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 and Paul S. Fischbeck

Summary:

The manuscript presents application of Quantile Regression (QR) to generate probabilistic river forecasts for short lead times. For the study, five predictors, i.e., *forecast, error 24, error 48, rate of rise 24* and *rate of rise 48*, are considered, and optimum configuration (i.e., predictor combination) is identified evaluating the QR technique for all possible 31 configurations. The experiment, identification of optimum QR configuration, is repeated for a combination of lead time and water level magnitudes (i.e., different exceedance probabilities), for all selected 82 gages of the North Central River Forecast Center (NCRFC) region. The optimum QR configuration exhibited pattern in terms of number of predictors with respect to low and high water levels, and short- and long-lead times, however, it is unique in terms of exact combination of predictors, for each location, and for a combination of lead time and water level magnitude. The optimum QR configuration is compared with two other QR configurations, i.e., one uses only NQT'ed forecast as predictor whereas the other, one-size-fit-all, uses four predictors excluding forecast, and results indicate better performance of the optimum QR configuration across all lead times and water levels. Performance of the best QR configuration is verified in the hindcast mode and it was concluded that as small as '3-yrs' sample size is enough to apply the QR technique to produce forecasts of high BSS values.

The manuscript shows possibility of providing probabilistic forecasts with a simple (post-processing) technique on a large spatial region. Importantly, the technique neither requires large resources nor modifications to the existing forecasting system. Thus, the study is an important contribution in hydrologic ensemble forecasting field. The manuscript reads reasonably good (particularly selection of predictors), however, it can be improved significantly in terms of organization, usage of words and tightening content of the manuscript. The results presented for all gages but lack a deep discussion on a basin level as well as detailed verification metrics. It is obvious that significant effort was went in, hence lots of figures. However, cutting down figures and content and maintaining consistency greatly improves readability and holds the reader attention. The authors addressed one of the reviewers major comment, i.e., cited previous literature. However, a few other major comments, e.g., physical insights for selecting predictors other than forecast, discussion for not having forecast in the best configuration, validation criterion, '3'-yr sample size and nonstationarity, need to be addressed. My comments are below mentioned, and some of these reflect previous reviewers' comments.

Comments:

1. <u>Regarding cutting down figures:</u> Currently results are presented for four different water levels (i.e., 10^{th}, 25^{th}, 75^{th}, and 90^{th} percentile of observed water levels) and for four decision-relevant flood

stages (i.e., action- minor-, moderate-, and major- flood stage). The selected '8' water levels are unique and have important role in decision making. However, given that the QR technique generated forecasts did not differ significantly for all selected '8' water levels (hence no detailed discussions in the manuscript with respect to categories), it makes sense to present results for '4' water levels instead '8' water levels. The four water levels can be two extreme low and extreme high water levels. Presenting results only for '4' water levels reduces number of figures at least by '9' without information loss. Currently, the manuscript has '22' figures, which is quite large. It is good to have figures, but multiple figures telling the same information makes the reader to distract and confuse. Therefore, it is strongly recommended to reduce the figures.

   a. Figures 7 and 8 (similarly, Figures 9 and10) can be combined into one single file presenting information for four different water levels
   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 authors can use a simple plot of with more panels or CDF plot, but the text needs to be interpreted accordingly. Note that decision-relevant flood stages are compared using only two methods whereas other flood stages are compared using three methods, however, it should not be a problem with an explanation.
   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.
   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)

2. Combining two sections (3.2.1 and 3.2.2) and two sets of figures, Lines 367-370: What is the significant difference between "a" and "b", i.e., 'best and worst most often' and 'best BSSs on average'? Related content and two sets of figures (7 & 8 and 9 & 10) tell the same story with little differences, therefore, it is suggested to cut down the content and figures (can be replaced with one figure four water levels), and make the point simple and easy to understand, i.e., QR configuration with more number of predictors yield better skills forecasts for medium to large water levels compared to very large water levels.
3. Reg. excluding of the forecast from the best configuration, lines 406-407: "*The combinations including the forecast perform less well than those that exclude it*." 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. 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. 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.

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 . 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.

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.
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; (ii) On what basis predictors other than forecasts are selected (which is one of the previous reviewers' comment as well), the authors explanation that the predictors yield good results for one location, hence adopted for all locations does not provide information; (iii) Current and the most recent water levels provide aid forecast because of persistence in streamflows (??) (iv) In 'error 24 or error 48' terms, '24 or 48' refers to the forecast lead time? It is suggested to provide more information what these terms mean and how they come aid of regression to improve forecast. The term 'error 24/48' in the regression equation reflects training technique based on forecasts' recent performance, and having discussion in this aspect provides light on selection of predictors; 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. The authors response is that they do know details of the current forecast, however, contacting the River Forecast Centers (RFC) personnel makes sense.
6. Data section, line 317: 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. The RFCs, in general, provide the requested information so it is

suggested to contacting them for basic details of the current forecast; (ii) 'Data' section to be a separate main section; (iii) it suggested to comment on general characteristics of streamflow (or water levels) including seasonality, snow-driven vs. convective rainfall dominated streamflows, etc.; (iv) 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.

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; (ii) 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; (iii) 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.
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 (ii) What is mean by river gage as predictor, is that refers to basin area? What values are considered as predictor?; (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?; (iv) Provide additional information on lead time (is it one- or twenty four-hours) and event thresholds?; (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. 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.

