first general comment.

292,19: $2^5 = 32$, but one of these (no fcst, err, rr, at all) would not result in climatology, which is the baseline for BSS.

Exactly, that is why that combination is not included, so that there are 31 combinations. The combination you describe would mean that the model had no variables, but only a constant.

292,23: *above?*

Correct, agreed.

293,5: *at the river AT LOCATION X exceeds*

Good point. Agreed.

293,9: *Why only use these quantiles? Maybe as well calculate for every percentile, no? Especially if you are interpolating after the fact, this may have a positive effective on the predicted exc probs*

As I have written in response to your specific comment 7, I envisioned this technique to be used by companies like 3Tier where Wood works/worked. The choice to predict only these percentiles is the result of a cost-benefit consideration. The computation would take ~5 times longer, if we included all percentiles, which would not be justified by the marginal benefit in my opinion.

293,10: *This paragraph would benefit from an equation, to make sure that it is unambiguously clear what you are doing. If it helps: you may find the equations in our Lopez-Lopez paper useful.*

I started implementing what you suggested. But I came to think that those formulas make the paper unnecessarily much longer with limited benefit to clarification. Responding to a suggestion by the other reviewer I added the following part:

"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 (**Error! Reference source not found.**). 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
 $a_{xx,\tau}, b_{\tau}$ – configuration coefficients; forced to be zero if the predictor is excluded from the joint predictor that is being studied.”

293,11: *use of the term model for each of the estimated quantiles is potentially confusing here. I would just refer to quantiles.*

I see what you mean. This is the new sentence: “Each predicted --> estimated contributes one point to that distribution.”

293,16: *This is irrelevant here: (1) You’ve made the point before, and (2) by construction, the Brier Score assesses the quality of event probabilities rather than the quality of the probability distributions.*

ok

I deleted those two sentences. See also my response to your general comment 4.

293,23: *Not sure what “across all the days” means – does the statement pertain to sample size?*

Yes, it means that I use the forecast for all days in the verification dataset to calculate the BSS. New sentence: “To be able to compare various configurations, the Brier Skill Score is determined based on forecast exceedance probability for all days in the verification dataset.”

294,5: *four decision-relevant flood stages*

Change ok

294,12: *“four event thresholds” (may as well list the number thereof as you are doing this for all other items as well)*

Updated: “The result is 31 B_{ok} for 82 river gages for four different lead times (one to four days) and for eight event thresholds (i.e., flood stages or percentiles of the observed water level).”

295,7:

(1) It would be interesting (though not strictly required, I think) to analyze whether basin size affects forecast quality.

Well, we didn't analyze basin size, but did look into the characteristics of the river gages in the regression in Table 4 (now Table 2). Figure 21 (now Figure 23) illustrates that poorer forecast performance is correlated with being located upstream a river or close to confluences. The position of the gage along the river relates to watershed size. In my opinion though, the sub-average performance depends less on basin size. Rather, **at the upstream gages** the model is not able to "see" a flood wave coming down the river and at confluences of rivers the hydrology is more complex.

(2) Are all 82 gages/basins you consider independent or do some constitute subbasins of others?

Again, as you can see, the gages is situated **yes but if they're on the same river then they're not independent and one should take care in interpreting**

(3) Not required, but maybe you could show an ecdf of basin size to visualize how basin size is distributed.

I added ecdf in the Data section.

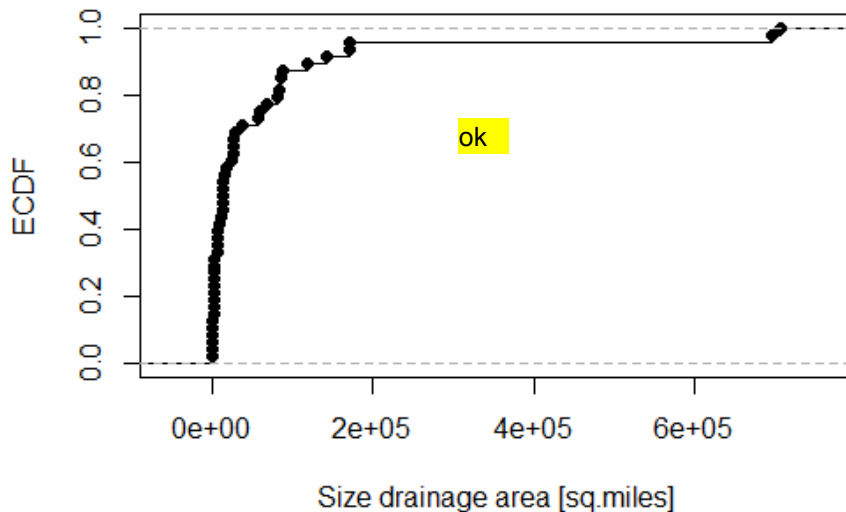


Figure 4: Empirical cumulative density function (ecdf) of sizes of drainage area for the river gages that are being forecasted daily by the NCRFC.

