Response to Reviewer #1

February 4th, 2015

Revision of 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 Reviewer,

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

General Comments

1) The authors are apparently unaware of the first presentation of QR, which pertained to an American river, and predated the Weerts et al (2011) paper by several years. Wood et al (2009) is a citable conference presentation and is available online through the Amer. Met. Soc. (note the paper currently cites one conf. presentation). It is notable because the presentation also presents the rationale for using river rise as a predictor in QR, and demonstrates the application to operational river forecasts. This paper claims repeatedly to be the first application in an American context, which it is not given the earlier work, and also claims to introduce the concept of the additional predictors. I recommend that the paper recognize both Wood et al (2009) and Weerts et al (2011) as introducing the QR method for streamflow post-processing (until another earlier ref. can be found!), recognize the Wood et al inclusion of predictors such as river rise, and remove the framing of Weerts et al as the 'original' method versus this papers 'new additions'. The authors make a substantial contribution in their detailed examination of river rise together with the new predictor – trailing error – and the use of QR to estimate exc. probs. The presentation by Wood et al. is great, because it is exactly the type of application that motivated me to do this research. I think, it is very valuable that small business provide uncertainty estimates as long as the NWS doesn't. Additionally, they can provide more localized services when there is a need/market for it.

Thank you for referring me to that presentation, I was definitely not aware of it. It is a pity that that presentation is so hard to find and watch. It took a while before I found the right browser to watch it; on the university network I wasn't able to watch it at all.

Throughout the paper, I added in references to the presentation, removed the "first application" references, and reworded the "original" additions vs. method "with additions".

2) Though the authors highlight several interesting characteristics about the varying performance of predictor combinations, they currently offer little physical explanation for outcomes such as the (1) forecast itself being a poor predictor in some cases, or (2) multiple predictors faring worse at high thresholds. Physical reasoning would help dispel the possibility of simple overtraining, or perhaps mis-aligned training given the sample. I think the paper needs a stronger physical or at least statistical discussion to provide insight into the cause of such findings.

I can give more statistical discussion:

(1) The forecast is not a poor predictor. He just cannot be combined well with the other predictors. I explain more in response to your comment 298,7.
(2) I discuss overfitting in response to your comment 294,7.

3) The paper argues in several places that the exc. prob. forecasts are somehow 'more useful' for decisions than confidence intervals on forecasts (a widely used output). This arguably depends on the user. The position is taken to bolster the author's claim of an 'advance', but it's unnecessary – both are useful, and the author's can simply note that they have taken a different tack than in earlier uses.

I removed this claim throughout the paper.

4) The results section is somewhat long, and I think the paper could still be effective if the figures and tables were trimmed somewhat – but I leave this to the author to decide.

Given that this isn't a print journal, I would like to keep the descriptions of all types of analysis that we have done. I think, they all add a new insight and their descriptions are concise. Additionally, the other reviewer has asked for more figures rather than less.

Specific Comments

282,2 – awkward first sentence: 'further develops [QR]'? or just 'applies', or perhaps 'further develops an application of QR'. I don't think QR itself is being further developed. also, suggest rephrasing ". . .to predict flood stage exceedence probabilities based on post-processing single-value flood stage forecasts."

I revised sentence to be:

"This study applies quantile regression (QR) to the prediction of flood stage exceedance probabilities based on post-processing single-value flood stage forecasts."

282,5 – it was not the first, actually – see comment below for 285,6.

True. I deleted that sentence.

282,8 – suggest avoiding references in the abstract. Also, this statement is not correct – see comment on 285,6 below – the first implementation did use additional variables. The Weerts

implementation was far more comprehensive, leading to an article, and also added the nice feature of flow normalization as an innovation to the approach.

I agree. I revised this section of the abstract to be:

"Besides the forecast itself, this study uses the rate of rise of the river stage in the last 24 and 48 hours and the forecast error 24 and 48 hours ago as predictors in QR configurations. When compared to just using the forecast as independent variable, adding the latter four predictors significantly improved the forecasts, as measured by the Brier Skill Score (BSS)."

282,17 – I suggest adding one more sentence to the abstract to state the value of the approach – ie, that it helps quantify forecast uncertainty for the outputs of a deterministic forecasting process, which is currently common practice in many national flood forecasting services.

Good idea. I made this the second sentence in the abstract:

"A computationally cheap technique to predict forecast errors is valuable, because many national flood forecasting services, such as the National Weather Service (NWS), only publish deterministic single-value forecasts."

283,3 – "quantify 'forecast' uncertainty"

Added "forecast" there.

283,13 – perhaps mention that the HEFS system described in Demarge also includes a method for post-processing total uncertainty.

Please see my response to your comment 284,10 below.

283, 23 – 'serves as' – perhaps, but who knows? It's never been verified. Better to say 'may serve as'

The other reviewer suggested removing this section on QPF forecasts, so I did.

284,3 – this is true in the eastern US – in the west, ensemble forecasts go out as long as 2 years. This figure could be trimmed to reduce paper length.

The other reviewer suggested removing this section on outlooks, so I did.

284,10 – NWS also has a technique called HMOS which is applicable to postprocessing single value forecasts. HEFS also includes the EnsPost module, which post-processes total forecast uncertainty, and these both should be mentioned.

Thank you for pointing me to HMOS, I had not considered it yet. I added the text below. On a side note, I really don't understand why NWS does not publish the valuable uncertainty information produced by HMOS alongside the deterministic forecast, even though it is a standard product of CHPS. Do you know?

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

284,13 – again, 'further developed'? What does this mean exactly? perhaps just use 'applied' or clarify what aspect of R. Koenker's method is being 'further' developed.

That part is not essential to the sentence. So I shortened the sentence to be:
"In contrast to an ensemble approach such as HEFS, the statistical post-processing in this paper does not distinguish between sources of uncertainty, but studies the overall uncertainty in a lumped fashion."
I also included the following disclaimer:
"The study does not add to the mathematical method of quantile regression itself."

285,12 – this view is a bit narrow; certainly many users are concerned with low flow thresholds as well, and in any case, confidence bounds on forecasts are directly relatable to risk of threshold crossing (high or low).

Please see answer to general comment 3.

285,6 – QR for streamflow post-processing was introduced both by Wood et al (2009) and Weerts et al (2011). The former reference described what was likely the first application of QR to streamflow in the 'US American context', and possibly anywhere: Wood, AW, M Wiley and B Nijssen, 2009, Use of quantile regression for calibration of hydrologic forecasts, 23rd Conf. on Hydrology, Phoenix, AZ, Amer. Meteor. Soc., 11.3 [available online at: http://ams.confex.com/ams/89annual/wrfredirect.cgi?id=10049] Wood et al. described using QR to provide confidence limits for deterministic forecasts of the Lewis River in Washington State (e.g., Figure 1). The work emphasized the need for determining the QR error models as a function of the rise rate of the river as well as lead time (e.g., Figure 2), and then demonstrated the application. An earlier version of this presentation had been given by the same author at the 2008 HEPEX workshop in Delft, NL on Hydrological Ensemble Post-processing Methods, and this was acknowledged as the inspiration for Weerts et al (2011). It is likely that the work was not submitted to a journal because the authors worked in the private sector, where publication is typically less encouraged than conference presentation. Incidentally, the Wood et al streamflow

QR work had in turn been inspired by the application of QR for calibrating temperature forecasts, as described by Hopson and Hacker (2008), as well as by applications in the wind forecasting industry. Hopson, TM and JP Hacker, 2008, Combined approaches for en-semble post-processing,19th Conference on Probability and Statistics, New Orleans, LA, Amer. Meteor. Soc., 3.1 [available online at: http://ams.confex.com/ams/88Annual/wrfredirect.cgi?id=7501]

I revised this bit to be the text below. Additionally, I re-worded all references to the "original" approach throughout the paper. Please see also my answer to general comment 1.

> "This paper further develops one of the techniques mentioned above: the Quantile Regression approach to post-process river forecasts first introduced by Wood et al. (2009) and further elaborated by Weerts et al. (2011) and López López et al. (2014). The Weerts study achieved impressive results in estimating the 50% and 90% confidence interval of river-stage forecasts for three case studies in England and Wales using QR with calibration and validation datasets spanning two years each."

285, 23 – given the previous comment, this statement is incorrect and should be removed. The paper should recognize the earlier work and related ideas therein.

Done.

285,25 – this paragraph summarizes results, and seems out of place. Better to state that QR is conditioned on several factors in the study, and say what those are and why they are considered, than to the tell the outcome (here) of doing so.

I agree. I re-wrote that paragraph:

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

286,10 – having established earlier that Weerts, and I suggest also Wood, introduced QR, it is not necessary to return to it repeatedly in the paper (eg 286, 15, 19 etc). Overall, I think the paper should de-emphasize the verbiage about 'additions' and 'further development' in contrast to an 'original method', especially since the rise-conditioned error approach actually was the first method introduced at a national scientific meeting. Instead, just emphasize what has been done, as it is good work, and the paper can stand on its efforts alone, without requiring the label of being 'new' or 'first'.

Thank you for the compliments. Please see my answer to general comment 1. The paragraph now reads as follows:

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

286,20 – I would just write here that the work combines elements of Weerts et al and Wood et al, and also does [b] and [c] (though take out the word 'more' – not needed, and perhaps debatable).

Please see my answer to general comment 1. The paragraph now reads as follows:
"The use of quantile regression to estimate the error distribution of river-stage forecasts has first been introduced by Woods et al. (2009) for the Lewis River in Washington State. Later, Weerts et al. (2011) applied it to river catchments in England and Wales. 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, 21 – Here and throughout the rest of the paper, please reframe the presentation of Weerts et al (2011) as the 'original' implementation focusing only on the forecast as predictor, with an 'addition' being the use of other predictors or conditioning factors – as this addition is quite clearly described in the earlier Wood et al (2009). Both work should be recognized, as they are citable/viewable by the field, and assigning the term 'original' to the second reference is misleading. Your paper, as noted above, makes other valuable contributions in addition to exploring these ideas, and does not need to work so hard to distinguish itself. Perhaps call the Weerts version the 'forecast-based' or 'W11' approach, versus multiple predictor approaches, or any other labeling that seems better.

Please see my answer to general comment 1. The paragraph now reads as follows:
"When applying QR to river forecasts, Weerts et al. (2011) transformed the forecast values and the corresponding forecast errors into the Gaussian domain using Normal Quantile Transformation (NQT) to account for heteroscedasticity. Detailed instructions to perform NQT can be found in Bogner et al. (2012)."

289, 16 – again, I object to the characterization that exceedence prob. is 'more useful' for decisionmaking than confidence intervals. This really depends on the decision, and I have actually more often, in forecast office settings, heard users ask about confidence than risk of exceedence, though again, it depends on the use. There is no reason to argue this point in the paper. Both uses of the uncertainty are valuable, and I support the authors focusing on the risk

of exceedence predictand, and stating that is 'also important' or even 'more useful for some users'. But the assertion that it is somehow categorically more useful is needlessly provincial, and can be removed.

Please see my answer to general comment 3. I changed this sentence to:

"First, to be able to optimize model performance it is best to choose a single measure."

Section 2.3 – as per earlier comments, suggest retitling this 'Inclusion of additional independent variables'. Please reference Wood et al (2009) as described earlier in recognizing the value of including rise rate and lead time as variables (this can be done obliquely, eg, ". . .as noted earlier, rise rate and lead time have been previously shown to be informative independent variables. We assess these factors as well as . . ." etc. Also, please give more detail (ie, an update of equations 1 &/or 2) to show mathematically how the additional predictors were included.

The section now reads:

"2.3 Identifying the best-performing sets of independent variables

The challenge is to identify a well-performing set of predictors that is both parsimonious and comprehensive. Wood et al. (2009) found rate of rise and lead time to be informative independent variables. Weerts et al. (2011) achieved good results using only the forecast itself as predictor. Besides these variables, the most obvious predictors to include are the observed water level 24 and 48 hours ago, the forecast error 24 and 48 hours ago (i.e., the difference between the current water level at issue time of the forecast and the forecast that was produced 24/48 hours ago), or the time of the year, e.g., using month or season as categorical predictors. Additional potential independent variables are the water levels observed up- and downstream at various times, the precipitation upstream of the catchment area, and the precipitation forecast. However, requesting the corresponding precipitation and precipitation forecast requires an extensive effort or direct access to the database at the National Climatic Data Center (NCDC).

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

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_τ – configuration coefficients; forced to be zero if the predictor is excluded from the joint predictor that is being studied.

Table 1: Joint predictors. "

293,16 – again, this is a needlessly narrow view, as what aspects of the forecast PDF are required entirely depends on the decision model to which the forecasts may be input. For hydropower optimization, for instance, the full PDF of the forecast would be desired, and is 'decision relevant' input. I think all but the last sentence of this paragraph should be removed, and the remaining sentence added to the preceding paragraph.

Please see my answer to general comment 3. That motivation was superfluous in the method section anyways, so I shortened the sentence:

"Then, we calculate the probability with which various water levels (called event thresholds hereafter) will be exceeded."

294,7 – (1) this is clearly quite a lot of work (which would lessen its operational applicability), and a somewhat brute force approach to determining the best functions. Can the authors suggest any more expedient alternatives to the more or less 'trial and error' search for the best predictor combinations? Is there an analogue to 'stepwise regression' here, perhaps? In stepwise fit approaches to MLR, there is typically a stopping criterion that discourages the addition of new predictor variables – would any similar measure be useable here? (2) Later comments in Section 3.2.2 suggest that there may be overfitting with larger predictor sets. (3)Overall, the results presentation is quite long, although the figures do support a range of conclusions of the paper, and most are of interest. I suggest the authors look for chances to remove a few of the figures

and/or tables, which may be overkill, especially if they jointly support a conclusions, but I leave it up to them.

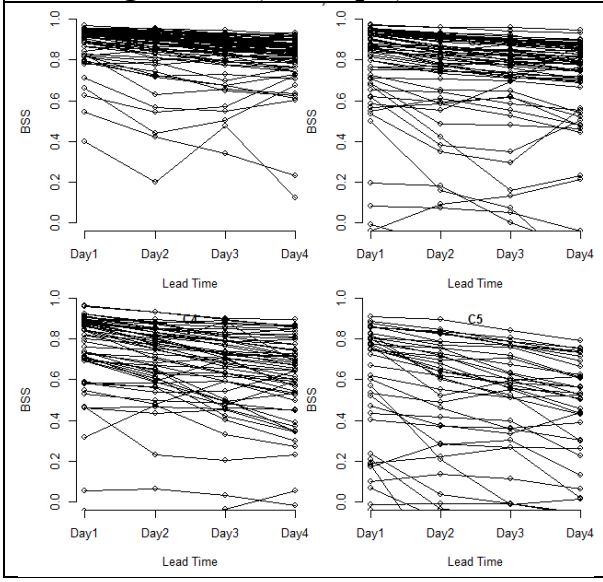
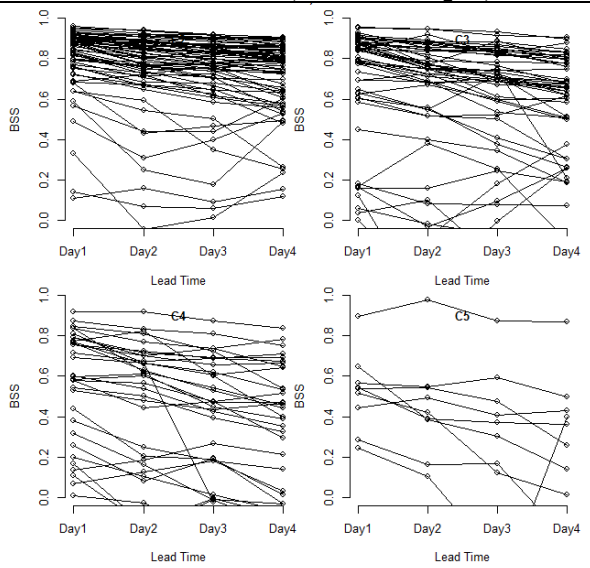
(1) Yes, this is definitely a brute force approach. Of course, stepwise regression would be a much more elegant method. However, we felt that applying stepwise regression to QR in a mathematically responsible way would require too much of our time and resources. Additionally, we are not aware of a R-package or example for stepwise QR. Considering the costs and the benefits of figuring implementing stepwise QR, we felt that we could get a good idea of how different sets of independent variables compare by using the theoretically much simpler brute force approach. Our ambition was to improve the application of QR to estimating forecast errors, rather than further developing QR itself. Stepwise QR would probably warrant the subject of a stand-alone paper, rather than be a detail in a paper on river forecasting. We included this suggestion in the section "Further Work":

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

(2) Regarding overfitting, especially for extreme events, I added the following text in section 3.2.2.:

"For moderate and major flood stage, combinations with fewer independent variables rank higher on average. The most likely explanation is that extreme events like major and moderate flood stage are infrequent. After all, major flood stage equals 90^{th} to 100^{th} 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.

Training dataset (in-sample)	Verification dataset (out-of-sample)
	

(3) Finally, please refer to my answer to the general comment 4.

298, 7 – (1) Please provide greater insight into why the inclusion of the forecast itself might degrade the performance of the post-processed forecasts. (2) Elsewhere, findings that, eg, more variables lead to worse performance at higher stages, also bear more physical explanation. What aspect of the variables could make them damaging to the high threshold models? (3) Also, it's not entirely clear that figures 11-12 support the assertion that "Without a transformation into the normal domain, the forecast does not provide a lot of information for the QR model" – giving metrics of these relationships (r^2 for instance) may help show that in fact, they are significantly different with normalization. There is a lot of scatter in both figs 11 & 12.

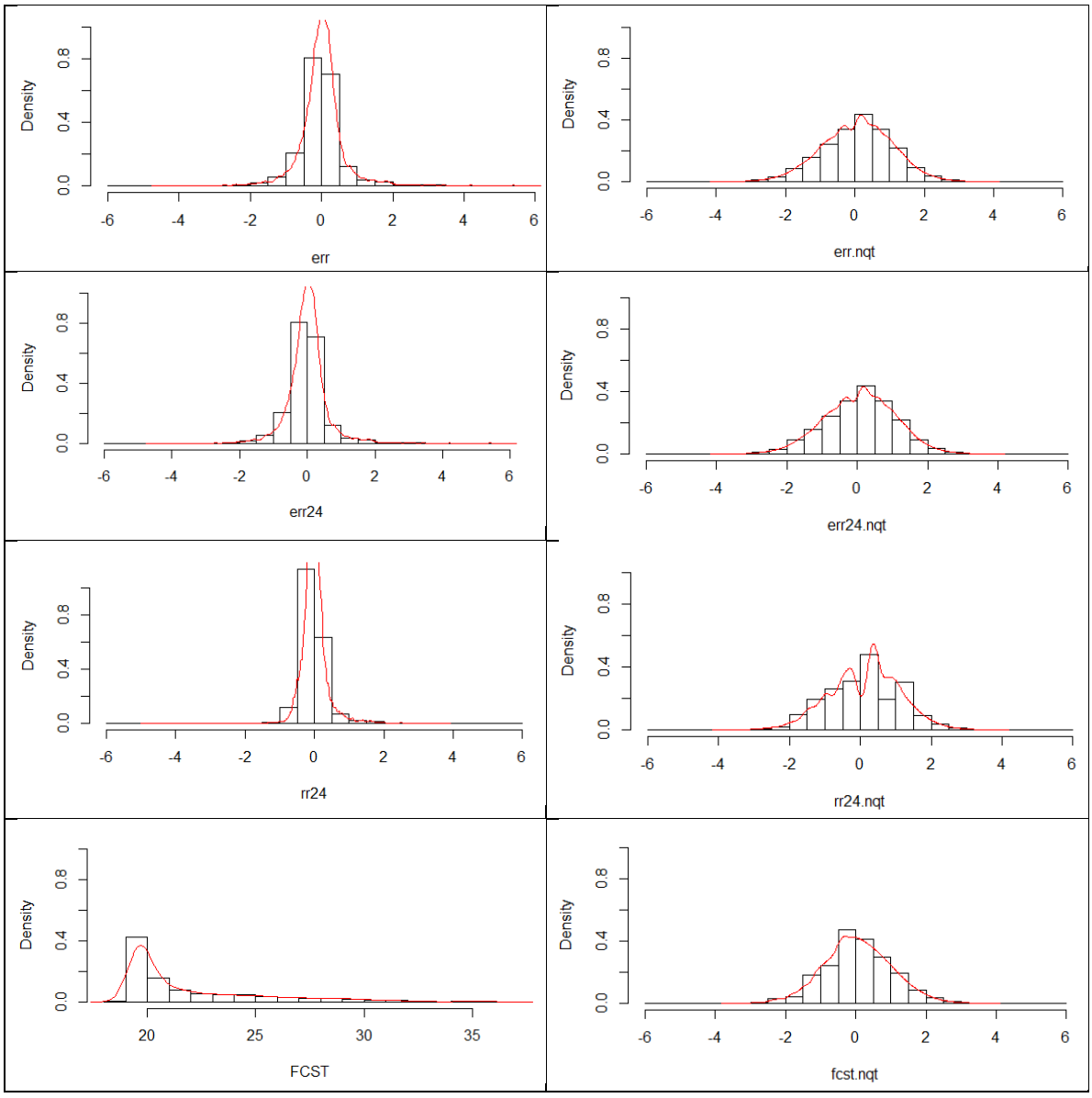
(1+3) I clarified the explanation a bit:

> "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. "

The variables other than the forecast have very similar distributions (see plots below). Mapping them onto normal distributions (i.e., "smearing them out more equally"), removes the

bit of trend that is visible in the untransformed scatterplots. However, the forecast is very differently distributed. Transforming the data does obviously not change the physical relationship between these variables, but it does change the statistical relationship bringing out a relationship between forecast and forecast error that was not seen before.

Interestingly, Weerts et al. originally transform the forecast to account for heteroscedasticity, but it turns out that quantile regression would have been much more difficult, if not impossible, without NQT. So accounting for heteroscedasticity made the approach possible at all.



Showing the R2 of the relationships would not cover what I am trying to say. Even after NQT, linear regression would not detect a correlation between forecast and forecast error. But there are trends in the percentiles. They “stick out” from the origin in a fan-like fashion.

And yes, there is a lot of scatter in these scatterplots. This partly explains the mediocre performance of the forecast as shown in Figure 13-17.

(2) Please see my response to comment 294,7 (2).

300,7 – I may be misinterpreting the figures (19,20), but it appears that length of record does matter (longer is better) somewhat more than the authors suggest, and more for the lower thresholds, which is surprising – I'd think those were better represented in any length record than high extremes, given the typical skew of flow distributions. Please comment or provide a more nuanced assessment.

I was not referring to Figure 19 and 20 here, but rather to the regression summarized in Table 4 (now Table 2). That regression generalizes the results that were illustrated in Figure 19 and 20 for just two gages. The size of the dataset was the only independent variable that was not statistically significant in this regression, meaning that the size of the dataset therefore has at the very least much less impact than gage, lead time, and event threshold etc. on forecast performance. However, your comment is still valid: Event thresholds are statistically significant in that regression, meaning that forecasts for low thresholds perform less well. As a likely explanation, I mention that the forecast error is very small for low water levels, so that there is little variability to run a regression on. You are right though, that forecast for low event thresholds seem to be particularly disappointing when the training set was short. While I did not explicitly study these interactions, I qualified my statements a bit. Here are the following relevant, updated text experts:

> "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."

301,22 – as per earlier comments, Wood et al (2009) preceded this study in the American context, and further argued for and demonstrated the use of the 'additional' variables of both river rise and lead time. Please adjust text appropriately.

Please see my answer on general comment 1. I deleted that sentence and updated the Conclusions section accordingly. For the sake of brevity, I did not copy the whole Conclusions section into this letter. Please refer to the new version of the paper.

301,26 – Instead: "This work confirms a prior finding that including additional predictors such as rise rates in the past 24 and 48 h benefits the resolution of the resulting probabilistic forecasts. In the first comprehensive assessment of various combinations of. . ., we found that . . . "

This paragraph 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 (2009) finding that QR error models should be a function of rate of rise and lead time. The configuration with the forecast as the only independent variable, as studied by Weerts et al. (2011), produced estimates with high reliability. Including the other four predictors mentioned above mainly increases the resolution."

302,10 –It's inaccurate to call these 'the new independent variables' as rise rate was used earlier.

Updated:

"When forming a joint predictor, the independent variables rates of rise and forecast errors do not combine well with the forecast itself."

302,14 – it's not clear why these variables do not lend themselves to transformation – please be more specific and speculate as to why you are finding this. Are they distributed such that the transformation reduces their correlation with the predictand? It's an interesting result, but not intuitive why it should be.

Also, see my answer to comment "298, 7".

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

Response to Reviewer #2

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 Strategies 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 independent variables would be useful, but that that data is hard to come by.

(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 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 exceedence or non-exceedance. Likewise, not all users are interested in extreme events per sé. This comes back a couple of times in the paper.

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*
○ *'configuration' rather than 'model' (to avoid confusion withunderlying 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, youmay want to consider omitting it altogether.

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 availability of data is often the reason why I chose certain model configurations. I want to make clear, that the data IS accessible, if anybody wishes to continue this study, but that 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 RFC produced "outlooks"*

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 exceedence probabilities only. Why not, however, use a range of verification metrics? See, for example, some of the recent Brown and Seo papers as well as some of my own work (where the verification approach was inspired on the Brown/Seo papers).

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).

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

Okay, I changed that.

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 given there. I place those descriptions there, because otherwise the reader has to go back to the Method section. I thought that given the brevity of the 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.