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; (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; (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; (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; (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.

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.
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.
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; (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.

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? 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. Also, content in lines 409-412 suggested to be improved; (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. (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?
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.
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.
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 (Regonda et al., 2013)." - The HMOS technique provides an ensemble of streamflows from which variety of statistics and products including exceedance probabilities are estimated.

17. Lines 475-480: Suggested to improve. 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.; (ii) not much difference in ROC values is obesrved 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; (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.

18. Lines 377-378: "...as if extreme events are characterized by different processes at different gages". 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.
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.
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.
21. Figures:
    - Figure 2, suggested to overlap basin delineation with river network and river names
    - 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.
    - 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.

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.
- Line 28: "due to the infrequency of extreme events and " - suggested to remove 'of extreme events'.
- 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."
- 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.
- Line 27-31: could be improved
- Lines 36: Those addressing --->Those accounting.
- Lines 35-44: Suggested to improve it.
- 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.
- Line 43: "less resource- intensive" ??
- 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 sources of uncertainty, but rather quantifies the total uncertainty"
- 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.
- Line 76:"... meta-analysis ..." - what its meaning?
- Line 80: "This paper further develops one of the ...." - This paper applies one of the...
- Lines 91-93: Should not be 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.
- 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.

- 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.
- 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...."
- Lines 119-120: "...linear quantile regression..." what is mean by linear QR?
- 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.
- Line 138: "...on this study.." is it refers to Weerts et al or Bogner et al.?
- Line 140: "omitting NQT" --> without NQT
- Line 140: "They find that " --> They found that
- Lines 146-147: "...a fixed effects models..." It is not clear.
- 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.
- 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.
- 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.
- 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.
- Lines 184: In equation (3), specify explicitly the terms corresponds to reliability, resolution and uncertainty.
- Lines 211-226: Suggested to be improved
- Lines 213-218: Interpretation related to resolution is unclear.
- 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.
- Lines 227-228: "A forecast possesses skill, i.e., ..." suggested to modify the sentence.
- 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.
- Lines245: "...lead time", How a lead time considered as predictor? Or the authors mean developing configurations specific to a lead time?
- 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.

- 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.
- 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"
- 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.
- Line 280: "The output of our QR application" -- remove 'our', unless plan to take a patent :)
- 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 (?)
- Lines 284-286: Repeated, see lines 144-145 and lines 262-265.
- Lines 292-294: How much additional time it will take if exceedance probabilities of closer intervals are considered?
- Lines 296-298: Redundant, remove it.
- Lines 300-301: "...interpolating to determine the excedance.." is it repetition of what mentioned in the lines 293 or different?
- 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.
- Lines 303-305: Known information, redundant.
- 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.
- Line 322: "for a sufficient number of days" - not clear what is meant by this.
- Line 322: "... not inflow forecasts ..." - does it mean water level forecasts?
- Line 323-324: Is this technique applicable to only water level forecasts, not for streamflow forecasts?
- 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?
- Lines 335-336: instead using 'probably', the authors may want to cite the references or instances.
- Line 346: Is "0.41 feet" small amount?
- 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?
- Lines 364-365: Comment on over estimation of forecasts for a few locations, i.e., left side of black vertical line of Figure 6.
- Lines 384: "...further out one..." --> at long lead times

- 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.
- 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.
- 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.
- Line 395: "...the higher that set of predictors will rank [high] on average". Is not [high] is missing from the sentence.
- Line 403: "...four or more variables." Maximum number of predictors that are present in a combination are five, therefore, change it to five.
- 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.
- 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;
- 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.
- 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."
- 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.
- 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.
- Lines 439 - 441: " ... river gage deviation (Figure 15, 16)." What is mean by river gage deviation?
- Lines 442: "...this last conclusion..", what is it? It is not clear.
- Lines 443: "...does improve the BSSs considerably deviation" what it means? I think the authors might be referring range of BSSs.
- 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.
- 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.
- 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?

- Lines 577-578: "..do not combine well with the forecast" --> "no information is added from the forecast"
- Lines 579-581: The NQT related content is not clear. see earlier comment.
- 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.
- Lines 591-593: This suggests importance of ensemble forecasting, shouldn't be in the Introduction instead at the end of the manuscript?
- Lines 595-599: Repeated elsewhere, and suggested to shorten.
- 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.
- 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.
- Lines 606-610: Limitations with data access should go in the data section, or elsewhere but not in the conclusions/future work section.
- Lines 612-613: why it is not considered in this study when it is easy to consider it?
- 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.
- 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.
- Lines 617-618: redundant, already mentioned earlier in the Conclusion section
- Lines 618-: "Further study should investigate..." move it to Line 599.
- 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
- Lines 621-625: The content should be either in the 'Data' and 'Introduction'. Some of the content is repeated.
- 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.

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.
- 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?
- 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.
- The authors response to the reviewer's comment 298, 7(2) which deals calculation of correlation coefficient is not clear.
- The reviewers' comment, 302,14, is not addressed

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 ?
- "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.
- "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.

- *"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.
- "*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.
- *"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.
- "*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.
- 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.
- *"300,8: I think it mea"ns 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
- "*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.

- "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.
- *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.
- *"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.