295,9: *upstream of*

Added the "of" **ok**

295,13: *I see why you want to include both SI and Imperial units, but do realize that it doesn't contribute to the readability (if that's a word) of the manuscript.*

Since we are talking **ok** about the U.S. in this paper, I deleted the km units.

295, footnote: *References should be included in a bibliography, not in a footnote.*

Footnotes were removed throughout the paper.

296, 13: *I am guessing **ok** that the relative error in terms of streamflow rate could be quite high.*

True. In this paper, I worked with water levels because that is the unit forecasts are published in for these gages. For the sake of brevity, I chose only to report the absolute values in Table 2 (now an ecdf figure), because those seemed more decision-relevant to me.

296, footnote: *i.e., there is a process with a considerable effect on your variable of interest which is not actually included in your model, or not modeled according to what happens in reality.*

True. Humans are much more difficult to predict than hydrology. It would be interesting if for example the price of electricity would be a good predictor of streamflow, because it drives dam operation to some extent.

297,2: *A table would be useful, as I'm not confident I understand what it is you are doing here.*

Isn't Figure 7 the table you are looking for? I also changed the sentence: "For each lead time (i.e., one to four days) and the eight event thresholds (i.e., 10th, 25th, 75th, 90th percentiles as well as the four flood stages), we counted at how many river gages each joint predictor resulted in the highest and the lowest BSS."

297,3: *"combination of variables" is better, as "variable combination would imply that "variable" is an adjective that qualifies the noun "combination".*

Changed throughout paper.

297,9: *flatter?*

Yes, change.

297,12: *"thus" implies statistical significance. Is there evidence to support this?*

New sentence: "This *some hydrologic analysis could maybe go some way towards explaining why this is the case becomes to include more data in the configuration.*"

299,2: *a one-size, not a one-size*

Changed.

299,16: *Pls consider not using the term variables, but instead predictors. This prevents possible confusion with the noun/adjective and also unambiguously makes clear that we are talking about the configuration of the...*

Changed. Updated throughout paper.

299,21: *If resolution increases while maintaining high reliability then yes, your contingency table will look better and hence the derived metrics will improve also.*

Yes, of course. I find picturing the improvement along those metrics useful (Figure 18, now Figure 20), because other researchers might have been working with those, rather than the BSS. And if I picture them, I have to mention them in the text. I did change the word "dimensions" to "metrics".

299,23: *Descriptions of verification metrics and their interpretations belong in a dedicated subsection in "approach" section (or similar). In any case, I would not describe these in the "results" section.*

Please see my answer to *matter of taste i suppose.* t 4.

300, 5: *I'm not sure I fully understand this sentence. Are you training ("calibrating") the models on one single year and then applying ("validating") these models to all remaining years? The figures don't really clarify this either. I thought I understood the approach from the plots, but the caption confuses me.*

That is not correct. I hope the new sentence clarifies it: "Each year between 2003 and 2013 was forecast by configurations trained on that year, i.e., the forecasts for that year were computed using only data from that year, clear now. however, i do think you should then calculate uncertainty bounds for your BSS 2013. Then, the BSS for that year (e.g., 2005 or 2013) was computed."

My recommendation is to either (i) do a leave-one-year out analysis, or (ii) simply compare joint predictor, predictand distributions.

(i) train on all available data except one year, on which you apply the calibrated models. Vary the validation year so that after x iterations, you'll have applied your model on all years in your dataset. Then calculate your verification metrics.

(ii) The success of QR, or any post-processing technique for that matter, depends on predictor, predictand relations remaining 'as is' during training and validation years. By directly checking this assumption, you can predict whether or not QR will do well. I do realise that this check may be cumbersome if you have many predictors.

See my answer to your general comment 7. The objective here is to test how robust the technique is to the stationarity assumption. This point is clear, I added: "We were particularly interested in testing how many years of training data are necessary to achieve satisfactory forecasting results."

300,8: *I think it means that for the years chosen, stationarity *can* be assumed. If there were no stationarity, your post-processing would have performed poorly.*

That is not correct. If I can include fewer years in my training dataset and still achieve good results, I rely less on stationarity. Maybe it's a matter of definition. A basic assumption underlying any post-processing technique is that the system is stationary. If it is not, the results will be less reliable. See also my answer to your comment 25.

300,9: *That depends on how you're configuring your post-processor. If you have a large database, then the QR calibration is unlikely to be affected by a few extreme events. The way around this is to calibrate QR on a sub-sample of data only, say on the top 10% of observations and associated forecasts and additional predictors.*

Well, just focusing on a subset of your observations does not increase your number of data points. The QR already looks at percentiles, so it is not very sensitive to outliers anyways. But your estimation of the 10th percentile for example will be better if you have more data points to fit your model to. I.e., even if you just look at a sub-set, you would want as many data points as possible in it, because any regression benefits from more data points.