"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 re-visit 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. Coincidently, 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 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 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:

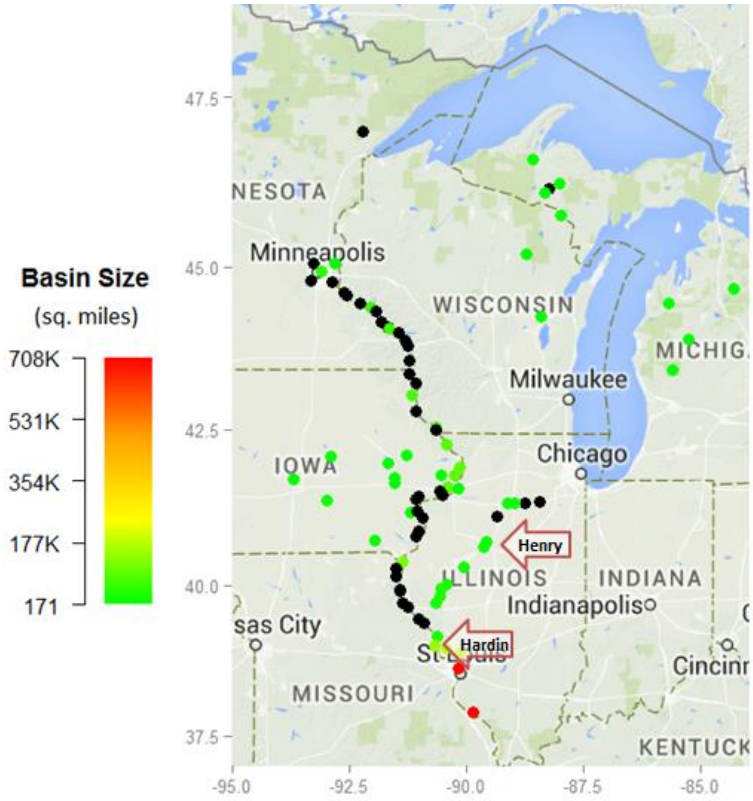
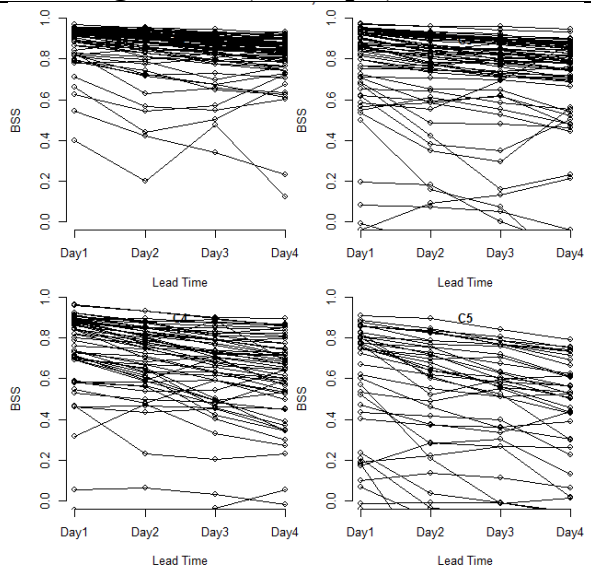
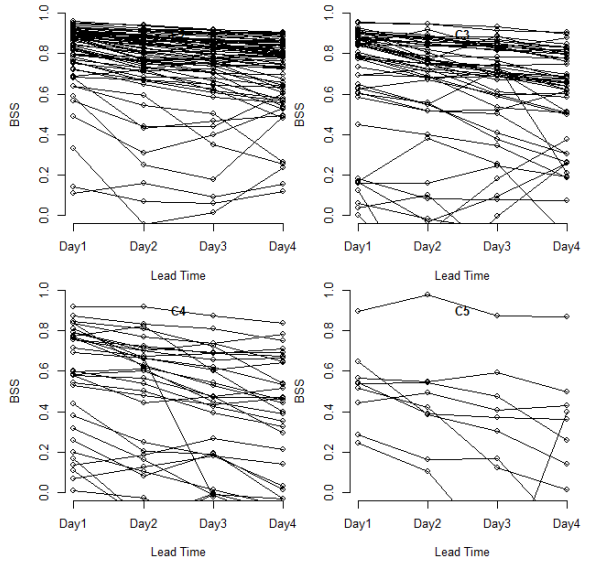


Figure 3: River gages for which the North Central River Forecast Centers publishes forecasts daily. Henry (HYNI2) and Hardin (HARI2) are indicated by the upper and lower red arrow respectively. For gages indicated by black dots the basin size is missing.

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.

Training dataset (in-sample)	Verification dataset (out-of-sample)
	

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 on the least amount of data possible. That would not only save computation time, but alleviates the problems of non-stationarity as a consequence of climate variability and climate change and human intervention. I think, if possible stationarity 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.

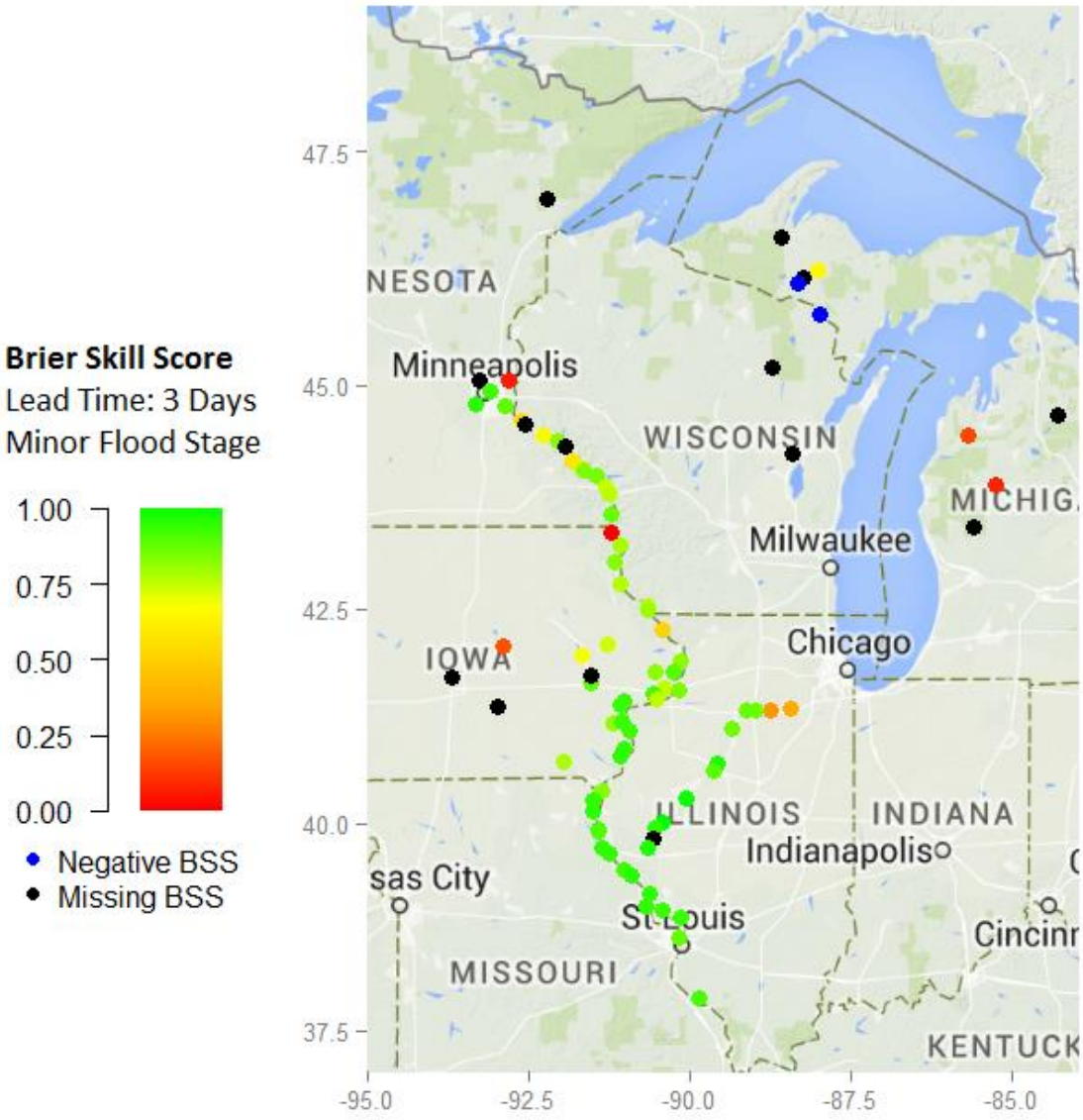
…

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 figures 10 and 11, this can be done for the vertical axes also. In R: par(mar = c(.5,0,0,0)) and then plot(…, xaxt="n") for plots where you can omit horizontal axis.

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: "The National Weather Service has chosen to quantify the most significant sources of uncertainty using ensemble techniques (Demargne et al., 2013)."

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 to define our configurations."

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

Correct. Changed 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: "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 forecast output is the exceedance probability. The performance measure only evaluates that final output.

285,14: Can you substantiate that claim with evidence or a reference?

I removed this claim throughout the paper.

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

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

286,1 & 286,6:

Reacting to a comment by the 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 performance measure, and discusses the performed analyses and data."

286,19: Omit "the".

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 do think that *effectively* QR predicts a percentile while OLS predicts a mean. In any case, I find that 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 valuable for many applications, not just hydrology.

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 first reviewer, I omitted all footnotes.

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

True. But the R-package is based on 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 accordingly.

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 paper.

292,9: "additional potential independent variables"

Changed.

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, changed.

293,5: at the river AT LOCATION X exceeds

Good point. Added.

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_τ – 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 percentile 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.

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

Changed.

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 BSSs 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 in Figure 21 (now Figure 23) and as I describe in the Data section, half of the gages is situated along the Mississippi and the Illinois River.

(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.

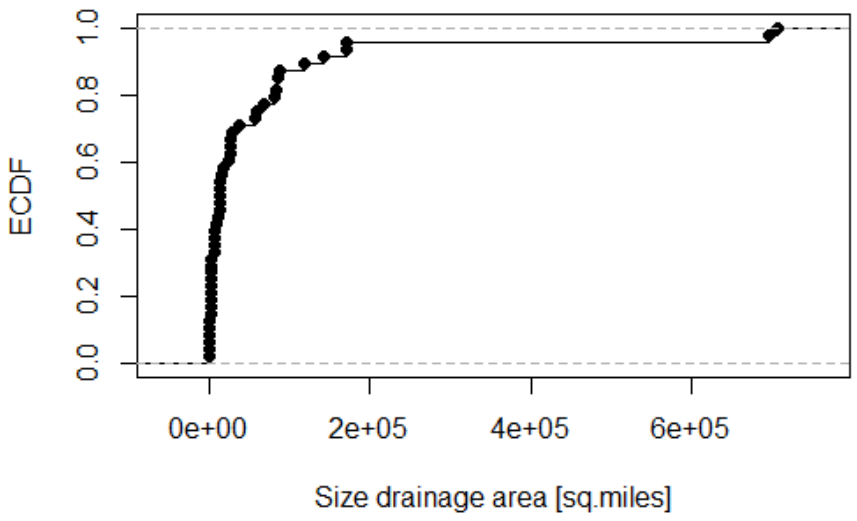


Figure 4: Empirical cumulative density function (ecdf) of sizes of drainage area for the river gages that are being forecasted daily by the NCRFC.

295,9: upstream of

Added the "of".

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 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 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., 10^{th}, 25^{th}, 75^{th}, 90^{th} 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 the paper.

297,9: flatter?

Yes, changed.

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

New sentence: "This suggests that the further out one is forecasting, the more important it 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 your specific comment 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 however many years of archived forecasts were available in that year, i.e., the forecasts for 2005 produced by a model trained on less data than those for 2013. Then, the BSS for that year (e.g., 2005 or 2013) was computed."

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. To make this point 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 the stationarity assumption. Stationarity would be much more important, if I needed twenty years of data to produce a skillful forecast. The first few of those twenty years are likely to be less representative of the coming year. Think for example of progressing urbanization. See also my answer to your specific comment 11.

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 10^{th} 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 reliabiliby.

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 'satisfatorily'

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 uncertainty. Added "forecast".

*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 always limited by the scarcity of data. 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.

Done.

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

Called "Joint Predictors" 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 various 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.

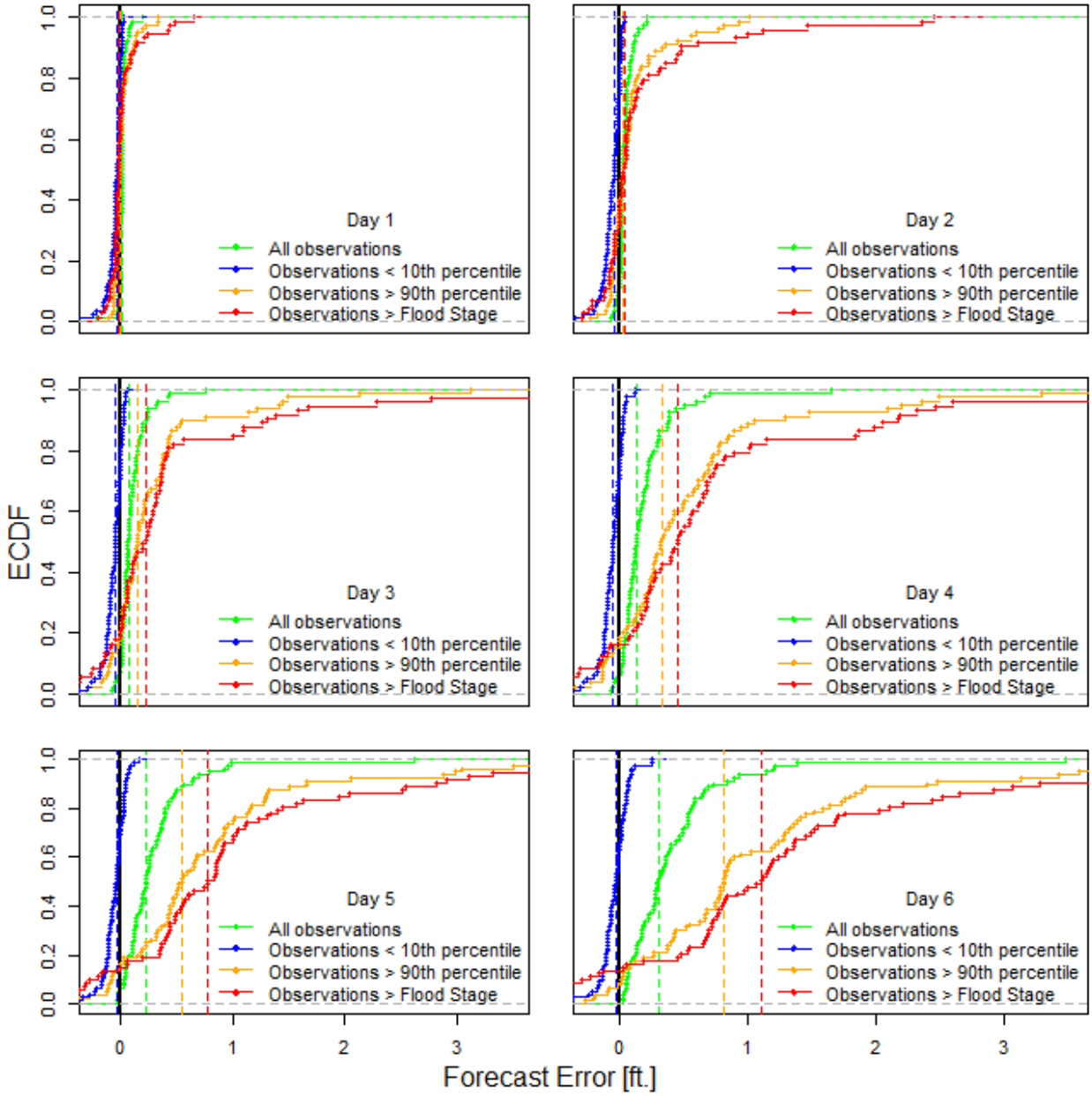


Figure 6: Empirical cumulative distribution function (ecdf) of forecast error at 82 river gages for six lead times. Vertical lines show the median forecast error of the corresponding subset.

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].

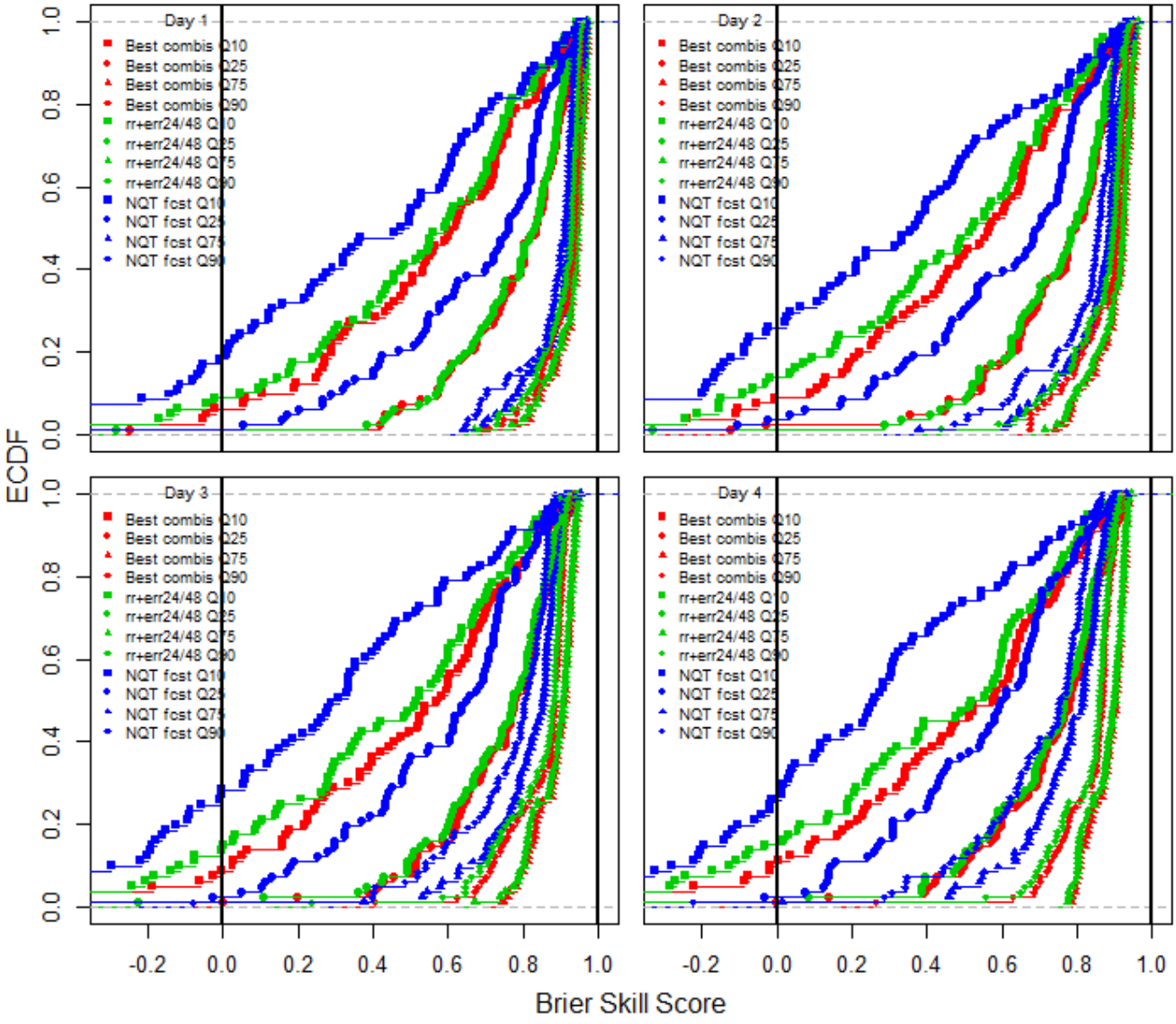
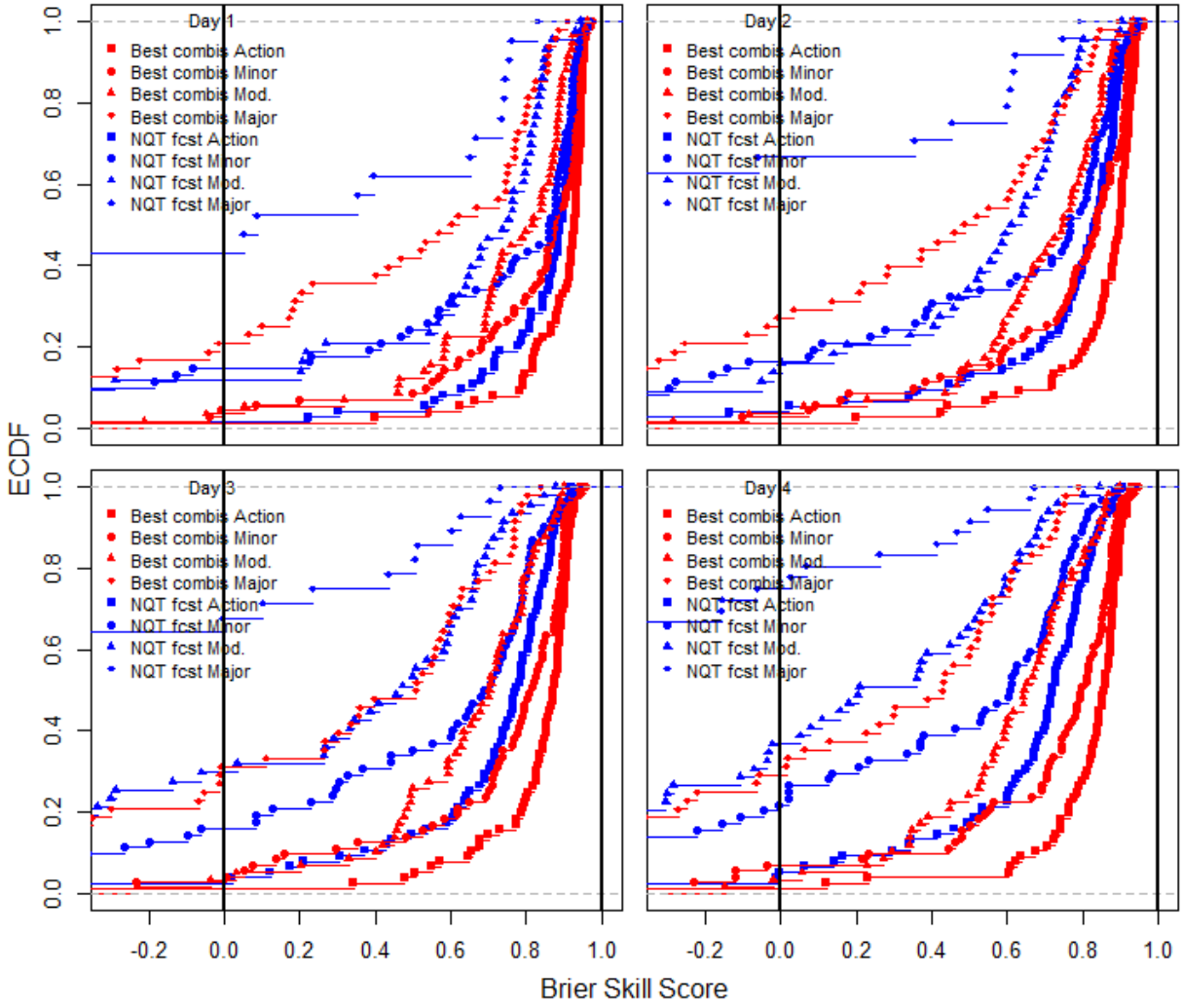
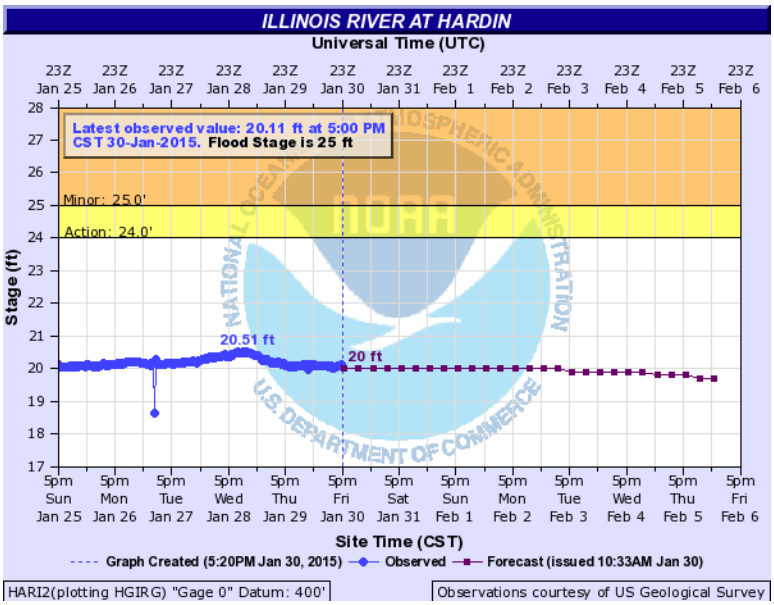


Figure 19: Empirical cumulative density functions of three QR configurations predicting exceedance probabilities of the Action, Minor, Moderate, and Major Flood Stage: the configuration using the transformed forecast as the only independent variable [NQT fcst]; the best performing combination for each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



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

Omitted.

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

Omitted.

