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.

<u>Specific comments</u>

<u>Introduction:</u>

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.

<u>2.2 Brier Skill Score:</u>

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
ō – 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.”

<u>2.5 Data:</u>

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:

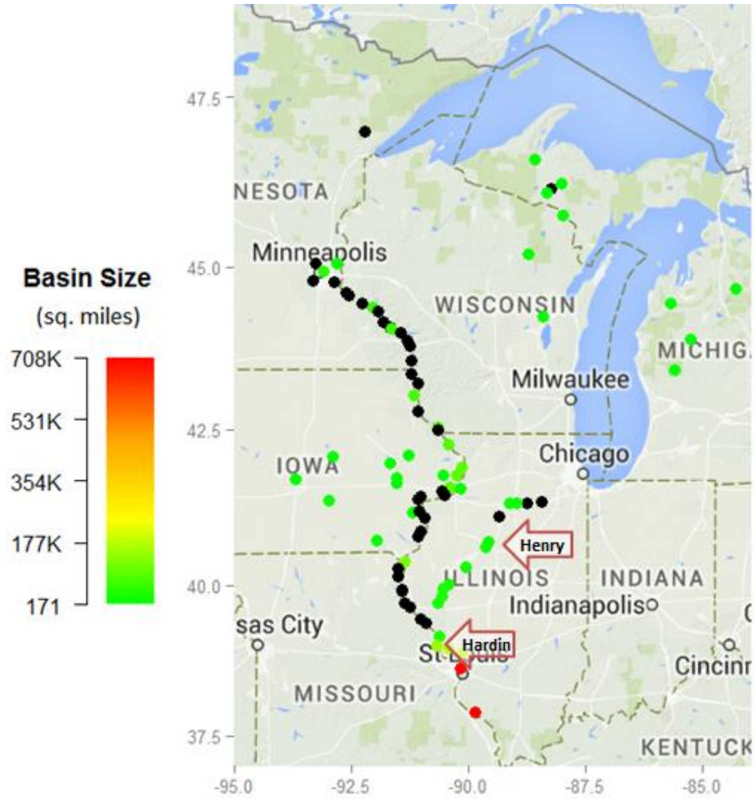


Figure 3: River gages for which the North Central River Forecast Centers publishes forecasts daily. Henry (HYNI2) and Hardin (HARI2) are indicated by the upper and lower red arrow respectively. For gages indicated by black dots the basin size is missing.

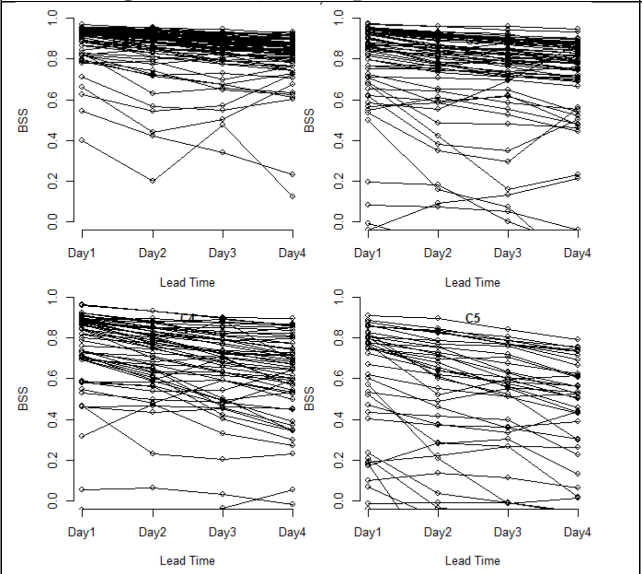
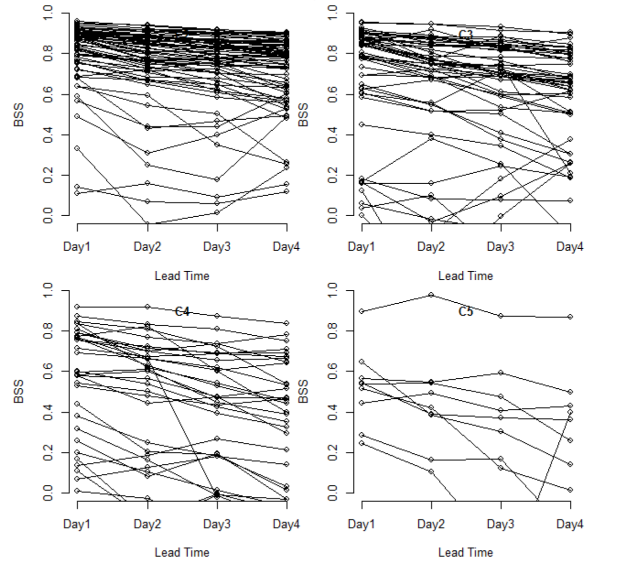
3.2.2 Best performing combinations

