Letter to the Editor

July 27th, 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 editor,

Please find the revised paper attached. The sixteen pages of comments by Reviewer 1 resulted in some major improvements:

- I supplemented the visual analysis with paired t-tests to verify that my visual observations are statistically significant. This decreased the number of figures in the paper as well.
- As requested by both reviewers, I extended the hydrological interpretation, especially in the sensitivity analysis. However, the selection of predictors was constrained by data availability etc., so that pretending that these choices were based on hydrological considerations would be plainly incorrect. I did further elaborate on the selection of predictors though to make this point clearer.
- The reviewers were concerned that my findings might not be true for other performance measures. So, I added the continued ranked probability score into the discussion where it made sense to show that that is not the case.

I hope that you find that these changes to have satisfactorily addressed the reviewers' concerns. If there are additional changes that you believe are needed, please let me know.

Regards,

Frauke Hoss

Response to Reviewer #1

July 27th, 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.

General Comments

1.a. Figures 7 and 8 (similarly, Figures 9 and10) can be combined into one single file presenting information for four different water levels.
Those figures cannot be combined. Figure 9 presents water levels defined by percentiles, Figure 9 those defined by action levels. The action levels correspond with different percentiles for each gage. Combining the figures would make this difference too easy to miss. So instead of combining them, I cropped the figures to two displays each. I did the same for Figures 7 and 8.

1.b. Figures 13, 14 and 15, and Figures 16 has same information, latter being the CDF plot. Similarly, Figures 17 and 18, and Figure 19, latter being the CDF plot. Presenting results only for four water levels, does not cause loss of information content and decreases number of figures by '7' ….

The CDF plots (figures 16 and 19) were requested by a reviewer. Each of them includes the information of three figures, but I think they are much harder to see the trends in the CDF plots than in the other figures.

I now included t-statistics for to test whether BSSs increase statistically significantly, so I have removed Figures 13-16. Figures 17-19 are staying to visualize some of the information in the table with t-statistics.

1.c. Figures 5 and 6 represent same information, i.e., how forecast error varies with respect to lead time and with respect to streamflow magnitude, and suggested to keep only one figure.

Yes, that is true. Figure 6 was requested by a reviewer. Again, I would argue that Figure 5 is much easier to read. I could move the CDF plots to an appendix if that would make everyone happy.

1.d. The two sets of figures, i.e., 7 & 8 and 9 &10, essentially represent the same information, and suggested to remove one of them (detailed comments mentioned below).

See answer to comment 1.a

2. *Combining two sections (3.2.1 and 3.2.2) and two sets of figures.*

I omitted the frequency analysis in favor of tables showing the improvement in BSS to be statistically significant.

3. *Reg. excluding of the forecast from the best configuration*

One would expect significant amount of information in the current forecast relative to any other variable(s), however, to some extent, contrary and not very intuitive is observed, i.e., predictors excluding the forecast yield high BSS scores.

Correct. That is what makes this finding interesting.

I would say 'to some extent' because configuration with forecast is not resulting very bad forecasts, but relatively smaller BSS values compared to the configuration without forecast.

I included a new table with regression results showing that joint predictors with forecast included perform *significantly* less well. See my response to your comment on line 397.

Decomposition of BSS values show the difference is due to BSS-resolution, however, it is not clear how much significant are these values to claim that the technique performed better without forecast. It is important that this to be analyzed and discussed in detail because issuing a forecast involves lots of resources, and coming up with a suggestion that says to exclude the current forecast and use yesterday or day before yesterday forecast means something is not right with the current forecast.

It is not due to the BSS-resolution that the forecast as a predictor cannot be combined easily with the other predictors, if that is what you are saying. The causality is the other way around. The rise rates and past forecast errors improve the resolution.

As I have explained in the text, there is a mathematical reason why the forecast cannot be easily combined with the other predictors. They simply have very different distributions. Please see the plot below that the errors and rates of rise are normally distributed while the forecast has a highly skewed distribution. The physical reasons for that should be very obvious.

Due to the different distributions, the forecast is only a good predictor after NQT transformation, while the other predictors do better without NQT transformation. I specifically included Figures 11 and 12 to illustrate this point.

The fact that forecasts do not go well together with the other predictors says nothing about the quality of the forecast and all the efforts that have gone in it. Nothing is “wrong” with the forecast. It is just a mathematical reality that it cannot be combined easily with the other predictors.

Predictors based on recent water levels and recent error patterns may add additional information other than what is present in the forecast, however, it is not clear how these

products are able to supply information that is already present in the current forecast and interestingly counteract with the forecast.

It is also incorrect to say that those other predictors "counteract" the forecast. Like I have written, they just cannot be combined for mathematical reasons. I did write that this is due to their mathematical distribution.

Your statement that error rates are "already present in the current forecast" is incorrect. The error rates include *more* information than the forecast, because to compute the error the observed water level is needed.

This needs to be analyzed and discussed in detail because of its implications. The authors explanation in lines 409 - 412 and using of NQT plots is not clear (also, it is not explained clearly why did the authors use NQT'ed variables when regression was developed using original variables). A very similar was comment made by reviewers (e.g., 298, 7 (1)) and I think this is to be answered.

I included NQT transformation in the discussion, because it is the method as first introduced by Weerts et al. I will make sure that this is mentioned when introducing the Weerts study (see next comment).

As requested, I added more explanation:

"The combinations including the forecast (indicated by gray vertical lines in Figure 7 and Figure 8) perform significantly better than those that exclude it (Table 2). This disadvantageous impact of forecast as an independent variable is less pronounced for very high or low event thresholds (Table 2a). Including the forecast into the joint predictor is even beneficial for major flood stages (Table 2b), when joint predictors with less rather than more variables perform better.

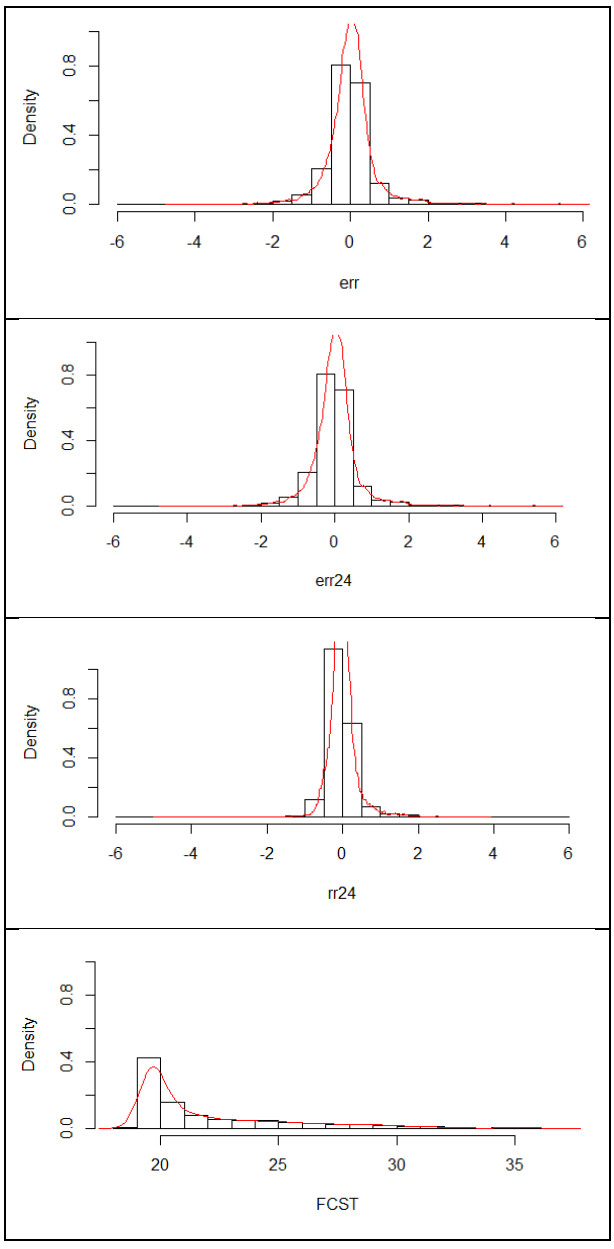
The forecast is difficult to combine with the other four predictors (err24/48, rr24/48), because their statistical distributions are different. Unlike the dependent variable (forecast error), the forecasts are highly skewed towards the left, because low water levels occur more frequently. Due to its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

In contrast, errors and rise rates are already approximately normally distributed. There are no quantile trends visually detectable anymore after the other four predictors have been subjected to NQT (Figure 11 b-e). In sum, forecast performance in this study is better without NQT, because four of the five independent variables were approximately normally distributed already. Further research is necessary to reconcile predictors with different distributions. Possible solutions could be to define QR configurations for subsets of the transformed dependent and

independent variables or to experiment with subjecting only some, but not all predictors to NQT."

A similar statement is made in this NOAA user manual, on the bottom of page 8: http://www.nws.noaa.gov/oh/hrl/general/HEFS_doc/HEFS-0.3.2_EnsPost_Users_Manual.pdf

Please see the plot below for the distributions. I did not include them in the paper, because the paper already has many figures:



4. *Review of past literature, lines 85, 110-111,135-143: It appears that Wood et al., Weerts et a., and Lopez Lopez et al., are the key studies, and relatively later two studies are mentioned (not detailed) at different places of the manuscript. Given the importance of these studies and*

fact that they are only '3', it is suggested to detail these past studies at one single place, probably in the Introduction, with respect to common points, differences, limitations, advantages and results, and at the of the Introduction section in one or two lines how the proposed study differs with past studies. In this way, by the end of the Introduction section, the reader will have a clear idea what is being done and what to expect. Also, having all past studies at one single place will avoid repetition at other places of the manuscript.

Good idea. This is the text I added:

"This paper further develops one of the techniques mentioned above: the Quantile Regression approach to post-process river forecasts first introduced by Wood et al. (2009) and further elaborated by Weerts et al. (2011) and López López et al. (2014). In a comparative analysis of four different post-processing techniques to generate confidence intervals, the quantile regression technique was one of the two most reliable techniques (Solomatine and Shrestha, 2009), while being the mathematically least complicated and requiring few assumptions. After Wood et al. (2009) presented the proof-of-concept for the Lewis River in Washington State at a conference, Weerts et al. (2011) published a formal study of quantile regression to compute confidence intervals for river-stage forecasts. Weerts et al. (2011) achieved impressive results in estimating the 50% and 90% confidence interval of river-stage forecasts for three case studies in England and Wales using QR with calibration and validation datasets spanning two years each. When applying QR to river forecasts, Weerts et al. (2011) transformed the deterministic forecasts and the corresponding forecast errors into the Gaussian domain using Normal Quantile Transformation (NQT) to account for heteroscedasticity. Building on Weerts et al. (2011) study, López López et al. (2014) compare different configurations of QR with the forecast as the only independent variable, including configurations without NQT and preventing the crossing of quantiles. They found that no configuration was consistently superior for a range of forecast quality measures (López López et al., 2014). "

5. *Selection of predictors and identification of the best set of predictors: The authors mentioned several predictors as most obvious potential predictors, however, did not select lead time and water levels at upstream locations, which are fairly easy to implement - this needs to be explained*
   - *On what basis predictors other than forecasts are selected, the authors explanation that the predictors yield good results for one location, hence adopted for all locations does not provide information*
   - *Current and the most recent water levels provide aid forecast because of persistence in streamflows (??)*
   - Revised section:

     "The challenge is to identify a well-performing QR model with a set of predictors that is both parsimonious and comprehensive. Wood et al. (2009) found rate of rise and lead time to be informative independent variables. Weerts et al. (2011) achieved good results

using only the forecast itself as predictor. Besides these variables, the most obvious predictors to include are the current water levels and those observed 24 and 48 hours ago, and the forecast error 24 and 48 hours ago (i.e., the difference between the current water level at issue time of the forecast that the error distribution is being predicted for, and the forecasts that were produced 24 and 48 hours earlier to predict the current water level). Additional potential independent variables are the water levels observed at gages up- and downstream at various times, the precipitation upstream of the catchment area, and the precipitation forecast.