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 it is:

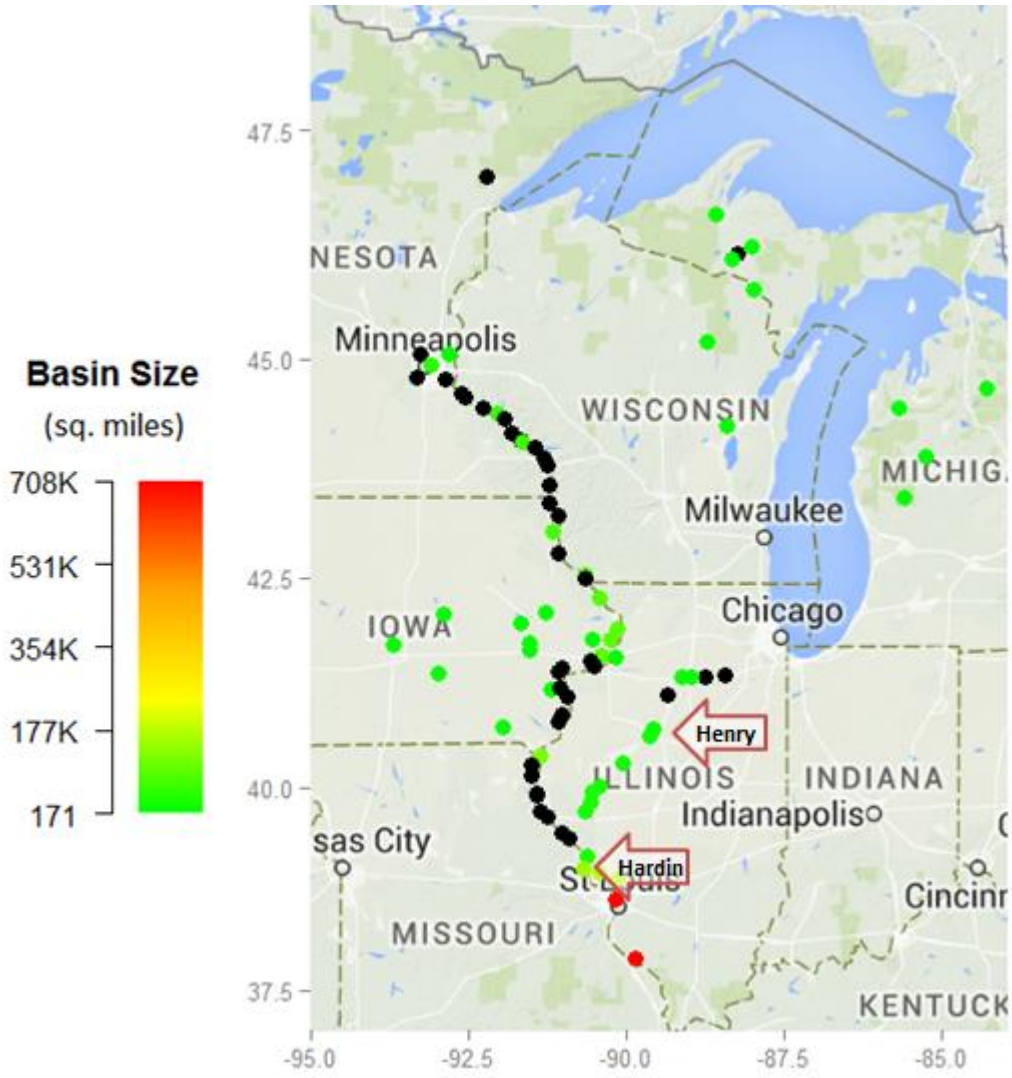


Figure 3: River gages for which the North Central River Forecast Centers publishes forecasts daily. Henry (HYNI2) and Hardin (HARI2) are indicated by the upper and lower red arrow respectively. For gages indicated by black dots the basin size is missing.

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.

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

Did so for all figures.

322: (1) omit repetitive labels where possible; (2) would also recommend zooming in on coulds, 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 cropped 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.

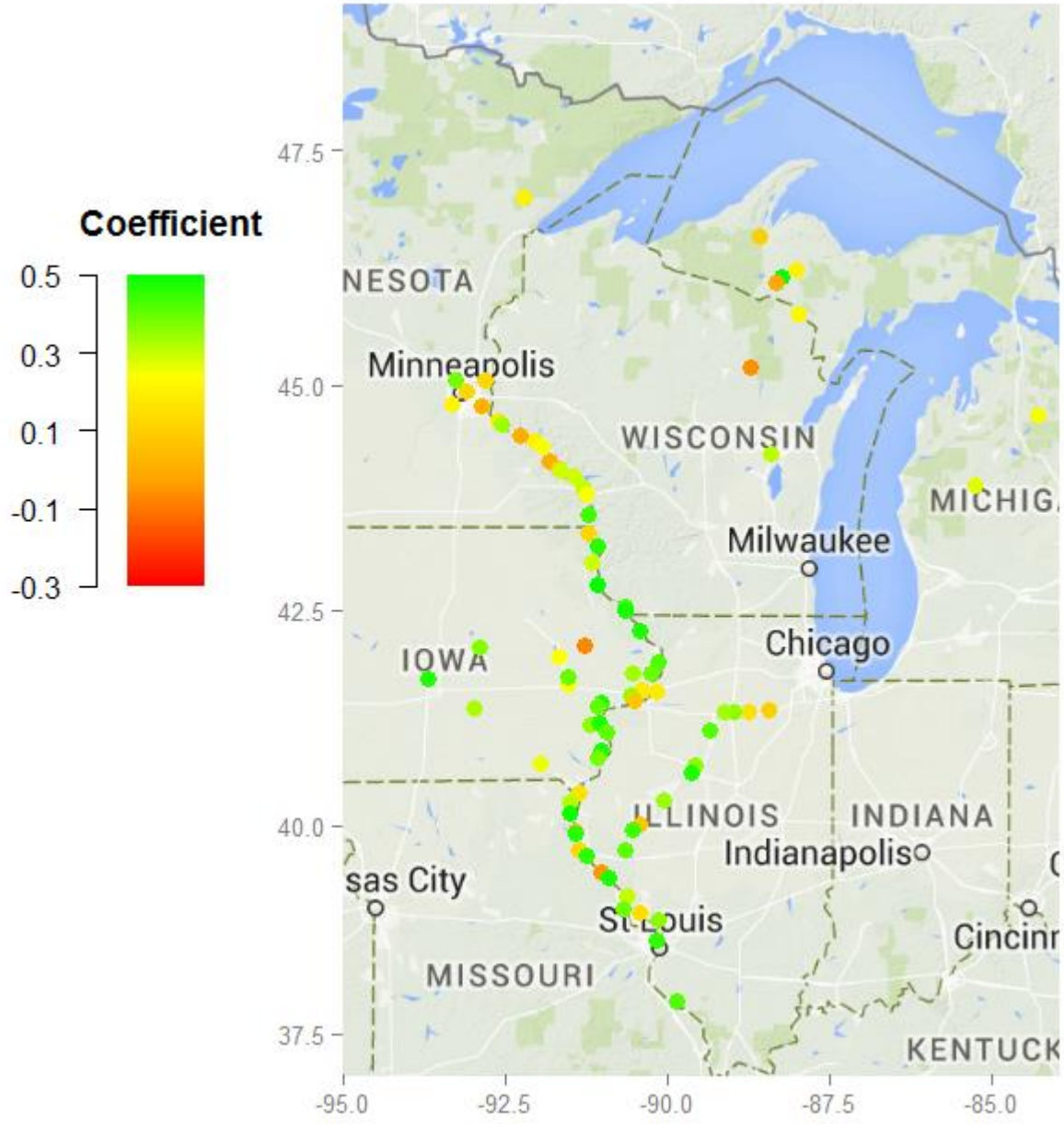


Figure 23: Geographical position of rivers. Colors indicate the regression coefficient of each station with the Brier Skill Score as dependent variable.

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

~~Article Submission~~Revised Article *Hydrology and Earth System Sciences*

Title:

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

Authors:

Frauke Hoss, Paul S. Fischbeck

Affiliation:

Carnegie Mellon University

Department of Engineering & Public Policy

5000 Forbes Avenue

Pittsburgh, PA 15213

Corresponding Author:

Frauke Hoss: fraukehoss@gmail.com

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

Abstract

This study ~~further develops the method of~~applies quantile regression (QR) to the prediction of~~predict~~ flood stage exceedance probabilities ~~of flood stages by post-processing forecasts~~based on post-processing single-value flood stage forecasts. A computationally cheap technique to predict forecast errors is valuable, because many national flood forecasting services, such as the National Weather Service (NWS), only publish deterministic single-value forecasts. ~~Using data~~The study uses data from the 82 river gages, for which the ~~National Weather Service's~~NWS' North Central River Forecast Center issues forecasts daily. ~~, this is the first QR application to U.S. American river gages.~~ Archived forecasts for lead times up to six days from 2001-200~~0~~13 were analyzed. ~~Earlier implementations of QR used the forecast itself as the only independent variable .~~ Besides the forecast itself, t~~T~~his study ~~adds~~ uses the ~~rise rate~~rate of rise of the river stage in the last 24 and 48 hours and the forecast error 24 and 48 hours ago ~~to~~ as predictors in ~~the~~ QR ~~model~~configurations. ~~Including those~~When compared to just using the forecast as independent variable, adding the latter four ~~variables~~ predictors significantly improved the forecasts, as measured by the Brier Skill Score (BSS). Mainly, the resolution increases, as the forecast-only ~~original~~ QR ~~implementation~~ configuration already delivered high reliability. Combining the forecast with the other four ~~variables~~ predictors results in much less favorable BSSs. Lastly, the forecast performance does not strongly depend on the size of the training

dataset, but on the year, the river gage, lead time and event threshold that are being forecast. We find that each event threshold requires a separate ~~model~~ configuration or at least calibration.

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

1 Introduction

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

There are two types of approaches to ~~quantify~~ estimate forecast uncertainty (e.g., Leahy, 2007; Demargne et al., 2013; Regonda et al., 2013)~~(e.g., Leahy, 2007; Demargne et al., 2013; Regonda et al., 2013)~~: Those addressing ~~certain~~ major sources of uncertainty individually in the output, e.g., input uncertainty and hydrological uncertainty, and those taking into account all sources of uncertainty in a lumped fashion. Both approaches have their advantages. Modelling each source separately can take into account that the different sources of uncertainty have different characteristics (e.g., some sources of uncertainty depend on lead time, while others do not). This approach is likely to result in better performing, more parsimonious

~~model~~configurations. On the downside, ~~it~~ the approach is expensive to develop, maintain and run. As an alternative, the lumped quantification of uncertainty is a less resource-intensive approach (Regonda et al., 2013)~~(Regonda et al., 2013)~~.

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

~~As of today, only the "outlooks" produced by the Ensemble Streamflow Prediction part of the NWS River Forecasting System are probabilistic, i.e., quantify uncertainty: an exceedance curve for a period of three month and bar plots for each week of a three months period, see and . These graphs can be used to determine with which probability each river stage will be exceeded in those weeks or three-months period. Although the short-term weather forecasts for the next~~

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

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

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

The ~~Figure 1~~2: ~~Probabilistic long-term forecast as published by the NWS for Commerce, OK on 14 December 2012: Exceedance curve for three months period. (Not available for Hardin, IL). Source: http://water.weather.gov/ahps2/hydrograph.php?wfo=tsa&gage=como2~~

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

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

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

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

In contrast to ~~the~~ an ensemble approach ~~chosen by the NWS~~such as HEFS, the statistical post-processing ~~method that is further developed~~ in this paper ~~– quantile regression –~~ does not distinguish between sources of uncertainty, but studies the overall uncertainty in a lumped fashion. ~~This choice is motivated by the fact that the total predictive uncertainty, rather than its different sources, are relevant for decision-making . To further strengthen the main advantage of this method, i.e., requiring relatively little resources,~~ To make this approach useful for actors with limited resources, we exclusively use publicly available data to ~~build our models~~define our configurations.

Most previously developed post-processors to generate probabilistic forecasts share the overall set-up but differ in their implementation. ~~Explanatory~~ Independent variables such as the forecasted and observed river stage, river flow or precipitation, and previous forecast errors are used to predict the forecast error, conditional probability distribution of the forecast error or other ~~metrics~~ measures of uncertainty for various lead times (e.g., Kelly and Krzysztofowicz, 1997; Montanari and Brath, 2004; Montanari and Grossi, 2008; Regonda et al., 2013; Seo et al., 2006; Solomatine and Shrestha, 2009; Weerts et al., 2011)~~(e.g., Kelly and Krzysztofowicz, 1997; Montanari and Brath, 2004; Montanari and Grossi, 2008; Regonda et al., 2013; Seo et al., 2006; Solomatine and Shrestha, 2009; Weerts et al., 2011)~~. ~~Among others,~~ T~~t~~hese ~~method~~techniques differ ~~in their mathematical methods~~in a number of ways, including their sub-setting of data, and the output ~~metri.c.~~ Please see Regonda et al. (2013)~~(2013)~~ and Solomatine & Shrestha (2009)~~(2009)~~ for a summary of each ~~method~~technique. In a meta-analysis of four different post-processing ~~method~~techniques to generate confidence intervals, the quantile regression

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

This paper further develops one of the ~~method~~techniques mentioned above: the Quantile Regression ~~method~~ approach to post-process river forecasts first introduced by Wood et al. (2009) and further elaborated by Weerts et al. (2011)~~(2011)~~ and López López et al. (2014). ~~.~~ The Weerts~~at~~ study achieved impressive results in estimating the 50% and 90% confidence interval of river-stage forecasts for three case studies in England and Wales using QR with calibration and validation datasets spanning two years each. This paper combines elements of the studies mentioned above. ~~–~~In some aspects, our approach differs from ~~the original approach by Weerts et al. and López López et al. .~~those three studies. We predict the ~~probabilities that flood stages are exceeded~~exceedance probabilities of flood stages rather than uncertainty bounds~~.,~~ ~~because the former are more relevant to decision-making. In an attempt to balance missed alarms and false alarms, decision-makers are likely to resort to the best estimate (i.e., the deterministic forecast) rather than basing actions on the 50% or 90% confidence interval. Additionally, predicting the probability of an event corresponds with other forecasts with which users have much experience, e.g., the probability of precipitation. Morss et al. found in a survey of the general U.S. public that most people are able to base decisions on those forecasts.~~ Additionally, we are fortunate to have a much larger dataset than the three earlier studies ~~,~~consisting of archived forecasts for 82 river gages covering 11 years ~~available~~. The study does not add to the mathematical technique of quantile regression itself.

In this paper, the QR ~~method~~technique is applied to the 82 river gages of the North Central River Forecast Center (NCRFC) encompassing (parts of) Illinois, Michigan, Wisconsin, Minnesota, Indiana, North Dakota, Iowa, and Missouri.[3]

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

~~The method is further developed by demonstrating the benefit – measured by an increase in Brier Skill Score (BSS) – of including the rise rate~~rate of rise~~s~~ of water levels in past hours, and the past forecast errors ~~as independent variables into the quantile regression~~. The performance of these joint predictors has been measured and compared using the Brier Skill Score (BSS).~~.~~ ~~For extremely high water levels the variable combination has to be customized for each river gage. For those, sets of few independent variables work best. Variable combinations for other event thresholds should include as many dependent variables as possible. Using the same combination for all of them works satisfactorily. Furthermore, it is found that the forecast – the only independent variable in the original QR method – is difficult to combine with the other dependent variables. Last, the method is shown to be robust to the size of the training dataset. However, the forecast performance does vary significantly across locations, lead times, water levels, and forecast year.~~ This exercise has been repeated for various water levels and lead times. Additionally, the robustness of the resulting QR configurations across different sizes of training datasets, locations, lead times, water levels, and forecast year has been assessed.

The paper is structured as follows. The Method section ~~summarizes the additions that this paper makes to the quantile regression method introduced by Weerts et al. . It~~ reviews ~~the~~

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

~~method~~quantile regression, ~~explains the additions,~~ introduces the performance ~~metric~~measure, and discusses the ~~computations~~ performed analyses and data. The Results section first reviews the overall forecast error for the dataset. ~~It then compares the proposed method to the original quantile regression as demonstrated for river gages in Wales and England .~~ It then describes the results of identifying the best-performing set of independent variables. Finally, it discusses the robustness of the ~~proposed method~~studied QR configurations. The fourth and last section presents the conclusions and proposes further research ideas.

2 Method

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

In the following, ~~the~~ quantile regression itself and ~~,~~ the ~~proposed addition to the method~~analysis to identify the best-performing set of independent variables ~~, and the undertaken computations~~ are explained.

2.1 Quantile Regression

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

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

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

~~In its original application to river forecasts by~~ When applying QR to river forecasts, Weerts et al. (2011)~~(2011)~~ transformed~~,~~ the forecast values and the corresponding forecast errors ~~are transformed~~ into the Gaussian domain using Normal Quantile Transformation (NQT) to account for heteroscedasticity. Detailed instructions to perform NQT can be found in~~, as instructed by~~ Bogner et al. (2012)~~(2012).~~ ~~to account for heteroscedasticity.~~ Building on this study, López

López et al. (2014)~~(2014)~~ compare different configurations of QR with the forecast as the only independent variable, including configurations omitting NQT. They find that no configuration was consistently superior for a range of forecast quality ~~metrics~~ measures (López López et al., 2014)~~(López López et al., 2014)~~. To be able to combine predictors ~~variables~~ of different nature, we ~~build a model~~ based our QR configuration on untransformed ~~variables~~predictors. The reason to do so will be discussed and illustrated later (see Figure 11 and Figure 12).

~~Using the transformed data,~~ A~~a~~ quantile regression is run for each lead time and desired percentile with the forecast error as the dependent variable and the forecast and other variables as ~~the~~ independent variables.[4] To prevent the quantile regression lines from crossing each other, a fixed effects model is implemented below a certain forecast value. Weerts et al. (2011)~~(2011)~~ give a detailed mathematical description for applying QR to river forecasts. Mathematically, the approach is formulated as follows (with and without NQT):

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

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

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

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

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

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

fcst(t)	– original forecast at time t	
$V_i(t)$	– the independent variable i (e.g., the original forecast) at time t	
$V_{i;NQT}(t)$	– the independent variable I transformed by NQT at time t	
$a_{i,\tau}$, b_τ	– ~~model~~ configuration coefficients	

The second part of the equations stands for the error estimate based on the quantile regression ~~model~~ configuration for each percentile τ and lead time. In Equation 1, that was used ~~in the original QR method proposed~~ by Weerts et al. (2011)~~(2011)~~, this estimation was executed in the Gaussian domain using only the forecast as independent variable. Our study mainly uses Equation 2, i.e., it does not transform the predictors and the predictand. All quantile regressions were done using the command *rq()* in the R-package "quantreg" (Koenker, 2013).[5]

2.2 Brier Skill Score

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

We use the Brier Skill Score – first introduced by Brier (1950) – to ~~compare~~ assess ~~the different versions of the~~ QR ~~model~~configurations ~~proposed in this paper. We chose to optimize~~

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

~~our QR models based on the BSS, first introduced by Brier~~ for ~~two~~ two reasons. ~~-~~First, to be able to optimize model performance it is best to choose a single measure. ~~First, for decision-making the probability with which a certain water level, e.g., a flood stage, is exceeded is more useful than confidence intervals. Second~~Second, out of the available measures the Brier Score is attractive, because it can be decomposed into two different measures of forecast quality (see Equation 3): Reliability and resolution. The third component is uncertainty, which is a hydrological characteristic inherent to the river gage. This uncertainty is different than the forecast uncertainty that the technique studied in this paper estimates. Besides the uncertainty that can be mathematically explained, it also includes natural variability. ~~Thus~~In sum, the BS' uncertainty term~~it~~ is not subject to the forecast quality. Equation 3 gives the definition of the (de-composed) Brier Score (e.g., Jolliffe and Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009)~~(e.g., Jolliffe and Stephenson, 2012; Anon, 2014; WWRP/WGNE, 2009)~~.[6]

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

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

with BS – Brier Score

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

N – number of forecasts
K – the number of bins for forecast probability of binary event occurring on each day
n_k – the number of forecasts falling into each bin
$\bar{o}_k$ – the frequency of binary event occurring on days in which forecast falls into bin k
f_k – forecast probability
$\bar{o}$ – frequency of binary event occurring
f_t – forecast probability at time t
o_t – observed event at time t (binary: 0 – event did not happen, 1 – event happened)

The Brier Score pertains to binary events, e.g., the exceedance of a certain river stage or flood stage. Reliability compares the estimated probability of such an event with its actual frequency. For example, perfect reliability means that on 60% of all days for which it was predicted that the water level would exceed flood stage with a 60% probability, it actually does so. ~~A forecast with~~ The reliability curve for the forecast representing perfect reliability would follow the diagonal in Figure 2, ~~–~~ i.e., the area in Figure 2a representing reliability would equal zero (Jolliffe and Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009)~~(e.g., Jolliffe and Stephenson, 2012; Anon, 2014; WWRP/WGNE, 2009)~~. The configuration by López López et al. (2014)~~(2014)~~ performs well in terms of reliability. When estimating confidence intervals, Weerts et al. (2011)~~(2011)~~ achieved good results especially for the more extreme percentiles (i.e., 10^{th} and 90^{th}).

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

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

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

Jolliffe and Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009)~~(e.g., Jolliffe and Stephenson, 2012; Anon, 2014; WWRP/WGNE, 2009)~~.

A forecast possesses skill, i.e., performs better than ~~random guessing or climatology~~the reference forecast (in this case climatology) , if it is inside the shaded area in ~~b~~Figure 2b. The Brier *Skill* Score (BSS) equals the Brier Score normalized by climatology to make the score comparable across gages with different frequencies of a binary event. Equation 4 defines the BSS' decomposition into the resolution and reliability components described above (Brown and Seo, 2013). [7] ~~-~~The BSS can range from minus infinity to one. A BSS below zero indicates no skill; the perfect score is one (e.g., Jolliffe and Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009)~~(e.g., Jolliffe and Stephenson, 2012; Anon, 2014; WWRP/WGNE, 2009)~~. All measures of forecast quality were computed using the R-package "verification" (NCAR, 2014).

Equation 4: Decomposition of Brier Skill Score

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

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

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

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

~~Intuitively, more information should lead to better prediction of the distribution of the forecast error, because the regression models would be based on more data~~The challenge is to identify a well-performing set of predictors that is both parsimonious and comprehensive. Wood et al. (2009) found rate of rise and lead time to be informative independent variables. Weerts et al. (2011) achieved good results using only the forecast itself as predictor. Besides these variables, t~~T~~he most obvious ~~variables~~ predictors to include ~~besides the forecast itself~~ are the observed water level 24 and 48 hours ago, ~~the observed rise in water level in the last 24 and 48 hours (called rise rate hereafter),~~ the forecast error 24 and 48 hours ago (i.e., the difference between the current water level at issue time of the forecast and the forecast that was produced 24/48 hours ago), or the time of the year, e.g., using month or season as categorical predictors. ~~Other~~ Additional potential ~~variables~~ independent variables are the water levels observed up- and downstream at various times, the precipitation upstream of the catchment area, and the precipitation forecast. However, r~~R~~equesting the corresponding precipitation and precipitation forecast requires an extensive effort or direct access to the ~~database.these latter variables are much more difficult to gather because of the way data is archived~~database at the National Climatic Data Center (NCDC).[8]

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

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

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

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

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

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

$$+ a_{err24,\tau} * err24(t) + a_{err48,\tau} * err48(t) + b_\tau$$

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

~~**To determine which set of variables preforms best in generating probabilistic forecasts, all 31 possible combinations of the forecast (fcst), the rise rate in the last 24 and 48 hours (rr24, rr48), and the forecast error 24 and 48 hours ago (err24, err48) were tested for 82 gages that the NCRFC issues forecasts for every morning (). Based on the Bier Skill Score, a metric of forecast quality explained below, it was determined which variable combination on average and most often leads to the best out-of-sample results for various lead times and water levels.**~~ **Table 1: Joint predictors.**

2.4 Computations

The output of our QR application to river forecasts is the probability that a certain water level in the river or flood stage is exceeded on a given day, e.g., "On the day after tomorrow, the probability that the river exceeds 15 feet at location X is 60%." This is done in two steps. First, a training dataset (first half of the data) is used to ~~build~~ define one quantile regression ~~model~~configuration for ~~each~~ each ~~of the following~~ percentile~~s:~~ π = [0.05, 0.1, 0.15, … , 0.85,

0.90, 0.95] and each lead time.~~.~~ The dependent variable is the water level. As described ~~above~~in Equation 5, the forecast itself, the ~~rise rates~~rates of rise and forecast errors serve as independent variables.

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

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

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

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

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

exceedance probability for a certain water level and calculating the BSS) were done for the 10th, 25th, 75th, and 90th percentile of observed water levels and the ~~decision-relevant~~ four decision-relevant flood stages (action stage, and minor, moderate, and major flood stage) of each gage. Flood stages indicated when material damage or substantial hinder is caused by high water levels. Therefore, the flood stages correspond with different percentiles at different river gages.

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

2.5 Data

The National Weather Service (NWS) ~~issues river-stage forecasts for ~4,000 river gages every day. Such~~'s daily ~~published~~ short-term river forecasts predict the stage height in six-hour intervals for up to five days ahead (20 6-hour intervals).[~~10~~] When floods occur and increased information is needed, the local river forecast center (RFC) can decide to publish river-stage forecasts more frequently and for more locations. Welles et al. (2007)~~(2007)~~ provides a detailed description of the forecasting process.

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

For this paper, all forecasts published by the North Central River Forecast Center (NCRFC) between 1 May 2001 and 31 December 2013 were requested from the NCDC's HDSS Access System (National Climatic Data Center, 2014; Station ID: KMSR, Bulletin ID: FGUS5).[11] In total, the NCRFC produces forecasts for 525 gages. ~~().~~ For 82 of those gages, forecasts have been published daily for a sufficient number of years, and are not inflow forecasts. The latter have been excluded from the forecast error analysis because they forecast discharge rather than water level. About half of the analyzed gages are along the Mississippi River (Figure 3). The Illinois River and the Des Moines River are two other prominent rivers in the region. The drainage areas of the 82 river gages average 61,500 square miles (minimum 200 sq.miles; maximum 708,600 sq.miles). Figure 4 shows an empirical cumulative density function of drainage areas sizes.