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):

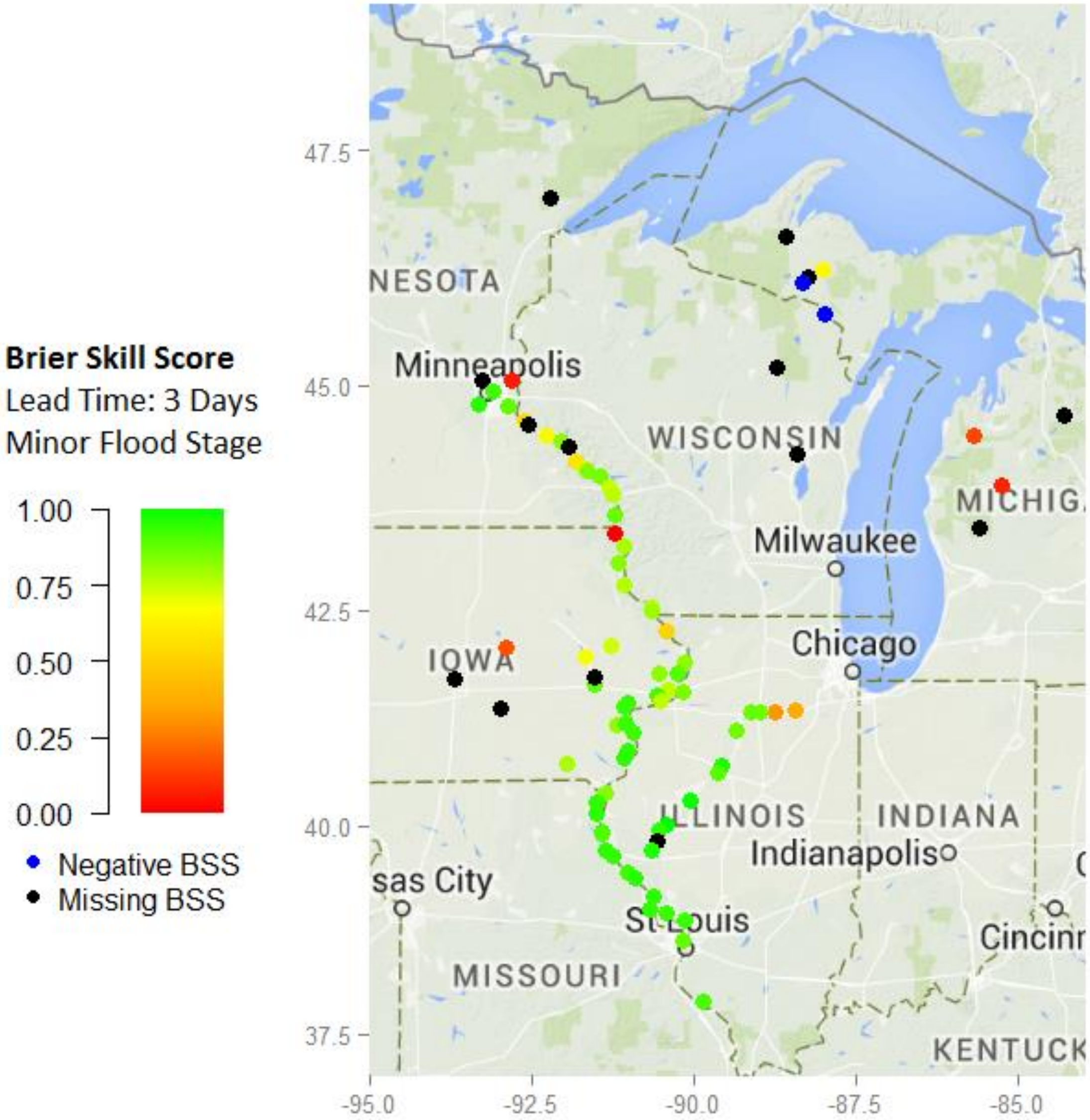
> "Figure 21 and Figure 22 show that training datasets shorter than three years result in very low BSSs for low event thresholds (Q10) at Henry and Hardin. For the other event thresholds, it barely matters for the BSS how many years are included in the training dataset. That is good news, …
>
> …
>
>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