Rates of rise and forecast errors were chosen to complement the forecast as independent variables for the following reasons. So instead of using it as an independent variable, separate QR models have been built for each lead time. After all, the best choice of independent variables might depend on lead time. Precipitation and precipitation forecast were not available for this study, because without direct access to the database at the National Climatic Data Center (NCDC) requesting that data is a very lengthy effort.

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

- *In 'error 24 or error 48' terms, '24 or 48' refers to the forecast lead time?*
No, and looking at your following description you did understand what I mean. It already says explicitly: “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 and

48 hours ago respectively)." I am not sure why this makes you think of lead time. Nonetheless, I extended it a bit:
"(i.e., the difference between the current water level at issue time of the forecast that the error distribution is being predicted for, and the forecasts that were produced 24 and 48 hours earlier to predict the current water level)."

- *One of the reviewer asked whether the selected variables are already used in the current forecast system or not? This is an important question to be answered because it will be interesting to see how these variables aid the forecast if already used in issuing the current forecast.*

  I asked the local RFC. Answer: "Currently, past forecast errors and rates of rise are not directly incorporated into the forecast model process thru a kalman filter, post-processor, or similar approach. Visually the forecasters can certainly see past performance of the streamflow and can manually make adjustments as appropriate to help account for some of these issues."

6. - *Additional details regarding forecasts (how these are generated, i.e., using QPF for 12 hours, limitations with the existing system, etc) and human influences, e.g., flow diversions and regulations, etc., need to be mentioned - this aids interpretation of forecast skill and uncertainty. Typically RFCs apply runtime modifications, i.e., MODs, on hydrologic model output. See HMOS manuscript for details.*
   - *it suggested to comment on general characteristics of streamflow (or water levels) including seasonality, snow-driven vs. convective rainfall dominated streamflows, etc.;* Honestly, this would make the paper a book. Looking at 82 different gages, it is impossible to detail all the human influences, etc. In my opinion, it is also superfluous if each and every paper on forecasting parrots the forecasting process and general characteristics of streamflow again. There is no added value to doing that, given that there are reference materials to look that up. I am attempting to publish a paper on the technique of quantile regression rather than a handbook on forecasting and hydrology.
   - *'Data' section to be a separate main section;*

     Done. I combined the first section of the Results section (former 3.1) with the data section and made it a separate Data section.

   - *Unlike the authors mentioned it is less likely that the selected two locations serve as an illustration for the study area for several reasons, e.g., these two stations neither upstream stations nor have large error. If possible, suggested to list out reasons that made to select the locations. A few stations, i.e., CIDI4, MORI2, MMOI2 and MIWI4, exhibited large forecast errors (See Figures 5 and 6) and showing results for these locations highlight importance of the QR technique.*

     We touched on that issue above already, see my response to your comment 5. Additionally, I would like to note that gages with extremely large errors are by definition

not representative of other gages. For example, gages with large errors are often right downstream of a dam. However, it is the stated purpose of this paper to develop a method that produces generally good results, rather than being tailored to extreme cases. So I chose gages that represented different conditions, but no extreme ones, rather than focusing on extreme gages as you suggest.

7. *Regarding validation criterion and conclusions regarding sample size and stationarity, lines 492-504:*

   The authors conclude that as small as '3' years of data for decent forecasts based on validation experiment in which BSSs are slightly increased or unchanged for increase in training years. The adopted validation criterion is much like real-time forecasting for the past (hindcasting), i.e., data is not considered beyond the forecast time even though it is available. Therefore, it does not guarantee similar performance for streamflow sequences other than observed. Note that the observed streamflows (or water levels) are one realization of infinite, and a technique would be called robust when it performs well for various situations. However, the adopted scheme is limited in that aspect. 'Leave a few years out cross-validation' scheme, a widely used scheme, simulates multiple conditions because the scheme allows to vary training length of data sets, and randomizes sequences of years, hence, the experiment offers stricter criterion and conclusions from the experiment may free from apparent biases;

   It is true that this scheme is not perfect. But let me explain the choice. First and foremost, I wanted to develop a method with practical value. So I wondered how much data was needed to make a reliable forecast with this method after there has been a human intervention or another change in the river. The leaving-one-year-out method does not represent this situation well. In that method, you use more data than might be available in reality. More data makes the statistical estimates better, but it does not answer the question I was trying to answer with this analysis. Furthermore, the-leaving-one-year-out approach does not take into account gradual changes. It is better to use consecutive years in this type of analysis that focuses on changes in the river. At the end, there will never a “guarantee” as you mention, regardless of the method used. I think, our misunderstanding is based on different understandings of the word “robustness”.

- *based on the validation results (as mentioned above, it is limited), the authors conclude that sample size as small as 3-years is good enough to yield skillful streamflow forecasts using the QR technique, and suggest using the technique in the presence of change in flow regime, i.e., non-stationarity. However, the authors' conclusion does not hold if tested locations do not exhibit any change in flow regime, i.e., stationary. In this regard, it is suggested to test locations against stationarity and trend;*

  Generally, research on US rivers has not found climate-related trends. However, individual rivers might exhibit trends over periods of time due to land use change or other human intervention. To identify those periods for 82 river gages is beyond the scope of this paper.

I am not sure why you think that my conclusion is not valid for stationary rivers. All I am saying is that you need very few years of data to make skilled forecasts. The assumption here is that those few years are representative for the following year, which implies either stationarity or slow change. When gradual changing is happening, the years further in the past are increasingly less representative of the year ahead.

- *The authors' response, citing of urbanization example, holds only when change in flow regime is smooth and not strong enough to influence (or reflect) future period of the record differently than the past (or calibration) period.*
  No, this is not what I am trying to say. I want to say that after an abrupt change you need only few years of data before you can make skillful forecasts again. And when there is gradual change you should limit yourself to few years of data, because the further in the past the data is the less representative the data is of the future when gradual change is happening.

To clarify all these misunderstandings, I extended the section:

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

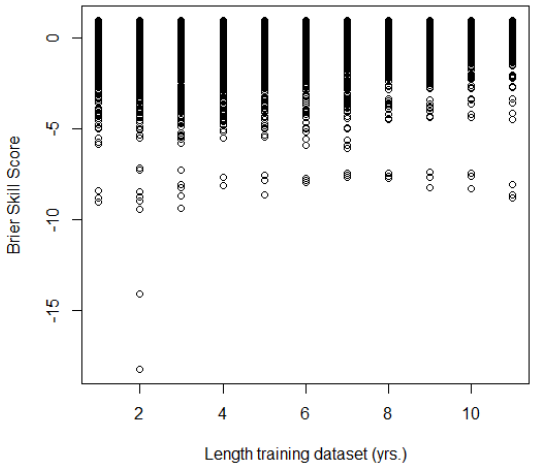
…

Figure 17 and Figure 18 show that at Henry and Hardin it barely matters for the BSS how many years are included in the training dataset. This finding is congruent with the fact that Weerts et al. (2011) were able to achieve outstanding results with the QR method using training datasets that were only two years long. Only needing short time series to define a skillful QR configuration implies (i) skillful forecasts can be produced not long after a step-change, and that (ii) the configuration parameters can be updated regularly so that gradually changing relationships between predictors etc. can be taken into account."

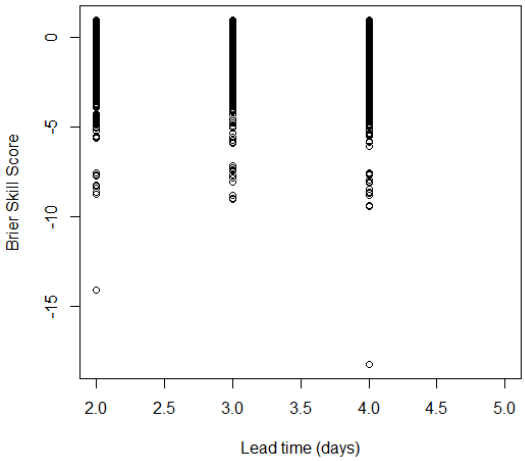
8. *Regarding fitting a regression between BSS and other variables, lines 512-541: A regression equation was developed for BSS as function of several variables (listed as river gages, forecast years, etc) to identify variables' influence on BSS. However, the section is not clear and request to provide the information for the following:*
   (i) Linear relationship is assumed between BSS and other predictors. Is not it possible for predictors to influence skill score nonlinearly? If so, the analysis and interpretation needs to modified.
   Good observation! Plotting the numerical dependent vs. the independent variables does suggest linear relationships, see below. As the regression indicates, there is no

relationship to be seen for the number of training years (Plot a). The relationship for lead time is barely visible as reflected by its small coefficient (Plot b).

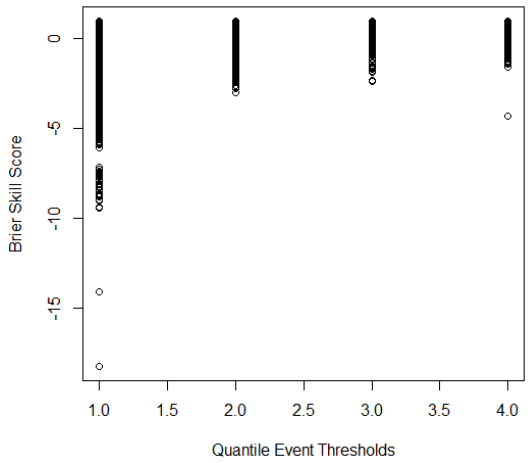
(a) Length training dataset



(b) Lead time



(c) Event thresholds



The plots show that these three independent variables seem to be ordinal. However, for lead time and training years, the distances between the "categories" is meaningful and zero is defined, so that they are really ratio variables, justifying linear regression.

I did make one change. On second thought, the distance between quantile event threshold is not equal: [Q10,25,75,90]. Plot (c) also suggest a somewhat concave relationship, rather than a linear one. So I included this variable as a nominal variable, i.e., as a factor, instead. Please see the new results in the table below (new Table 2). The overall findings didn't change. I changed the text where necessary.

	Coef	**St.Dev.**	
Intercept	-0.111	0.029	***

Event Thresholds			***
Q25	0.584	0.006	***
Q75	0.852	0.006	***
Q90	0.805	0.007	***
Forecast Years			***
2004	-0.259	0.019	***
2005	-0.083	0.017	***
2006	-0.136	0.017	***
2007	-0.123	0.016	***
2008	-0.205	0.016	***
2009	-0.128	0.016	***
2010	-0.141	0.016	***
2011	-0.127	0.016	***
2012	0.048	0.016	***
2013	-0.042	0.016	***
Lead Times	-0.021	0.003	***
Number of Years in Training Dataset	0.001	0.001	
River Gages			***
For the sake of brevity, the 82 river gages included in the regression as nominal variables have been omitted here.			
R^2		0.32	
Adjusted R^2		0.31	
P-Values: *** – <0.001; ** – 0.01: * – 0.05; . – 0.1			

(ii) What is mean by river gage as predictor, is that refers to basin area? What values are considered as predictor?

No. This does not refer to basin area. The river gages have been included as factors, i.e., nominal variables, to see whether it matters which gage is being predicted. It already says so in the text:

"A regression analysis was executed with the BSS score as the dependent variable and event thresholds, the river gages and forecast years as factorial independent

variables, and the lead time and number of training years as numerical independent variables."

(iii) What values considered when forecast year used as predictor? If year values 2001, or 2002, or... 2011 are used then how it is different from any random sequence of numbers?

What do you mean? In this case year numbers are labels for years and have been included in the regression as factors. I could have called the years "MMI", "MMII" as well. I think there might be misunderstanding what I mean with "factorial" and "numerical". So I changed those words to "nominal" and "ratio":

"A regression analysis was executed with the BSS score as the dependent variable and event thresholds, the river gages and forecast years as independent nominal variables, and the lead time and number of training years as independent ratio variables."

(iv) Provide additional information on lead time (is it one- or twenty four hours) and event thresholds?

Oh, sure. I clarified that:

"A regression analysis was executed with the BSS score as the dependent variable and event thresholds (Q10, Q25, Q75, Q90), the river gages and forecast years as independent nominal variables, and the lead time (one to four days) and number of training years as independent ratio variables."

(v) Why only one coefficient listed for flow threshold and number of training years, while multiple flow thresholds and different lengths of validation periods are considered in this study.

