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 has helped 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 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 #1.

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 and 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.

Moreover, 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.

In accordance with the above explanations, we added the following text in page 10185 (before the last paragraph):

> *"It should be noted that by design UNEEC uses a richer set of predictors than QR and a more sophisticated non-linear regression model, so the comparison may seem unfair. However, 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) showed that adding more predictors does not necessarily mean higher accuracy on unseen data. Parsimony (Box, Jenkins, and Reinsel, 2008) often leads to better generalization. Overall, selection of the most appropriate uncertainty processor for a specific catchment is a matter of compromise between its complexity and accuracy in consideration of the data availability and also the characteristics of the catchment, and we believe the findings of such a comparative analysis could be useful for the operational hydrology community."*

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.

To reply to this we have added the following sentence on page 10191 (after describing ARIL, Line 9):

> *"This problem could be helped by removing all observations above a certain threshold from the calculations (a suggestion from one of the reviewers of this paper); we leave this idea for further testing for the future research."*

Overall, we appreciate the referee's comment and to address it, have updated 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 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 is 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.

In the end of Section 2.2 (pages 10191-10192) we updated the last paragraph and now it reads as follows:

> *"There is no single objective measure of the quality of an uncertainty prediction method (since the "actual" uncertainty of the model (error pdf) at each time step is not known). Closer PICP is to the confidence level higher the trust in a particular uncertainty prediction method should be. In principle, a reliable method should lead to reasonably low values of MPI (and ARIL). However, a wide MPI does not mean that a method estimating prediction interval is inaccurate – it could simply mean that the main model is not very accurate and the high MPI shows that.*
>
> *PICP indeed evaluates if the expected percentage of observations fall into the predicted interval, and should be seen as an important average indicator of the predictor's performance. However, in case of high noise in the model error (aleatoric uncertainty) the fact that PICP is far from 90% could mean simply that none of the data-driven predictive models can capture the input-output dependencies and to predict quantiles accurately. For comparative studies however, PICP can very well be used: the method with PICP closest to 90% should be seen as the best (with some tolerance).*
>
> *It is also worth mentioning that all considered measures are averages so should be used together with the uncertainty bound plots which visual analysis reveals more information on the capacity of different uncertainty prediction methods during particular periods."*

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 with the part of this comment and realize a possibility for a better formulation and the necessity to provide quantitative data for validation periods rather than just referring to visual inspection of Fig. 11. In the revised manuscript, Table 2 is extended such that the computed PICP, MPI and ARIL values are also given for the two periods (highest peak and medium peak events shown on Fig. 11a and Fig. 11b). The relevant discussion on the results has been integrated to the existing discussion on Fig. 11.

We also considered to provide such estimates for the clusters in validation set (similar to how it is done for the clusters found by UNEEC during training). However, as explained below this comment, in our view, this does not have much sense since in operation these methods are applied for the individual data points at each time step of the model run, and not for a large set of data (so the "validation set" in reality does not exist).

So we think calculation of PICP, MPI and ARIL for the two periods (in Fig. 11) provide now enough information about the methods' performance for different periods, along with the overall estimates for the whole period.

In the Response to item (b) below we are providing the text from the new version of the 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 agree with the referee that there is no "contradictory relation" between PICP and MPI in Table 3. Here the term "contradictory" is not used correctly. We have modified this section, and now it (second part of 4.2.1) reads as follows:

> *"We have also compared performance of QR and UNEEC for each cluster found by UNEEC during training. Unlike for the whole data set (which is highly heterogeneous due to extremes in rainfall-runoff process), analysis for each individual cluster focuses on the more homogeneous data sets. Table 3 shows the corresponding PICP, MPI and ARIL. In general, it is difficult to decide which method is better – results are mixed. However there is one observation that can be made. For most clusters there is a dependency between PICP and MPI: typically the higher MPI corresponds to PICP being closer to the confidence level (90%). This may be explained by the fact that for narrow MPIs PICP would be under "pressure" and be lower (however it would be difficult to generalize). For example, for the high flow cluster (cluster 4) QR appears to be better in terms of PICP, whereas UNEEC ends up with very narrow MPI and this is probably the reason why its PICP could not reach 90% confidence level.*
>
> *The reported comparison was done for the clusters found by UNEEC during training, so for the validation set it was not done since clustering is not carried out during the validation phase. In principle a similar comparison can be also made for the homogeneous groups of data in the validation set, however this may not have much sense since in operation these*

methods are applied for individual data points at each time step of the model run, and not for a large set of data (so the "validation set" in reality does not exist)."

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) should be only seen as an additional indicator – which however does not necessarily characterize the performance fully. Here we completely agree with the reviewer that NUE cannot say much. In the revised manuscript we have removed calculation of NUE from Table 3.