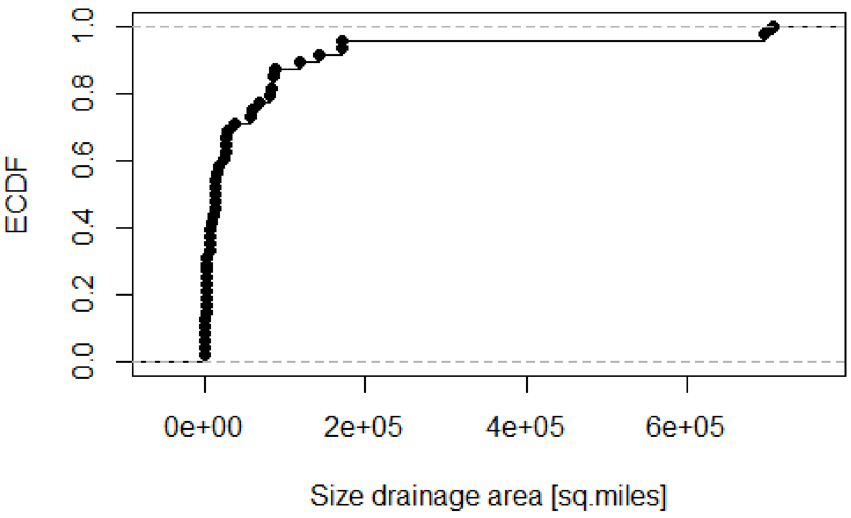


Figure 4: Empirical cumulative density function (ecdf) of sizes of drainage area for the river gages that are being forecasted daily by the NCRFC.

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., 10th, 25th, 75th, 90th percentiles as well as the four flood stages), we counted at how many river gages each joint predictor resulted in the highest and the lowest BSS."

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

Changed throughout 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 10th percentile for example will be better if you have more data points to fit your model to. I.e., even if you just look at a sub-set, you would want as many data points as possible in it, because any regression benefits from more data points.

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

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

301, 2: Table 3, maybe?

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

> "To generalize the result, the same analysis as just described for Hardin and Henry was repeated for all 82 gages. Following that, a regression analysis was executed with the BSS score as the dependent variable and the river gages and forecast years as factorial independent variables and the lead time, event thresholds, and number of training years as numerical independent variables (Table 2). The forecast performance was found to vary statistically significantly across all those dimensions except the number of training years. This results in a very wide range of Brier Skill Scores (Figure 22). Accordingly, for the user, it is particularly difficult to know how much to trust a forecast, if the performance depends so much on context. Likewise, this is case for the QR configuration based on the forecast only (not shown).
>
> A closer look at the regression coefficients (Table 2) provides interesting insights. For low event thresholds, the BSSs are much worse than for high thresholds. The QR configurations might be performing less well for low event thresholds, because the variance in the dependent variable – the forecast error – is smaller. After all, river forecasts have much smaller errors for lower water levels. The illustrative cases of Henry and Hardin, described above, indicate that using longer time series to predict exceedance probabilities of low event thresholds improves forecast performance.
>
> As expected, the BSSs slightly decrease with lead time. Regarding the forecast quality for each forecast year, the regression is slightly biased. The earlier years are included less often in the dataset with on average less years' worth of data in their training dataset, because, for example, unlike for the year 2013, ten years of training data were not available for the year 2006. Nonetheless, the regression indicates that 2008 was particularly difficult to forecast and 2012 relatively easy, i.e., they are associated with relatively low and high coefficients respectively (Table 2).
>
> The performance of the forecast additionally depends on the river gage. The coefficients of the river gages, included as factors in the regression, have been excluded from Table 2 for the sake of brevity. Instead, Figure 23 maps the geographic position of the river gages with the color code indicating each gage's regression coefficient. The coefficients are lower, and therefore the Brier Skill Scores are lower, for gages far upstream a river and those close to confluences. At least for the gages at confluences, the QR model could probably be improved by including the rise rates at the river gages on the other joining river into the regression."