As I mentioned above, I included event thresholds and number of training years as independent numerical/ratio variable in the regression. Obviously, ratio variables only have one coefficient that tells you how much the dependent variable increases for every increase of the independent variable. It does not make sense to make gages and forecast years ratio variables, because for example it does not make sense to say the Brier Skill Score increases or decreases with every additional gage. It *does* make sense to say that the Brier Skill Score changes proportionally with increasing lead time, etc. That is why I included them as rational variables. I addressed your justified concerns about the relationship being linear above.

It is not clear how it is useful this whole exercise, i.e., regression and indirectly identifying key variables; when a variable selected as predictor then it indicates that that particular variable has influence. Is not? The other reviewer expressed similar concerns, and I doubt how much a typical reader can follow the whole exercise.

I am not sure what you mean by saying that selecting an independent variable for regression means that the variable has influence. That is obviously not true. Not every independent variable that is included in a regression analysis is automatically

significant, that should be obvious. Number of training years turns out not to be significant, and that is the main point, I am making here.
I included an additional sentence for the "typical" reader that you mention:
"This regression is meant to identify the factors to which the forecast performance as measured by the BSS is sensitive to, i.e., which factors statistically significantly impact forecast performance."

9. *Verification:*

(i) *Ensemble forecasts generated from the optimum configuration are not evaluated in detail. Given that entire PDF is available, it is suggested to calculate single-valued and ensemble forecast based verification metrics for all lead times and important water levels is suggested. These include correlation coefficient, mean error, reliability plots. This is a basic verification to makes sure that the technique is not degrading the current forecast. Currently, results presented only for Day 3 and Q75, and is suggested to extend the analysis;*

To go through your suggestions, including reliability plots for each lead time and river gage would result in messy plots. Additionally, you yourself criticized having too many plots already. The correlation coefficient and mean error are not applicable for probabilistic forecasts. I could compute them for the median (best estimate) of my estimated distribution, but it would tell me nothing about the estimated distribution itself. In the current version, I include one additional metric- the CRPS suggested by you, see comment on lines 475-480 – that are applicable for probabilistic forecasts. I now mentioned CRPS throughout, but I decided to drop the ROC. Its small variability did not make it worth it to include it throughout the paper. Except for the reliability plots, this choice of verification measures is in line with those used by López López et al. (2014) and more than Weerts et al. (2011) published in the same journal.

Finally, I am guessing you are talking about Figure 20 only showing a lead time of three days for the 75th percentile? That is because all others look similar. I added a table showing the difference for the metrics in Figure 20 for all event thresholds and lead times. Figure 20 serves as an illustration of part of this table. See my response to your comment 21 – Figure 20.

(ii) *BSS values are already at its maximum value, i.e., "1", however, it does not mean that the forecasts are perfect in all aspects. There are other aspects to be considered and suggested to discuss;*

Which section are you talking about? In Figure 20, the BSS values do not equal one. I discuss different components of the BSS and included Figure 20 for the very reason of highlighting different aspects of forecast performance.

(iii) *Lines 594-595: "The studied QR configuration perform less well for longer lead times, for gages far upstream a river or close to the confluences,..." Typically,*

regression equations do not perform well for long forecast leads due to lack of forecast skill and for both extremes (i.e., small [large] values over [under] estimated). It is suggested to discuss results, particularly, in terms of forecast skill and uncertainty, with respect to location of a gage (upstream or confluence) given large number of locations in a single basin;
Granted, it is not new that forecast performance decreases with increasing lead time, but for the sake of comprehensiveness I included it. Also granted, that forecasts perform less well for extremes. So I clarified this part a bit.

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

For a discussion of gage location please see my response to your third comment on Reviewer 2's comments.

(iv) Raw forecast error was interpreted only for a few locations with respect to upstream or downstream of a dam, but given the area of the study and number of locations, other factors such as basin area, location of site, time of concentration and flow regulations come into the analysis and a discussion is expected to tell the story on a high level;

The focus of the paper is to develop the QR method. I only plotted raw error to sketch the context. Also, I plotted raw forecast error for *all* stations. I explained why some gages are outliers with abnormally large errors, this is when I mentioned dams. I did *not* include dam operation into the QR method for any gage. I cannot possibly study and discuss the site details of 82 gages and their effect on raw forecast error. I also think that the marginal value of doing so is minimal, as most gages show normal error ranges. Again, section 3.1 is meant to give context and does not contribute to developing the QR method further. I moved this section from the Result section to Data to make clear that this is context information.

(v) Figure 24 presents BSS values for combination of lead times, water level magnitudes, different lengths of training years, and certain combinations exhibited negative values - this needs to be discussed. Also, it is interesting to note that the authors never discussed this particular figure in the manuscript. The verification part, either more verification metrics or presenting for different lead days, flow levels and locations could be improved.

Figure 24 was removed for the benefit of a table showing the stats of a number of verification metrics.

10. *Display of forecasts: Unlike verification metrics, which summarizes the forecast performance, visualization of the forecast highlights a few important aspects of the forecast. In this regard, it is suggested to develop PDF plots for forecasts of different water levels (key events), for different lead times and for different locations (one of the following locations, i.e., MORI2, MMOI2, MIWI4, CIDI4) showing good-, bad- and average- performance of the QR technique.*

    While the suggested figures would most certainly provide additional information, but the paper already has a lot of figures, as you yourself have criticized. Plotting forecast distributions for various event thresholds, lead times and locations would mean making several figures. Additionally, you suggest plotting these for extreme cases, which are often dominated by dam operations upstream. Predicting those dam operations goes beyond the scope of the paper and is not part of the QR method presented in the paper. The plots for those abnormal cases would also not provide information that is valid for the other 78 gages. In sum, I decided not to include these additional figures.

11. *Reg. Introduction, lines 33-34: Yes, the forecast uncertainty is key, therefore, the prior sentences should discuss how forecasting has been issued (i.e., single-valued or deterministic forecast), limitations (being a single-value the decision maker should consider only one value, which often times wrong and forces to take wrong decisions) and how providing forecast uncertainty helps in terms of better decision making. The fact that the QR technique provides a probabilistic forecast strengthens the point to discuss. The 'Introduction' section could be improved. One of the previous reviewers also suggested incorporating forecast uncertainty in the section.*
    I think a reader interested in a paper on advanced river forecasting methods can be expected to know the basics of forecasting and forecast uncertainty. So a discussion of forecast uncertainty has little added value in my opinion. To address your other points, I rearranged and extended this section:

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

    Currently, the National Weather Service does not routinely publish uncertainty information along with their deterministic short-term river-stage forecast (Figure 1). Given the many sources and complexity of uncertainty and the lacking user experience, it is easy to see how forecast users find it difficult to estimate the forecast error. Additionally, users might only experience such an event once or twice in their lifetime, so that they have no experience

to what extent they can rely on forecasts in such situations. Including uncertainty in river forecast would therefore be valuable, just as has been recommended for weather forecasts in general (e.g., National Research Council, 2006). Hopefully, decision-makers would then consider the whole bandwidth of possible future water levels, rather than focusing on the best estimate that is currently being published."

12. *Reg. QR procedure, (i) line 108 and Line 111: "...to estimate the error distribution..." & " ...our predictand is the probability of exceeding flood stages ...." It is not clear what QR technique estimates, "estimation of probabilities of exceedance probabilities", "estimation of errors" and "estimation of different percentiles of water levels" are mentioned at different places. These are different quantities, however, understandably related. This needs to be clarified, that is, mention the procedure in a few steps step-by-step, and suggested to use same terminology through out to the manuscript to improve the readability;*

    I clarified this incident, see below, and wherever else necessary.

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

    (ii) Reg. choosing BSS: the authors first reason, "best to choose a single measure" is not accurate, and there are other verification metrics, e.g., CRPS, can be decomposed. Therefore, it is suggested to improve the content so that it will be clear to the reader for choosing BS in calibration.

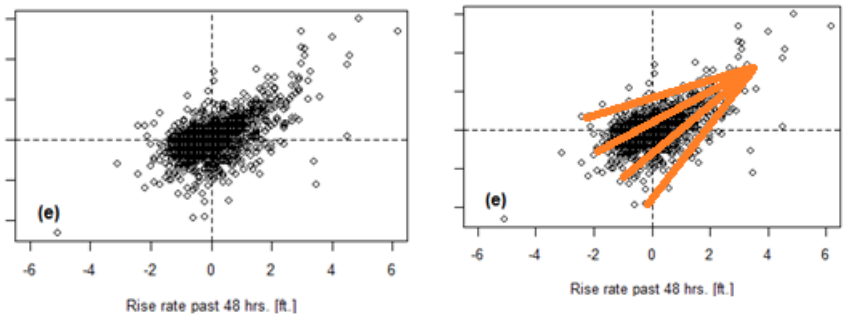
    To address your particular concern, I added a sentence:

    "We focus on the Brier Skill Score (BSS) – first introduced by Brier (1950) – to assess QR configurations for two reasons. First, to be able to determine the best set of predictors it is easiest to choose a single measure. *Second, the BSS allows us to study forecast performance at individual event thresholds.* Third, out of the available measures the Brier Score is attractive, because it can be decomposed into two different measures of forecast quality (see Equation 3)."

13. *Reg. Normal Quantile Transformation (NQT), (i) Figure 11 and 12 are interesting plots in the sense that they do not suggest any significant pattern between forecast error and other variables except 'error 24 hrs ago' in both original and NQT space. This warrants further*

discussion and highlights conditioning or categorization of the data (the authors referred it). However, it is interesting to note that the QR able to generate probabilistic forecasts with high BSSs, though, there is no visible pattern (attributed to power of regression techniques). A few questions: Did the authors develop these type of plots for all the '82' stations and analyzed? If so, how different are they?

No, we only plotted it for the two case studies. They look similar. Eighty-two gages are simply too many to plot. This could be subject to further research. However, I would note that quantile regression is different to linear regression in that it predicts percentiles. So a "visible pattern" looks very different for quantile regression than it does for linear regression. Here is an example to illustrate that point, a scatterplot with and without percentile trends:



Finally, this is not the point I was trying to make here. The point is that NQT does very different things to variables with different distributions.

Discussing in detail raw forecast pattern and how ensemble forecast via QR technique improves forecast in the single-valued as well as ensemble forecast sense adds important content. As of now the authors present results but does not inform on a basin level. This could be improved.

In essence, there are only two basins, the Illinois River and the Mississippi River which later join. Due to this lack of variability, there is little additional information to be had there. In Figure 23, I do talk about the impact of gage location.

Also, content in lines 409-412 suggested to be improved;

It is hard to improve things, if you don't tell me what is wrong. I already re-wrote that paragraph in response to your comment 3.

(ii) Lines, 141-142, "To be able to combine predictors of different natures...." Is not standardization or NQT scales down predictors of different range of values on to same scale and other benefits, for example, NQT to address heteroscedasticity. The content is not clear and does not answer the question why variables are not NQT'ed? If it is claimed that there is no pattern between forecast error and dependent variables, then the same argument holds for the relation in the original space as well. In addition, what about heteroscedasticity or some other regression assumptions? This needs to be discussed.

I changed this sentence to:

"We chose not to apply NQT, because four of five of our independent variables are already approximately normally distributed; only the forecast itself is not."

(iii) line 409, "Without a transformation into the normal domain, the scatterplot of forecast and forecast error does not show a trend" - Does not this imply applying NQT?

Yes, that is true. But applying NQT makes the other four independent variables worthless, so it is better to not use NQT in this study. As I have stated, this needs further research into questions such as, what happens if you only apply NQT to one independent variable instead of all. But this goes beyond the scope of this paper.

Note, I revised this section already in response to your comment 3.

14. *Regarding computing time, lines 292-294, the authors may want to mention the amount of computation time (providing basic details such as processor speed, etc) for a location, for all thresholds, for calibration as well as when applied in real-time, i.e., for a single day. Having this information in the conclusion section gives an idea what to expect if somebody wants to apply it.*

    Well, computation time is very arbitrary. The computations take much less time on my current laptop than on the one that I was using when doing research for this paper.

15. *Reg. lead time: Lines 10-11, 310, 314, 345: Forecast lead time of six days mentioned, however, in the data section it is mentioned as five days and results presented only for four days. Two points, i.e., (i) inconsistency in the statements and (ii) analysis only for four days when the data available for four days need to be addressed.*

    Well… I did have data for six days of lead time, but only analyzed four days. I analyzed the errors in Section 3.1 for six days, because it was computationally light. However, for the computationally heavy QR analysis I used only four days, which I clearly stated in line 310. I did change line 314 where it says five days.

