Response to review comments of Dr. Jullie Demargne on the manuscript "Alternative configurations of Quantile Regression for estimating predictive uncertainty in water level forecasts for the Upper Severn River" By Lopez, Verkade, Weerts and Solomatine, 2014

We would like to thank Dr Demargne for the time and effort spent to review our manuscript. Her comments and suggestions are very thoughtful indeed and will have a positive contribution the quality of this manuscript. According to your comments, we will try our best to revise and improve this manuscript. In the revised manuscript, the revision will include the following aspects:

General comments:

Comment 1:

In the introduction on pages 3812-3813, the authors should clearly mention the 2 main sources of uncertainty in hydrological forecasts: the meteorological uncertainty from the forecasts used as inputs to hydrological models (e.g., precipitation and temperature), and the hydrologic uncertainty, which also includes uncertainty from human influences (e.g., flow regulations). The authors should clarify that statistical post-processing could be used either to model the total uncertainty when using hydrological forecasts for its calibration or to model the hydrologic uncertainty when calibrating with hydrologic simulations, while the meteorological uncertainty is modeled separately, with precipitation and temperature ensembles for example. The pros and cons of these two approaches should be mentioned (see discussion in e.g., Regonda et al 2013 and Demargne et al. 2014).

Answer:

We will modify the beginning of the introduction in terms of the definition of the two mains sources of uncertainty: input uncertainty and hydrological uncertainty:

"Forecasting may reduce but can never fully eliminate uncertainty about the future. Hydrological forecasts will always be subject to many sources of uncertainty, including those originating in the meteorological forecasts used as inputs to hydrological models (e.g. precipitation and temperature), and in the hydrological models themselves (e.g. boundary conditions, model structure, model parameters and human influences). Informed decision making may benefit from ……"

We will clarify the possible uses of statistical post-processing techniques according to reviewer's comments:

"It can be used as either an alternative or additional step to hydrological ensemble forecasting. In many hydrological forecasting applications, postprocessing is used in combination with deterministic forecasts (but it can also be applied to ensemble hydrological forecasts if available; see, for example, Verkade et al. (in preparation)…"

Comment 2:

The section on the forecast verification strategy and the verification terminology used in the paper needs to be improved. When referring to forecast skill, the authors generally mean forecast quality, one aspect being the forecast skill when using a given verification metric

and a specific reference forecast. Referring to "forecast quality" is important since the authors analyzed also the reliability, sharpness and event discrimination of the forecasts. For the reference forecast used in skill computations, the selection of the reference forecast needs to be clearly defined. For the Relative Operating Characteristic Score (ROCS, see pages 3824 and 3835), the reference forecast is a climatological or unskillful forecast, for which the Probability of Detection is equal to the Probability of False Detection. For the other skill scores, the choice of the sample climatology as the reference (see page 3824) should be clearly explained (I would add the term "unconditional climatology"); the forecasts from the QR0 procedure could have been selected to show the performance gain compared to the most basic post-processing technique tested in the experiments. A probability threshold of 0.95 is used to stratify the sample of forecast-observation pairsalyze the forecast performance for the subsample corresponding to the 5% higher observed events. However the authors need to clarify that, for the Brier Score and the ROCS (see page 3824, 3833 and 3835), the probability threshold is used to define the binary event to be observed and forecasted.

Answer:

We will modify the manuscript according to the three main reviewer's suggestions in this general comment:

a) We will modify forecast skill by forecast quality where appropriate.
b) We will modify lines 13-15 in Page 3824 to clearly define the reference forecast considered for each verification metric as follows:
   "In the case of BSS and CRPSS, the reference score comprises that of the sample unconditional climatology; in case of the ROCS, the reference score is the ROCA associated with climatological or unskilled forecast, associated with an unskilled forecast which states that the probability of event occurrence is equal to the probability of non-event occurrence."
c) We will state that sample unconditional climatology has been chosen as the reference forecast for BSS and CRPSS (as it has been mentioned in the previous comment), although the forecasts from the QR0 Configuration could have been selected to show the performance gain compared to the most basic post-processing technique tested in the experiments:
   "…non-event occurrence. To show the performance gain compared to the most basic post-processing technique tested in the experiments, QR0 Configuration could have been selected as the reference forecast for BSS and CRPSS".
d) We will state that, for the Brier Score and the ROCS (see page 3824, 3833 and 3835), the probability threshold is used to define the binary event to be observed and forecasted.