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

This is actually already a different type of analysis, than the one you wanted to change to a leave-one-year-out-analysis. Even if I took your suggestion, I would still do this regression, to gain deeper insight into what causes the variability in BSS. The analysis before just visualized that there is variation, this regression studies this variation.

301,11: Why?

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

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

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

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

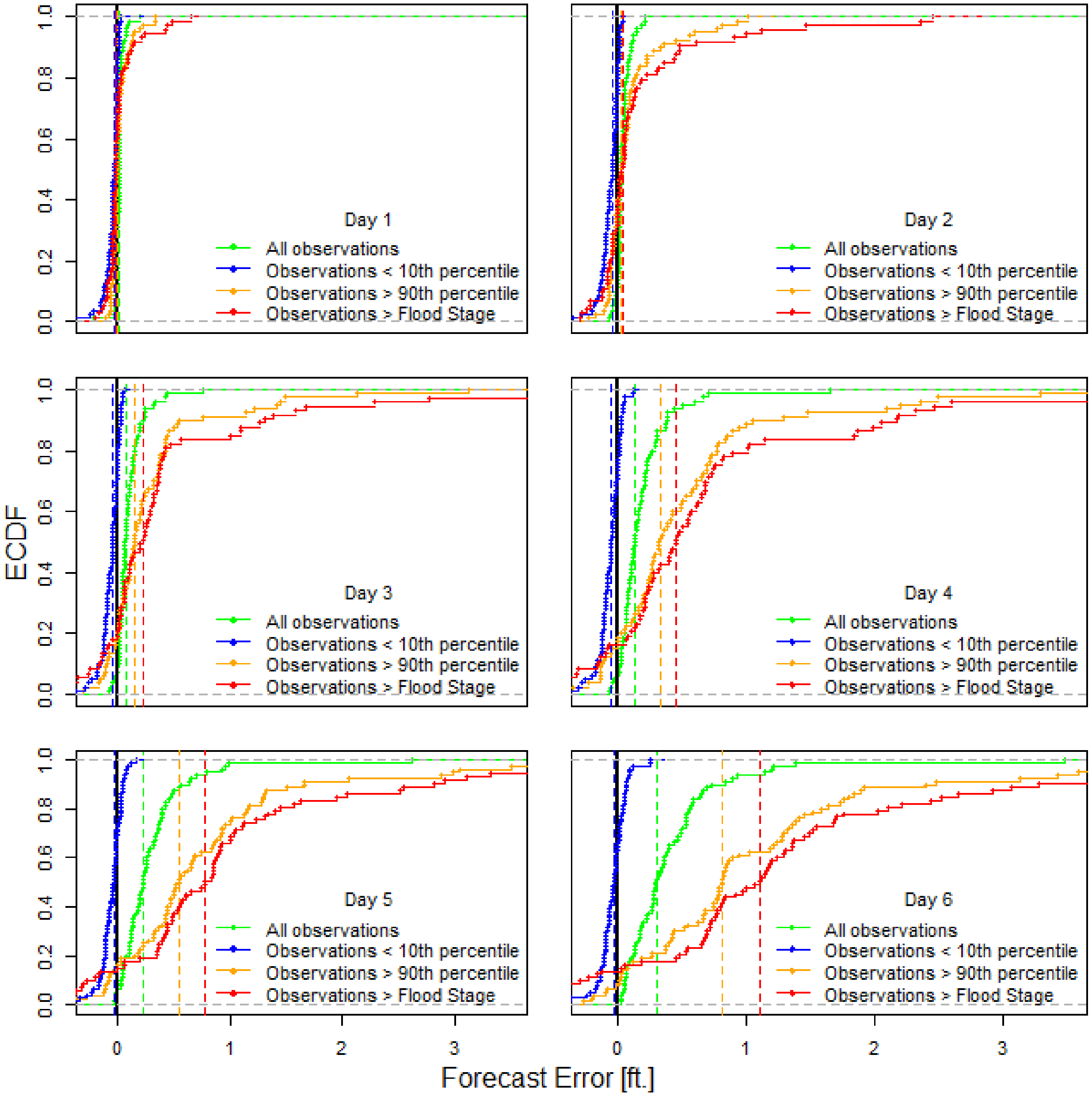


Figure 6: Empirical cumulative distribution function (ecdf) of forecast error at 82 river gages for six lead times. Vertical lines show the median forecast error of the corresponding subset.

