Response to review comments of Anonymous Referee #1 on the manuscript "Estimation of predictive hydrologic uncertainty using quantile regression and UNEEC methods and their comparison on contrasting catchments" *by Dogulu et al. 2014*

We would like to thankfully acknowledge Referee #1 for her/his thorough review and valuable comments. We believe that addressing these comments will help improving the quality of the original manuscript. We hereby respond to all the comments raised by the Referee #1 one by one.

General Summary

RC: *This paper is generally well written and addresses an operationally important problem such as the application of uncertainty processors in flood forecasting. In this context, the authors present a comparison between two existing methods for uncertainty assessment, Quantile Regression (QR) and UNEEC. Even though the main topic of the paper could be an informative contribution to the hydrological literature, overall the structure of the paper is quite confused, especially regarding the experimental setup and the analysis of the results. In particular, the latter is not enough in-depth and often contradictory. The comparison of the two methods is not carried on rigorously and the evaluation indexes used to compare the results of the two methodologies are often misinterpreted. In my opinion, these gaps preclude the paper to be a novel contribute to the hydrological literature and helpful in the choice between the two methods in operational applications.*

AC: We thank the referee for her/his remarks. We agree that the overall structure of the paper can be improved at several points. In the revised manuscript, we will more clearly distinguish between the sections on experimental setup and analysis of the results, and provide more concise explanations. The aspects regarding the depth and contradictoriness of the analyses are clarified below in response to the main comments provided by the referee.

Main Comments

(1)

RC: *The comparison of the two methods is not enough rigorous. In fact, since I do not see any limitations regarding the number and typology of predictors to be used in both methods, the comparison should have been done using the same predictors for both estimators; otherwise, the effect of a different information level used to force the estimators becomes more significant than the differences in the methods. QR is always used with the only model prediction as a predictor, while UNEEC includes observed values at previous time steps and state variables provided by the hydrological model (such as ground water lever and soil moisture deficit). The authors, in Chapter 5 (page 10208, lines 14-15), confirm this saying that introducing more predictors in the QR methodology could possibly increase the performance of QR, assuming that the conclusions of the comparisons are affected by the choice of different predictands. Moreover, it is not clear to me why the hydrological model prediction has not been used as a predictor in the UNEEC setup on both rivers, the authors did not explain this choice. In my opinion, the authors should have carried out the comparison using the same predictors or at least giving a convincing explanation for the choice of different predictors, otherwise the result of the comparison are obviously biased towards the estimator forced with the better information. The paper in its current form shows a misunderstanding of uncertainty assessment capability of the methodology and informative level of the predictors.*

AC: We appreciate this comment and realize we were not fully clear in the paper. Indeed the referee is right to point out the effect of different information levels used to force the estimators (i.e. the models). Yes, QR uses less input (predictors), and uses the linear model, and this makes it

different from UNEEC. Nevertheless, we believe that we have the full right to compare various methods reported in literature (similarly to the studies comparing hydrological models that use different inputs). Likewise, we compare two uncertainty prediction methods, with the aim of investigating how well a simpler method using less input data performs over a more complex method with more predictors (which can be less acceptable by practitioners in flood forecasting). Overall, selection of the most appropriate uncertainty processor for a specific catchment is a matter of compromise between its complexity & accuracy in consideration of the data availability and also the characteristics of the catchment. Therefore, we believe the findings of such a comparative analysis could be useful for the operational hydrology community.

More, we believe that it is quite risky to state that "*the result of comparison are obviously biased towards the estimator forced with the better information*". In theory this is right, but more predictors may not bring more information needed for accurate prediction. Only experiments can allow for stating that for each particular case. Our experience with data-driven models (and both QR and UNEEC are such) have shown that adding more and more predictors does not necessarily mean higher accuracy on unseen data. Parsimony (Box, Jenkins, and Reinsel, 2008) often leads to better generalization.