300,25: *The use of multiple predictors may result in overfitting of some kind, whereas using a single predictor reduces this risk.*

Yes, true. But I am not sure what you are referring to in that sentence/paragraph. I am saying that the same joint predictor can result a range of BSS across river gages, event thresholds, etc. That does not refer to the number of predictors in the configuration.

301, 2: *Table 3, maybe?*

No, Table 4 (now Table 2) actually. This paragraph describes the results of the regression described in the paragraph before. Table 4 (now Table 2) is the corresponding table for the regression. Mainly in response to the other reviewer, I updated this part a bit:

“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).

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.”

301,6: Please see my note about the 'leave one year out analysis'. That would omit the need for this -imho confusing- analysis.

This is actually already a different type of analysis, than the one you wanted to change to a leave-one-year-out-analysis. Even if I took your suggestion, I would still do this regression, to

gain deeper insight into what causes the variability in BSS. The analysis before just visualized that there is variation, this regression studies this variation.

301,11: Why?

Because adding 82 rows to the table (gages are categorical variables) would have made it a really long table. Plus, the visualization in Figure 23 (before Figure 21) adds the very interesting geographic component.

301, 14: Depending on basin size, could it be that for some basins, time of concentration is shorter than 48h or even 24h? In that case, the additional predictors pertaining to past error and rate of rise at those moments in the past will have little information.

True. See my answer to your comment 295,7 (1).

302,2: This conclusion cannot be based on your analysis. changing the configuration of the postprocessor doesn't necessarily mean that you're maintaining same levels of reliability.

Figure 18 (now Figure 20) shows no change in reliability. In reaction to comments by the other reviewer, the section now reads:

“When compared to the configuration using only the forecast, it was found that including rates of rise in the past 24 and 48 hours and the forecast errors of 24 and 48 hours ago as independent variables improves the performance of the QR configuration, as measured by the Brier Skill Score. This confirms Wood et al.’s finding that QR error models should be a function of rate of rise and lead time (Wood et al., 2009). 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 additionally increases the resolution.”

302,9: Define 'satisfactorily'

Replaced that sentence with: “Additionally, customizing the set of predictors to the event thresholds does not improve the BSS much.”

302,15: why not?

I clarified this part:

“The combinations including the forecast (indicated by gray vertical lines in **Error! Reference source not found.** and **Error! Reference source not found.**) perform less well than those that exclude it. Plotting the independent variables against the forecast error as the dependent variable makes the reason visible (**Error! Reference source not found., Error! Reference source not found.**). Without a transformation into the normal domain, the 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 predictors become uniform distributions after NQT transformation. There is no trend detectable anymore. Further research is necessary to reconcile these two types of predictors. A possible solution could be to define QR configurations for subsets of the transformed dependent and independent variable. ”

302,20: *see earlier note*

See earlier answer.

302,27: *uncertainty in... what?*

Forecast and I would change this in "predictive uncertainty"

303,5: *what about applying QR to *streamflow* forecasts instead?*

That is a good idea, especially since streamflow is what is actually calculated by the hydrological models. But the archived forecasts used in this study were in water levels and not available as streamflow. At this point, I was trying to explain why the technique does not perform well for low thresholds. Even expressed in streamflow, the variability in low streamflows is probably going to be less than for high streamflows.

303,6: *it's not scarcity of data per se, but the fact that joint distributions of predictors and predictands vary with regime (low flows, medium flows, high flows). since a single set of QR parameters was derived from the full sample, low-end or high-end application cannot be expected to do really well. this is a problem inherent to the use of post-processing techniques.*

Forecasting extreme events is not sure if my comment was understood. See my answer to your comment 300,9.

303,12: *what models? the predicted probabilities of water level exceedance?*

I meant the performance of the classification trees. I change the sentence to: "Trials with a different technique, classification trees, showed that the observed precipitation, the precipitation forecast (i.e., POP – probability of precipitation) and the upstream water levels significantly improve forecasting performance."

304,15: *Please refer to this as Wikipedia, 2014.*

Donok

308,1: *Combinations of variables. See earlier comment.*

Called "Joint Predictions" now.

308,2: (1) *what's the difference between the filled circles and the open circles?*

None. Just a visual help, so that you see that the first column does not continue in the second column. At the end of the first column, the joint predictor includes two variables, and in the beginning of the second column, it includes three variables.

(2) *the use of statistical models *without* the det forecast as an explanatory variable opens up a whole new set of considerations... maybe good to comment on this?*

I don't understand. Which considerations are you referring to? That you can include some variables that were of little benefit when you included the forecast? That the forecast does not combine well with the other predictors is a finding of the paper. I did not know that starting out. This table is part of the method section.