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].

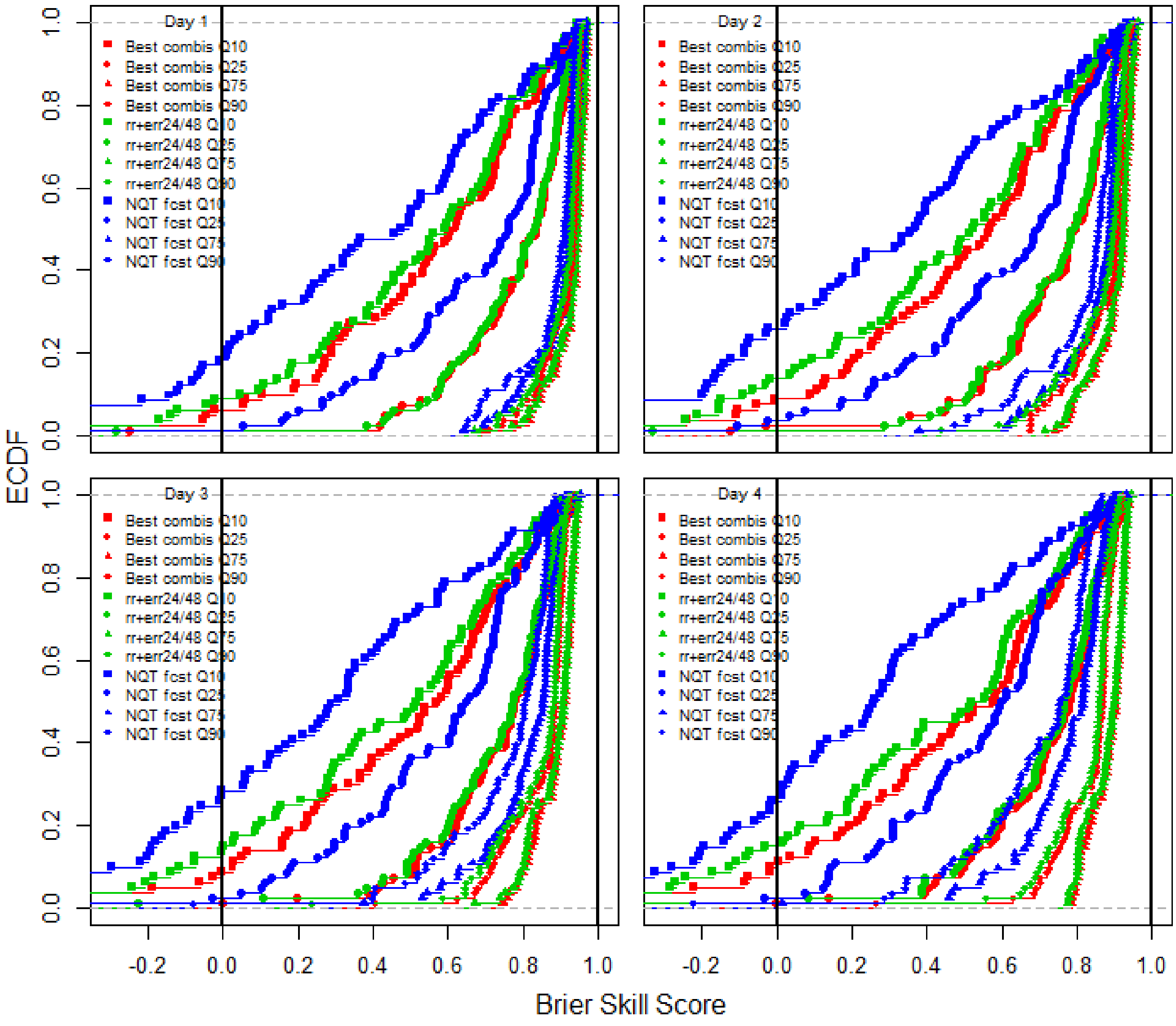