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

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

Two river gages serve as an illustration for the points made throughout this paper. Hardin, IL is just upstream of the confluence of the Illinois River and the Mississippi River (Figure 3). Therefore, it probably experiences high water levels through backwatering, when the high water levels in the Mississippi River prevent the Illinois River from draining. Henry, IL is

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

located ~200 miles ~~(~320 km)~~ upstream of Hardin, having a difference in elevation of ~25 feet. ~~(~7.6 m).~~ The Illinois River is ~330 miles ~~(~530 km)~~ long (Illinois Department of Natural Resources, 2011),[12] draining an area of ~13,500 square miles ~~(~35,000 km^2)~~ at Henry (USGS, 2015a)[13] and ~28,700 square miles ~~(~72,000 km^2)~~ at Hardin (USGS, 2015b).[14]

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

3 Results

3.1 Forecast error at NCRFC's gages

In general, the NCRFC's forecasts are well calibrated across the entire dataset. The average error, defined as observation minus the forecast, is zero for most gages. For lead times longer than three days, a slight underestimation by the forecast is noticeable. By a lead time of 6 days this underestimation averages 0.41 feet only (~~a, a~~Figure 5a, Figure 6). Extremely low water levels, defined as below the 10th percentile of observed water levels, are also well calibrated (Figure 5b, Figure 6). ~~(b, b).~~ However, when considering higher water levels the picture changes.[15] The underestimation becomes more pronounced, averaging 0.29 feet for three days of lead time and 1.14 feet for six days of lead time, when only observations exceeding the 90th percentile of all observations are considered (Figure 5c, Figure 6). ~~(c, c).~~ When only looking at

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

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

observations that exceeded the minor flood stages corresponding to each gage,[16] the underestimation averages 0.45 feet for three days of lead time and 1.51 feet for 6 days of lead time (Figure 5d, Figure 6). ~~(Figure 6d, Table 2d).~~ However, some gages, such as Morris (MORI2), Marseilles Lock/Dam (MMOI2) – both on the Illinois River – and Marshall Town on the Iowa River (MIWI4) experience *average* errors of 5 to 12 feet for water levels higher than minor flood stage. The gages MORI2 and MMOI2 are upstream of a dam. It is likely that the forecasts performed so poorly there, because the dam operators deviated from the schedules that they provide the river forecast centers to base their calculations on.

Figure ~~6~~5: Forecast error for 82 river gages that the NCRFC publishes daily forecasts for. In anti-clockwise direction starting at the top left: (a) Average error; (b) error on days that the water level did not exceed the 10th percentile of observations; (c) error on days that the water level exceeded the 90th percentile of observations; (d) error on days that the water level exceeded minor flood stage.

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

~~**Table 2: Error statistics for the forecast error a) of the whole dataset; b) on days that the water level did not exceed the 10th percentile of observations; c) on days that the water level exceeded the 90th percentile of observations; d) on days that the water level exceeded minor flood stage.**~~

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

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

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

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

3.2.1 Frequency Analysis

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

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

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

3.2.2 Best performing combinations on average

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

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

~~[17] Across the 82 stations, action stage corresponds with water levels between the 60th and 100th percentile.~~
~~[18] Across the 82 stations, minor flood stage corresponds with water levels between the 70th and 100th percentile.~~
~~[19] Across the 82 stations, moderate flood stage corresponds with water levels between the 80th and 100th percentile.~~
~~[20] Across the 82 stations, major flood stage corresponds with water levels between the 90th and 100th percentile.~~

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

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

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

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

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

3.2.3 Brier Skill Score

Figure 13 illustrates the BSS when using the forecast as the only predictor as studied by Weerts et al. (2011). Confirming Wood et al.'s findings (2009), additionally i~~I~~ncluding the ~~rise rate~~rate of rise and forecasts errors as independent variables into the QR ~~model~~configuration improves the Brier Skill Score (BSS) significantly . ~~. illustrates the BSS when using the model as originally introduced by Weerts et al. .~~ Using the best performing ~~variable combination~~joint predictors ~~instead,~~ gives an upper bound of the BSSs that can be achieved at best. This configuration increases the mean and decreases the standard deviation ~~(,).~~(Figure 14, Figure 16). The performance improves most where all ~~model~~ configurations perform worst: at the 10th percentile. Possibly, the configurations do not perform well for low percentiles, because the dependent variable – the forecast error – exhibits very little variance at those water levels, i.e., the average error is very small (Figure 16).[21] The decrease of the BSSs with lead time also becomes considerably less with this configuration. Additionally, a~~n~~ one-size-fits-all approach was tested to investigate, whether customizing the QR ~~model~~configuration to each river gage would be worth it. In this configuration, the ~~rise rates~~rates of rise in the past 24 and 48 hours and the forecast errors 24 and 48 hours ago serve as the independent variables (combination 30). It was found that this approach returns only slightly worse results than working with the best

~~[21] Possibly, the model~~**configurations** ~~do not perform well for low percentiles, because the dependent variable – the forecast error – exhibits very little variance at those water levels, i.e., the average error is very small (~~**Figure 6: Empirical cumulative distribution function (ecdf) of forecast error at 82 river gages for six lead times. Vertical lines show the median forecast error of the corresponding subset.**

~~Table 2).~~

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

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

~~**Table 3: Mean and standard deviation three QR configurations: the original using the transformed forecast only as independent variable; the best performing combination for each river gage (upper performance limit); rise rates in the past 24 and 48 hours and the forecast errors 24 and 48 hours ago as independent variable (one-size-fits-all solution).**~~

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

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

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

Figure 16: Empirical cumulative density functions of three QR configurations predicting exceedance probabilities of the 10^{th}, 25^{th}, 75^{th}, and 90^{th} percentile: the configuration using the transformed forecast as the only independent variable [NQT fcst]; the best performing combination for each river gage (upper performance limit) [Best combis]; rates of rise in the past 24 and 48

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

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

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

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

The fact that the Brier Score can be de-composed into reliability, resolution and uncertainty allows a closer look at which improvements are being achieved by including more ~~variables~~predictors than just the forecast. ~~Figure 18~~Figure 20 shows that the ~~original~~ forecast-only QR ~~model~~ configuration as studied by Weerts et al. (2011)~~(2011)~~ has high reliability (i.e., the reliability is close to zero). The Brier Score and the Brier Skill Score mainly improve when using ~~rise rates~~rates of rise and forecast errors as independent variables, because the resolution increases. This confirms the finding by Wood et al. (2009) that QR error models should be based on rate of rise (as well as lead time). The forecast quality improves along other ~~dimensions~~ metrics as well, i.e., the areas under the ROC curves and the ranked probability skill score (RPSS) increase. The first weighs missed alarms against false alarms and has a perfect score equal to one. The latter is a version of the Brier Skill Score. While the Brier Skill Score pertains

to a binary event, the RPSS can take into account various event categories. Its perfect score equals one (e.g., WWRP/WGNE, 2009)~~(e.g., WWRP/WGNE, 2009)~~.

Figure ~~18~~20: Comparison of the ~~original~~ forecast-only QR ~~model~~configuration (i.e., only transformed forecast as independent variables) and the one-size-fits-all approach (i.e., ~~rise rates~~rates of rise and forecast errors as independent variables) using various measures of forecast quality: Brier Score (BS), Brier Skill Score (BSS), Reliability (Rel), Resolution (Res), Uncertainty (Unc), Area under the ROC curve (ROCA), ranked probability score (RPS), ranked probability skill score (RPSS). Lead time: 3 days; 75th percentile of observation levels as threshold. The left figure zooms in on the right figure to make changes in reliability and resolution better visible.

3.3 Robustness

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

Figure 21 and Figure 22 show that training datasets shorter than three years result in very low BSSs for low event thresholds (Q10) at Henry and Hardin.~~show that for those gages,~~ For the other event thresholds, it ~~does not~~barely matters ~~for~~ the BSS how many years are included in the training dataset. That is good new~~news~~s, if stationarity cannot be assumed (Milly et al., 2008)~~(Milly et al., 2008)~~, a step-change in river regime has occurred, or forecast data have not

been archived in the past. In those cases, only short training datasets are available. Only needing short time series to define a skillful QR configuration implies that the configuration parameters can be updated regularly. This way, changing relationships between predictors etc. can be taken into account.

-However, the BSS varies considerably for what year is being forecast. The forecast performance varies greatly, especially for the 10^{th} and 25^{th} percentile of observed water levels. It is likely, that a very large dataset, including more infrequent events, would improve these results. However, most river forecast centers only recently started archiving forecasts in a text-format, so that even having ten years' worth of data is an exception. To illustrate that point, the National Climatic Data Center has archived data from 2001 onwards available in their HDSS Access System. ~~22~~

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

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

A closer look at the regression coefficients (Table 2) provides interesting insights. For low event thresholds, the BSSs are much worse than for high thresholds. The QR configurations might be performing less well for low event thresholds, because the variance in the dependent variable – the forecast error – is smaller. After all, river forecasts have much smaller errors for lower water levels. The illustrative cases of Henry and Hardin, described above, indicate that using longer time series to predict exceedance probabilities of low event thresholds improves forecast performance.

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

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

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

Figure ~~22~~20: Brier Skill Score for various forecast years and various sizes of training dataset across different lead times (colors) and event thresholds (plots) for Henry, IL (HNYI2). The filled-in end point of each line indicates the BSS for the forecast year on the x-axis with one year in the training dataset. Each point further to the left stands for one additional training year for that same forecast year.

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

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

4 Conclusion

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

~~Most importantly~~When compared to the configuration using only the forecast, it was found that including ~~rise rates~~rates of rise in the past 24 and 48 hours and the forecast errors of 24 and 48 hours ago as independent variables improves the performance of the QR ~~model~~configuration, as

measured by the Brier Skill Score. This confirms Wood et al.'s (2009) finding that QR error models should be a function of rate of rise and lead time. ~~Since the reliability was already high with the original QR method as proposed by Weerts et al. ,~~ The configuration with the forecast as the only independent variable, as studied by Weerts et al. (2011), produced estimates with high reliability. Including the other four predictors mentioned above mainly~~the new configuration mainly~~ increases the resolution.

For extremely high water levels, the combinations of independent variables that perform best vary across stations. On those days, combinations of fewer independent variables perform better than those that include more. The most likely explanation is that QR configurations based on large joint predictors result in overfitting the data. In contrast to these extremely high event thresholds, larger sets of ~~variables~~ predictors work better than smaller ones for non-extreme and low event thresholds. Additionally, customizing the set of predictors to the event thresholds does not improve the BSS much. ~~a one-size-fits-all approach (i.e. the rise rates and forecasts errors as independent variables) performs satisfactorily for those cases.~~

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

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

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

The ~~proposed~~ studied QR ~~approach~~ configurations perform~~s~~ less well for longer lead times, for gages far upstream a river or close to confluences, for low event thresholds and extremely high ones. The QR ~~model~~configurations might be performing less well for low event thresholds, because the variance in the dependent variable – the forecast error – is smaller. After all, river

forecasts have much smaller errors for lower water levels. In turn, for extremely high water levels, the scarcity of data decreases the ~~model~~configuration's performance.

Future Work

This~~e~~ ~~method~~techniques~~~~ can be further developed in several ways to achieve higher Brier Skill Scores and more robustness. First, more independent variables can be added. Trials with a different ~~method~~technique, classification trees, showed that the observed precipitation, the precipitation forecast (i.e., POP – probability of precipitation) and the upstream water levels significantly improve ~~models~~forecasting performance. Presumably, this is the case, because the ~~QPF~~ forecast used in this study includes the precipitation forecast ~~only~~ for only the next 12 hours. However, currently, the precipitation data and forecasts can only be requested in chunks of a month, three chunks per day, from the NCDC's HDSS Access System.[23] For a period of 12 years, requesting such data for several weather stations[24] is obviously time-consuming; n~~.~~ot least, because the geographical units of the weather forecasts bulletins do not correspond with those of the river forecast bulletins. Upstream water levels can easily be included after manually determining the upstream gage(s) for each of the 82 NCRFC gages. To improve ~~model~~ performance at gages close to river confluences, the upstream water level of the gages on the joining river should be included as well.

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

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

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

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

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

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

Acknowledgements:

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Bröcker, J.: Estimating reliability and resolution of probability forecasts through decomposition of the empirical score, Clim. Dyn., 39(3–4), 655–667, doi:10.1007/s00382-011-1191-1, 2012.

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

Ferro, C. a. T. and Fricker, T. E.: A bias-corrected decomposition of the Brier score, Q. J. R. Meteorol. Soc., 138(668), 1954–1960, doi:10.1002/qj.1924, 2012.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Morss, R. E., Lazo, J. K. and Demuth, J. L.: Examining the use of weather forecasts in decision scenarios: results from a US survey with implications for uncertainty communication, Meteorol. Appl., 17(2), 149–162, doi:10.1002/met.196, 2010.

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

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

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

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

R-Core Team: R: A language and environment for statistical computing., [online] Available from: http://ww.R-project.org/., 2014.

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

Seo, D. J., Herr, H. D. and Schaake, J. C.: A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction, Hydrol Earth Syst Sci Discuss, 3(4), 1987–2035, doi:10.5194/hessd-3-1987-2006, 2006.

Solomatine, D. P. and Shrestha, D. L.: A novel method to estimate model uncertainty using machine learning techniques, Water Resour. Res., 45, doi:10.1029/2008WR006839, 2009.

Stephenson, D. B., Coelho, C. A. S. and Jolliffe, I. T.: Two Extra Components in the Brier Score Decomposition, Weather Forecast., 23(4), 752–757, doi:10.1175/2007WAF2006116.1, 2008.

Weerts, A. H., Winsemius, H. C. and Verkade, J. S.: Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System

(England and Wales), Hydrol Earth Syst Sci, 15(1), 255–265, doi:10.5194/hess-15-255-2011, 2011.

Welles, E., Sorooshian, S., Carter, G. and Olsen, B.: Hydrologic Verification: A Call for Action and Collaboration, Bull. Am. Meteorol. Soc., 88(4), 503–511, doi:10.1175/BAMS-88-4-503, 2007.

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

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

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

Tables

Table 1: ~~Variable Combination~~Joint predictors

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

fcst = forecast; rr24, rr48 = ~~rise rate~~rate of rise in the past 24 and 48 hours;
err24, err 48 = forecast error 24 and 48 hours ago
The forecast error equals the difference between the current (i.e., at issue time of the forecast) water level and the forecast that was produced 24/48 hours ago.

Table 2: Error statistics for the forecast error a) of the whole dataset; b) on days that the water level did not exceed the 10th percentile of observations; c) on days that the water level exceeded the 90th percentile of observations; d) on days that the water level exceeded minor flood stage.

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

Table 3: Mean and standard deviation three QR configurations: the original using the transformed forecast only as independent variable; the best performing combination for each river gage (upper performance limit); rise rates in the past 24 and 48 hours and the forecast errors 24 and 48 hours ago as independent variable (one-size-fits-all solution).

	Q10	Q25	Q75	Q90	Q10	Q25	
	Day 1						Day 2
NQT-fcst	0.34 (0.52)	0.65 (0.36)	0.90 (0.07)	0.88 (0.08)	0.24 (0.57)	0.59 (0.35)	0.8
Best combi.s	0.54 (0.34)	0.78 (0.18)	0.93 (0.05)	0.91 (0.06)	0.49 (0.36)	0.74 (0.19)	0.9
Rise rate 24/48 +error 24/48*	0.49 (0.41)	0.77 (0.18)	0.92 (0.05)	0.93 (0.06)	0.42 (0.44)	0.73 (0.19)	0.9
	Day 3						Day 4
NQT-fcst	0.20 (0.61)	0.56 (0.33)	0.81 (0.10)	0.75 (0.15)	0.19 (0.55)	0.55 (0.31)	0.7
Best combi.s	0.47 (0.37)	0.74 (0.17)	0.89 (0.05)	0.85 (0.09)	0.46 (0.37)	0.73 (0.18)	0.8
Rise rate 24/48 +error 24/48*	0.40 (0.44)	0.72 (0.19)	0.88 (0.06)	0.84 (0.11)	0.39 (0.43)	0.71 (0.20)	0.8
	Action	Minor	Moderate	Major	Action	Minor	M
	Day 1						Day 2
NQT-fcst	0.81 (0.27)	0.42 (1.12)	0.38 (1.02)	-0.80 (2.07)	0.68 (0.59)	0.41 (0.90)	0.2
Best combi.s	0.86 (0.26)	0.78 (0.27)	0.73 (0.24)	0.36 (0.66)	0.82 (0.29)	0.73 (0.28)	0.6
	Day 3						Day 4
NQT-fcst	0.67 (0.37)	0.37 (0.87)	-0.09 (1.42)	-1.69 (2.24)	0.62 (0.35)	0.22 (1.00)	-0.
Best combi.s	0.81 (0.26)	0.71 (0.31)	-0.64 (0.23)	-0.19 (0.76)	0.79 (0.26)	0.69 (0.30)	-0.

* Combination 30

Table 2: Regression results

	Coef.	St.Dev.	
Intercept	-0.206	0.031	***
Event thresholds	0.265	0.003	***
Lead Times	-0.021	0.003	***
Forecast Years			
2004	-0.266	0.020	***
2005	-0.081	0.018	***
2006	-0.125	0.017	***
2007	-0.129	0.017	***
2008	-0.203	0.017	***
2009	-0.125	0.016	***
2010	-0.140	0.017	***
2011	-0.128	0.016	***

2012	0.056	0.017	***
2013	-0.054	0.016	***
Number of Years in Training Dataset	0.001	0.001	
River Gages			***
For the sake of brevity, the 82 river gages included in the regression as factors are omitted here.			
R^2		0.26	
Adjusted R^2		0.25	
P-Values: *** – <0.001; ** – 0.01; * – 0.05; . – 0.1			

Figures

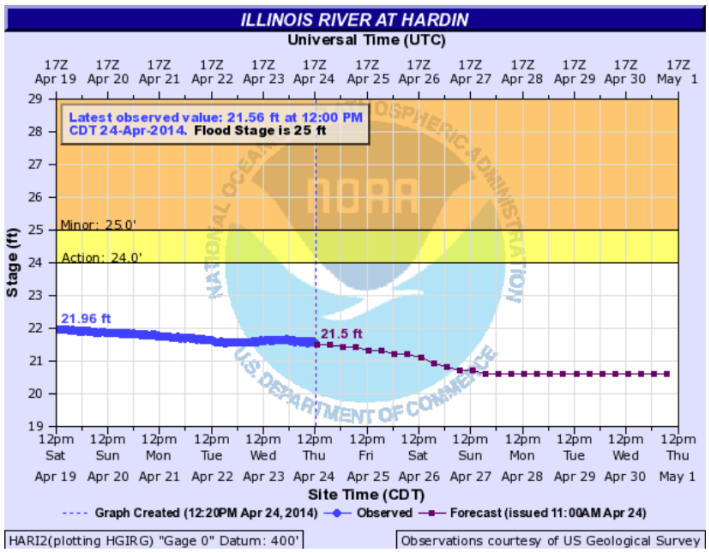
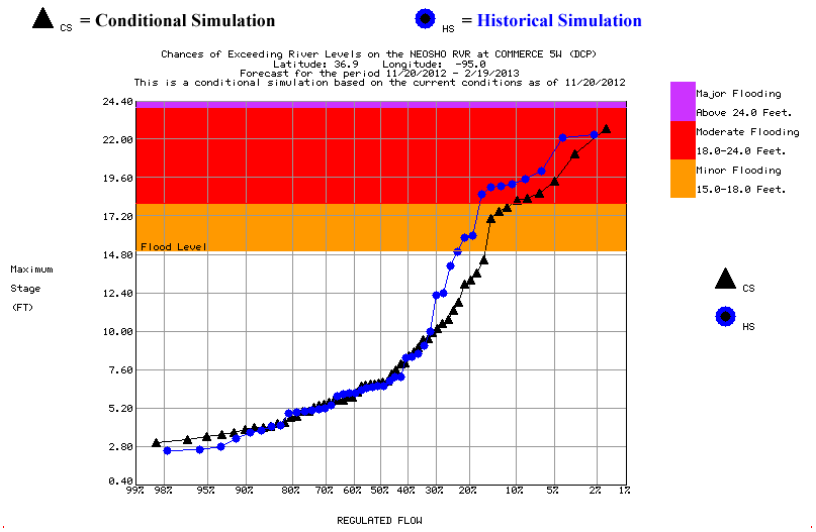
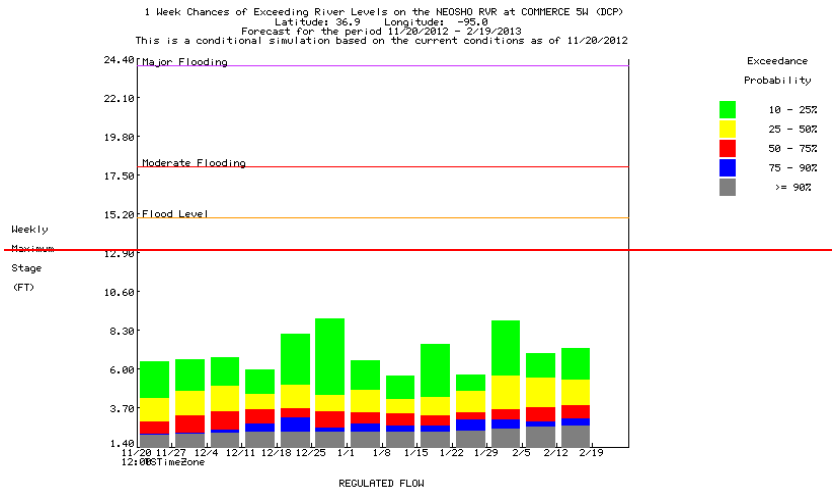


Figure 1: Deterministic short-term weather forecast in six hour intervals as published by the NWS for Hardin, IL on 24 April 2014.

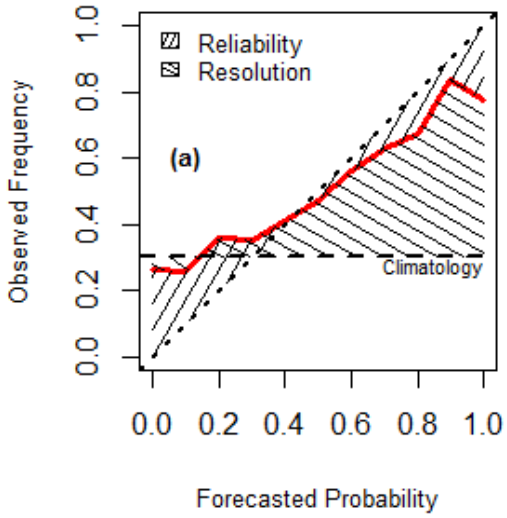
Source:http://water.weather.gov/ahps2/hydrograph.php?wfo=lsx&gage=hari2.

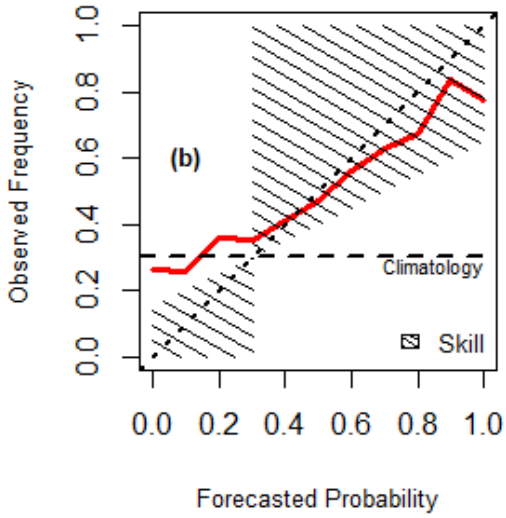


~~Figure 2: Probabilistic long-term forecast as published by the NWS for Commerce, OK on December 14th, 2012: Exceedance curve for three months period. (Not available for Hardin, IL). Source: http://water.weather.gov/ahps2/hydrograph.php?wfo=tsa&gage=como2~~



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





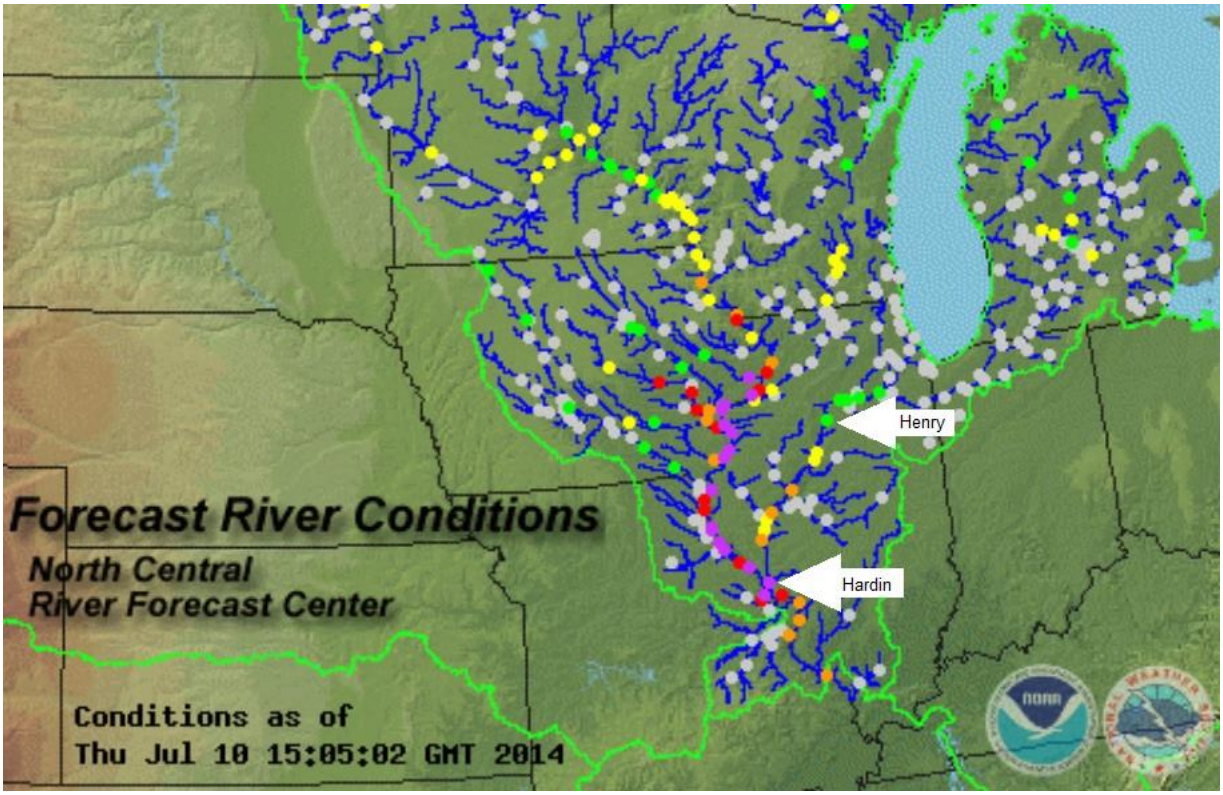


Figure 24: Theory behind Brier Skill Score illustrated for an imaginary forecast (red line): (a) reliability and resolution; (b) skill. In figure a, the area representing reliability should be as small, and for resolution as large as possible. The forecast has skill (BSS > 0), i.e., performs better than ~~random guessing~~the reference forecast, if it is inside the shaded area in the figure b. Ideally, the forecast would follow the diagonal (BSS=1). (Adapted from Hsu and Murphy, 1986; Wilson,

n.d.)~~(Adapted from Hsu and Murphy, 1986; Wilson, n.d.)~~.