Comment 3:

Some of the descriptions of the verification metrics should be moved to the verification strategy section to clearly introduce the various metrics used in the verification analysis. In particular, the discussion on page 3829 on the relative importance of reliability, sharpness and event discrimination should be introduced earlier, especially to stress out that improving sharpness at the expense of reliability is not desirable. Also the authors should mention that the event discrimination skill is generally important for decision makers interested in specific flood thresholds whereas reliability is crucial for modelers and forecasters since it could be improved via post-processing, which is the main focus of this paper.

Answer:

We will modify the introduction of the "Verification strategy" section according to the reviewer's comments as follows (moving some text from the results and analysis section):

"… .The old(-ish) adage has it that probabilistic forecasts should strive for sharpness subject to reliability (Gneiting et al., 2005): an improvement in sharpness at the expense of reliability is not desirable. In addition, decision makers may be interested in event discrimination skill for specific flood thresholds, for example. Forecasts were therefore assessed for reliability, sharpness and event discrimination, and a number of metrics were calculated.

The verification metrics include the Brier Score (BS), the mean Continuous Ranked Probability Score (CRPS) and area beneath the Relative Operating Characteristic (ROCA). Reliability was assessed using reliability diagrams….."

In this way, a more comprehensible introduction of reliability, sharpness and event discrimination skills is provided without the necessity of moving Appendix A to the main body of the manuscript.

Comment 4:

Regarding the estimation of the sampling uncertainty of the verification metrics, the authors should clarify which bootstrapping technique they used, provide a short description, mention the number of bootstrap samples, and give a reference (EVS uses the stationary block bootstrapping technique, see the EVS user's manual). The cross-validation approach mentioned on page 3819 should also be mentioned in the verification strategy since it is important to test the robustness of the post-processing techniques as one of the goals was to develop a parsimonious technique that could be easily implemented operationally.

Answer:

We will introduce more details about the estimation of the sampling uncertainty of the verification metrics as follows:

"… .Sampling uncertainties of the verification metrics were explored by bootstrapping. The stationary block bootstrapping technique was applied. This method constructs resample blocks of observations to form a pseudo time series, so that the statistic of interest may be recalculated based on the resampled data set (Politis and Romano, 1994). The minimum sample size was set to 50 and the number of bootstrap samples to use in computing the confidence intervals was set to 1000. The applied resampling method…"

We will include a sentence in the "Verification strategy section" focused on the importance of cross-validation approach to test postprocessing techniques:

"To understand and inter-compare the performance of different QR configurations, an extensive verification of forecast quality was carried out. The post-processing procedure separated calibration from validation hence the verification can be considered to be independent. Forecasts were assessed…"

Comment 5:

The verification results should be commented in more details to provide specific details (location, lead time, probability range) when one of the configurations outperforms the others, especially since not all readers are familiar with the interpretation of the verification plots. The authors usually concluded the analysis by saying that the differences are not significant (see pages 3828, 3829, 3930) but more specific information should be given about the differences/similarities when inter-comparing the confidence intervals of the verification metrics.

Answer:

According to the reviewer's comment, we will introduce an explanation in Section 3.3.2. Skill scores about performance differences between QR configurations. The following sentence will be added to the second paragraph in section 3.3.2:

"Many of the plotted results are very similar in that the distribution of verification metrics is very similar – both in terms of the median as well as the confidence bounds shown – across all leadtimes (columns) and values of P (horizontal axes). As the distributions are approximations – the verification pairs used are not strictly independent – a formal statistical hypothesis testing procedure cannot be used. Hence the interpretation is necessarily largely subjective."

In addition, as per the reviewer's suggestion, where appropriate, specific details (location, lead time, value of P) will be included in the text to further guide the reader in understanding the verification graphs.