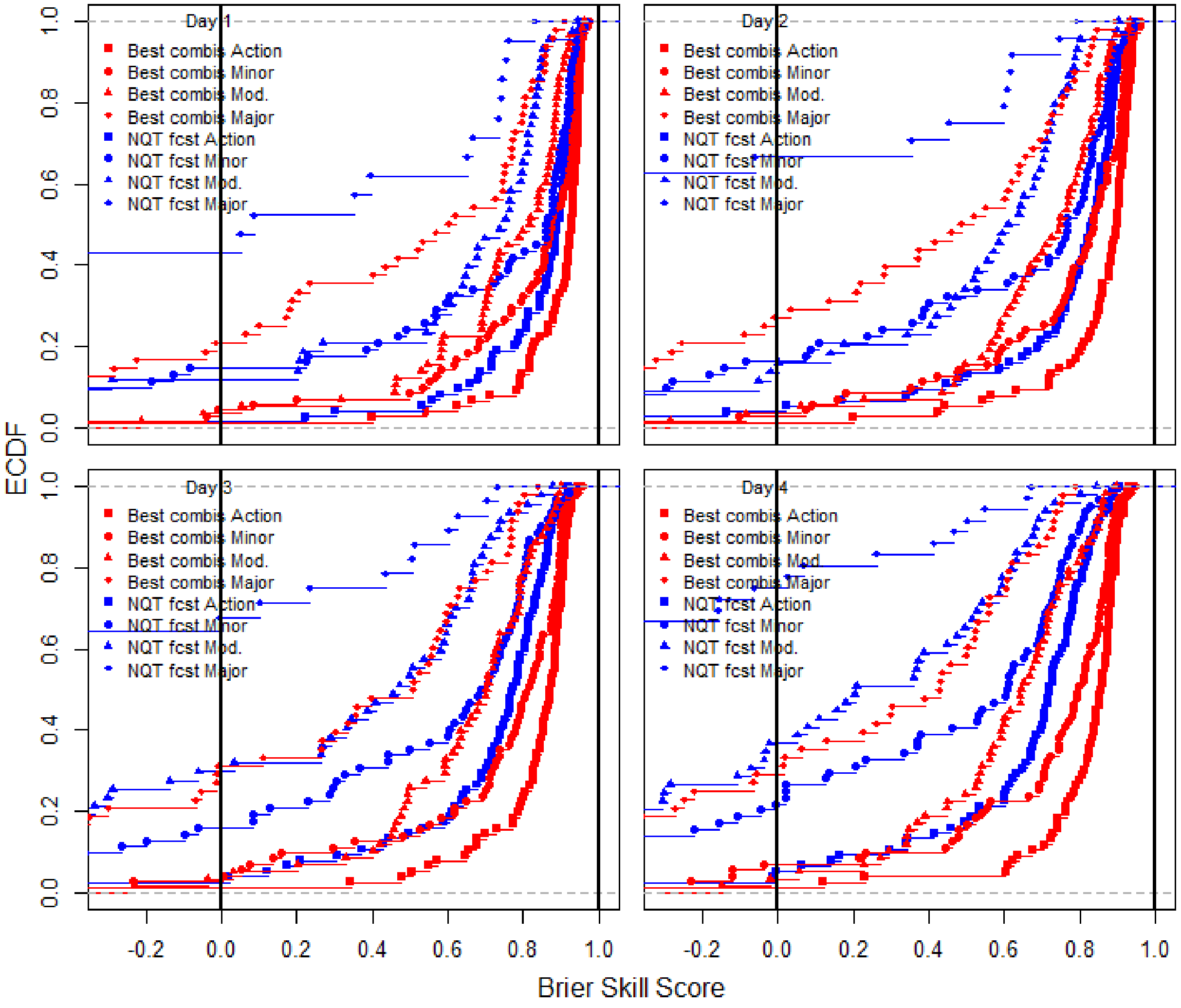
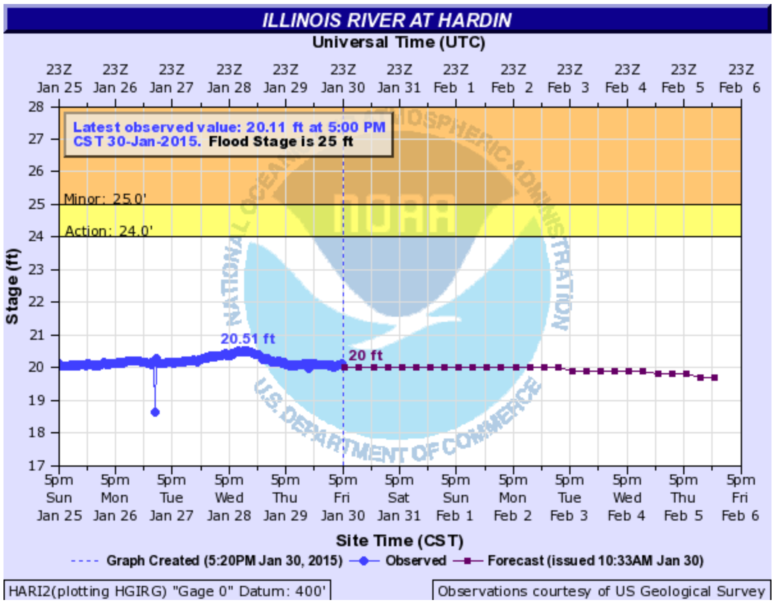