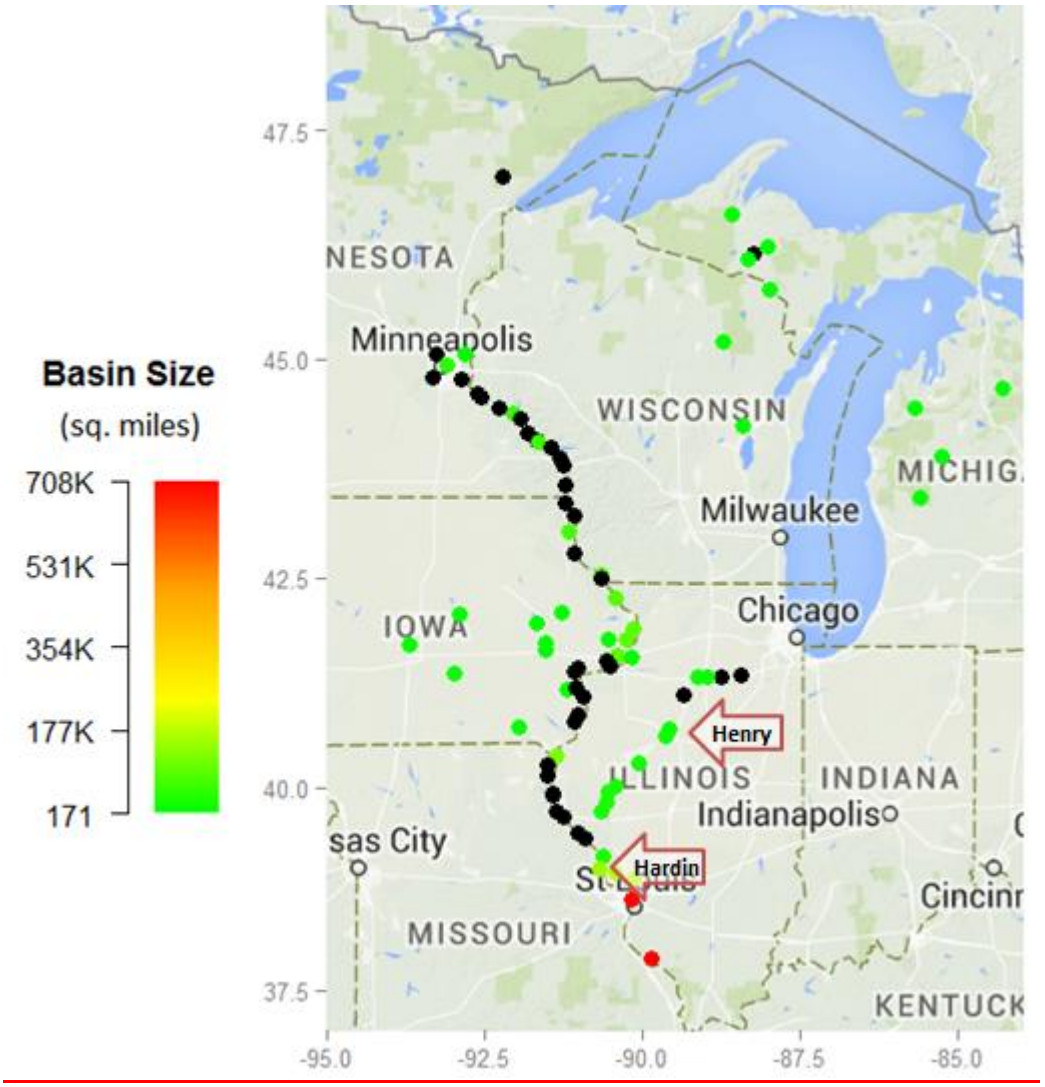


Figure 35: ~~Portion~~ River gages for which the ~~of the~~ North Central River Forecast Centers ~~river gages with~~publishes forecasts daily. Henry (HYNI2) and Hardin (HARI2) are indicated by the upper and lower red arrow respectively. For gages indicated by black dots the basin size is missing.~~Source:~~

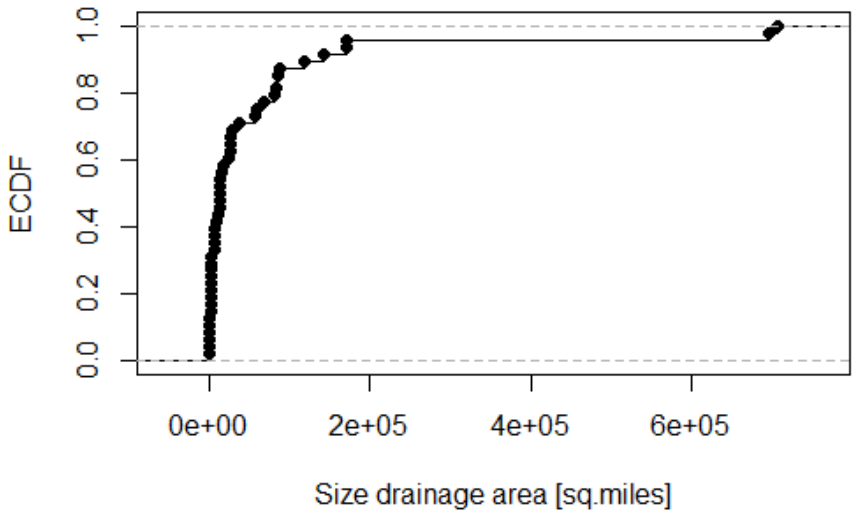


Figure 4: Empirical cumulative density function (ecdf) of sizes of drainage area for the river gages that are being forecasted daily by the NCRFC.

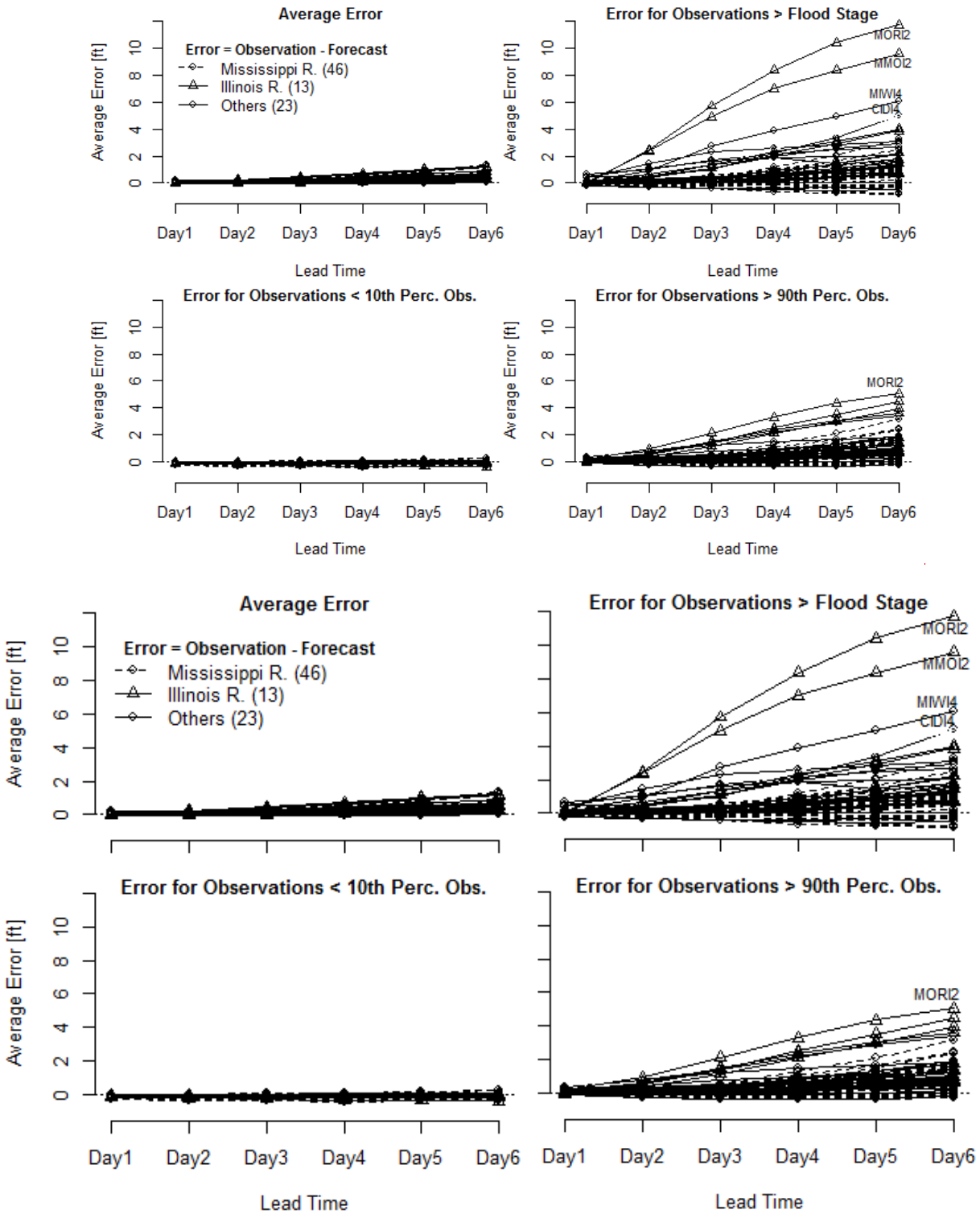

Average Error
Error = Observation - Forecast
Mississippi R. (46)
Illinois R. (13)
Others (23)
Average Error [ft]
0 2 4 6 8 10
Day1 Day2 Day3 Day4 Day5 Day6
Lead Time
Error for Observations > Flood Stage
MORI2
MMOI2
MIWI4
CIDI4
Error for Observations < 10th Perc. Obs.
Error for Observations > 90th Perc. Obs.

Figure 5~~6~~: Forecast error for 82 river gages that the NCRFC publishes daily forecasts for. In anti-clockwise direction starting at the top left: (a) Average error; (b) error on days that the water level did not exceed the 10th percentile of observations; (c) error on days that the water level exceeded the 90th percentile of observations; (d) error on days that the water level exceeded minor flood stage.

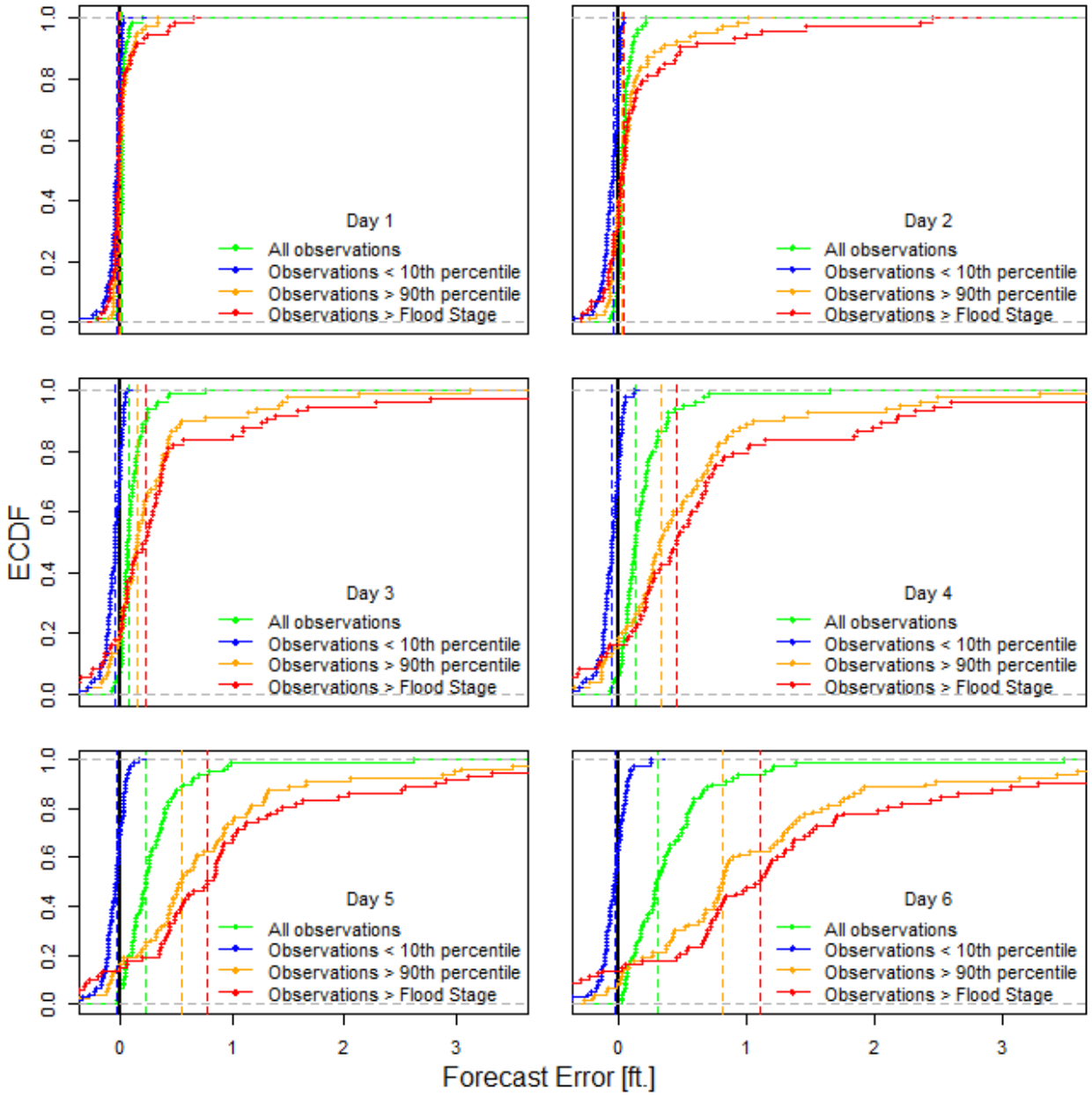
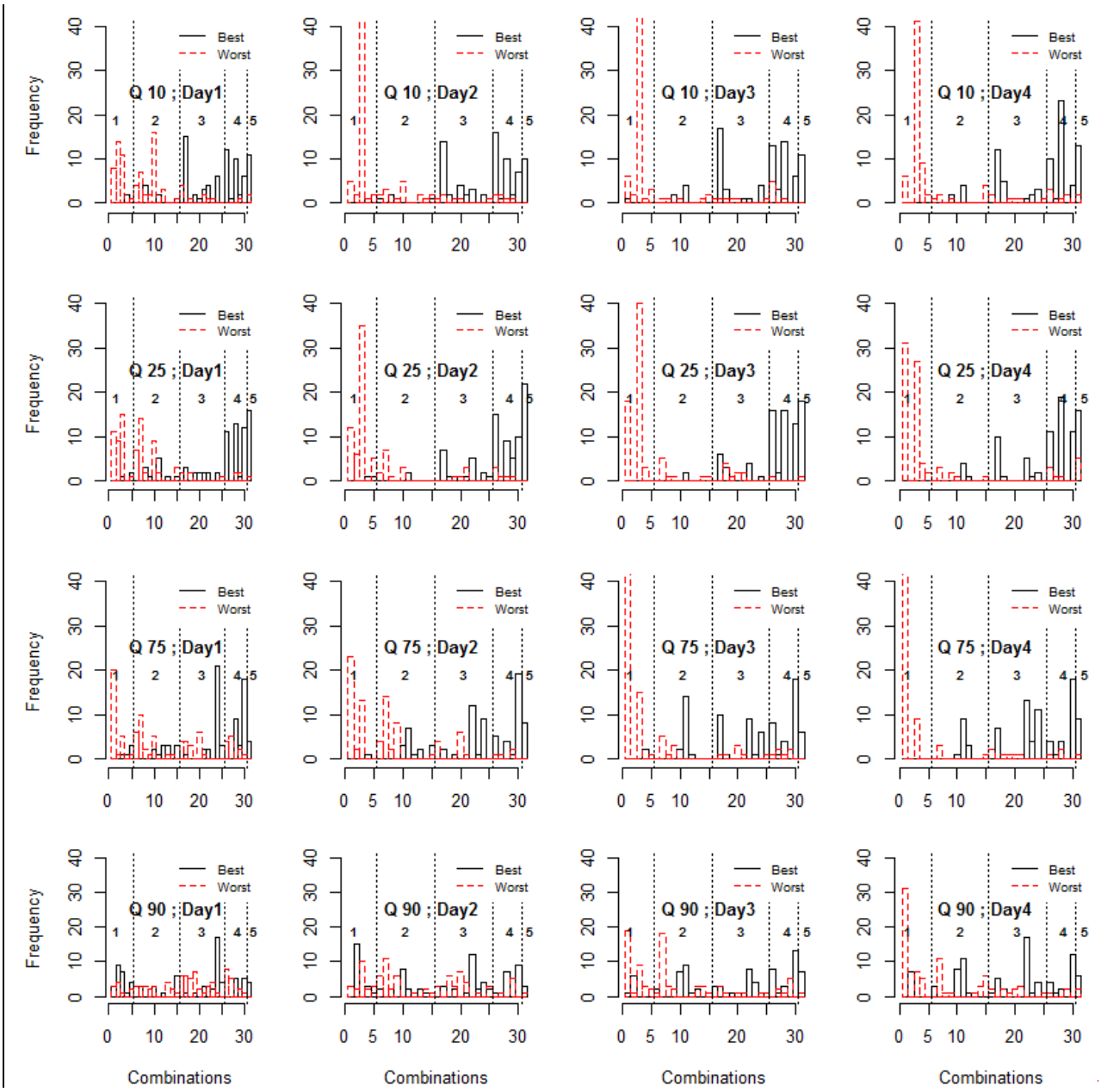


Figure 6: Empirical cumulative distribution function (ecdf) of forecast error at 82 river gages for six lead times. Vertical lines show the median forecast error of the corresponding subset.


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

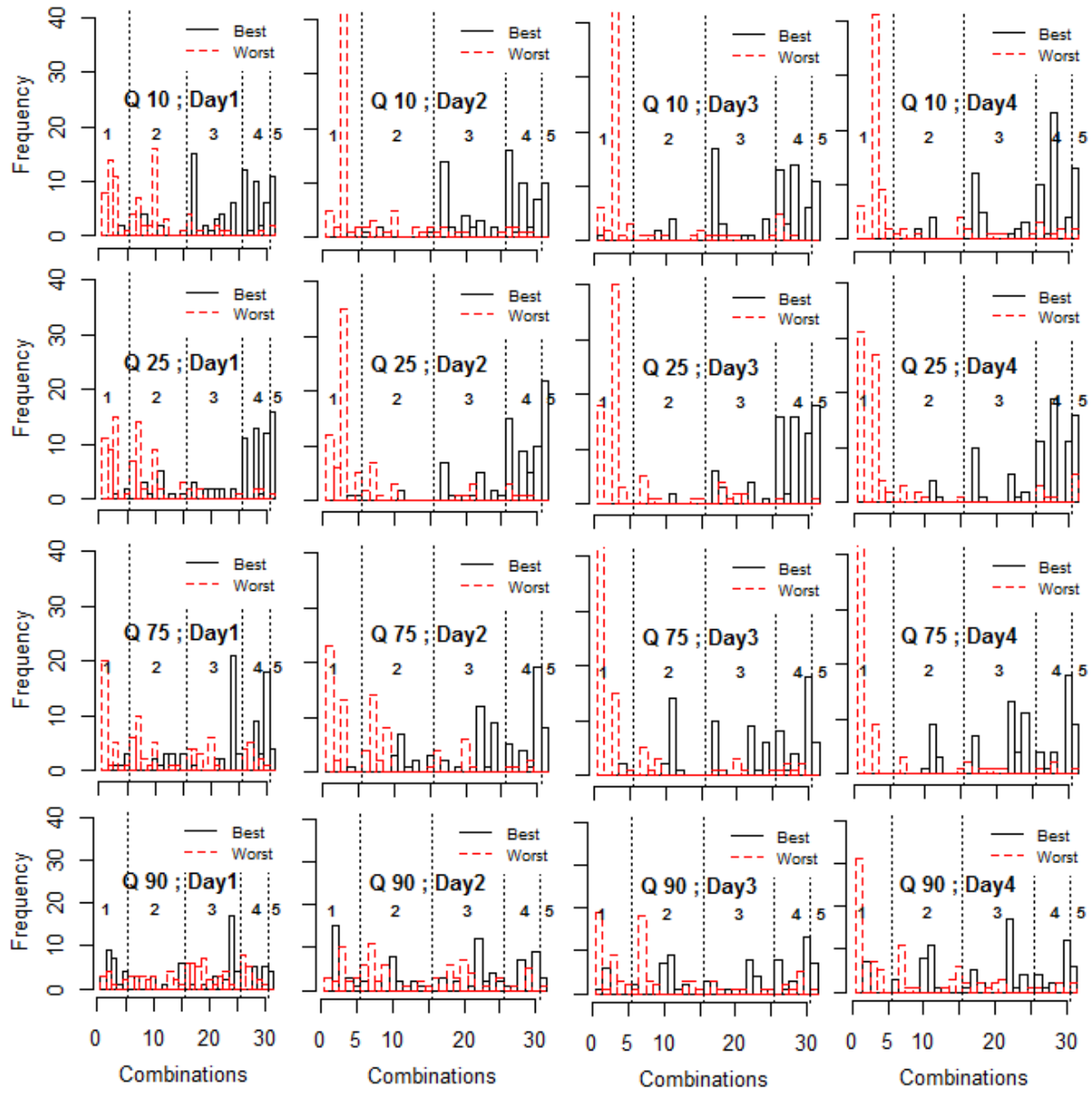
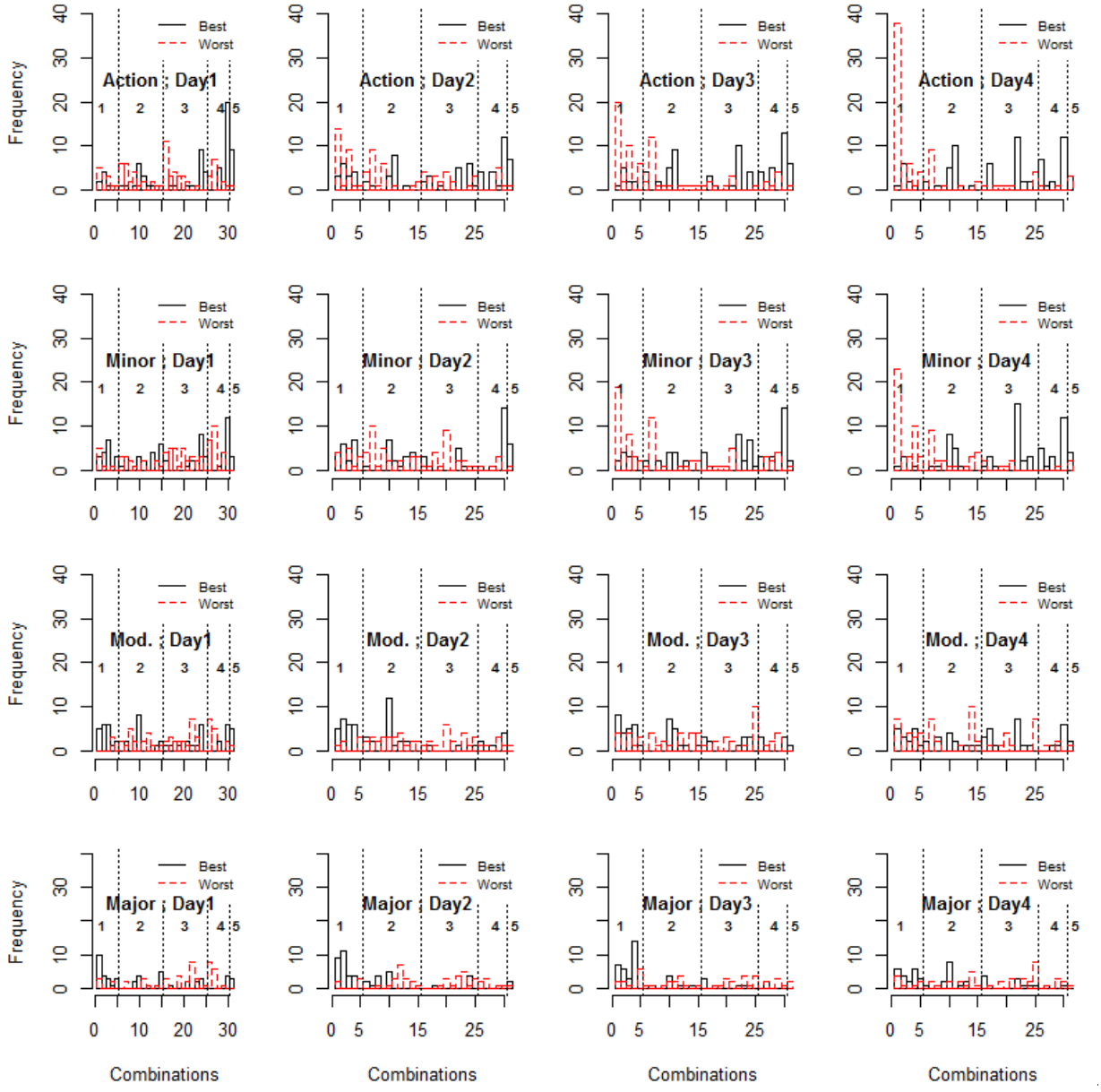


