Review comment on the paper entitled "Alternative configurations of Quantile Regression for estimating predictive uncertainty in water level forecasts for the Upper Severn River" by Lopez, Verkade, Weerts and Solomatine, submitted to Hydrology and Earth System Sciences in March 2014

This manuscript presents some inter-comparison results of four different configurations of a hydrological post-processor based on Quantile Regression (QR). This is an important research area as hydrologic post-processing is needed to reduce and quantify either the hydrologic uncertainty or the total uncertainty of hydrological forecasts and therefore provide reliable and skillful forecasts for decision making. The post-processor techniques are applied to water level forecasts produced by the UK's Midlands Flood Forecasting System (MFFS). Detailed verification results for 3 test basins and 3 different lead times are discussed and recommendations for selecting appropriate post-processing techniques for operational forecasting are given.

The paper is well written; it includes substantial verification results and appropriate references. I recommend this paper to be published as it documents potential enhancements of the operational hydrological forecast system MFFS used in the UK and the proposed post-processing techniques could be relatively easy to implement in other operational forecasting systems.

I have the following minor comments to help improve the quality of the manuscript.

General comment

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).

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 pairs to analyze 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.

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.

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.

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.

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).

Specific comments

- Page 3812, line 26: would add a reference to uncertainties from the meteorological forecasts and human influences (see general comment).
- 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.
- Page 3814, line 25: would clarify what is meant by quantile crossing.
- Page 3815, lines 6-7: replace forecast skill by forecast quality (see general comment).
- 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.
- Page 3817, lines 18-19: would rephrase “Flood risk management is supported by the MFFS”.
- Page 3819, line 16: would use “UK Environment Agency” instead of the acronym.
- 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.
- Page 3823, line 11: would add a comment on having large enough sample size for each subgroup of data.
- Page 3823, line 14: use “verification of forecast quality”, not skill (see general comment).
- 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).
- Page 3829, lines 4-5: the comment on sharpness and reliability should come earlier in the presentation of the verification strategy.
- 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.
- 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?).
- 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)”.

- Page 3835, line 14: change “adjusting for randomness” since POD=PoFD corresponds to an unskillful climatological forecast.
- Page 3843, Figure 1: change the legend for the catchments and urban areas.

Suggested additional references

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.

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