Figure 19: Empirical cumulative density functions of three QR configurations predicting exceedance probabilities of the Action, Minor, Moderate, and Major Flood Stage: the configuration using the transformed forecast as the only independent variable [NQT fcst]; the best performing combination for each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

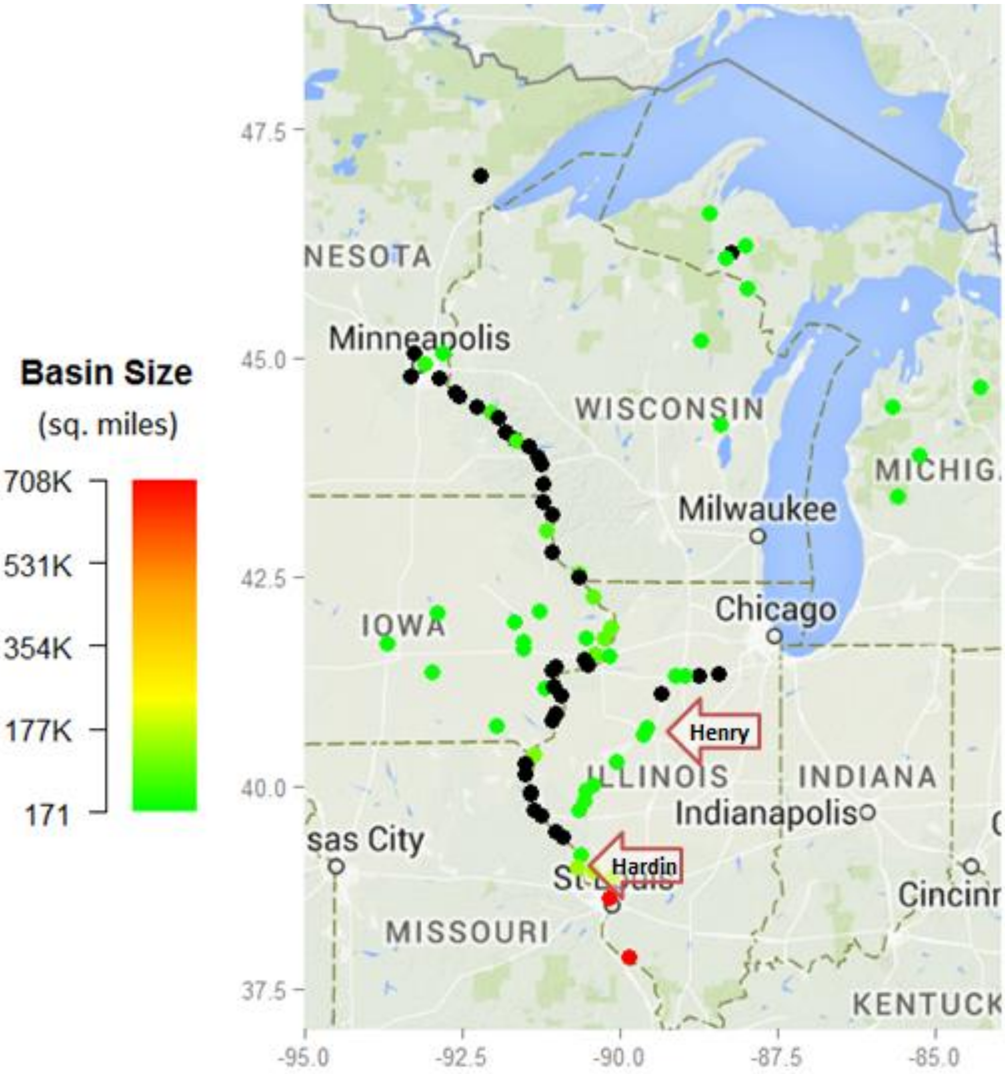


Figure 3: River gages for which the North Central River Forecast Centers publishes forecasts daily. Henry (HYNI2) and Hardin (HARI2) are indicated by the upper and lower red arrow respectively. For gages indicated by black dots the basin size is missing.