Comments 6 and 7:

In the conclusions section, the authors should discuss the following additional points:

- As in all statistical post-processing techniques, this work requires a long and consistent calibration dataset to include enough extreme events; consistency concerns any potential hydrologic/hydraulic changes of the river system (including regulations, diversions), as well as potential changes in the forecasting system (including changes of the forcing inputs or boundary conditions); however information about these changes may not be available to reproduce hindcast dataset consistent with the current/real-time application of the forecasting system; for example real-time modifications of regulated rivers are generally not archived and therefore cannot be reproduced in the hindcasting process; the lack of consistent and long dataset could be a barrier to develop and implement any statistical post-processing, or at least significantly reduce its impact on the quality of post-processed hydrological forecasts.

- On page 3832, the authors should add a comment on increasing the data requirements when estimating additional parameters or when introducing additional data stratification; other more data-demanding techniques may not be a practical choice when only limited sample size is available. The authors should also mention the HEPEX intercomparison project on different post-processing techniques to offer guidance on the best practices (see van Andel et al. 2012).

Answer:

A note about data requirements in post-processing techniques will be added in the relevant paragraph in the introduction, where post-processing techniques are described. The reader shall be reminded of this in the final section where stratification is discussed.

Addition just before the last sentence on page 3813: "Post-processing techniques poses requirements on the data used for calibration. The data record should be sufficiently long as to include extreme events. As post-processing techniques aim to describe the statistical relationship between forecasts and observations, both should be consistent in time as to ensure that this relationship does not change. In the absence of such a consistent data record, there is a serious risk of a significant reduction of post-processed forecasts."

Addition just before "Another option" on p3832, l2: "Both the addition of predictors as well as stratification, however, introduce additional data requirements that may not be met, and in the absence of which the quality of post-processed forecasts may be reduced. In those cases, alternative techniques may be considered; a recent article by Van Andel et al (2012) discusses various techniques in the context of the HEPEX intercomparison experiment."

Specific comments

Comment 8:

- Page 3812, line 26: would add a reference to uncertainties from the meteorological forecasts and human influences (see general comment).

Answer:

We will make the correction according to reviewer's suggestion, included in general comment 1.

Comment 9:

- Page 3813, line 10: would add "in the forecasting routines to capture the atmospheric uncertainty". Note that the HEFS, given as an example, does include a statistical post-processor developed by Seo et al. (2006); this should be clarified as it could be misleading for a reader not familiar with HEFS.

Answer:

We will modify the last part of the paragraph as follows:

"… in the forecasting routines to capture the meteorological uncertainty. An overview of applications and best practices was given by Cloke and Pappenberger (2009). More recent applications include the Environment Agency's National Flood Forecasting System – NFFS – (Schellekens et al., 2011) and the US National Weather Service's Hydrologic Ensemble Forecast Service HEFS (Demargne et al., 2014). Note that HEFS also includes a statistical post-processor (developed by Seo et al., 2006)."

Comment 10:

- Page 3814, line 25: would clarify what is meant by quantile crossing.

Answer:

We will include a sentence with the meaning of quantiles crossing:

“The problem of quantiles crossing (in which developed quantiles cross and yield predictive distributions that are not, as a function of increasing quantiles, monotonously increasing) was addressed by omitting the domain…”

Comment 11:

- Page 3815, lines 6-7: replace forecast skill by forecast quality (see general comment).

Answer:

We will modify forecast skill by forecast quality where appropriate, following referee’s advice.

Comment 12:

- Page 3816, lines 12-13: would clarify whether the WWV2011 configuration corresponds to any of these configurations or how it differs with the 4 tested configurations.

Answer:

The description of WWV2011 QR Configuration is described in Page 3814, lines 19-28 and Page 3815. We will modify lines 13-14 in page 3816 to clarify that WWV2011 QR Configuration is very similar to QR2 Configuration: “Quantile Regression in normal space” in the present manuscript:

“... (iii) QR derived on time series that have been transformed into the Normal domain (similar to wwv2011 QR configuration, with the exception of how the quantile crossing problem is addressed), and (iv) a piecewise linear derivation of QR models. A priori, …”

Comment 13:

- Page 3817, lines 18-19: would rephrase “Flood risk management is supported by the MFFS”.

Answer:

We will modify the sentence according with reviewer’s suggestion.