Concerning the suggestion of using the model output as yet another input to UNEEC (along with the observed Q), we are taking it on board to test in the further studies. It has to be seen if indeed inclusion of this variable would improve the model performance. However in this study we, unfortunately, cannot test this idea since it will mean to rerun all experiments for which we do not have resources.

Overall, we appreciate the referee's comment and to address it, and to make the text clearer, we will update the manuscript accordingly.

(2)

RC: *In Section 2.2 (page 10191, lines 24-27), the authors claim that "none of the presented measures allow for accurate comparison between different methods of uncertainty prediction and should be therefore seen only as indirect indicators of method's performance". In my opinion, this statement is incorrect. In fact, the PCIP is often enough to evaluate the correctness and performances of an uncertainty estimator because it verifies whether the estimated uncertainty distribution is correct (i.e., includes the right amount of observed data) or not. Once this is proved, the other indexes (MPI and ARIL) can be used to understand how wide the uncertainty is and if it may be reduced using different predictors. The authors also point out correctly that ARIL may be affected by misleading values when the streamflow is 0 or very small. In order to evaluate how much ARIL is affected by these values the reader should have a better idea of the streamflow distribution of the case studies, but the authors only provide the mean flow making the interpretation of ARIL very difficult. Moreover, when the streamflow is close to 0 the uncertainty is usually pretty small (compared to the average value of the uncertainty band width), so it would have been helpful to screen out these values, which do not have a significant impact in the analysis, when computing the index. The wrong interpretation of the indexes led to some arguable conclusions:*

AC: The authors agree with this the referee's comment and realise that manuscript should be much clearer on this point; the necessary modifications have been made to the manuscript.

At the same time, we would like to explain why we would like to downplay a bit the role of PICP as a universal indicator of the model U performance.

Residual uncertainty prediction (RUP) methods build a data-driven predictive model U based on the model M residual errors (RE) allowing for predicting the *pdf* of RE (or some of its quantiles). Real distribution ("observed data") of the RE is unknown, so we cannot know if U represents this *pdf* accurately, i.e. U cannot be accurately validated against the real *pdf* (observed data).

In the cases considered in this paper, U predicts only two quantiles of this *pdf* - e_5 and e_{95}.

In operation, outputs of model M ($\hat{y}$) and U (*pdf* or quantiles) can be combined: the error *pdf* is shifted to have $\hat{y}$ to coincide with the median (becoming thus *pdf* of the uncertain model output, corrected by the estimates of the error), or in case of two quantiles,

$q_5 = \hat{y} + e_5$
$q_{95} = \hat{y} + e_{95}$

If U predicts only two quantiles, there is a measure of the average quality of model U across the whole data set – PICP. It was used in this paper and in our earlier publications. However it is an average measure – it cannot be calculated for every time step (so it is "weaker" than e.g. RMSE which is an average of individual errors). PICP allows to check how much of data is inside the (90%) prediction interval (to be close to 90% is good). Actually for any q% quantile qi such test can be made (to check what share of data is below this quantile qi; q% is good).

Now, why do we think that using PICP has to be done with care, and that it is not an ultimate measure for judging about the performance (quality) of the model U?

Let's consider case 1 when model M is accurate and not biased, so that error is low and has close-to-zero mean. In this case the predicted PI will be "around" the observed data and PICP will be close to 90%.

Let's consider case 2 when model M is really bad, has high random errors (noise), its variance is also high, and its *pdf* has non-zero mean (bias). Such data may present a difficult challenge for any machine learning method (model U), and its accuracy for this reason may be low, and hence PICP far from 90%. In this case it is difficult to say what is the reason – data with a lot of noise (random components), or the method used. It has to be taken into account that if data is noisy then most machine learning methods may not discover the input-output relationship.