(3) *are any of the errXX and rrXX values used in the hydrological models used to produce a fcst? If so, please mention this and comment on what this means.*

I don't know, I do not have access to the NWS models. The HMOS post-processor only uses streamflow at **10k** us time steps as explanatory variables: page 3, http://ac.els-cdn.com/S0022169413003958/1-s2.0-S0022169413003958-main.pdf?_tid=09b4b0ba-a80c-11e4-be2a-00000aab0f6c&acdnat=1422573218_6d0fa1b246a9bedfdafc04a172e794f5

309: Personally, I would show this information as a set of six ECDFs (one for each lead time considered) in a four-plot figure (one for each sample/subsample)

Good idea. **thanks- i like the new plots**

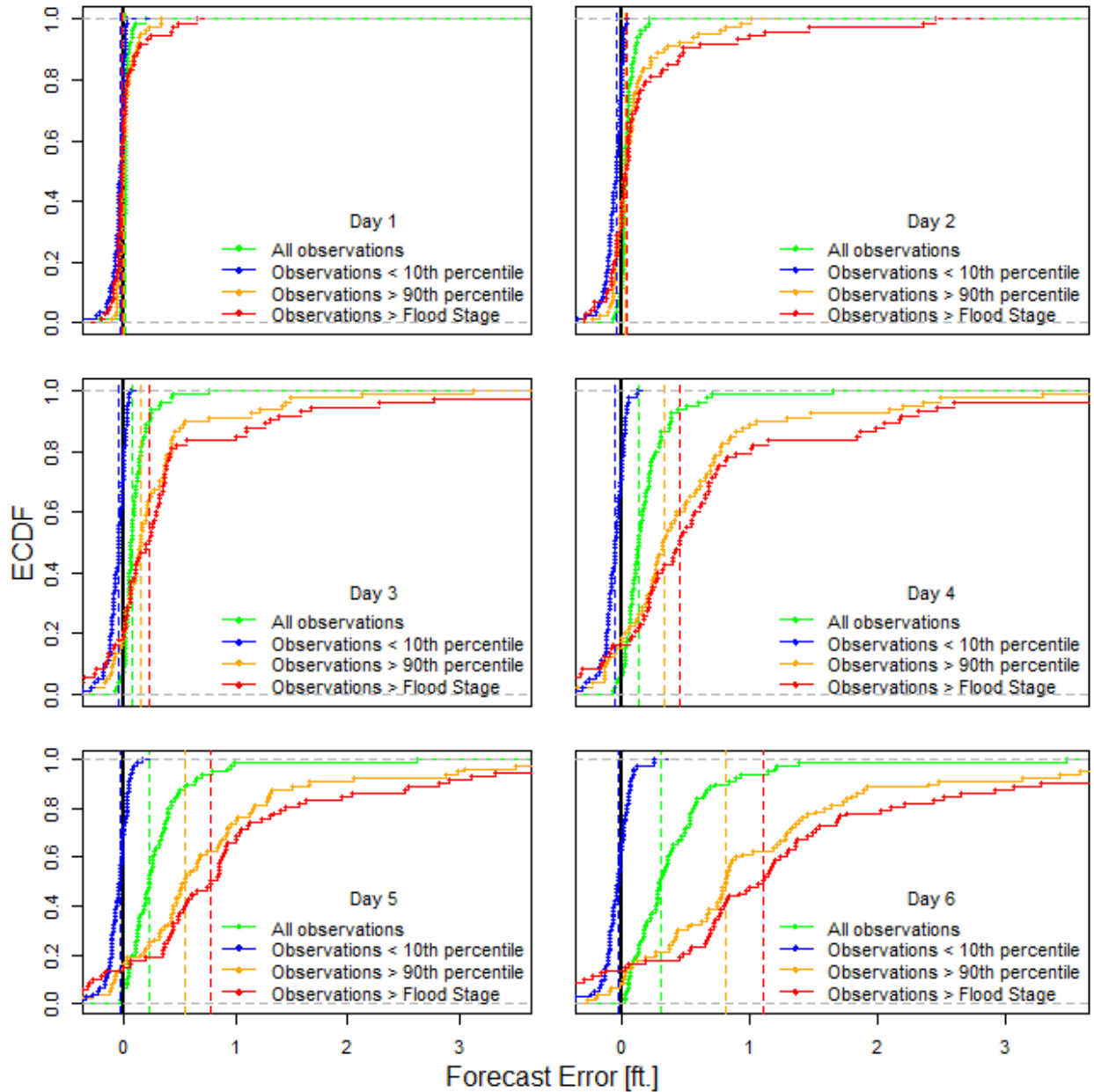


Figure 6: Empirical cumulative distribution function (ecdf) of forecast error at 82 river gages for six lead times. Vertical lines show the median forecast error of the corresponding subset.