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.

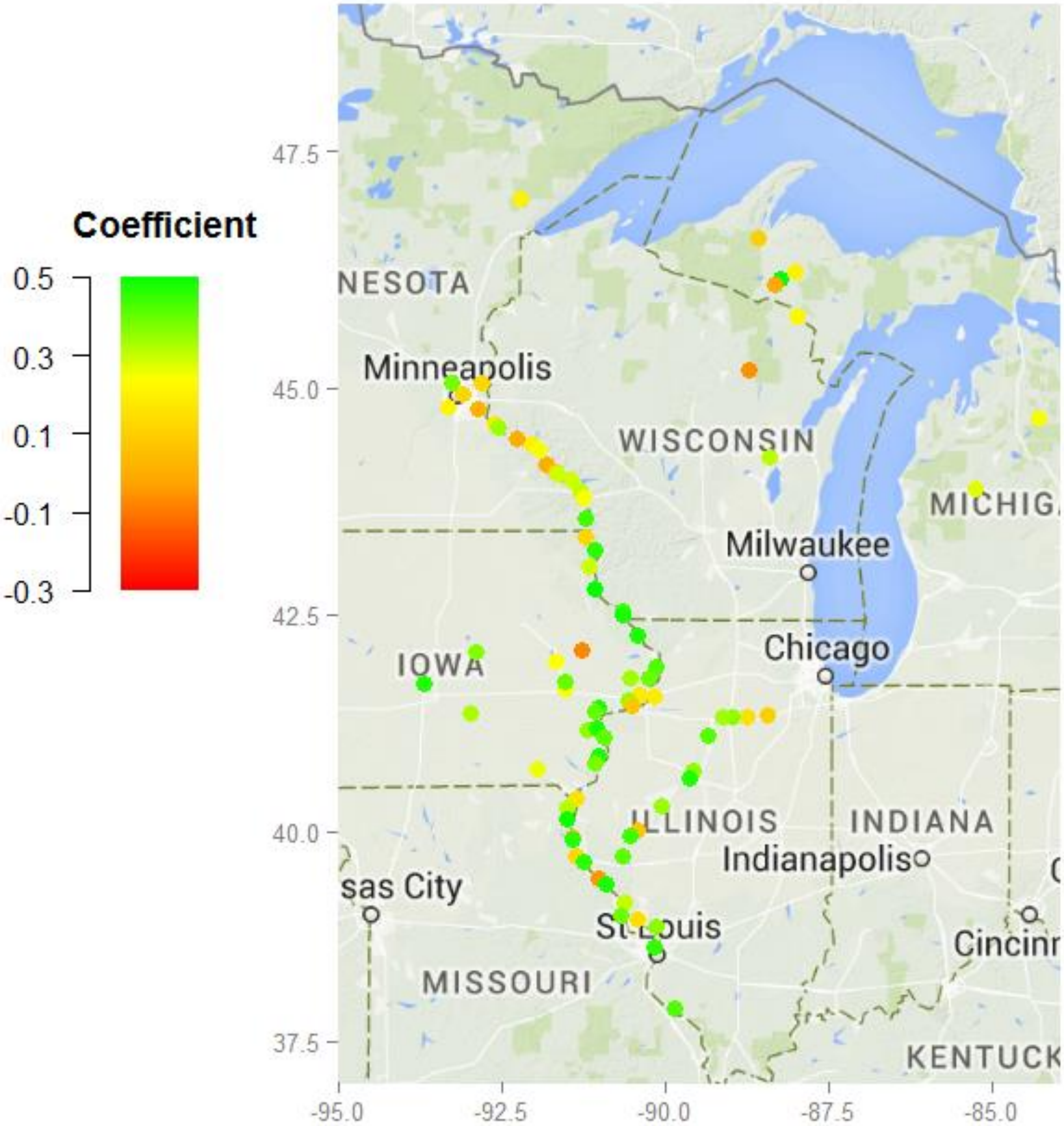


Figure 23: Geographical position of rivers. Colors indicate the regression coefficient of each station with the Brier Skill Score as dependent variable.

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