16. *Line 52-63: has significant inaccuracies, suggested to go through the relevant papers, and a few comments here:*
    - *(i) Line 53-54: "HEFS includes two types of post-processors." - If the authors are referring to the HMOS and EnsPost, then it is not correct. See Demargne et al., 2014; neither content nor the Figure 1, which is a schematic of HEFS, mentions the HMOS. If the authors consider MEFP then needs to be mentioned explicitly.*
    - *(ii) Line 54-57: "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" - The HMOS technique corrects bias and estimates the forecast uncertainty associated with the single-valued forecast. I am not sure what the authors meaning of 'evaluates the uncertainty of each ensemble'. The HMOS uses single-valued forecasts and generates ensembles; although the HMOS can be applied on ensembles it was not done in its current settings.*
    - *(iii)Line 57-58: "while Hydrologic Ensemble Post-Processing (EnsPost) corrects bias and lumps the set of ensembles into one uncertainty estimate (Demargne et al., 2013; Seo, 2008)." - what is meaning of "lumps the set of ensembles into one uncertainty estimate" ?*

This needs to be modified. The EnsPost parameters account for hydrologic uncertainty, therefore, when EnsPost applied on HEFS generated ensembles, newly adjusted (or modified) ensembles together account for hydrologic uncertainty. Here is an excerpt from the Demargne et al., 2014 "In the HEFS, the EnsPost (Seo et al. 2006) accounts for the collective hydrologic uncertainty in a lumped form. Since MEFP generates bias-corrected hydrometeorological ensembles that reflect the input uncertainty, EnsPost is calibrated with simulated streamflow (i.e., generated from perfect future meteorological forcings) without any manual modifications of model states and parameters. The hydrologic uncertainty is, therefore, modeled independently of forecast lead time. The postprocessed streamflow ensembles result from integration of the input and hydrologic uncertainties and hence reflect the total uncertainty. "

(iv) *Lines 59-63: "HMOS performs a similar task as the QR approach presented here, but with two major differences." - (because the HMOS approach is relatively old:), the sentence (should) read " The proposed QR approach is similar to the HMOS approach however, different in following two ways:"*

(v) *Lines 60-61: "First, it relies on linear regression based on streamflows at various times as predictor, instead of using QR with several types of independent variables. " - the authors may want to be more clear, there are multiple facts to be mentioned: both studies use different regression techniques, i.e., the HMOS uses a simple linear regression whereas the proposed technique is based on quantile regression, and differ in terms of predictands (streamflow vs. water levels) and predictors (or independent variables), i.e., the HMOS technique uses recent observed flows, current flows and QPF information, and categorizes forecasted streamflows into multiple groups for which separate regression modes are developed. Nevertheless, both techniques develop separate equations for each lead time.*

(vi) *Lines 61-63: "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 □Rego□da et al., □□□□□.□ - The HMOS technique provides an ensemble of streamflows from which variety of statistics and products including exceedance probabilities are estimated.*

Revised section:

"HEFS includes a post-processor, the Hydrologic Ensemble Post-Processing (EnsPost). It models the hydrological uncertainty by estimating the probability distribution for each of the ensemble members which have been produced with varying input to account for input uncertainty (NWS-OHD, 2013). The Experimental ensemble forecast service (XEFS) additionally features the more parsimonious Hydrologic Model Output Statistics (HMOS) Streamflow Ensemble Processor, which estimates the total uncertainty (input and hydrological uncertainty) of single-valued streamflow forecasts based on conditional probability distributions (U.S. Department of Commerce/NOAA, 2012).

…

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

17. *Lines 475-480: Suggested to improve.*

(i) *The sentences are vague, "The latter is a version of the Brier Skill Score"; "Its perfect scores equals one". The authors would want to say that the RPSS value of one indicates a perfect categorical forecast instead 'perfect forecasts'; maximum value of RPSS is one.*

Section was removed in order to use CRPS more throughout the discussion, also see my response to your comment 9.

(ii) *not much difference in ROC values is observed for both combinations, however, for one combinations the values are pretty tight. It is not a safe assumption that the reader knows RPS and RPSS, therefore, it is suggested to list and brief the verification metrics. Otherwise, do not know what to expect and hard to follow;*

Section was removed in order to use CRPS more throughout the discussion, also see my response to your comment 9. An introduction of CRPS was included in the Method section:

“To verify that the results hold up for verification measures other than the BSS, we additionally use the Continuous Ranked Probability Score (CRPS). The BSS assesses forecast performance for one point on the forecast distribution, i.e., one event threshold. In contrast, the CRPS, defined by Equation 5, measures the forecast performance for the forecast distribution as the whole. Therefore, the CRPS cannot detect whether the forecast does better or worse in the tails. Instead, it is a measure of the forecast’s overall performance. The CRPS’ perfect score equals zero (e.g., Jolliffe and Stephenson, 2012; WWRP/WGNE, 2009).

Equation 5:

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

with CRPS – Continuous Ranked Probability Score

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

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

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

(iii) *CRPS is superior to RPS in the sense that it considers the entire PDF, therefore, unless the forecast to be verified for specific categories it is suggested to replace RPS with CRPS.*

You are right, I changed that:

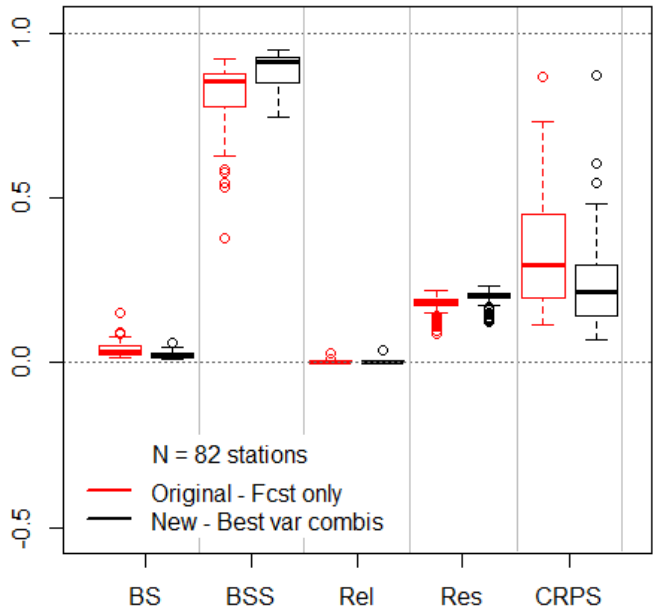


Figure 16: Comparison of the forecast-only QR configuration (i.e., only transformed forecast as independent variables) and using the best-performing joint predictor at each gage along various measures of forecast quality: Brier Score (BS), Brier Skill Score (BSS), Reliability (Rel), Resolution (Res), and continuous ranked probability score (CRPS). Lead time: 3 days; 75th percentile of observation levels as threshold.

18. *Lines 377-378: "...as if extreme events are characterized by different processes at different gages".*

(i) *This needs to be supported as it is contradictory to the intuition. In general, one would think that all stations in a basin dominated by a single mechanism. More over many stations of the study are along a major course of the river;*

(ii) *Different set of predictors in a basin might suggest either absence of information content in all predictors (not much different from random variables) or all predictors equally good.*

Or it means you cannot generalize across river gages for high water levels, which is my argument. It is not true that all predictors are equally good, they perform very differently at different gages. I don't want to gloss over that. It is also not true that the predictors do not provide information, because as I show throughout the paper – particularly Figure 17 and 18 – that this method delivers good results and the additional independent variables do improve BSSs.

Studying the QR method with a focus on extreme water levels (>90th percentile) is enough work to write another paper on it.

Anyways, I omitted the frequency analysis in the new version for the benefit of showing that improvements in BSSs are statistically significant. .

19. *Reg. one-size-fits-all: (i) Lines 435-441: It is not clear the benefit of exercise "one-size-fits-all". Based on Fig. 7- 10, it is clear that there is a no specific combination that performs best all the time, therefore, this idea, one-size-fits-all, is less likely encouraged. Even though the experiment yields similar skill scores, it is less likely that one will use same combination for all locations because of the larger study area and importantly, each location needs to be calibrated separately. Therefore, as long as identifying the predictor combination is viable in terms of computing resources, nobody opts for one-size-fits-all. However, it is a good scientific exercise but requires more discussion instead simply presenting results. Therefore, it might be better to remove this section unless detailed discussion followed; (ii) Lines 489-491: It is not clear why one-size-fits-all QR configuration is analyzed for various training data lengths when it's performance found to be not good as the best QR configuration - this needs to be explained.*

(i) I rewrote this section:

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

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

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

However, using the best joint predictors results in much better performance for major flood stages than the one-size-fits-all approach. The average difference between average BSSs amounts to 0.21 to 0.38 (Table 5). Given that a BSS for a forecast with skill ranges between one and zero, this is a substantial difference. In sum, the same joint predictor can be

used for all river gages without much loss in performance, except for extremely high water levels.

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

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

”

(ii) You are right, I updated text and figure:

“The impact of the length of the training dataset on the configuration’s performance measured by the BSS was assessed for *the best joint predictor* (i.e., rates of rise and forecast errors as independent variables for all gages) for Hardin and Henry on the Illinois River. Each year between 2003 and 2013 was forecast by QR configurations trained on however many years of archived forecasts were available in that year, i.e., the forecasts for 2005 is produced by a model trained on less data than those for 2013. Then, the BSS for that year (e.g., 2005 or 2013) was computed. ”

20. *Regarding process complexity:*

(i) Lines 432-434: if the forecast error does not have much variability and magnitude of the forecast error is small, then, at least intuitively, less random noise and easy to model, but the technique did not perform well. Suggested to explain;

(ii) Lines 445-446: without much analysis, attributing to small sample is not a good idea because the process might be complex too;

(iii) Line 506-507: Exceedance probabilities for 10th and 25th percentile comprise a lot of data, however, the performance is not at expected levels particularly for the 10th percentile. Forecasting water levels of low exceedance probabilities has another

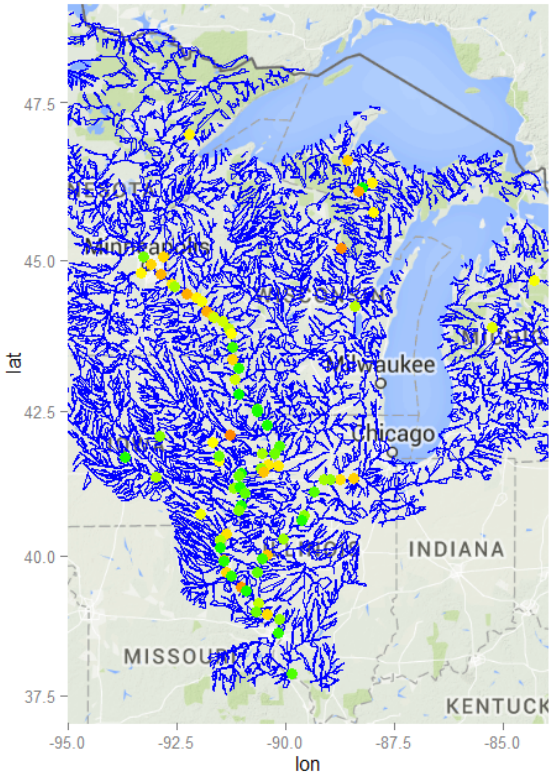
limitation, therefore, large data not always mean good skill. This message needs to be passed.
Revised section:

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

21. *Figures:*

 Figure 2, suggested to overlap basin delineation with river network and river names

Plotting all the rivers makes the plot rather messy:



Instead, I made the color scale logarithmic to have more variance in color:

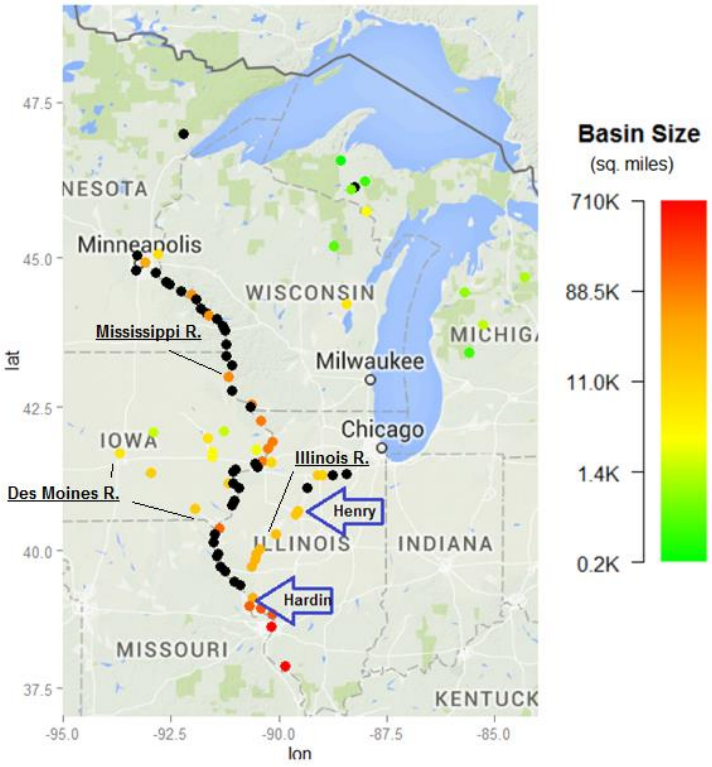
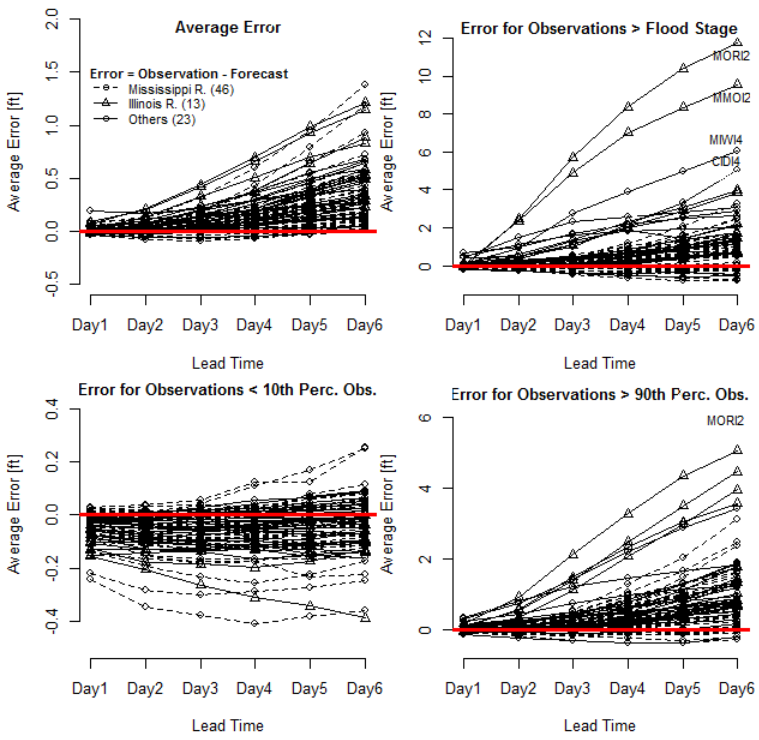


 Figure 5, suggested to vary range of y-axis for figures in the left column such that error pattern can be seen. Errors are large for more than a few locations for obs>90th percentile. This needs to be discussed.

New Figure 5 (now Figure 4):



The larger errors across all gages for obs>90th is already being discussed in the text:

"When only observations exceeding the 90th percentile of all observations are considered, the underestimation becomes more pronounced, averaging 0.29 feet for three days of lead time and 1.14 feet for six days of lead time (Figure 4c, Figure 5). When only looking at observations that exceeded the minor flood stages corresponding to each gage, the underestimation averages 0.45 feet for three days of lead time and 1.51 feet for 6 days of lead time (Figure 4d, Figure 5)."

This is followed by a description of the most extreme cases, but I cannot describe every other case with large error individually.

 Figure 20, It is not clear why results presented only for Q75 and for lead day '3'. It is suggested to comment on othe2r percentiles and for other lead days, and present results for at least one of the action flood stages. A four panel figure (without zoomed values) may provide more information.

I included a table instead, so that you have all the information. I kept Figure 20 as an illustration:

"The fact that the Brier Score can be de-composed into reliability, resolution and uncertainty allows a closer look at which improvements are being achieved by including more predictors than just the forecast. Table 4 summarizes the results of paired t-tests comparing the forecast-only and the best performing joint predictor for each gage for the components of the BSS.

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

Event Thresh.	Lead Time	Brier Score	Brier Skill Sc.	Reliabil.	Resol.
Q10	**1 Day**	-.012***	.20***	-.002***	.008***
	2 Days	-.014***	.25***	-.002***	.010***
	3 Days	-.016***	.28***	-.002***	.012***
	4 Days	-0.17***	.27***	-.001*	.013***
Q25	**1 Day**	-.018***	.13***	-.003***	.013***
	2 Days	-.023***	.16***	-.002***	.018***
	3 Days	-.027***	.18***	-.003***	.021***
	4 Days	-.031***	.20***	-.002***	.025***
Q75	**1 Day**	-.005***	.03***	.000	.011***
	2 Days	-.011***	.05***	-.000 **.**	.015***
	3 Days	-.016***	.08***	-.000	.021***
	4 Days	-.025***	.12***	-.000	.028***
Q90	**1 Day**	-.003***	.03***	-.000**	.013***
	2 Days	-.005***	.06***	-.000*	.015***
	3 Days	-.010***	.10***	-.000	.019***
	4 Days	-.015***	.15***	-.000*	.025***

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

The Brier Score and the Brier Skill Score mainly improve, because the resolution increases when using the best-performing set of independent variables at each gage (Table 4). Visualizing the improvement in forecast performance for a lead time of three days and the 75th percentile threshold (Q75), Figure 16 illustrates that the forecast-only QR configuration as studied by Weerts et al. (2011) has high reliability (i.e., the reliability is close to zero). So reliability improves statistically significantly for lower water levels (Q10, Q25), but the magnitude of improvement in reliability is by one order smaller than the improvement in resolution (Table 4). “

Minor Comments

Line 25: "...ignorance of the potential forecast errors" - Do the authors mean bad forecasts or ignoring the forecast uncertainty, suggested to be more clear.

Revised sentence: “River-stage forecasts are no crystal ball; the future remains uncertain. The past has shown that unfortunate decisions have been made, because of users’ unawareness of the magnitude of potential forecast errors (Pielke, 1999; Morss, 2010).”

Line 28: "due to the infrequency of extreme events and " - suggested to remove 'of extreme events'.

Sure, why not: “Unfortunately, it is exactly in those situations that forecast are the most uncertain, i.e., forecast errors are the largest, due to the infrequency and the subsequent scarcity of data.”

Lines 5-6: The QR technique (can be) applied to predict other than flood stage exceedance probabilities, therefore, the sentence needs to be modified to reflect it. Also, the technique uses variables other than single-valued flood stage forecasts as predictors, and it needs to be mentioned. "This study applies Quantile Regression (QR) to predict various water level, including flood stage exceedance probabilities using combinations of forecasts and observed water levels."

Revised sentence: “This study applies Quantile Regression (QR) to predict exceedance probabilities of various water levels, including flood stages, with combinations of deterministic forecasts, past forecast errors and rates of water level rise as independent variables.”

Lines 6 -9: *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.□ - The technique might be dealing the forecast errors, however "predicting forecast errors" is not intuitive. One would say that instead doing forecast of water level and then forecasting 'forecast error', it is better to combine them or other questions would popup. In this regard, from a readability perspective it is better to say that forecast uncertainty is estimated or provide one more line why 'forecast error' is important.*

Revised sentence: “A computationally cheap technique to estimate forecast uncertainty is valuable, because many national flood forecasting services, such as the National Weather Service (NWS), only publish deterministic single-value forecasts.”

Line 27-31: could be improved

See my response to your comment 11.

Lines 36: Those addressing --->Those accounting.

That would put “accounting” twice into one sentence. Not pretty.

Lines 35-44: Suggested to improve it.

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

Line 41 & Lines 244-245: ".....set of predictors that is both parsimonious and comprehensive": Use of parsimonious: Is not parsimonious technique/model refers to a model that uses fewest variables and yields desired performance? Neither 'predictors that is both parsimonious..' nor 'parsimonious configuration' is clear to me.

Well, a configuration is one of many possible realizations of a model (e.g., the QR model in this case). Here is the revised version: “The challenge is to identify a well-performing QR model with a set of predictors that is both parsimonious and comprehensive.”

Line 43: "less resource- intensive" ??

Revised: “The alternative, i.e., the lumped quantification of uncertainties, is a less demanding in development and computation run-time, but glosses over many of the finer details of uncertainties (Regonda et al., 2013).”

Lines 64-66: “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.” The authors might want to change the sentence something like “… does not account/model different types of source of uncertainty, but rather quantifies the total uncertainty”.

Revised: “In contrast to an ensemble approach such as HEFS, the statistical post-processing in this paper does not model different sources of uncertainty, but rather quantifies the total uncertainty in a lumped fashion.”

Lines 66-67: "....actors with limited resources.." decision makers (instead actors). Remove limited resources, it is vague definition, how it is defined, often times downloading data and reformatting is requires good resources, instead, say, a simple approach that uses relatively less resources, etc.

Omitted.

Line 76:"... meta-analysis ..." - what its meaning?

Revised: “In a comparative analysis of four different post-processing techniques to generate confidence intervals, the quantile regression technique was one of the two most reliable techniques (Solomatine and Shrestha, 2009), while being the mathematically least complicated and requiring few assumptions.“

Line 80: "This paper further develops one of the ...." - This paper applies one of the...

No, it actually develops it further, rather than just applying it.

Lines 91-93: Should not be in the "Data" section?

Okay, deleted. It already said the same thing in the data section.

Lines 94-100: Content in these lines suggests what is being done in this paper in detail, this needs to be either cut short to one or two sentences and merged with the last paragraph of the Introduction or should be moved to Methods section.

This paragraph is important to help the reader decide whether the paper is relevant to him and should read on after the introduction. Nonetheless, I shortened it a bit:

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

Lines 101-106: The standard practice is that each section briefed in one or two sentences so that the reader will have an idea what to expect and what to read.

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

Lines 101: "The Method section reviews quantile regression, introduces the performance measure, and discusses the performed analyses and data." - (i) the QR regression is not reviewed at least from mathematics; (ii) why 'data' part discussed in the 'Method' section when it has a separate section.

Revised: “The Data section describes the used data and reviews the overall forecast error for the dataset. The Method section introduces quantile regression and the performance measures, and discusses the performed analyses.”

I combined the first section of the Results section (former 3.1) with the data section and made it a separate Data section.

Lines 112-113: suggested to modify sentence, i.e., "The study tests the robustness of the technique by calculating and analyzing its performance across different locations, lead times...."

That sentence ceased to exist as a consequence of your comment 12.

Lines 119-120: "...linear quantile regression..." what is mean by linear QR?

Well, exactly what it says: that the regression is linear. You could predict percentiles also with non-linear regression.

Lines 129-134: This should be in the "Introduction" section to show wide range of applications of QR technique, otherwise dilutes and distracts what the authors want to convey.

Well, this is the introduction section for the method. To me the introduction of a paper is meant for the reader to decide whether the paper is relevant to him. To put the introduction of the QR method into the Introduction section would make the Introduction section bloated and would lead to an unnecessary disconnect of information about QR.

Line 138: "...on this study.." is it refers to Weerts et al or Bogner et al.?

Revised: “Building on Weerts et al. (2011) study, López López et al. (2014) compare different configurations of QR with the forecast as the only independent variable, including configurations without NQT and preventing the crossing of quantiles.”

Line 140: "omitting NQT" --> without NQT

See revised sentence above.

Line 140: "They find that " --> They found that

Revised: “They found that no configuration was consistently superior for a range of forecast quality measures (López López et al., 2014).”

Lines 146-147: "...a fixed effects models..." It is not clear.