So, the accuracy of the model U may depend on the quality of model M. That is why we think PICP does not necessarily reflect the quality (performance) of the certain machine learning method used by U. (That is, PICP far from 90% could mean simply that model M errors are close to random, so in principle it is not possible to train any model U to predict them.) For comparative studies however, when various types of U are compared, PICP can be used: the method with PICP closest to 90% should be seen as the best (with some tolerance).

However, again, we recognise that we have downplayed the value of PICP too much, and it will be now corrected.

MPI can be seen as a complementary measure to judge about the quality of U, since it only indicates the width of *pdf* (i.e. an indicator of the model M error variance), and not of the model U ability to represent the *pdf* of this error.

The authors also consider the referee's suggestion about the update of ARIL ("screen out these values, which do not have a significant impact in the analysis") as a potentially workable idea, which can be taken up in the further studies.

a) Page 10202, lines 14-15. *The authors say that "QR produces unnecessarily wider uncertainty bounds for medium peaks in validation". The fact that the uncertainty band is unnecessarily wider should be proved showing the PCIP for the cluster including medium events. The authors do not show these indexes for the validation period so the reader can only rely on Table 3, which shows the indexes for the training period. According to this table, QR has a PCIP often lower than 90% or very close and only for cluster #3 it is significantly higher, but also UNEEC for that cluster gives a high value of PCIP. From Table 2, QR shows lower PCIP values during validation that those computed during training, so I suppose (maybe wrongly) that the same happen for most of the clusters. This would lead to think that the PCIP in validation is always lower than 90% for all the clusters and this would be in contrast with what the authors claim about the unnecessarily wide bounds.*

AC: The authors agree that this is a conclusion which is too firm and has to be formulated better. This interpretation is done based on the visual analysis of 90 % prediction intervals, and can be misunderstood. The authors intend to say that QR provides wider uncertainty intervals for

medium peaks, higher values of MPI, compared with UNEEC, without a significant improvement of the PICP values. The relevant explanations will be clarified in the revised manuscript.

b) Page 10202, lines 20-22. *The authors write that from Table 3 it is possible to verify a contradictory relation between PCIP and MPI, since the latter is higher when the former is closer to 90%. In Table 3 I do not see any contradictory relation, because for every cluster MPI is higher when PCIP is higher and this just implies that a wider uncertainty band includes a higher number of observed values.*

AC: The authors will modify the mentioned statement according to the results. Higher MPI values are related with higher PICP values for both analysis carried out with the whole data set and for the different clusters. This can be shown in Tables 2, 3 and 4.

c) Page 10203, lines 3-4. *The authors say that for the cluster of high flow the NUE index must be considered to correctly compare QR and UNEEC and they conclude that UNEEC performs better because it yelds a higher NUE value. This statement is misleading because it seems that the authors compute NUE for better analyzing the cluster #4, but then they point out general conclusion for every cluster based on that index. Moreover, the author themselves claimed in page 10191, lines 21-23 that a higher NUE does not imply a better performance and I completely agree with this statement. However, they are now contradicting their words using a higher NUE as simple evidence of the UNEEC better performance, without considering, for example, the fact that for clusters 4 and 5 the PCIP given by UNEEC is very far from 90% if compared to that given by QR.*

AC: The authors thank the referee for pointing this out. Indeed the PICP is the primary indicator and values for QR are better. NUE (Nasseri and Zahraie, 2011) is an update of PICP (ARIL is added in denominator) and indeed has to be seen as an additional indicator. It shows however that UNEEC is better than QR. In the revised manuscript this point will be properly addressed.

d) Page 10203, lines 9-11. *The low values of MPI in Yeaton catchment do not surprise me mainly because the mean flow (as reported in Table 1) is much lower than that of the other catchments and, in smaller part, also for the fact that the hydrological model is more accurate. I would rather use ARIL for this analysis, because it accounts for the flow magnitude. Actually, if one considers ARIL the situation is different, Yeaton has the worst value for 24-hr lag time and it has a value higher than that of Llanerfyl for all the others lag times.*