Figure 7: Histograms of ~~variable combination~~joint predictors returning the best and worst Brier Skill Scores across 82 river gages. Each row of histograms refers to an event threshold defined as a percentile of the observed water levels, and each column to a lead time. The dotted vertical lines in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


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

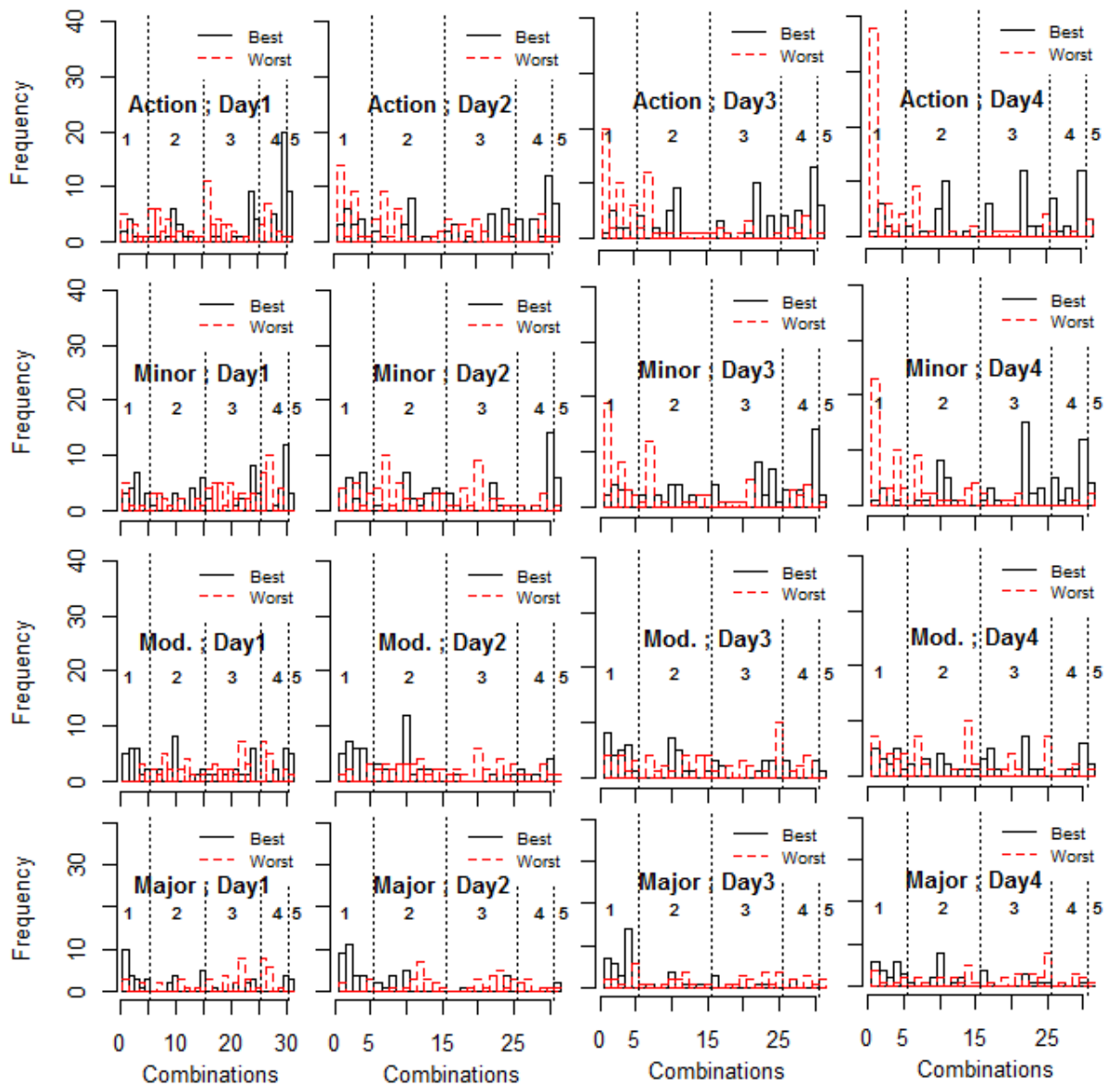
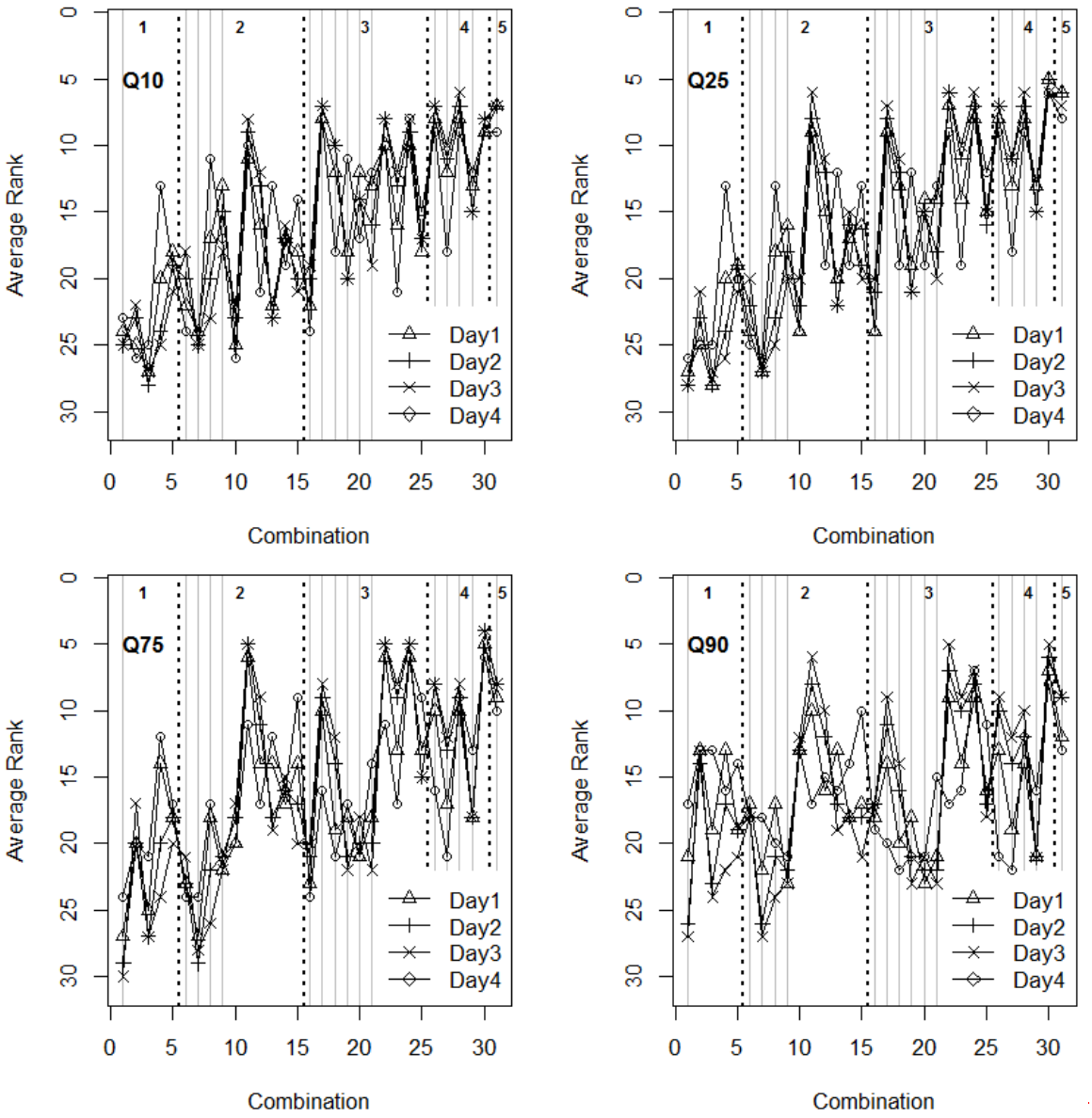


Figure 8: Histograms of ~~variable combination~~joint predictors returning the best and worst Brier Skill Scores across 82 river gages. Each row of histograms refers to a flood stage, and each column to a lead time. The dotted vertical lines in the histograms distinguish ~~variable combination~~joint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


Q10
Q25
Q75
Q90
1
2
3
4
5
Average Rank
Combination
Day1
Day2
Day3
Day4
0
5
10
15
20
25
30

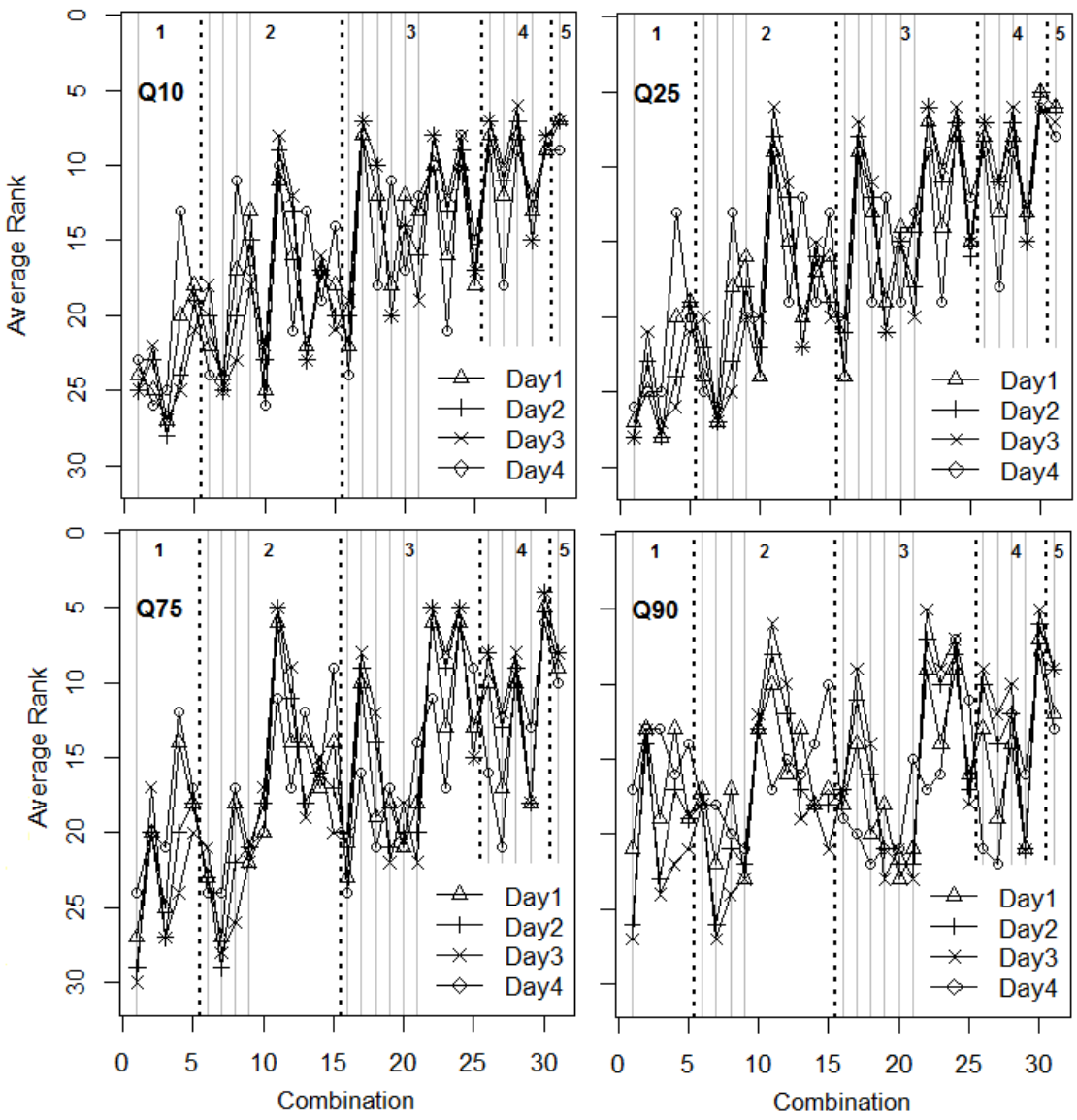
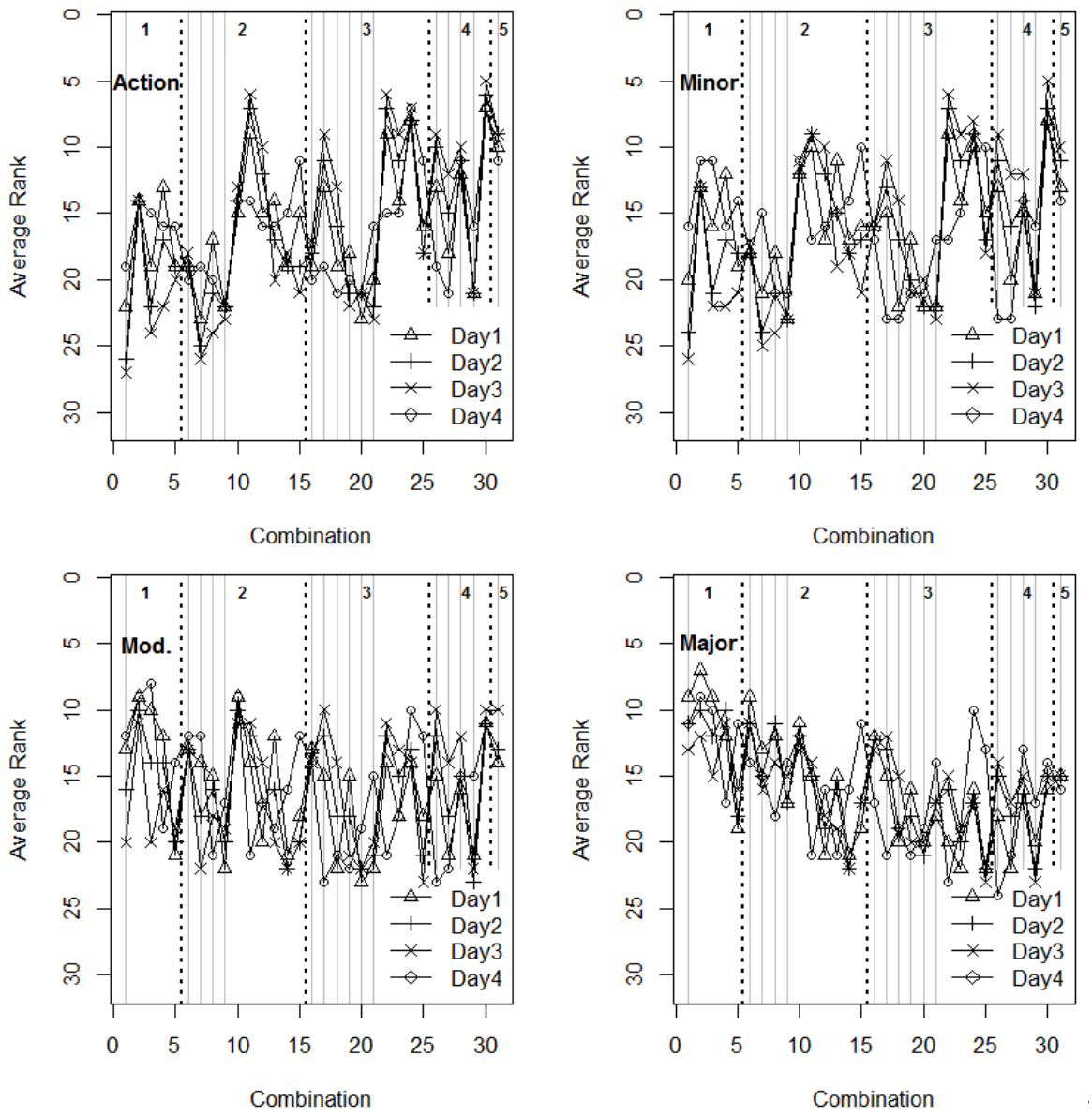


Figure 9: Average rank for each ~~variable combination~~joint predictor for one to four days of lead time and four percentiles of observed water levels. Vertical gray lines indicate ~~variable combination~~joint predictors including the forecast.


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

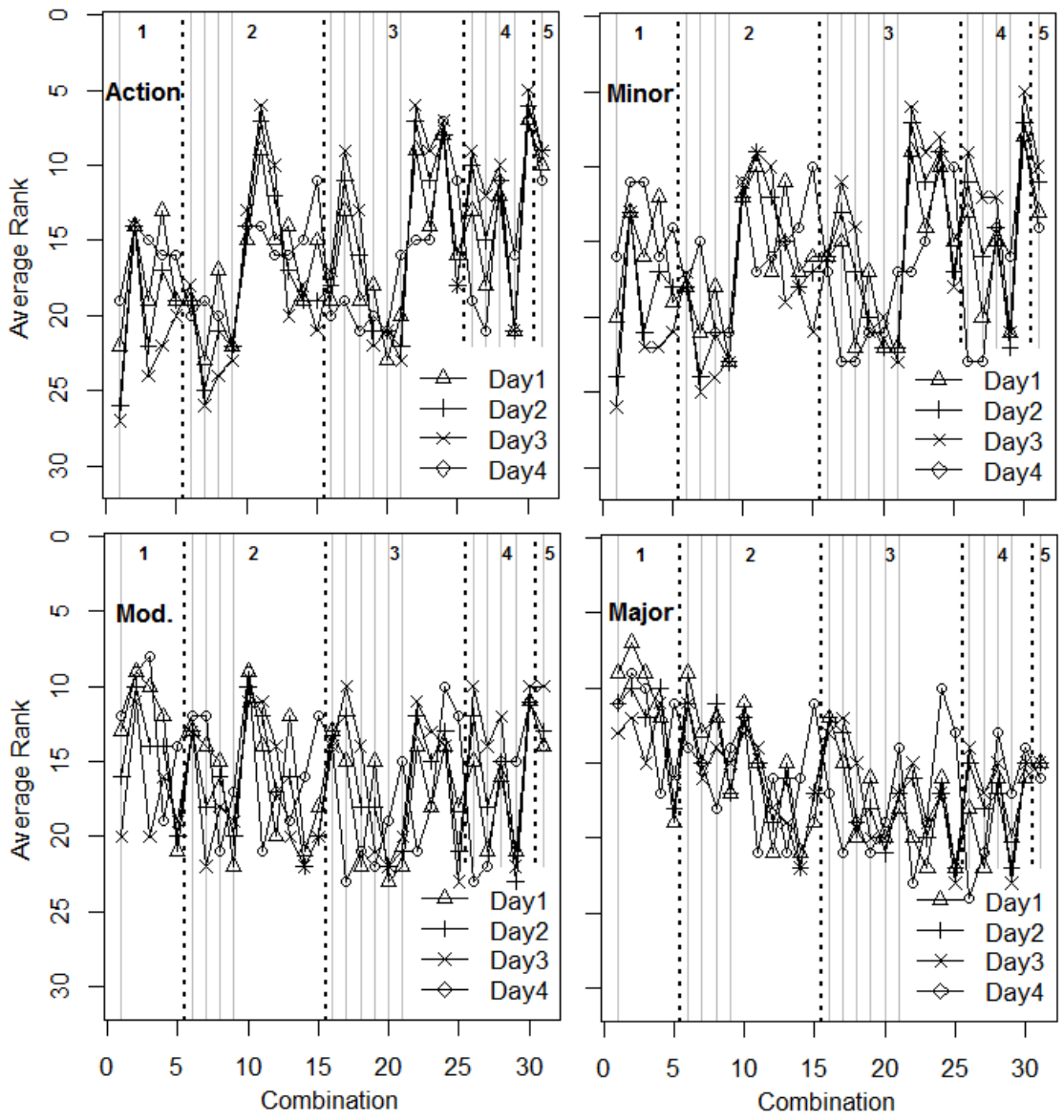
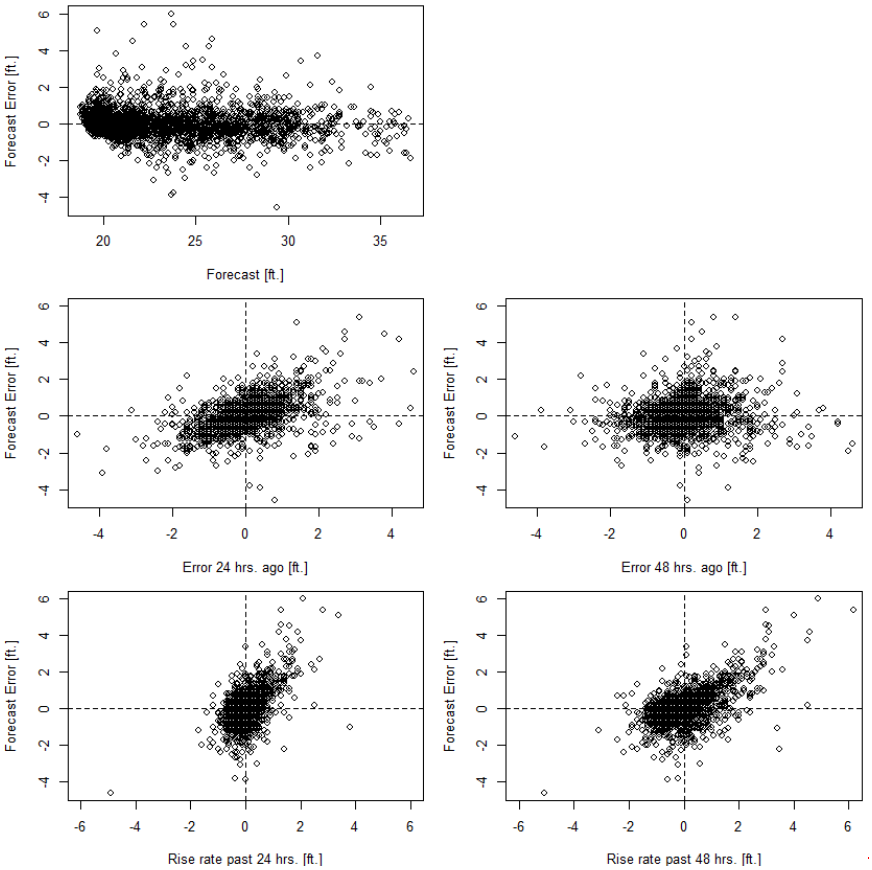


Figure 10: Average rank for each ~~variable combination~~joint predictor for one to four days of lead time and four flood stages. Vertical gray lines indicate ~~variable combination~~joint predictors including the forecast.