That is commonly known and widely used mathematical method:
https://en.wikipedia.org/wiki/Fixed_effects_model

Lines 150-153: Suggested to present equation generalizing for multiple variables, so that both Equations 1 and 2, i.e., with NQT and without NQT, respectively, maintain consistency in interns of information content and can be rewritten for as many variables as the reader wants.

Revised:

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

Lines 160: the second part of the error corresponds to error, however, it is not clear from both equations how error for different percentiles look like or formulated. Suggested to be addressed.

What do you mean? The quantile regression (in the square brackets) estimates the error distribution, i.e., the percentiles of the error. The forecast is then added to the error distribution to get a distribution of forecasts. Revised:

"The second part of the equations stands for the error estimate based on the quantile regression configuration for each error percentile τ and lead time."

Lines 175: "....optimize model performance it is best to choose a single measure." Having a single verification metrics in the objective function simplifies interpretation, however, it is not correct to say that it is best to choose a single measure. The EnsPost technique (Seo et a.,) uses combination of two measures.

Revised: ". First, to be able to determine the best set of predictors it is easiest to choose a single measure."

Lines 179-183: It is suggested to provide more insights on what is 'uncertainty', what contributes to this uncertainty, and how it is different from forecast uncertainty.

Revised: "The third component is uncertainty. This type of uncertainty describes the uncertainty inherent in an event caused by natural variability. It is narrower than forecast uncertainty, because the latter additionally includes the uncertainty that is caused by imperfections of the forecast model, i.e., the variables that could explain some of the uncertainty have not been identified or correctly parameterized yet."

Lines 184: In equation (3), specify explicitly the terms corresponds to reliability, resolution and uncertainty.

Revised:

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

Reliability Resolution Uncertainty

Lines 211-226: Suggested to be improved

Revised: "Resolution measures the difference between the predicted probability of an event on a given day and the historically observed average probability. For example, imagine a gage where flood stage has historically been exceeded on 5% of the days in a year. If every day at that gage the probability of exceeding flood stage is forecasted to be 5%, the resolution of those forecasts would be zero. After all, the difference between the predicted frequency and the historical average is zero. So a forecast with higher resolution is better. (e.g., Jolliffe and Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009). In Figure 2, the curve for a forecast with good resolution would be steeper than the dashed line that represents the historically observed frequency (climatology). It follows that forecasters should strive to maximize the area in Figure 2a representing resolution. In absolute terms, the resolution can never exceed the

uncertainty inherent to the river gage, as represented by the third term in Equation 3 (e.g., Jolliffe and Stephenson, 2012; Wikipedia, 2014; WWRP/WGNE, 2009).”

Lines 213-218: Interpretation related to resolution is unclear.

See revised paragraph above.

Lines 223-226: "The latter likewise quantifies how much better than the reference forecast....". Is not ROC widely used to verify ability of a technique in terms of discrimination of events.

I deleted the reference to the ROC.

Lines 227-228: "A forecast possesses skill, i.e., ..." suggested to modify the sentence.

Revised: “A forecast performs better than the reference forecast (in this case the historically observed frequency), if it (the red line) is inside the shaded area in Figure 2b. Then the forecast is said to have “skill”.”

Lines 230-231: The equation 4 allows to analyze and interpret forecast performance in different aspects, and is suggested to discuss instead simply providing it. Having an equation and not discussing distracts from the main content.

Revised: “The Brier *Skill* Score (BSS) equals the Brier Score normalized by the historically observed frequency, i.e., the resolution and reliability terms are being divided by the uncertainty term (Equation 4). In contrast to the Brier Score, this makes the Brier Skill Score comparable across gages with different frequencies of a binary event.”

Lines245: "...lead time", How a lead time considered as predictor? Or the authors mean developing configurations specific to a lead time?

See my response to your comment 5.

Lines 258-260: "It was also found that season and months are not significant ..." Does it mean the authors performed some kind of analysis and came to the conclusion. If so, what kind of analysis was performed, and might want to provide some information in the discussions section.

See my response to your comment 5.

Lines 260-261: "Probably, the time of the year is reflected in the observed water levels...". It is suggested to develop a plot of mean monthly streamflows to supports the authors' assumption. It is not required to present the graph in the manuscript but they can develop these plots for all the locations and see whether do they see any seasonality to support their assumption. Most likely, all locations in this basin are dominated by winter snow and spring snowmelt.

See my response to your comment 5.

I think a plot is superfluous here. It is well-known that in spring time, when the snow melts, water levels are higher and more volatile in those regions.

Line 266: "..which joint predictor on average and most often leads to the best out-of-sample results.." It is not clear the meaning of " ..average and most often..." and "out-of-sample"

I am not sure what is unclear about this, they are all standards words/expressions.

Revised: “Based on the Bier Skill Score, it was determined which joint predictor delivers on average the best out-of-sample forecast performance for various lead times and water levels..”

Line 279: "Computations " consists of calculation of parameters for various QR configurations and BSS needs to be discussed. Change section name either to estimation of parameters or merge with the previous section.

I changed the previous title to “Choice of independent variables” and this one to “Computational Process”.

Line 280: "The output of our QR application" -- remove 'our', unless plan to take a patent :)

Done.

Lines 280-281:" The output of QR application to river forecasts is the probability ..." suggested to be modified. It implies that the final output is probabilities not error values (?)

Revised: “The final output of the computational process is the probability that a certain water level in the river or flood stage is exceeded on a given day.”

Lines 284-286: Repeated, see lines 144-145 and lines 262-265.

I think this is a question of readability. Mentioning information exactly once rather every time it is relevant makes the paper harder to understand.

Revised: “To recap, depending on configuration (Table 1) the forecast itself, the rates of rise and forecast errors serve as independent variables. “

Lines 292-294: How much additional time it will take if exceedance probabilities of closer intervals are considered?

Well, every time you cut the intervals into half the computing time doubles, because the number of percentiles you predict doubles.

Lines 296-298: Redundant, remove it.

Done.

Lines 300-301: "...interpolating to determine the exceedance.." is it repetition of what mentioned in the lines 293 or different?

Yes, but here I am clarifying, what I mean with the “last computational steps.”

Lines 299-303: can be simplified, i.e., the technique verified for eight exceedance probabilities, which of four are....and the other four correspond to different stages of flood.

That would omit the reason why I am doing that and give the incorrect impression that the entire computational process is repeated for eight event thresholds.

Revised: “To study whether the various combinations of predictors perform equally well for high and low thresholds, these last computational steps (i.e., interpolating to determine the exceedance probability for a certain water level and calculating the BSS) were repeated for eight event thresholds: the 10^{th}, 25^{th}, 75^{th}, and 90^{th} percentile of observed water levels and the four decision-relevant flood stages (action stage, and minor, moderate, and major flood stage) of each gage.”

Lines 303-305: Known information, redundant.

To my knowledge, I do not mention this anywhere else in my paper. It is very crucial information that flood stages correspond with different percentiles at different gages. See also my response to your comment 1a.

Lines 308-309: Make it direct and simple. The robustness of the technique was tested analyzing technique's performance for 82 gage locations using different lengths of data sets for five different lead times.

Done.

Line 322: "for a sufficient number of days" - not clear what is meant by this.

Revised: “For 82 of those gages, forecasts have been published daily for at least two years, and are not inflow forecasts.”

Line 322: "... not inflow forecasts ..." - does it mean water level forecasts?

No, it means inflow forecast. How many cubic feet of water will flow in the reservoir rather than what will the water level be. That is what the sentence following that one says: “The latter have been excluded from the forecast error analysis because they forecast discharge rather than water level.”

Line 323-324: Is this technique applicable to only water level forecasts, not for streamflow forecasts?

No. Water level is readily convertible to streamflow and vice versa, if you have the discharge rating curve for each gage. Converting the streamflow forecasts would have been a lot of work, plus the conversion tables are associated with a lot of uncertainty themselves. Finally, many inflow forecasts are close to dams, so they are not representative for the entire sample.

Line 327: I understand that the one of the reviewer suggested CDF plot of basin areas, however, interpretation of the plot makes the plots more valuable. It appears that the two most downstream gages have large areas, does it have any implication in the analysis of results?

Naturally, the drainage area of a gage increases the further downstream you go. So I actually think that this CDF plot is stating the obvious. Because you know how much water is coming down the river, i.e., how much water the parts of the drainage area more upstream have yielded,

the downstream river gages are easier to forecast, unless a river confluence or dam are just upstream. I make that point with Figure 23.

Lines 335-336: instead using 'probably', the authors may want to cite the references or instances.

Well, that is a very common thing to happen when two rivers join. River engineering 101. Revised: "Therefore, it can experience backwatering, when the high water levels in the Mississippi River prevent the Illinois River from draining."

Line 346: Is "0.41 feet" small amount?

Well, compared to the other errors shown in the figure 0.41 feet is not much. I removed the "only" from the sentence though.

Lines 357-359: How reliable is the assumption, is it possible to verify? Is this behavior seen for other locations that are of upstream of a dam? How did the QR technique fair on these locations?

Verified. That´s what the people at the RFC in Tulsa told me. Studying the QR method for gages close to dams could be the next paper, but is beyond the scope of this one.

Revised: "*Possibly,* the forecasts performed so poorly there, because the dam operators deviated from the schedules that they provide the river forecast centers to base their calculations on."

Lines 364-365: Comment on over estimation of forecasts for a few locations, i.e., left side of black vertical line of Figure 6.

I am not sure what you are looking for other than what is already in the text. It is pretty obvious that there is a clear bias towards underestimation. Compared to underestimation, overestimation (left side of black vertical line) is negligible. No news in terms of magnitude of error either. The forecast performs worse for higher water levels.

Lines 384: "...further out one..." --> at long lead times

Sentenced ceased to exist.

Lines 383-384: In regression, often times, adding additional predictors increases skill, therefore, this needs to be verified making sure that the increase in BSS is significant.

Okay, I included paired t-tests in the section discussing the improvement in BSS:

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

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

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

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

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

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

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

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

	1 Day				2 Days				3 Days				4 Days			
	Diff.	**T-stat.**	**Df**	**p-val.**	**Diff.**	**T-stat.**	**Df**	**p-val.**	**Diff.**	**T-stat.**	**Df**	**p-val.**	**Diff.**	**T-stat.**	**Df**	**p-val.**

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

”

Line 394: (i) Mention explicitly how the BSSs ranked, the lower the BSS smaller the rank or the other way?; (ii) values @ y-axis are decreasing, mention it explicitly.

(i) Sentence added: “For each river gage, the combinations have been ranked by BSSs. The best performing combination was ranked first, the worst performing 31st.”

(ii) Captions Figure 9 and 10 extended: “The y-axis is reversed, so that an increasing trend indicates increasing performance.”

Line 397: "For action stage and minor flood stage, a slightly increasing trend is still visible" – How significant is this increasing trend? However, similar conclusion can be drawn from the Figure 8 and suggests merging of two sets of Figures 7 & 8, and 9 & 10.

As to merging figures, see my response to your comment 1.

I added a table to further investigate the trend and show that it is significant:

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

For extremely high water levels, this trend favoring larger joint predictors gradually reverses (Figure 8). The trend remains statistically significant, but its coefficient decreases for higher event thresholds (Table 2a) until it changes signs for major flood stages (Table 2b). A

possible explanation is that combinations with more variables suffer from overfitting for extreme event thresholds characterized by data scarcity.

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

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

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

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

(a) PERCENTILES of observed water levels

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

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

(b) FLOOD STAGES

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

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

Figure 9: Average rank for each joint predictor for one to four days of lead time and two percentiles of observed water levels. Vertical gray lines correspond to the configurations that include forecast as one of the predictors. The y-axis is reversed, so that an increasing trend indicates increasing performance.

Figure 10: Average rank for each joint predictor for one to four days of lead time and two flood stages. Vertical gray lines correspond to the configurations that include forecast as one of the predictors. The y-axis is reversed, so that an increasing trend indicates increasing performance."

Line 395: "...the higher that set of predictors will rank [high] on average". Is not [high] is missing from the sentence.