310: You'll have realised by now that I'm quite keen on seeing full empirical distributions rather than summary values only ;). Again, I would consider presenting this information as ecdfs rather than as tables.

Here you go:

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].

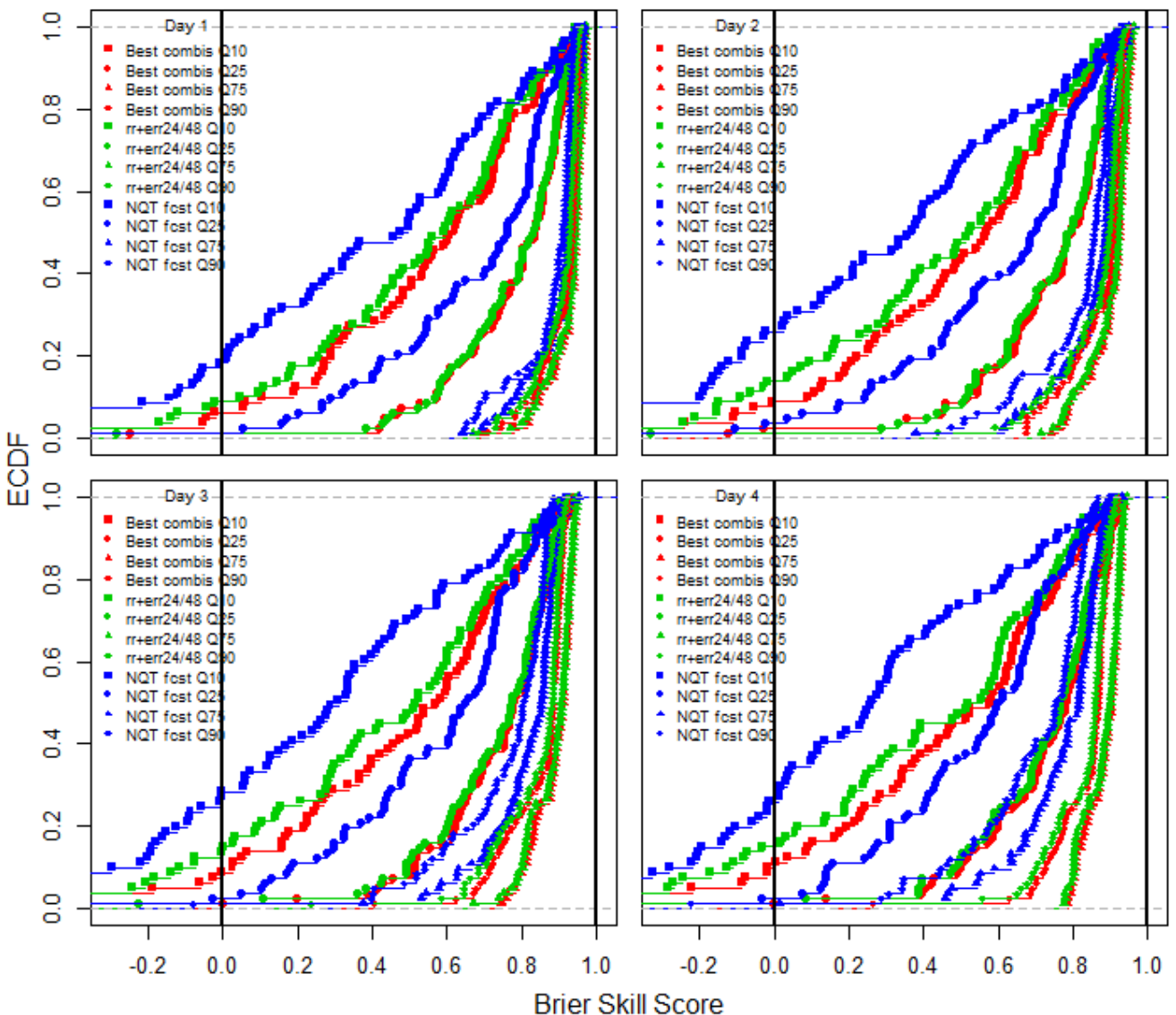
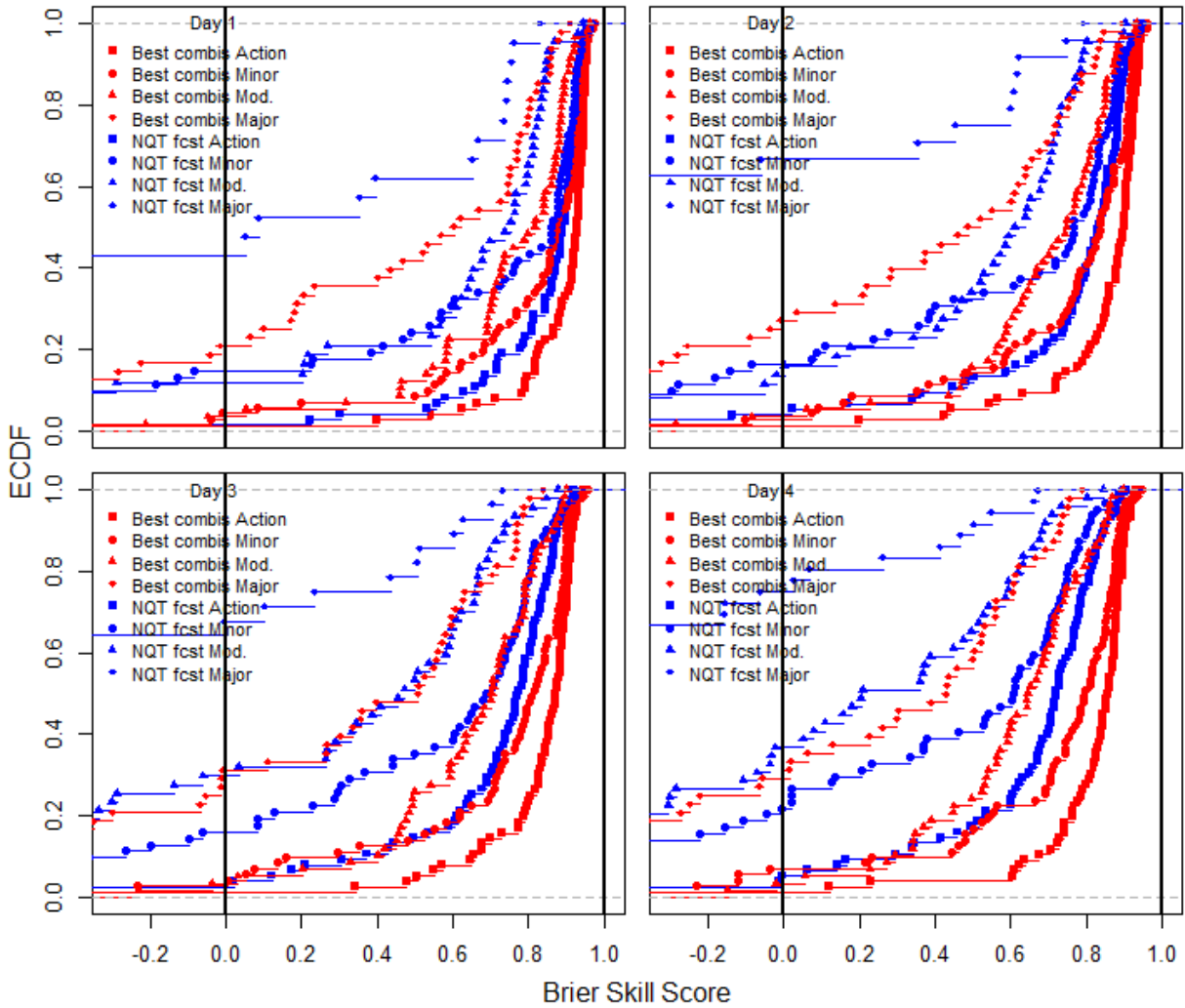
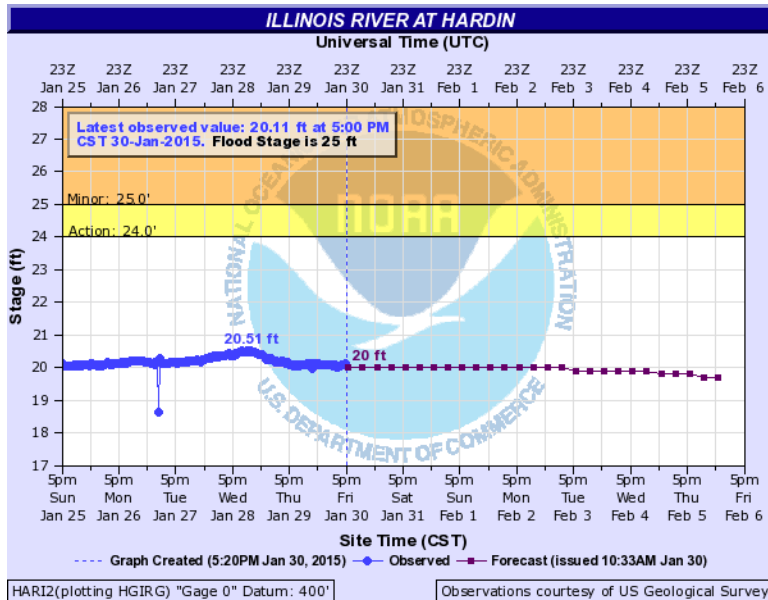


Figure 19: Empirical cumulative density functions of three QR configurations predicting exceedance probabilities of the Action, Minor, Moderate, and Major Flood Stage: the configuration using the transformed forecast as the only independent variable [NQT fcst]; the best performing combination for each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



312: Why download this not-so-exciting April forecast in October?