AC: On page 10203, the authors include a discussion about the uncertainty methods performance based on MPI and ARIL. Both measures are taken into account. According with the referee comment this will be formulated clearer.

e) Page 10203, lines 25-26. *The author claim that QR does slightly better than UNEEC in Yeaton. I would rather say that UNEEC performs very poorly on this catchment considering the extremely low values of PCIP when the 50% uncertainty band is considered.*

AC: The authors will consider the comments of the referee and they will modify the sentence to: "*UNEEC performs poorly on this catchment considering the extremely low values of PCIP when the 50% uncertainty band is analyzed.*"

f) Page 10204, lines 6-7. *I agree with the authors regarding the 90% uncertainty band, but I disagree for the 50% band since UNEEC gives very low PCIP for the lower time lags.*

AC: The authors will modify lines 6 and 7 in page 10204 about the performance of QR and UNEEC at Llanyblodwel station. Both methods are equally capable of providing uncertainty estimates reasonably well, regarding the 90% uncertainty band (as measured by both MPI and PICP). In terms of the 50% uncertainty band, QR performs slightly better considering that UNEEC provides very low PICP for the lower lead times.

g) Page 10204, lines 8-9. *It does not seem so clear to me, especially for the 50% band.*

AC: The authors will modify the manuscript according to the reviewer's comment. Differences in QR and UNEEC methods performances are very small. QR and UNEEC methods perform similarly at Llanerfyl station. (We agree with the reviewer that UNEEC is not outperforming QR, values of MPI and PICP are pretty similar).

h) Page 10205, lines 13-19. *From Figure 14 it is almost impossible to see that UNEEC prediction intervals are wider. However, the explanation of the reason why UNEEC provides wider intervals is not clear to me.*

AC: The authors will include the "zoomed image" at some time periods for medium water levels for Yeaton, Llanyblodwel and Welshbridge catchments in Figure 14. That would allow the reader to follow the analysis and discussion of this figure more clearly. We will work on the explanation as well.

i) Page 10206, lines 24-26. *This sentence is not clear at all. Slightly better values of PCIP compared to cluster 2 or to QR? Why can they be attributed to lower MPI?*

AC: Indeed a better formulation will be provided.

j) Page 10208, lines 1-5. *I do not agree with the authors when they claim that there is no basis for comparison of different uncertainty estimators. In my opinion, PCIP is the basis, if it is far from the expected value the estimator is not reliable and useless in real application. In case both estimators gives good PCIP values, then the PCIP for different conditions must be analyzed (e.g. for different clusters) to check if the correctness of the estimator is preserved for different situations. If still the methods give similar results, then the MPI and ARIL can be used to identify the better methodology.*

AC: The authors agree with the reviewer comment and according to it, the statement in lines 1-5 at page 10208 will be removed. (Please also see explanation to the very first comment about PICP.) In this study, the approach explained by the reviewer has been actually followed: as an initial validation measure of the uncertainty methods performances, PICP has been used. It has been computed and analysed for the complete range of water levels and for different clusters. As later steps in the validation, MPI and ARIL has been considered. This approach allows providing information about the reliability and sharpness of the forecast for its real application.

(3)

RC: *The evaluation of the methods performance should mainly focus on the PCIP and Q-Q plots (Laio and Tamea, 2007) of both calibration and validation data. The authors often show only analysis of training data (Tables 3 and 4; Figures 5, 6, 7, 8, 9 and 10) and sometimes only of validation data (Figure 12 and 13). Only in Table 2 training and validation periods are showed together. A comparison of both periods is necessary to evaluate the ability of an estimator to evaluate the uncertainty of new/unknown data, which is fundamental in real time applications. Moreover, in Figures 7b and 10b only the cluster with the distribution closest to normal is showed; I would rather show the cluster with the distribution furthest to normal (or at least both) to better understand the origin of the error in the uncertainty assessment.*

AC: *About the Q-Q plots:* We are not sure how we can use Q-Q plots to evaluate the performance of the methods - because the actual values of the quantiles (for a given time step) are unknown.