No. The higher is even included in the sentence expert you provide. It is a "The more …, the higher … " sentence structure. More examples of such sentences here (https://en.wiktionary.org/wiki/the#Adverb):

Adverb

the (*not* comparable)

1. With a comparative or more and a verb phrase, establishes a parallel with one or more other such comparatives.

   The *hotter,* ***the*** *better.*

   The *more I think about it,* ***the*** *weaker it looks.*

   The *more money donated,* ***the*** *more books purchased, and* ***the*** *more happy children.*

   It looks weaker and weaker, ***the*** *more I think about it.*

2. With a comparative, and often with *for it*, indicates a result more like said comparative. This can be negated with *none*.

   *It was a difficult time, but I'm **the** wiser for it.*

   *It was a difficult time, and I'm none **the** wiser for it.*

   *I'm much **the** wiser for having had a difficult time like that.*

Line 403: "...four or more variables." Maximum number of predictors that are present in a combination are five, therefore, change it to five.

No, I mean ≥4 rather than >4 or =5.

Lines 402-405: All figures from 7 through 10 are based on average values of all basins and does not specify a particular combination. Therefore, the part of the statement that says "...based on the same joint predictor of four or more ..." is not clear.

I don't understand. The x-axes of those four figures are the specific combinations as defined in Table 1. So the particular combinations are definitely specified. I clarified in the sentence:

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

Lines 409-412: (i) It is not clear how patterns in the NQT space provide information on the regression that developed in non-NQT space;

See my response to your comment 3.

Line 415: "Vertical gray lines indicate joint predictors including the forecast" --> Vertical gray lines correspond to the configuration that includes forecast as one of the predictors.

Revised: "Vertical gray lines correspond to the configurations that include forecast as one of the predictors."

Line 430-431: "...mean and decreases the standard deviation (Figures 14 and 16)." Add "..deviation of skill scores". ---> "....the configuration yields similar range of large BSS scores."

Revised: "Compared to using only the forecast as an independent variable, using the best combinations of forecast, rates of rise and past forecast errors as predictors at each gage not only increases the mean BSS, but also decreases the standard deviation of skill scores across gages, i.e., performance becomes more consistent (Figures 13 and 14)."

I don't know what that last part of the comment refers to.

Line 434: Figure 16 cited for average error, whereas Figure 16 consists of BSS values. BS corresponds to error in forecasts probabilities, but not BSS.

Changed that reference to Figure 5.

Lines 435-439: The content can be simplified in less number of sentences, i.e., "Additionally, a one size- fits-all, which uses predictors excluding the forecast (combination 30), was tested to investigate whether customizing the QR configuration...". Lines from 437-439 can be safely deleted, this information is already presented.

Adopted the one sentence and deleted the other as suggested.

Lines 439 - 441: " ... river gage deviation (Figure 15, 16)." What is mean by river gage deviation?

It seems like "deviation" is a stray word. I deleted it.

Lines 442: "...this last conclusion..", what is it? It is not clear.

It means exactly what it says. What I am about to say is only relevant for the last of all the findings I have just describe. I changed "conclusion" to "finding", hopefully that helps.

Lines 443: "...does improve the BSSs considerably deviation" what it means? I think the authors might be referring range of BSSs.

Somehow, this is another stray word. Strange. I deleted it.

Lines 508-511: The content is out of context, i.e., data, accessibility and other details suggested to be mentioned either in the Data section or as limitation in Conclusion section.

No, that is not out of context. I am saying that results could be improved with more data, but I qualify that statement by mentioning that more data is usually not available.

Line 559-562: Needs to be modified. The proposed study does not develop QR application, rather it adds useful information in terms of identifying useful predictors for the study location.

In my opinion, by including multiple predictors, it does develop the method further. Former published studies relied on one predictor only. I don't know develop QR, but I develop it *further.* It says so in the text.

Lines 566-567: (i) " ...QR error models should be a function of rate of rise and lead time" - 'should' in the sentence mandates the use of rate of rise and lead time in the QR technique, but QR technique can be developed with other predictors too. Is not? (ii) Why lead time is not considered as predictor in the study?

(i) Revised: "This confirms Wood et al.'s (2009) finding that rate of rise is a valuable predictor for QR error models."

(ii) See my response to your comment 5.

Lines 577-578: "..do not combine well with the forecast" --> "no information is added from the forecast"

That is not true. It is not that the forecast does not add information. Rather its distribution is so different (very skewed) from those of the other predictors (normally distributed) that they do not combine well.

Lines 579-581: The NQT related content is not clear. see earlier comment.

See my response to your comment 3.

Lines 589-591: "This means that the danger remains..." Not clear, in general, the forecast user thinks the forecast is going to happen unless it is proven to be not a good forecasts from various aspects.

Revised: "When forming a joint predictor, the independent variables rates of rise and forecast errors do not combine well with the forecast itself, because the forecast has a skewed distribution, while the other predictors are approximately normally distributed. The forecast becomes an excellent predictor for linear quantile regression after NQT. However, the other four variables lose their value as predictors when subjected to NQT, because their original distribution is already approximately normal. Therefore, it is difficult to combine predictors with different distributions. A possible solution could be to define QR configurations for subsets of the transformed data or to experiment with only subjecting some of the predictors to NQT."

Lines 591-593: This suggests importance of ensemble forecasting, shouldn't be in the Introduction instead at the end of the manuscript?

I am talking about the uncertainty in uncertainty estimation here, so second order uncertainty here. This should be part of the conclusion, because the results have shown how much uncertainty there is in uncertainty estimates.

Lines 595-599: Repeated elsewhere, and suggested to shorten.

Yes, but a conclusion is the summary of the most important findings, so by definition it repeats findings.

Lines 601:602: (i) The BSS values are almost close to "1", so, I am not sure what is mean by high brier scores. The authors may referring low event thresholds, at long lead times and for some locations, and it is suggested to be specific as it is Conclusion section; (ii) "..more robustness" - not sure meaning of the sentence.

I am not sure where you see BSS consistently equal to one. Quite on the contrary, I think the analysis shows that there is room for improvement. There was a whole section talking about robustness, so I think you should know by now what it is and how it is lacking. You are asking me to repeat the conclusion here to motivate why there should be future work.

Lines 602-605: I would rather delete it, and tell these are the possible ways. The initial experiments may turn out to be false when evaluated in detail, therefore, unless evaluated in detail it is not suggested to draw conclusions and mention.

Revised: “Observed precipitation, the precipitation forecast (i.e., POP – probability of precipitation) and the upstream water levels are promising candidates, because the forecast used in this study includes the precipitation forecast for only the next 12 hours.”

Lines 606-610: Limitations with data access should go in the data section, or elsewhere but not in the conclusions/future work section.

I am talking about the predictors that I am proposing for further research, rather than predictors that I have used in the study. I think it is valuable for the reader to know that it is difficult to get the data to do what I propose.

Lines 612-613: why it is not considered in this study when it is easy to consider it?

Because you have to manually match them, which is work-intensive: “Upstream water levels can easily be included *after manually determining the upstream gage(s) for each of the 82 NCRFC gages*.”

Lines 605-606: "Presumably, this is the case, because the forecast used in this study includes the precipitation forecast for only the next 12 hours." --> This should be mentioned in the 'data' section as well, how these forecasts are generated, etc with forecasts aspects, limitations.

No, I don’t think so. I am suggesting using these as predictors and I am motivating why I think so. I did not include these predictors in the current paper, so this sentence would be misplaced in the data section.

Lines 614-616: "early trials" --> "initial experiments"; the technique is very sensitive to the training data set, whereas in this study it is mentioned that '3-years' data is good enough. Both are contradictory statements.

No, that is not true. Clustering variability is a very different mathematical approach than linear quantile regression. What is true for one method is not necessarily true for the other. Nonetheless, I deleted that sentence as it is speculative.

Lines 617-618: redundant, already mentioned earlier in the Conclusion section

Revised: “Further study should investigate, why the QR approach works less well for low than for high event thresholds, and identify possible solutions.”

Lines 618-: "Further study should investigate..." move it to Line 599.

No. Line 599 is the conclusion where I summarize my most important findings. The section we are talking about is called “Future Work”, so that is where a sentence starting with “Further study should…” belongs.

Lines 619-620: "The current study focused on extremely high event thresholds..." The authors considered a wide range of thresholds, and it should be in the Methods section or as a summary in the first paragraph of the Conclusion, but not at the end of the manuscript

By looking at flood stages, I study high water levels much more intense than low water levels. It is in the "Future Work" section to indicate that the current study leaves work to be done there. I am not sure why you want this in the Method section.

Lines 621-625: The content should be either in the 'Data' and 'Introduction'. Some of the content is repeated.

I am repeating myself. This is the "Future Work" section detailing what work still needs to be done and giving arguments why it needs to be done. This most certainly does not belong into the Introduction nor into the Data section.

Lines 626-629: Redundant. Given only '31' combinations, I think it is better to calibrate the model for each combination and then evaluate instead going for stepwise QR. I understand the other reviewer's comment, therefore, having information on 'computing time' answer a few questions. It is not clear how 'stepwise regression' provide better safeguards against over fitting, which is much more at the discretion of model/technique developer. In general, adding additional predictor improves skill whether it is in brute force or step-wise regression.

Isn´t "*calibrate the model for each combination and then evaluate*" what I have done in this study? Or what do you mean?

Regarding computing time, see my response to your comment 14.

Stepwise regression would alleviate overfitting, because it would remove superfluous predictors. Stepwise regression is more elegant than brute force, because in a brute force approach all combinations are computed and the best determine. In stepwise regression, a statistical criterion measuring the added value of a variable decided whether that variable should be included in the model or not.

Finally, it is not true that adding an additional predictor is always beneficial. Among others, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) explicitly account for that (both not easily available for QR, if I am correctly informed.).

Reviewer 1 comments:

Comment on 294, 7 #2: Over fitting of the technique particularly for large flows is mentioned, which mean in real-time the technique's performance is not guaranteed. Implication of this suggested to be discussed.

What do you mean? First of all, performance is never guaranteed. Second, the paragraph is making the argument that configurations with less independent variables perform better for extreme water levels. Overfitting is given as the reason why configurations with more independent variables do not perform as well. I revised the paragraph a bit to make this clearer:

"For moderate and major flood stage, combinations with fewer independent variables rank higher on average. The most likely explanation is that combinations with more variables suffer

from overfitting. The infrequency of exceeding major and moderate flood stage means data scarcity that can lead to overfitting when using more predictors. ”

Figure in response to comment 298, 7: It is interesting to note that plots of err and err24 are almost similar, what these plots suggest?

Of course the distributions are the same. The forecast error that was err24 one day is err48 the next day. So looking at the distribution across all days, those are the same numbers and thus the same distributions, but when forecasting any given day they are different.

The authors highlighted the importance of the NQT, i.e., "..., 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.", however, the authors developed QR technique without NQT'd variables.

That is true when using forecast as an independent variable, but it is not the case for the variables that we have used. Also see my response to your comment 3.

The authors response to the reviewer's comment 298, 7(2) which deals calculation of correlation coefficient is not clear.

In my response to comment 298,7 (2) I refer to my answer to comment 294,7 (2), which I already discussed above. None of them mentioned a correlation coefficient, so I don’t know what you are referring to. If you meant 298,7 (1+3), please see my response on your comment 13 to learn more about visually detecting quantile trends. That should clarify why correlation coefficients are not helpful in this case.

The reviewers' comment, 302,14, is not addressed.

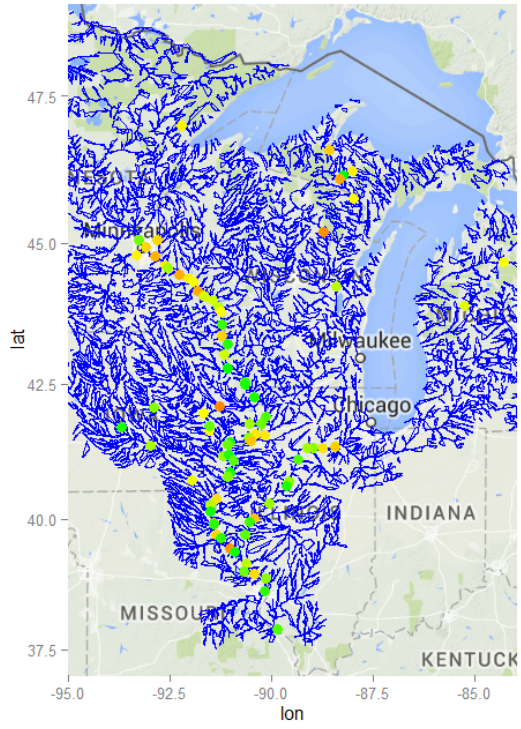
The comment you are referring to reads. “*It’s not clear why these variables do not lend themselves to transformation – please be more specific and speculate as to why you are finding this. Are they distributed such that the transformation reduces their correlation with the predictand? It’s an interesting result, but not intuitive why it should be.*” Please, again see my response to your comments 3 and 13. The other variables are already approximately normally distributed, so NQT doesn’t have much value for them.