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

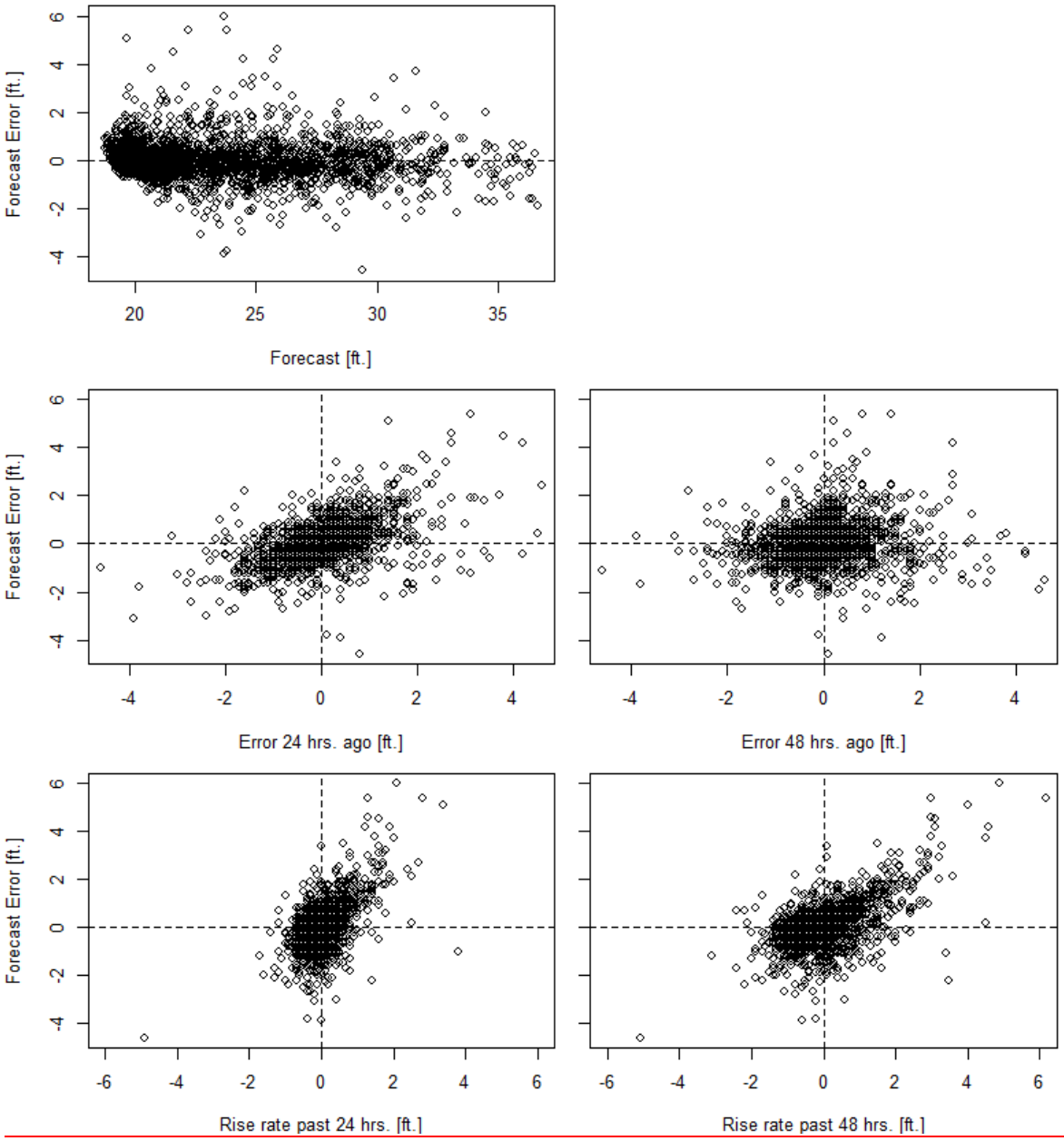
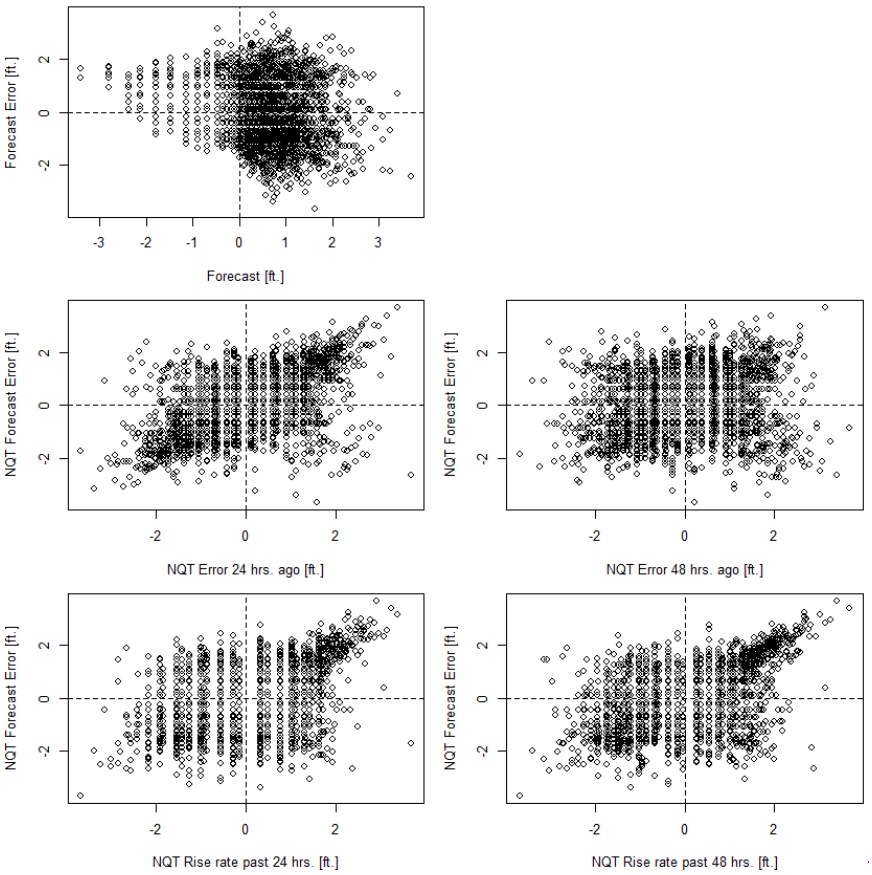


Figure 11: Independent variables plotted against the forecast error for Hardin IL with 3 days of lead time. First row: Forecast; second row: past forecast errors; third row: ~~rise rates~~rates of rise.


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

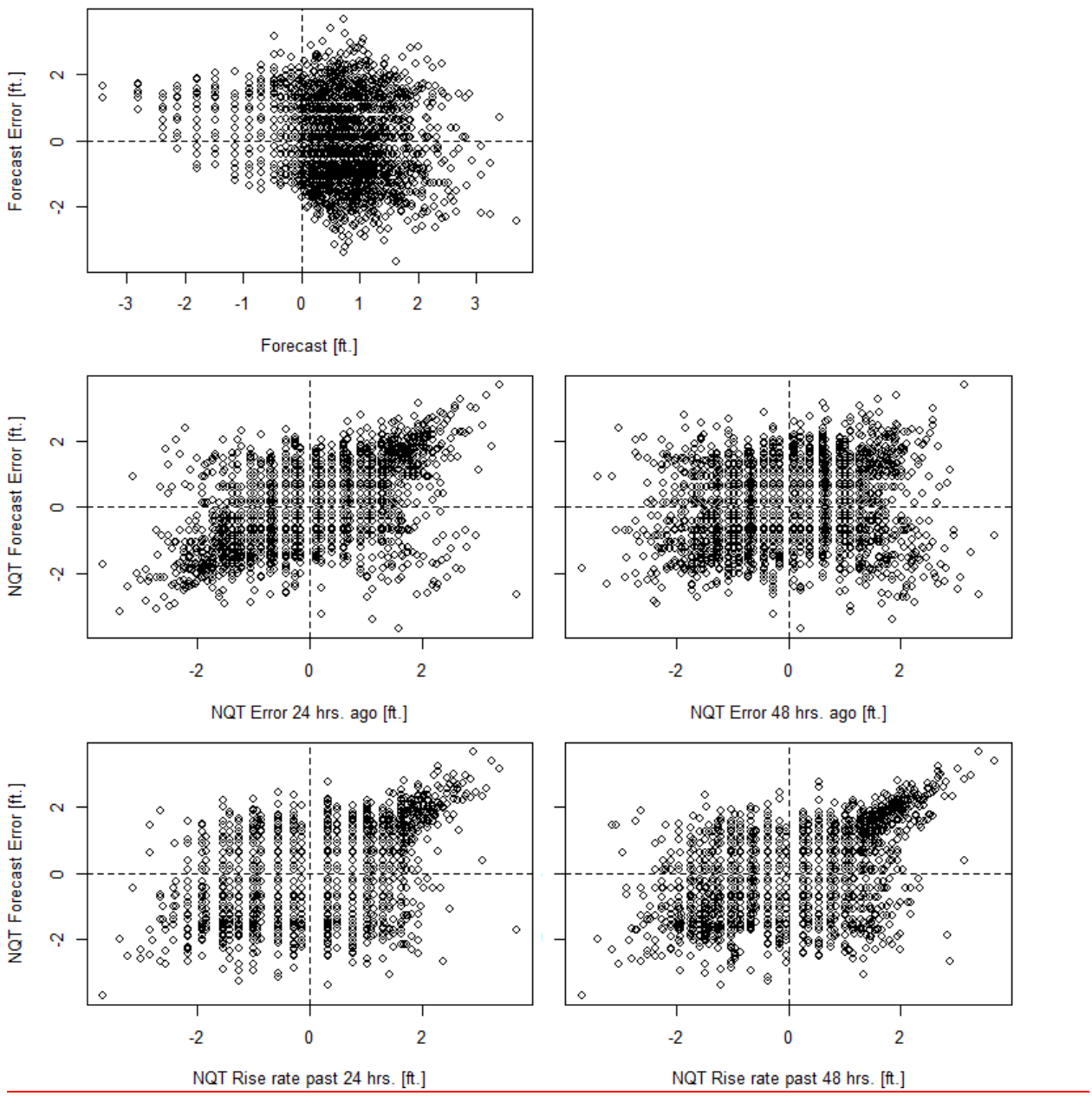


Figure 12: Independent variables after transforming into the Gaussian domain plotted against the forecast error for Hardin IL with 3 days of lead time. First row: Forecast; second row: past forecast errors; third row: ~~rise rates~~rates of rise.

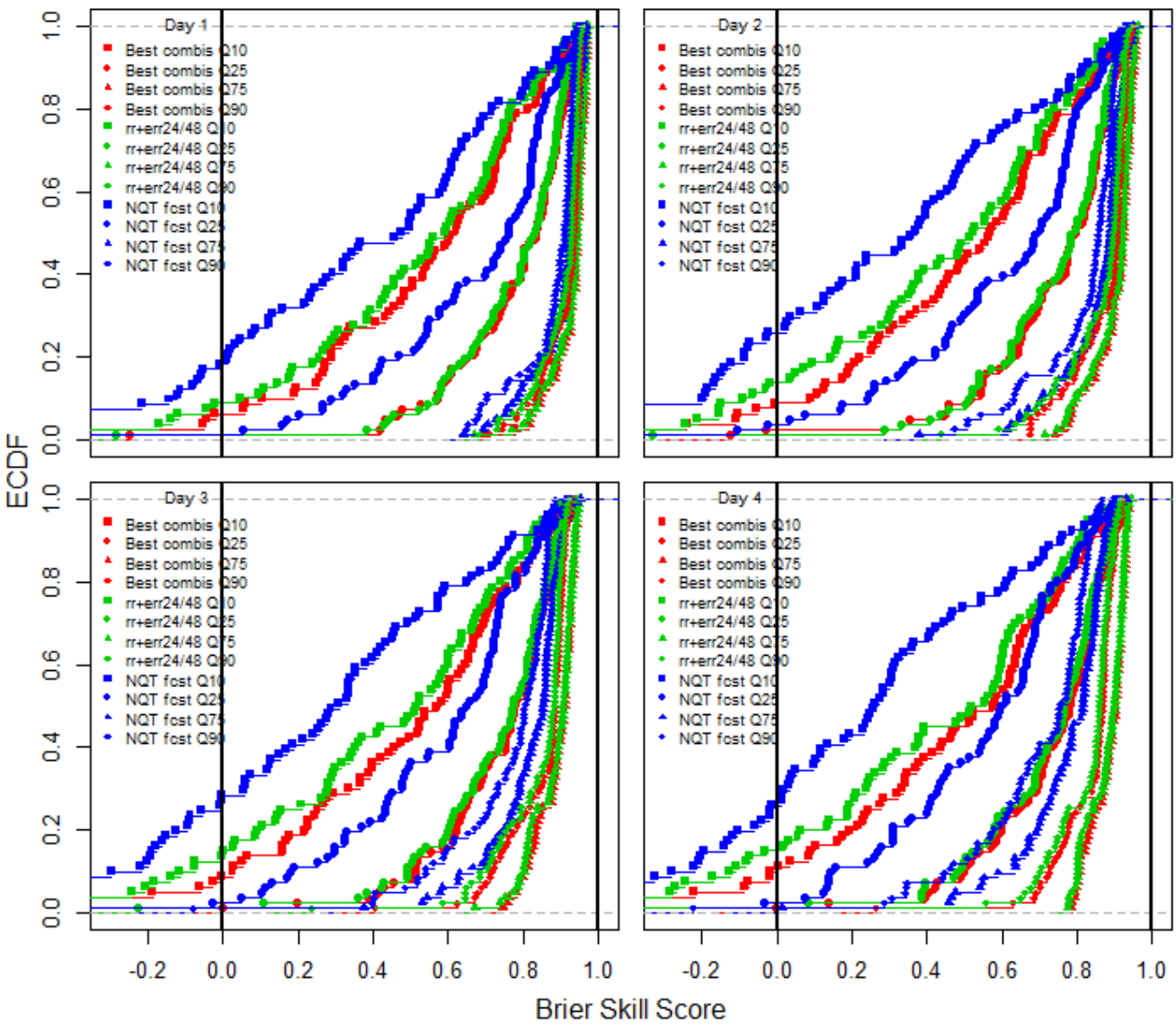


Figure 16: Empirical cumulative density functions of three QR configurations predicting exceedance probabilities of the 10th, 25th, 75th, and 90th percentile: the configuration using the transformed forecast as the only independent variable [NQT fcst]; the best performing combination for each river gage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

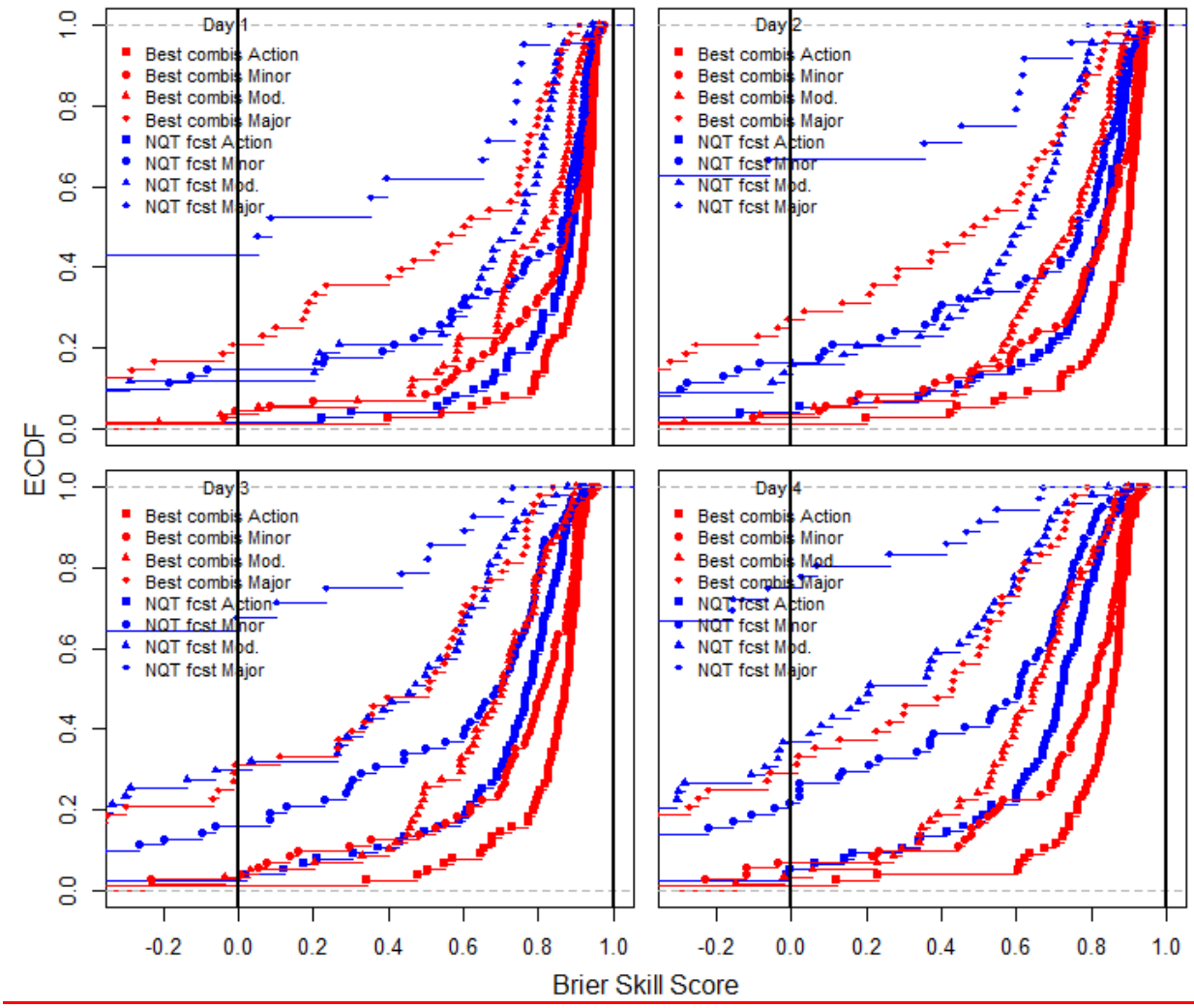
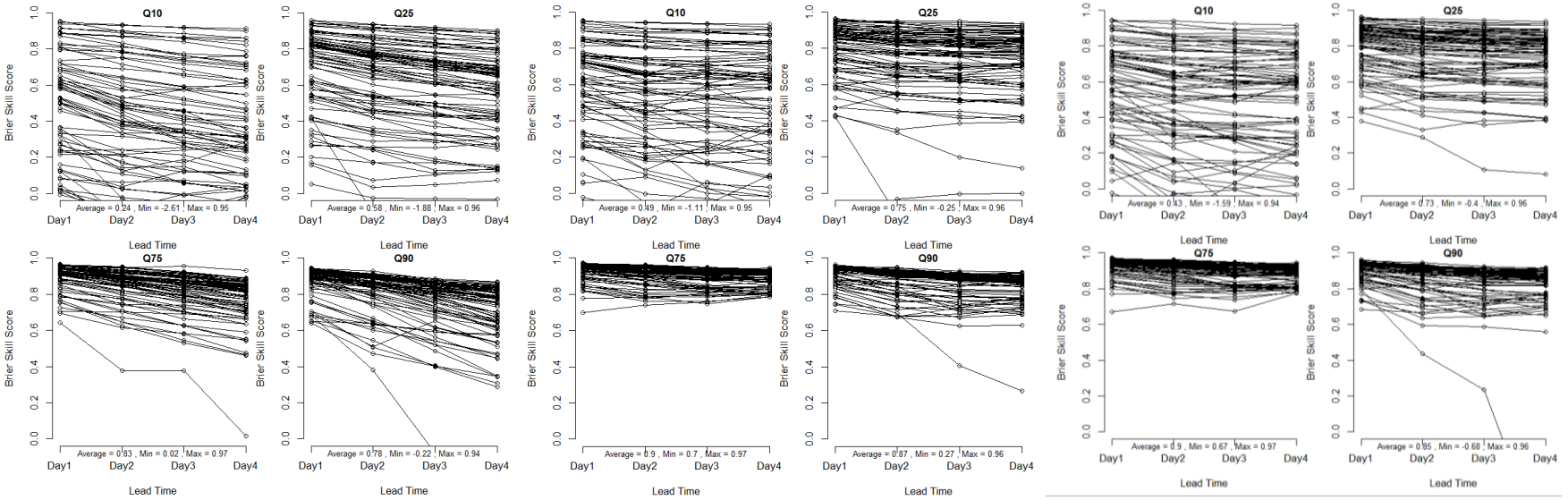


Figure 19: Empirical cumulative density functions of three QR configurations predicting exceedance probabilities of the Action, Minor, Moderate, and Major Flood Stage: the configuration using the transformed forecast as the only independent variable [NQT fcst]; the best performing combination for each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


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

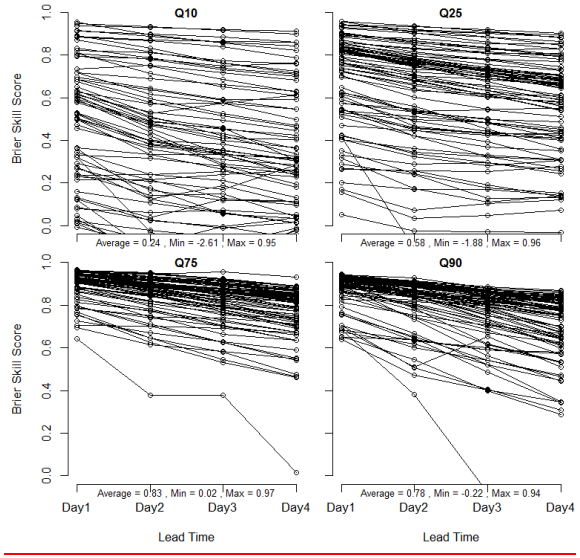


Figure 13: Brier Skill Scores of the ~~original~~ forecast-only QR ~~model~~configuration (i.e., using the transformed forecast as the only independent variable) for four lead times and percentiles of observed water levels.

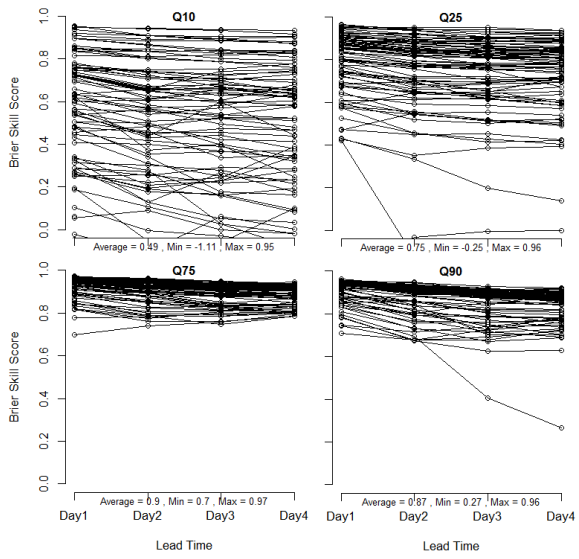


Figure 14: Brier Skill Scores for four lead times and percentiles of observed water levels using the best ~~variable combination~~joint predictor for each river gage as independent variables in the QR ~~model~~configuration.

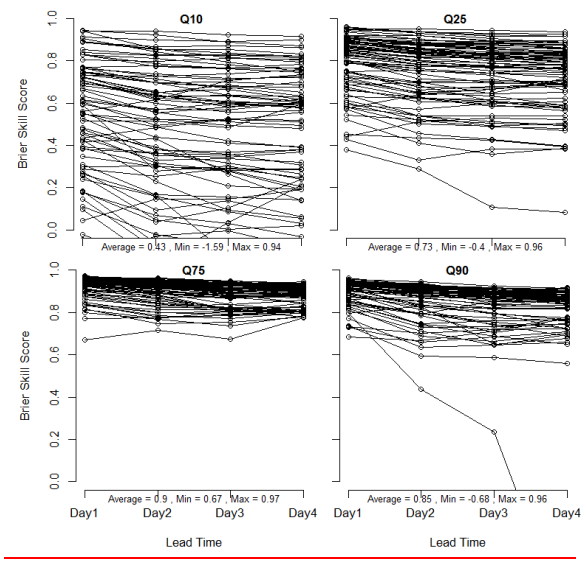
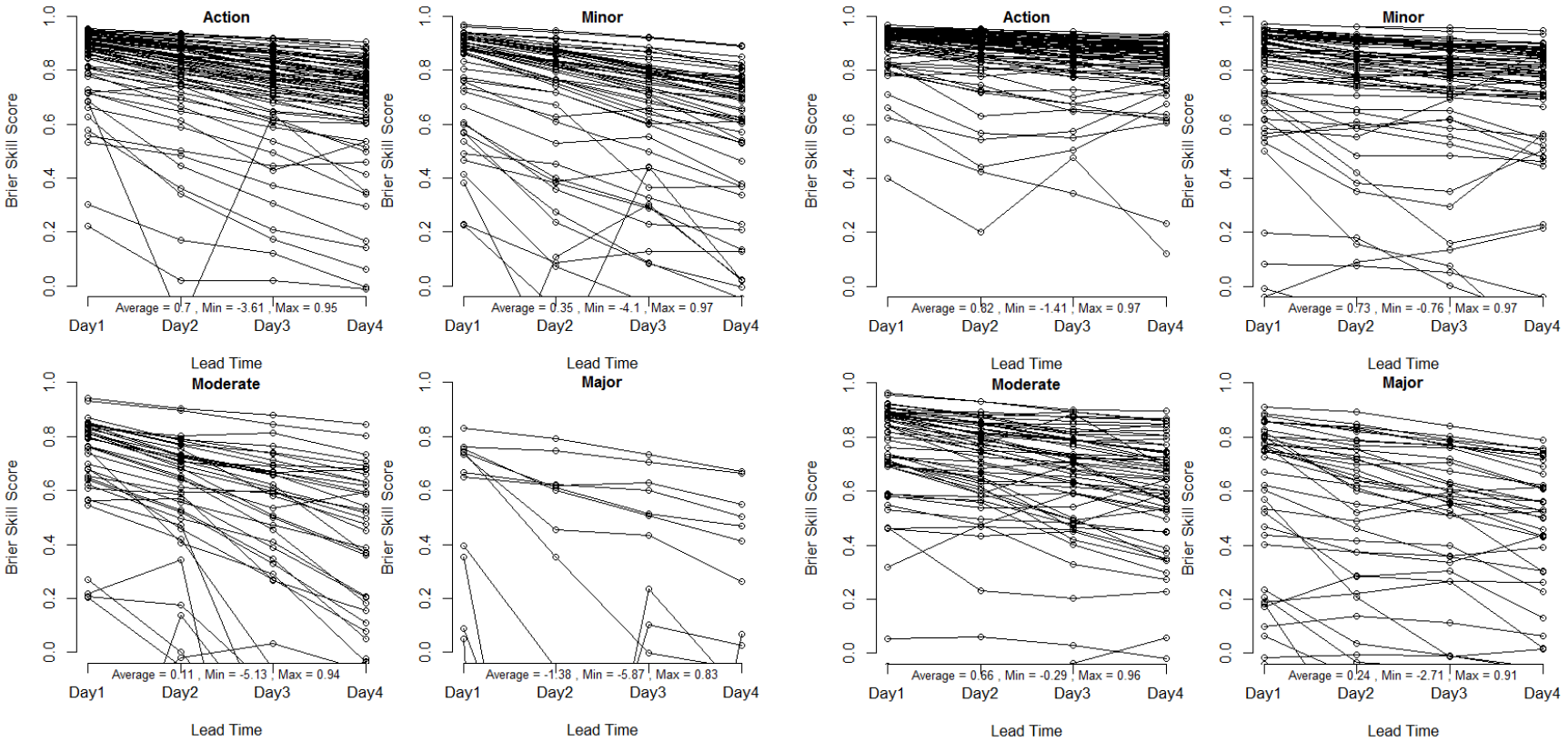


Figure 15: Brier Skill Scores for four lead times and percentiles of observed water levels using a one-size-fits-all approach (i.e., rr24, rr48, err24, err48) for the independent variables in the QR ~~model~~configuration.