Comment 14:

- Page 3819, line 16: would use “UK Environment Agency” instead of the acronym.

Answer:

We will modify the sentence according with reviewer’s suggestion.

Comment 15:

- Page 3822, lines 20-24: I would recommend including a reference to Bogner et al. (2012) for their discussion of sample size and extrapolation issues with the NQT technique.

Answer:

We will include the reference suggested by the reviewer:

"Back-transformation is problematic if the quantiles of interest lie outside of the range of the empirical distribution of the untransformed variable in original space. In those cases, assumptions will have to be made on the shape of the tails of the distribution (see Bogner et al., 2012 for a more elaborate discussion)".

Comment 16:

- Page 3823, line 11: would add a comment on having large enough sample size for each subgroup of data.

Answer:

We will include a sentence with the reviewer's comment:

"…Multiple, mutually exclusive and collectively exhaustive domains were identified based on a visual inspection of the data and considering to have large enough sample size for each subgroup of data. As this selection more or less coincided with two splits at the 20^{th} and 80^{th} percentile, thus three sub-domains were defined, comprising 20%, 60% and 20% of the data respectively."

Comment 17:

- Page 3823, line 14: use "verification of forecast quality", not skill (see general comment).

Answer:

We will modify forecast skill by forecast quality where appropriate, following reviewer's advice.

Comment 18:

- Page 3824, lines 18-22: clarify that the probability threshold is also used to define the binary event to be observed and forecasted for the BS and the ROCS (see general comment).

Answer:

We will modify "Verification strategy" section according to reviewer's comments.

Comment 19:

- Page 3829, lines 4-5: the comment on sharpness and reliability should come earlier in the presentation of the verification strategy.

Answer:

We will move lines 4-5 in Page 3829 to "Verification strategy section" in Page 3823 and 3824, following the reviewer's advice.

Comment 20:

- Page 3832, lines 22-23: this should be used to introduce the different forecast attributes in the verification strategy section for readers who are not familiar with forecast quality attributes.

Answer:

We will keep lines 22-23 in Page 3832 and we will produce a similar introduction to the different forecast attributes in "Verification strategy section" in Page 3823 and 3824, according to the reviewer's suggestion.

Comment 21:

- Page 3833, line 14: would add "For a given binary event (such as being above a flood threshold)". The authors should clarify the notations (why switching to the notations X and Y since S and H were used before?).

Answer:

We will add the reviewer's comment and we will modify the notations to be the same in the whole manuscript (X → S and Y to → H).

Comment 22:

- Page 3835, lines 6-9: would add "For a given binary event (such as Q>q), … (PoFD) for several probability thresholds. For each probability threshold, POD…decision rule (a probability threshold above which the discrete event is considered to occur)".

Answer:

We will include the reviewer's comments.

Comment 23:

- Page 3835, line 14: change "adjusting for randomness" since POD=PoFD corresponds to an unskillful climatological forecast.

Answer:

We will modify the sentence as follows:

The ROC score is a skill score that relates the area under the curve (AUC) of the forecast considered to the AUC associated with an unskilled forecast where probability of event occurrence and probability of event non-occurrence are equal, i.e. 50%.

Comment 24:

- Page 3843, Figure 1: change the legend for the catchments and urban areas.

Answer:

We will modify the legend for the catchments and urban areas in Figure 1, as to actually coincide with the colours used in the map.

Additional references to be included

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

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. Journal of the American Statistical Association, 89(428), 1303-1313.

Seo, D. J., Herr, H. D., & Schaake, J. C. (2006). A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. Hydrology and Earth System Sciences Discussions, 3(4), 1987-2035.

van Andel, S.J., A. Weerts, J. Schaake, and K. Bogner, 2012: Post-processing hydrological ensemble predictions intercomparison experiment, Hydrol. Process., 27, 158-161. doi: 10.1002/hyp.9595

Verkade, J.S., Brown, J.D., Davids, F., Reggiani, P., Weerts, A.H., in preparation. Estimating predictive hydrological uncertainty by dressing deterministic and ensemble forecasts; a comparison, with application to Meuse and Rhine. Journal of Hydrology.