Because it is not October. If these plots are being archived, I cannot access them. Today's is boring, too: **yes, it is. consider omitting maybe**



313: These spring outlooks aren't topic of this paper, are they? Omit!

Omit **ok**

314: These long term forecasts aren't topic of this paper, are they? Omit!

Omitted **ok**

315: incorrect: outperforms the reference forecast, in this case 'climatology' which is not a random guess.

New caption:

“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 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.)”

316: I would rather see a map of all 84 forecasting locations used, and with information about the July 10 conditions omitted.

Okay, here is **ok**

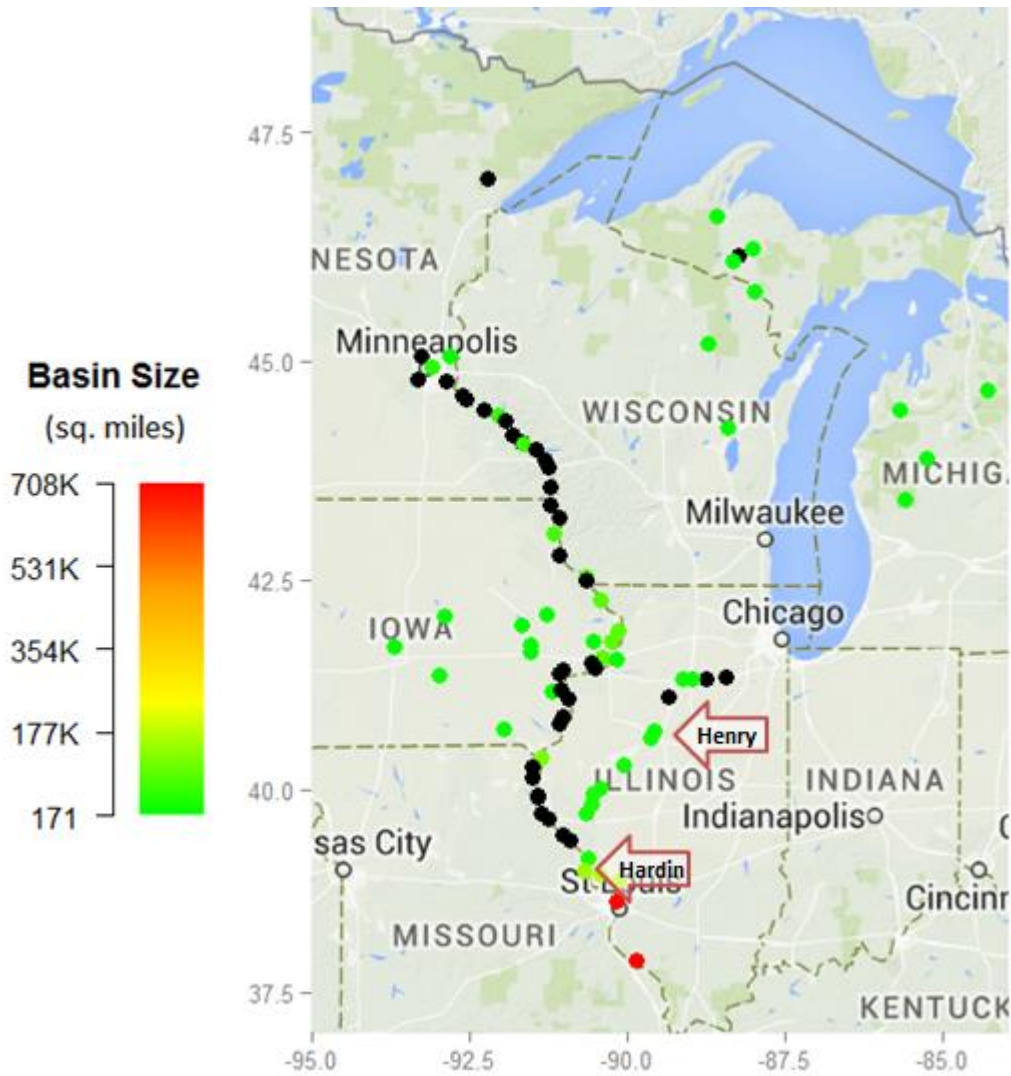


Figure 3: River gages for which the North Central River Forecast Centers publishes forecasts daily. Henry (HYN12) and Hardin (HARI2) are indicated by the upper and lower red arrow respectively. For gages indicated by black dots the basin size is missing.

317: This comment applies to various graphs: as both horizontal and vertical axes are identical, I would omit the axis labels on hor axes of top two plots, and axis labels on vert axes of two right-hand plots. You can then enlarge the actual plots.