*About training-validation issue***:** It is true that evaluation of both the training and the validation periods is important for understanding the ability of any model. We are constrained by the size but in the revised manuscript we will address this issue.

(4)

RC: *Section 3.2 is very confused. It is not always clear if the authors are referring to the case study of Brue or to the Upper Severn catchments. The description of the experimental setup is mixed up with few analyses of the hydrological model performances, which are not necessary for the purpose of the section (Pages 10198-10199, lines 25- 3). The description of the choice of the predictors for the Upper Severn catchments is not very clear and linear as it should be. Some sentences (e.g. "low soil moisture is more likely attributed to higher rainfall rates") show a lack of effort in making this section easily understandable.*

AC: Agreed. We will modify Section 3.2 in accordance with the referee's comments.

About Pages 10198-10199, lines 23-3: We still feel the explanation is needed. We have provided these four sentences because we felt the necessity of giving the relevant explanations for Fig. 5, which is presented and discussed only in here (Section 3.2). We will work on better formulations.

About the description of the choice of predictors for the Upper Severn catchments: Please see the response to the Main Comment (1).

About the sentence "low soil moisture is more likely attributed to higher rainfall rates": We agree English could have been better. It will be revised, possibly like this: "*Positive correlation between GW and model residuals can be explained by the fact that high groundwater levels are associated with excessive precipitation during which model error is higher in magnitude. High soil moisture deficit, on the other hand, indicates that there has been no excessive precipitation and the soil is not filled up with infiltrated water. High evaporation rates (causing soil to dry up) can also result in high soil moisture deficit. It should be noted that the latter is less likely to happen for the Upper Severn catchments considering the prevailing climate in the region. Accordingly, lower soil moisture deficit is linked with excessive precipitation events such that soil moisture deficit is negatively correlated with the model error*."

Minor Comments

(1)

RC: *Page 10187, lines 17-20. The authors do not explain why they used the variant called QR1.*

AC: We agree with the referee that a better justification should have been provided.

The idea was to compare two uncertainty methods with different approaches. Two main reasons for this:

i. According to results in Lopez Lopez et al. 2014: "sharpness and reliability may vary across configurations, but there are none results in a more favorable combination of the two. Intercomparison showed that reliability and sharpness vary across configurations, but in none of the configurations do these two forecast quality aspects improve simultaneously. Further analysis shows that skills in terms of BSS, CRPSS and ROCS are very similar across the four configurations."

ii. Taking into account the similarity between the performances of the four QR configurations, the "simplest one" is considered. The idea was to compare two methods which are very different: one that considers more information, several predictors and it is based on a cluster approach (UNEEC), and the second one, which consider only one predictor, water level and do not make any kind of clustering.

Considering these two aspects, the choice had to be made between QR0 and QR1. Due to the fact that QR1 has the same configuration than the classical one with the incorporation of the solution algorithm for the crossing problem, QR1 is selected for the comparison with UNEEC.

2)

RC: *Page 10196, lines 6-7. The sentence is not clear to me, maybe "are obtained" should be replaced with "is obtained".*

AC: The suggestion of the referee is indeed very correct. There is only one single regression model for estimating quantile τ of observed discharge. We will replace "are" with "is" on line 7.

3)

RC: *Page 10197, lines 18-19. The sentence "low soil moisture is more likely attributed to higher rainfall rates" does not make much sense.*

AC: Please see above the response to Main Comment (4).

4)

RC: *Page 10198, line 4. "et" should be replaced by "e" or "et-i"*

AC: The referee is right. This will be changed in the revised manuscript.

References

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C.: Time series analysis: forecasting and control (4th ed.), p.16, John Wiley & Sons, Inc., Hoboken, New Jersey, 2008.