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: Indeed, the mean flow (water level) and accuracy of the hydrological model influences the results in terms of PICP etc. We have reformulated the results and discussion, pointing this out and making them much more concise (since most things are clear from the figures directly) – the new text for Sec. 4.2.2 (page 10203) reads as follows:

"For these catchments, in order to reflect performance for different lead times better, we are using the graphical representation of results. [Note that the order of Fig 12 and 13 is changed, the numbering has been corrected in the revised manuscript accordingly.]

Fig. 13 shows the PICP values plotted against the MPI for the validation period. The most important general conclusion is that both methods show excellent results in terms of PICP for 90% confidence level. For the 50% CL the results seem to be worse, especially for UNEEC – but the reader should take into account that for the low lead times the hydrological models are very accurate, hence MPI is extremely narrow (especially for 50% CL) and it is no surprise PICP cannot be accurately calculated. Further, for the 90% CL, the following can be said: for Yeaton QR does slightly better than UNEEC; for Llanyblodwel both methods are equally good; for Llanerfyl: UNEEC method is a bit better than QR.

For the further analysis, Fig. 12 presents MPI and ARIL values for the 90% confidence level on validation data set. It can be seen that with the increase of the lead time, the forecast error obviously increases, and the values of both indicators follow. In view of the (high) model accuracy, the relatively low MPI values in Yeaton catchment are not surprising for both methods. Overall, the results are mixed: for some catchments QR is marginally better, for others – UNEEC."

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: Indeed, this is true. The new text makes this clear – please see above the answer to item (d).

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: Indeed, this is true. The new text makes this clear – please see above the answer to item (d).

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

AC: Indeed, this is true. The new text makes this clear – please see above the answer to item (d).

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: In the revised manuscript Figure 14 now includes also the "zoomed image" at some time periods for medium water levels for Yeaton, Llanyblodwel and Llanerfyl catchments. We hope that this would allow the reader to follow the discussion on this figure more clearly. We have also provided calculation of MPI for the medium water levels.

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 we agree. This section has been rewritten and now reads as follows:

> *"Table 4 shows the values of validation measures (PICP, MPI, and ARIL) for each cluster (obtained during training, for 90% CL) for Llanyblodwel catchment (lead time = 6 hrs). For flood management the cluster 2 (4.6% of all data) – with the high groundwater levels, and hence potentially corresponding to flood conditions – could be the most interesting one. In UNEEC, the largest MPI value was obtained for this cluster with a relatively bad PICP value compared to other clusters. Similar to UNEEC, the largest MPI was obtained for this cluster with QR method also. Both methods provide equally bad PICP values. Giving a wider uncertainty band than UNEEC on average, QR is less capable of estimating reasonable prediction limits for very high groundwater levels. This is also supported by its greater (12%) ARIL value compared to UNEEC.*
>
> *PICP and MPI values for the cluster 4 should be mentioned as well. This cluster represents the situations with the very low water levels, very low groundwater levels, and very high soil moisture deficit, and constitutes 16.6% of the whole data. In comparison to UNEEC, QR provides a PICP value very close to 90% CL despite its slightly lower MPI. Thus, one can say that UNEEC fails in providing reliable uncertainty estimates for the extreme condition associated to very low water and groundwater levels. This can be due to the effect of using state variables as predictors. All in all, state variables are model outputs and they cannot reflect real catchment conditions truly especially when the (hydrological) model is not very accurate. That is particularly true for the extreme events considering that models mostly fail in simulating such events.*
>
> *Overall, UNEEC is worse than QR on for one cluster but better or equal on all others, however, in general, both methods in terms of PICP show reasonably good results."*

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 the 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. In the revised manuscript:

- Fig. 6, 8, 9: the results are presented for validation too.
- Fig. 12, 13: the results are presented for training as well.
- Fig. 7, 10: the probability plot for the cluster with the distribution furthest to being normal is included. Accordingly, we have also extended our discussion.
- Table 3 and 4: Please see the relevant text from the new version of the manuscript provided in response to the Main Comment #2(b):

  "The reported comparison was done for the clusters found by UNEEC during training, so for the validation set it was not done since clustering is not carried out during the validation phase. In principle a similar comparison can be also made for the homogeneous groups of data in the validation set, however this may not have much sense since in operation these methods are applied for individual data points at each time step of the model run, and not for a large set of data (so the "validation set" in reality does not exist)."

- Figures 5, 7, and 10: We present these figures concerning the clusters only for the training period. This is mainly due to the methodological basis that defines the framework of the UNEEC method. In UNEEC, the clustering approach is employed only during the training period. Please also see the explanation given for the Table 3 and 4.

(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. Section 3.2 has been modified accordingly.

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). However, the authors are willing to remove the explanation if it has to be removed.

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

About the sentence "low soil moisture is more likely attributed to higher rainfall rates": We agree that the use of English language could have been better. It was revised and now reads as follows:

> "*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 favourable 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 considers only one predictor, water level and does 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. The manuscript has been revised in light of the explanations provided here.

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. In the revised manuscript "are" is replaced with "is".

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 is 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.