Reviewer 2 comments:

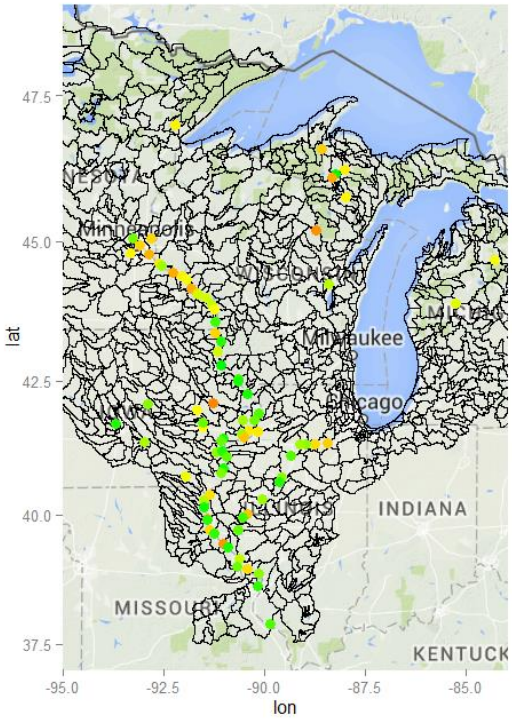
- **"How were the independent variables chosen: " -** the authors response is that these variables found to be good predictors for two other locations, hence used for all location. It is suggested to provide additional insights on why rates of rise and fall are chosen to start with ?
  Answer: See my response to your comment 5.
- **"I included the basin sizes in the figure, because those are in my opinion more relevant for this study than the delineations:"** - Except two downstream locations of the network, all locations are approximate of the same size, hence they are in some kind of green color and line in CDF plot is very steep. However, at least half of study locations

do not have basin area, hence coloring circles do not add much. Moreover, not detailed discussion of skill with respect to size of the basin is presented. For these reasons, it is not clear why the authors feel basin sizes/basin area is relevant. Basin delineations provides basin area information, particularly for locations that do not have basin area information, in this regard, it is suggested to overlap basin delineations on the current maps with color circles.

Answer: Changed color scale, see my response to your comment 21 – Figure 2. Basin delineations and river networks do not provide much visibly detectable information either:



Rivers



Basins

- **"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."** - Is it a reasonable assumption that locations off from the major course of a river can be treated as upstream locations? Not necessarily though, is not?. Basin delineations greatly assists in this aspect. Locations at the beginning of the main stem and in the north east of the basin (i.e., off from the main stem) exhibited either low BSSs or negative BSSs - this needs to be discussed, i.e., what factors went against forecast skill. High BSS values are seen for most of the downstream

locations, is increase in skill due to decrease in variability of water levels? For downstream locations, the flow is routed from upstream locations, which did not exhibit high skill scores. Does it mean most of the skill is coming from recent observed flows, needs to be discussed. The reviewer's comment related to time of concentration influence is suggested to be discussed.

Answer:

First of all, those color points do not represent BSSs. They are coefficients from the regression indicating how much better or worse than average (i.e., the regression intercept) the QR method performs at each gage. So negative values do not mean that the technique has no skill.

"The performance of the forecast additionally depends on the river gage. The coefficients of the river gages, included as factors in the regression, have been excluded from Table 6 for the sake of brevity. Instead, Figure 19 maps the geographic position of the river gages with the color code indicating each gage's regression coefficient. The coefficient indicates the method's performance at the particular gage as compared to the average performance. The coefficients are lower, and therefore the Brier Skill Scores are lower, for gages far upstream a river, off the main stream, and those close to confluences.

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

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

- ***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."** -- The

authors' response is not clear and is suggested to detail what exactly is being calculated in the Methods section.
Answer: As can be seen from equations 1 and 2, forecast error is the dependent variable of quantile regression. One more computational step is necessary to compute the exceedance probability from there. I tried to make that even clearer in the text:

"In the second step, these QR configurations are used to predict percentile by percentile the distribution of forecast error for each day in the verification dataset (the second half of the dataset). Effectively, for each day in the verification dataset, a discrete probability distribution of forecast errors is predicted. Adding the single-value forecast to the forecast error distribution results in a distribution of predicted water levels. Each predicted percentile π contributes one point to that distribution.

Then, we calculate the probability with which various water levels (called event thresholds hereafter) will be exceeded. The probability of exceeding each water level is computed by linearly interpolating between the points of the discrete probability distribution that was computed in the previous step. Next, the Brier Skill Score is determined based on predicted exceedance probability for all days in the verification dataset."

- ***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.** Not sure about the authors response. The QR is one type of probabilistic forecasting techniques and in this study the QR applied in the hydrology context.
Answer: Okay, that answer seems a bit off in hindsight. But the rationale for probabilistic forecasting is mentioned in the introduction:

"Currently, the National Weather Service does not routinely publish uncertainty information along with their deterministic short-term river-stage forecast (Figure 1). Given the many sources and complexity of uncertainty and the lacking user experience, it is easy to see how forecast users find it difficult to estimate the forecast error. Additionally, users might only experience such an event once or twice in their lifetime, so that they have no experience to what extent they can rely on forecasts in such situations. Including uncertainty in river forecast would therefore be valuable, just as has been recommended for weather forecasts in general (e.g., National Research Council, 2006). Hopefully, decision-makers would then consider the whole bandwidth of possible future water levels, rather focusing on the best estimate that is currently being published. "

- **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." -- ROC is widely used in the hydroclimatology discipline as well, suggest to explore the papers.

Answer: I omitted the ROC in favor of using the CRPS more throughout the result section.

- ***"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."** -- It might be good to have a configuration that falls back onto climatology, although I doubt it giving good results particularly for flood forecasting; Using climatology as one of the predictors has been in the practice.

  Answer: What do you mean with a configuration that falls back onto climatology? To use the historical average as a forecast? That by definition would be a forecast with zero resolution, thus not a very good one. Climatology as a variable does not make any sense either. The QR model is parameterized for each gage separately. There is no variance in climatology for a single gage, so it cannot be an independent variable.
- **Regarding 293, 9:** That authors cited cost-benefit and lots of computing time for not to consider all percentiles, however, given the availability of efficient computing resources and algorithms, the argument may not hold unless the technique takes lots of resources. In this regard, it might be good to provide time estimates.

  Answer: Well, unfortunately the argument did hold for my situation as a PhD student. Including all percentiles would have taken forever and it I am still of the opinion that it would not have been worth the effort. I think the results as compared to interpolating between percentiles [5,10,15,20,…] would have not been much better. Of course, this can always be improved. But to make this paper feasible I had to make some choices that I am not going to change at this late stage. Plus, I don't think that I am the only one having limited computing resources. So I do think that my argument holds that a method requiring less computation time is accessible to more users. On computing time estimates, please see my response to your comment 14.
- ***"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."** - The authors explanation holds as long as change is smooth, however, same does not hold if the trend/non-stationarity is strong and change is drastic right after the calibration. Nevertheless, the authors conclusion is based on small sample and based on a validation test that has limitations.

Answer: See my response to your comment 7.

- ***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).** The authors response in 295, 7 (1) suggests change in forecast performance along the course of the river, however, does not discuss in terms of time of concentration/adding past error, etc.
  Answer: In my mind, concentration time is a function of size of drainage area/basin size. Drainage area increases when moving more downstream. So upstream gages have less drainage area. One of my conclusions of Figure 23 is that upstream gages are harder to forecast. One reason could be the faster concentration time, see my earlier response.
  It is correct that for gages with faster concentration time, maybe the error 6 or 12 hours ago would be a good predictor to include. However, studying for which gages that is true, goes beyond the scope of this paper. You could easily write a whole new paper on it. Given that most gages are on large rivers like the Mississippi and Illinois, choosing 24 and 48 hours was the more reasonable thing to do.
- **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.** - Suggested not to draw conclusion from initial experiments which may or may not change.
  Answer: Omitted. See my response to your comment on line 602.
- ***308,2:" (1) what's the difference between the filled circles and the open circles?"*** - Suggested to use either filled or open circles to remove the confusion.
  Answer: Classical case of pedantry, but if it makes you happy:

Combi	fcst	err24	err48	rr24	rr48	Combi	fcst	err24	err48	rr24	rr48
1	●					16	●	●	●		
2		●				17	●	●		●	
3			●			18	●	●			●
4				●		19	●		●	●	
5					●	20	●		●		●
6	●	●				21	●			●	●
7	●		●			22		●	●	●	
8	●			●		23		●	●		●
9	●				●	24		●		●	●
10		●	●			25			●	●	●
11		●		●		26	●	●	●	●	

12	●			●	**27**	●	●	●		●
13		●	●		**28**	●	●		●	●
14		●		●	**29**	●		●	●	●
15			●	●	**30**		●	●	●	●
					31	●	●	●	●	●

fcst = forecast; rr24, rr48 = rise rate in the past 24 and 48 hours;

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

- ***"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:"** – The authors could have contacted the RFC personnel to understand how typically forecasting is done as well as to provide insights on why forecast being excluded from the optimum combination and whether errXX or rrXX is already used in the forecasting.
  Answer: See my response to your comment 5.

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

Regards,

Frauke Hoss, Paul Fischbeck

Response to Reviewer #2

July 27th, 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,

Thank you for reviewing the paper yet again. I did work through all your comments and made changes where necessary. As suggested by the editor, I will only go in on your main comment, your concern about " the lack of 'hydrological reasoning' to explain (a) the choices made in configuring the post-processor of hydrological forecasts as well as (b) to interpret results found."

(a) Choice of predictors:

I hope the revised section below, motivates my choice of independent variables to your satisfaction. For more information, please additionally refer to my response to comment 5 of the other reviewer.

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

Rates of rise and forecast errors were chosen to complement the forecast as independent variables for the following reasons. So instead of using it as an independent variable, separate QR models have been built for each lead time. After all, the best choice of independent variables might depend on lead time. Precipitation and precipitation forecast were not available for this study, because without direct access to the database at the National Climatic Data Center (NCDC) requesting that data is a very lengthy effort.

Forecasts and observed water levels were readily accessible from NCDC databases. Rates of rise and forecast errors can be derived from those two. As will be shown in section 4.3, it is mathematically challenging to combine independent variables with different distributions

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

(b) Interpretation for results

I revised the Result section quite intensively in response to the other reviewer. However, here it is the section that you were mainly referring to, I think:

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

The forecast performance was found to vary statistically significantly across all tested dimensions, except the number of training years (Table 6). This results in a very wide range of BSSs (Figure 13 and 14). Accordingly, for the user, it is particularly difficult to know how much to trust a forecast, if the performance depends so much on context. Likewise, this is case for the QR configuration based on the forecast only (not shown).

Table 6: Regression results sensitivity analysis

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

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

Regarding the forecast quality for each forecast year, the regression is slightly biased. The earlier years are included less often in the dataset with on average less years' worth of data in their training dataset, because, for example, unlike for the year 2013, ten years of training data were not available for the year 2006. Nonetheless, the regression indicates that 2008 was particularly difficult to forecast and 2012 relatively easy, i.e., they are associated with relatively low and high coefficients respectively (Table 6).

The performance of the forecast additionally depends on the river gage. The coefficients of the river gages, included as factors in the regression, have been excluded from Table 6 for the sake of brevity. Instead, Figure 19 maps the geographic position of the river gages with the color code indicating each gage's regression coefficient. The coefficient indicates the method's performance at the particular gage as compared to the average performance. The coefficients are lower, and therefore the Brier Skill Scores are lower, for gages far upstream a river, off the main stream, and those close to confluences.

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

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

Figure 19: 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