Did so for all figures. ^{ok}

318: Recommendations: (1) omit duplicate axis labels where possible; (2)

Did so for all figures. ^{ok}

322: (1) omit repetitive labels where possible; (2) would also recommend zooming in on could, at expense of extreme values

(1) Did so for all figures.

(2) I actually would prefer not cutting of the extreme values, keeping the plots symmetric and where applicable with the same axis limits.

325: What are 'perfect variables'?

Sorry, that picture should have been clipped like all others. It is now.

332: if you really must include this figure then please consider using a colorscale that better clarifies differences between the locations.

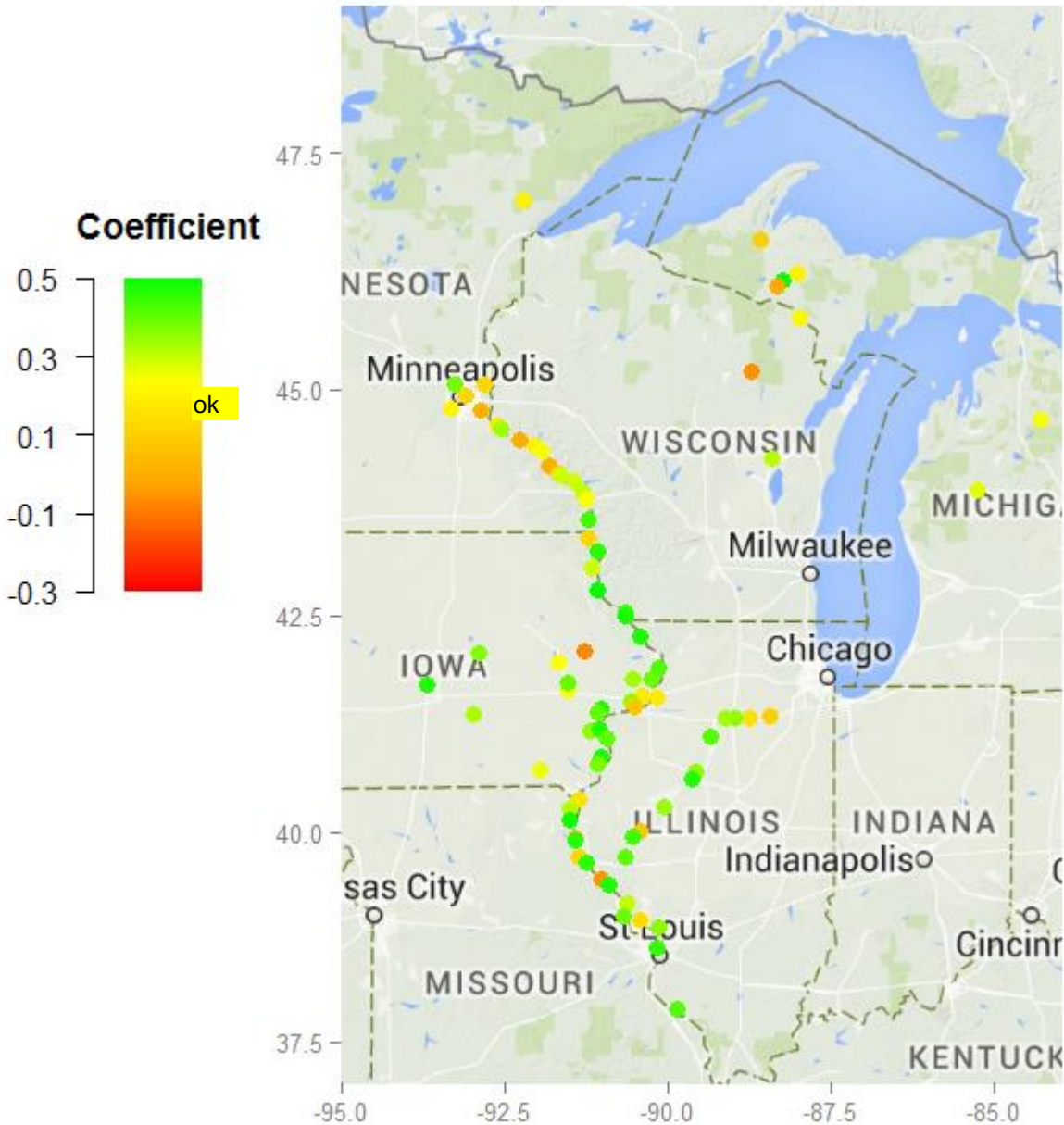


Figure 23: Geographical position of rivers. Colors indicate the regression coefficient of each station with the Brier Skill Score as dependent variable.

We hope that you find that these changes to have satisfactorily addressed the reviewer's concerns. If there are additional changes that you believe are needed, please let us know.

Regards,

Frauke Hoss, Paul Fischbeck