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

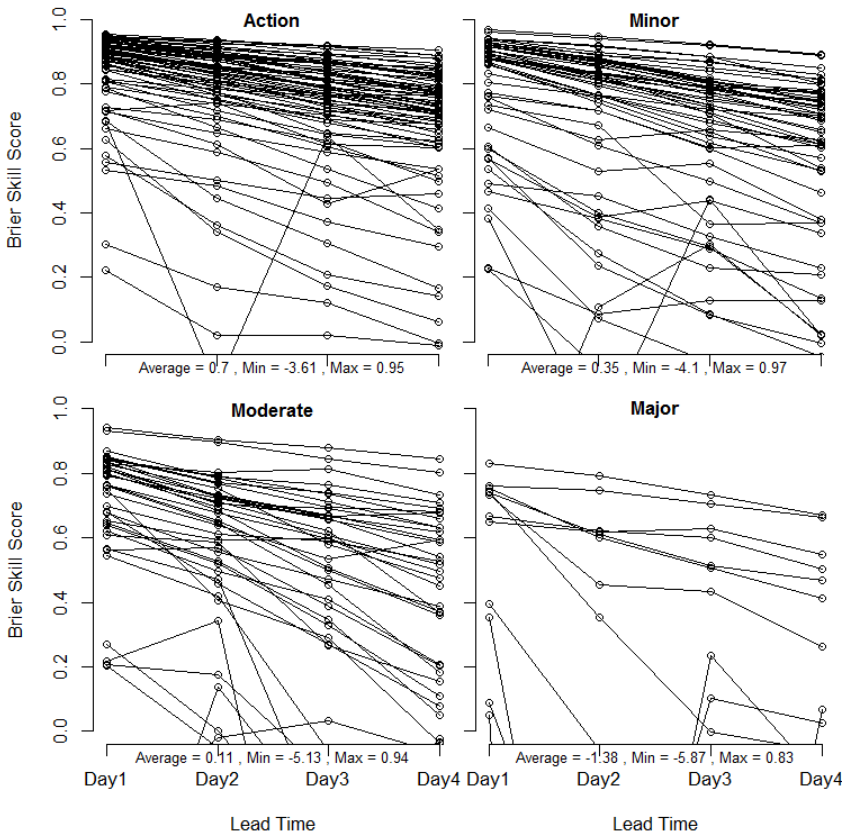


Figure 1~~7~~6: Brier Skill Scores of the forecast-only ~~original~~ QR ~~model~~configuration (i.e., using the transformed forecast as the only independent variable) for four lead times and flood stages.

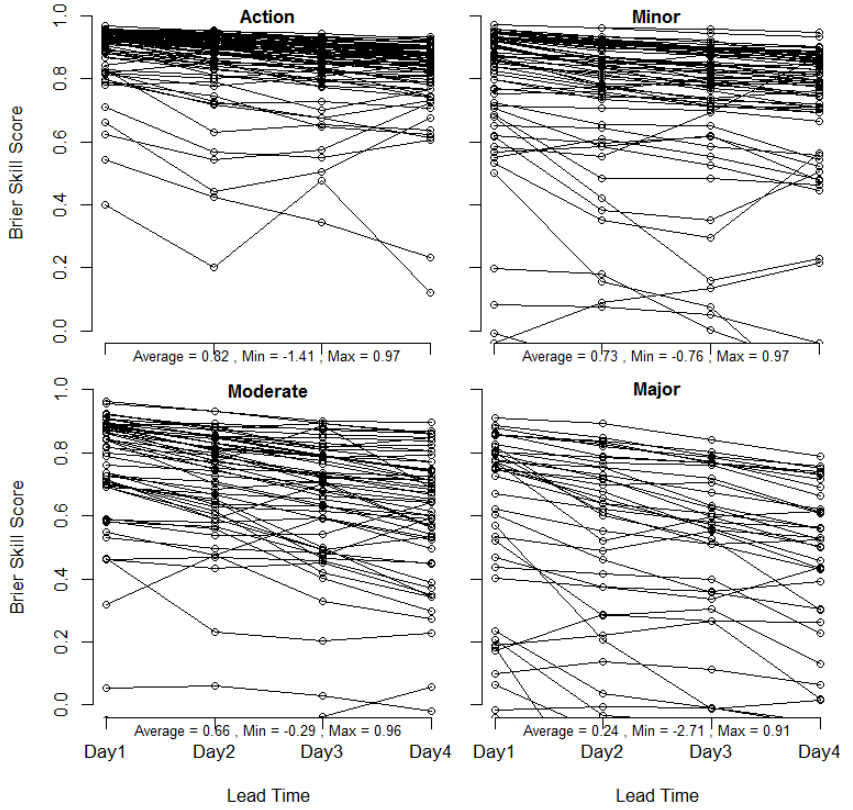


Figure 1~~8~~7: Brier Skill Scores for four lead times and flood stages of observed water levels using the best ~~variable combination~~joint predictor for each river gage as independent variables in the QR ~~model~~configuration.

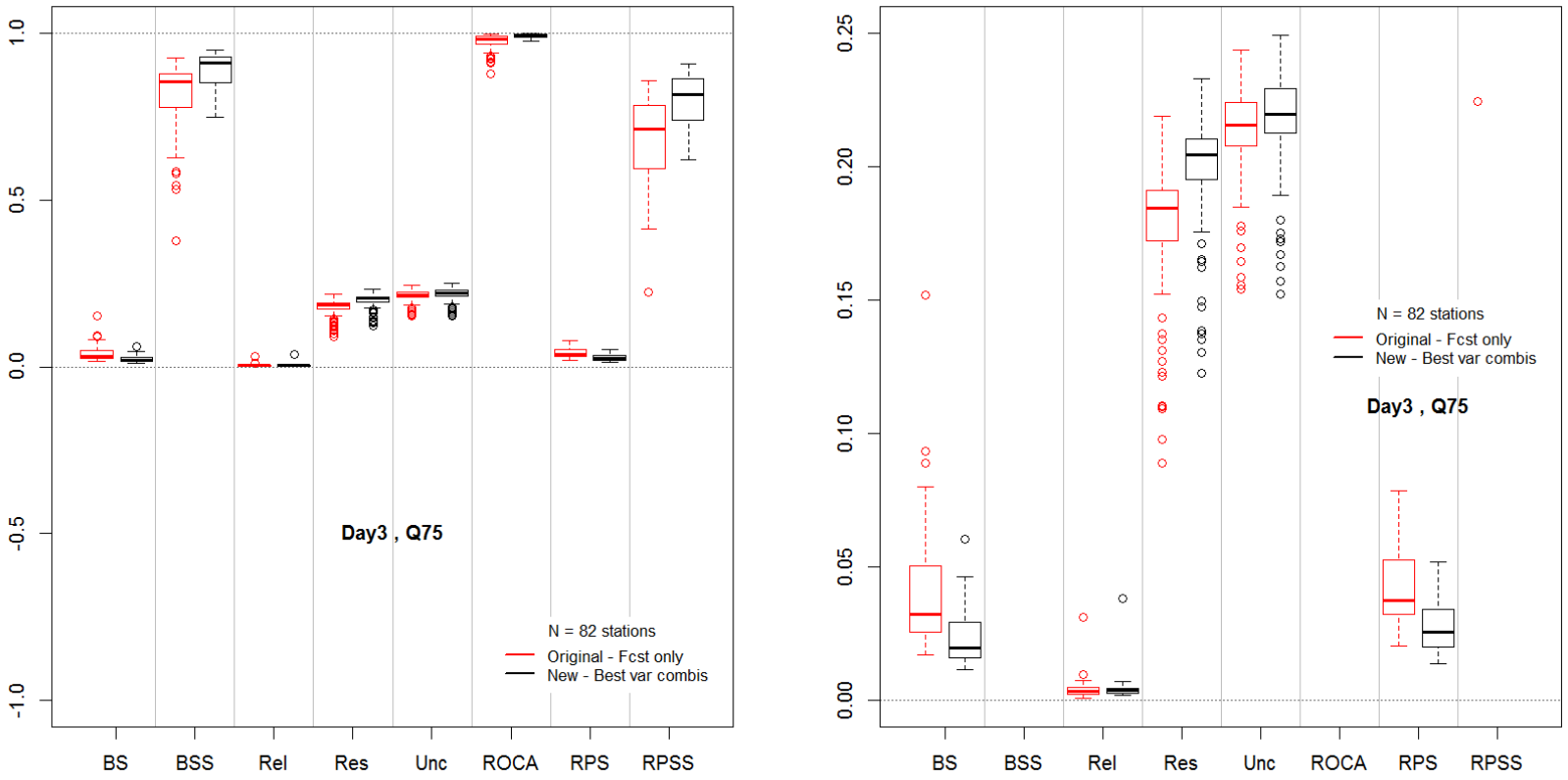
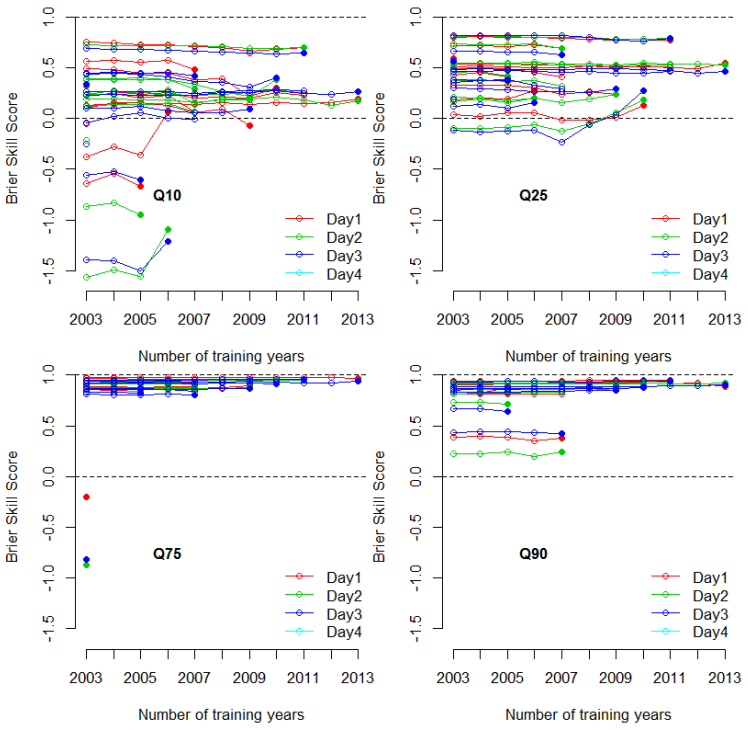
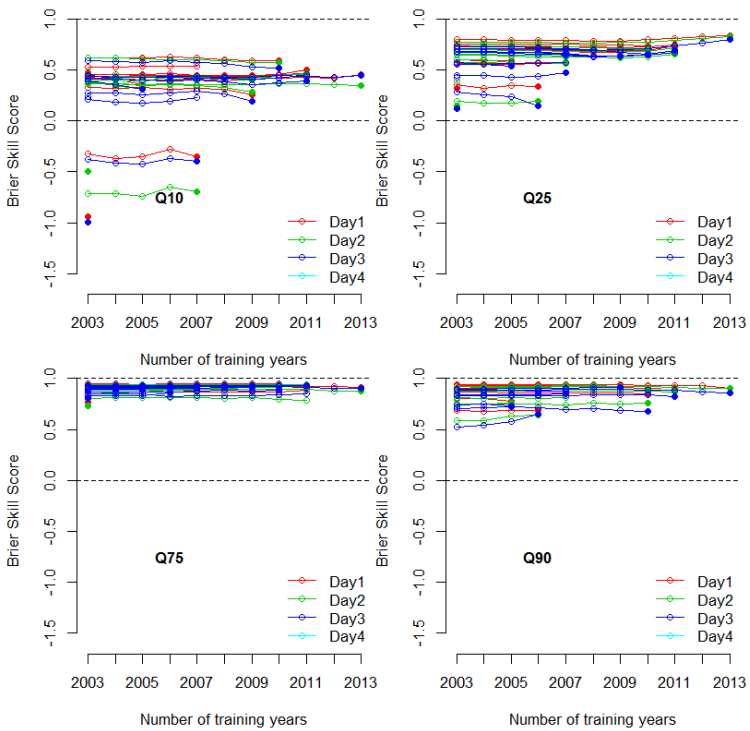


Figure 20~~18~~: Comparison of the forecast-only ~~original~~ QR ~~model~~configuration (i.e., only transformed forecast as independent variables) and the one-size-fits-all approach (i.e., ~~rise rates~~rates of rise and forecast errors as independent variables) using various measures of forecast quality: Brier Score (BS), Brier Skill Score (BSS), Reliability (Rel), Resolution (Res), Uncertainty (Unc), Area under the ROC curve (ROCA), ranked probability score (RPS), ranked probability skill score (RPSS). Lead time: 3 days; 75th percentile of observation levels as threshold. The left figure zooms in on the right figure to make changes in reliability and resolution better visible.


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


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

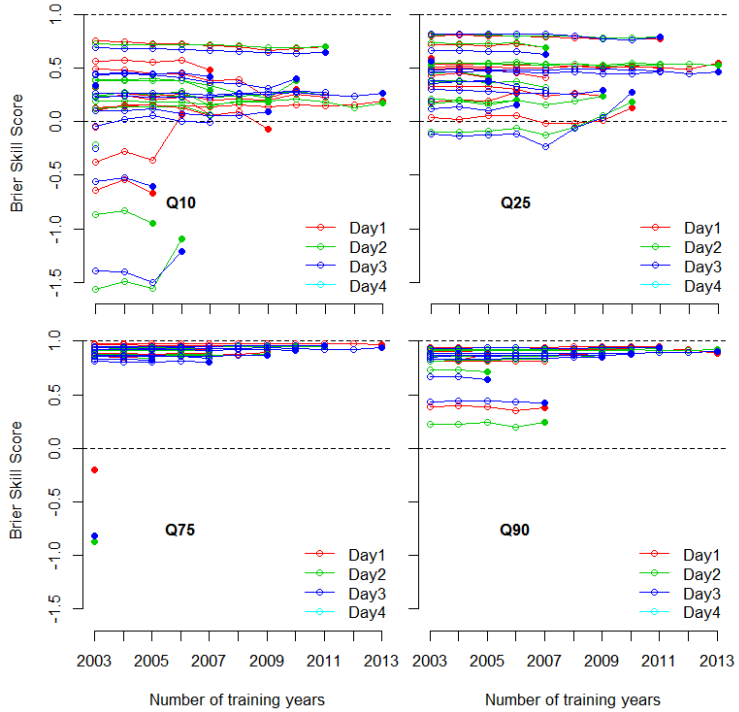


Figure 21~~19~~: Brier Skill Score for various forecast years and various sizes of training dataset across different lead times (colors) and event thresholds (plots) for Hardin, IL (HARI2). The filled-in end point of each line indicates the BSS for the forecast year on the x-axis with one year in the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

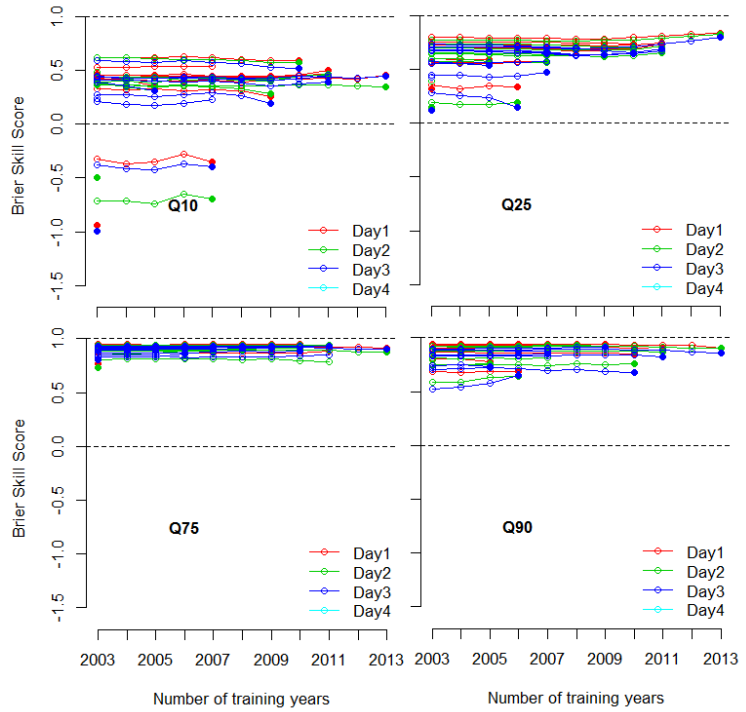
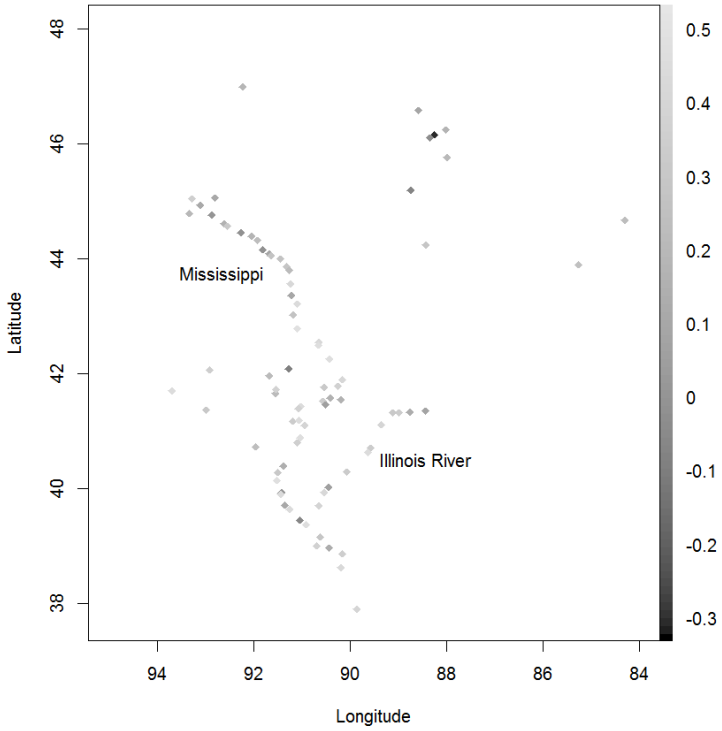


Figure 2~~0~~: Brier Skill Score for various forecast years and various sizes of training dataset across different lead times (colors) and event thresholds (plots) for Henry, IL (HNYI2). The filled-in end point of each line indicates the BSS for the forecast year on the x-axis with one year in the training dataset. Each point further to the left stands for one additional training year for that same forecast year.


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

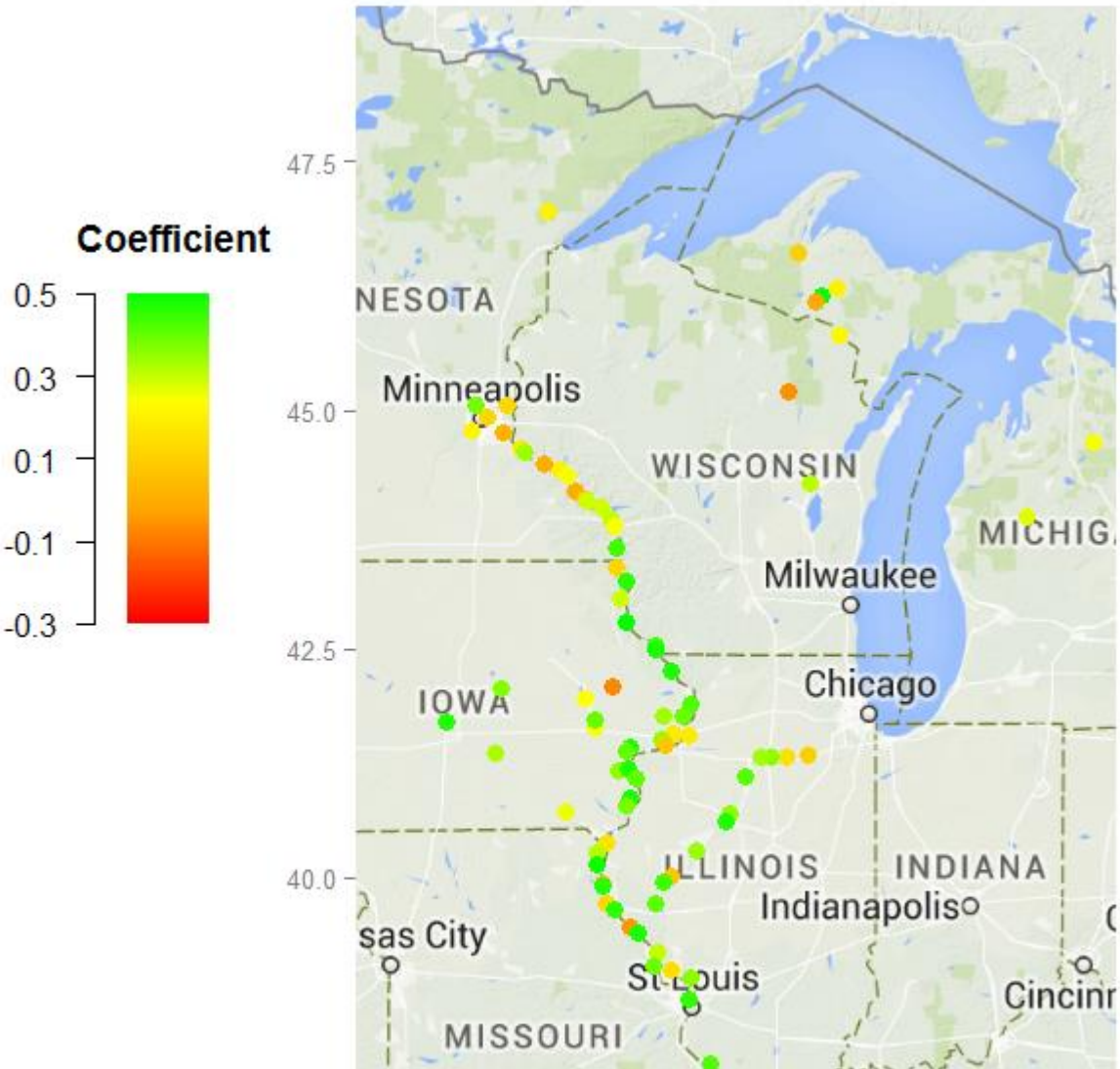

Coefficient
0.5
0.3
0.1
-0.1
-0.3
47.5
45.0
42.5
40.0
NESOTA
Minneapolis
WISCONSIN
MICHIG
Milwaukee
Chicago
IOWA
ILLINOIS
INDIANA
Indianapolis
sas City
St Louis
Cincinr
MISSOURI

Figure 2~~3~~1: Geographical position of rivers. Colors indicate the regression coefficient of each station with the Brier Skill Score as dependent variable.

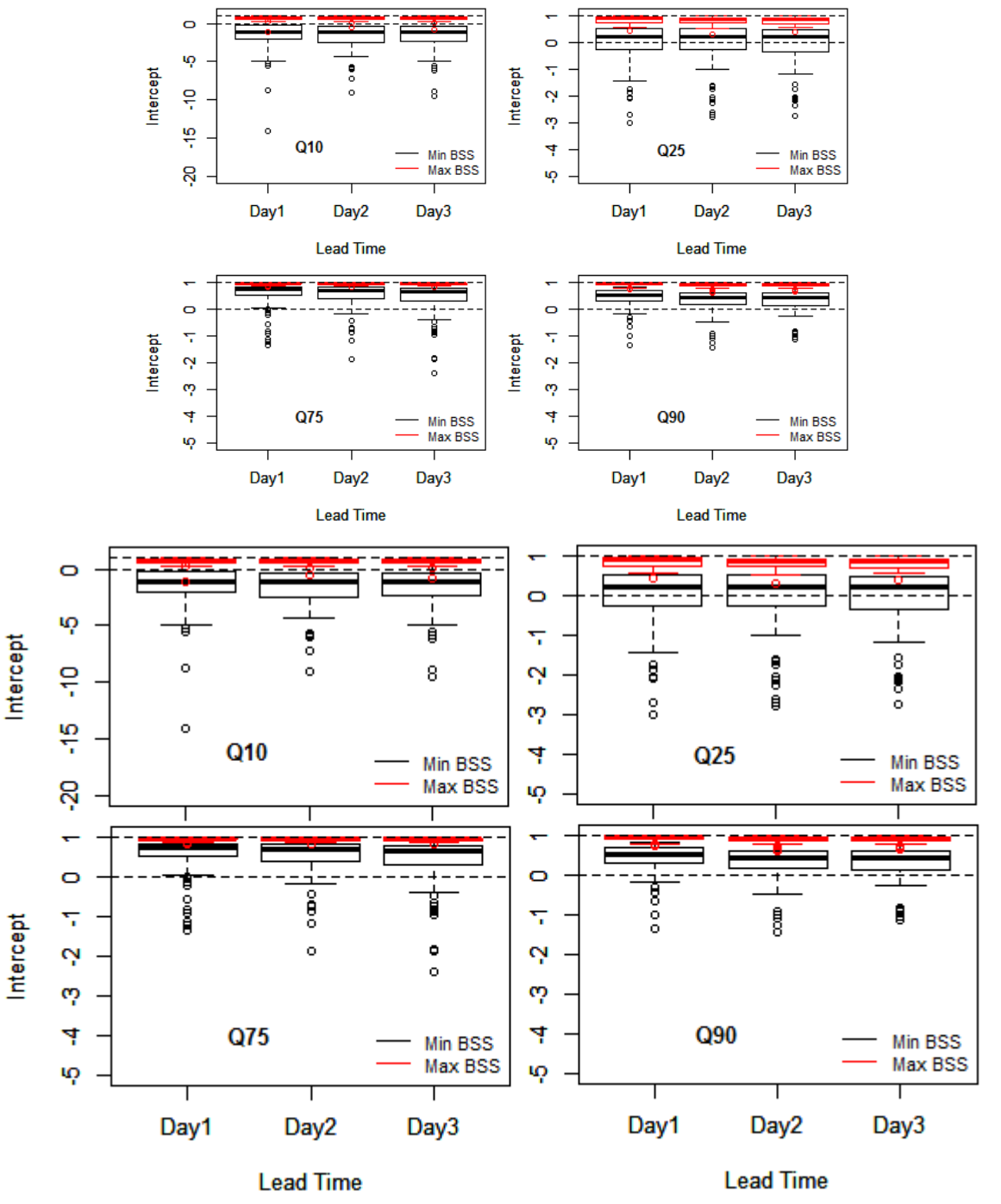


Figure 2~~4~~2: Minimum (black) and maximum (red) Brier Skill Scores for various lead times and event thresholds across locations, size of training dataset and forecast years.