Answers to Referee #1 comments

The authors would like to thank Referee #1 for his efforts and remarks that will allow the authors to noticeably improve the original version of the manuscript.

1) **C:** *The authors are correct to point out in the introduction (Pg 9221 Line 27) that is the exceedance of the actual not predicted water level that results in damage. However in all the techniques used they appear to substitute the observed water level (or worse discharge) for the actual value. Much as the model value isn't the actual value neither is the observed value. This need to be noted, comparison could be drawn to other Bayesian analysis which (at least theoretically if not in practise) allow the actual value to be predicted (;). Prediction of the observed value is still a useful exercise but may require a different interpretation in a risk context.*

Both Referee #1 and Referee #2 underline that, although the authors point out that the actually occurring water level is the one that causes damages, they use either the observed water level or the estimated discharge as the "predictand".

It is definitely true that observations do not coincide with reality, a measurement error must always be taken into account, nevertheless the authors would like to discuss in detail the choice of the predictand.

1) Water level measurements are affected by relatively small errors (with standard error of the order of 2-3 cm), and it is psychologically fundamental to use them as measures unaffected by measurement errors both because flood decisions have always been essentially based on these measures and because their errors have very small effect on the decisions, given the larger effects of the other sources of uncertainty.

2) Discharge measurements are generally unavailable in real time, although there is a recent tendency to using microwave surface velocity measurements, which could also improve discharge estimates in real time.

3) Classical discharge estimates based on water level measurements and the use of steady state rating curves are affected by errors that may reach 30% (Di Baldassarre and Montanari, 2009) due i) to extrapolation beyond the range of observations used for calibrating the rating curve as well as ii) to the effect of the potential formation of loop, the effect of which is sometimes reduced by correcting formulas, such as for instance the Jones formula or others (Dottori et al., 2009).

4) Hydrological model forecasts are essentially based on predicted discharges, while hydraulic model forecasts can be both in terms of water levels and/ or discharges. In any case they are affected by a wide range of errors.

Therefore, when dealing with real time flood forecasting, the "filtered" water level is in theory the most appropriate quantity to be used as the "predictand". Please note that we are here talking of a "filtered "quantity because the real occurrence will never be known, but one can reduce the measurement errors by using the classical Kalman Filtering technique (Kalman, 1960; Kalman and Bucy, 1961), which combines measurements and models.

Nonetheless, it is the belief of the authors that, in practice, the “observed water levels” are the “best” operational quantity to be retained as “predictand”: the errors are small and the belief of the “decision makers” is very high, while this is not so for the “filtered” quantities than are “estimated” and not “measured”.
Therefore, whenever possible, one should use the “observed water levels” as the “predictand to be used in any flood predictive uncertainty processor.

In the eventuality that water levels are not available or when one needs to predict inflows to a reservoir or a water detention area, “corrected” and “filtered” discharges should always be used. In other words prior to their use as predictands for the calibration of the hydrological uncertainty processors, improved discharge estimates must be produced both by accounting for the shape of the cross section and by taking into account the loop formation. This will eliminate most of the water level dependent “biases” , while the elimination of the random errors must again be approached by filtering techniques.

In terms of “predictors”, when available from a hydraulic model, the best choice would be the forecasted water levels. Otherwise one can either convert the predicted discharges into predicted water levels using a “corrected” rating curve, as mentioned above, or just use the predicted discharges when this is not available, since the effect of the conversion errors from discharge to levels, may not significantly affect the “order” of the predicted variables, which is what essentially dominates the NQT conversion into the Normal space.

The authors take the point and 1) will introduce this discussion in the paper, 2) will introduce another example, this time an operational one based on water levels, 3) will justify the use of discharges in the case of the Baron Fork river example, due to the fact the no water level observations or rating curve observations were made available to the participants to the DMIP2 Project.

2) **C:** *The probabilistic threshold paradigm is a natural consequence of producing probabilistic forecasts. Figures 1 & 2 are informative of the challenges in interpreting probabilistic forecasts when issuing warnings and these are worth keeping to help illustrate the latter results. However, since no attempt is made to define a method for giving warnings based on probabilistic forecasts, I believe this section could be substantially shortened. Section 1.2 is also poorly referenced, even mentioning the two texts on probabilistic decision making referenced on Page 9221 would be a start!*

A: Section 1.2 will be shortened and references will be added into the manuscript, also reflecting the comments of Referee #2 and the *other remarks* of Referee #1.

3) **C:** *Section 2 reviews a number of Bayesian statistical methodologies for expressing prediction uncertainty. All of the methods reviewed appear to work on the principle of building a probabilistic model of residuals of a given model. In this context the work of Kennedy and O’Hagan () may be of interest. Also this is just one way of assessing predictive uncertainty, alternative methods e.g. (; ;) which are more often used in calibration could be applied to generate Bayesian predictive uncertainty bounds. This should be mentioned and comparisons drawn.*

A: We certainly agree with Referee #1 that nowadays several bayesian uncertainty processors are available. The authors decided to mention just HUP and BMA because of their deeper acquaintance with these methodologies since they applied them to some study cases. However, the suggested methodologies will be cited in this paper omitting to include a deeper analysis and they will be taken in account in future researches.

4) **C**: *Unlike the outlined formulation of the HUP (Section 2.1) the model conditional processor (Section 2.3) makes no use of auto-correlation in the residuals. It would appear that at times these are significant (Fig 3 & 4) and could be of use in forecasting (depending upon the lead time). The authors should explain why this is not included. Note that the formulation used in the multimodel case (Section 2.4) would appear to allow for inclusion of lagged residuals as alternative 'models' and the later work (Section 2.4.3) for altering properties of the temporal pattern of residual dependency with output.*

A: As correctly outlined by Referee #1, the multi-model case allows for the inclusion of several models, independently from their nature and structure; therefore, also an Auto-Regressive (AR) model could be easily included, as in fact it was used in Todini (2008) where it was shown that MCP produced a conditional expected value which had a smaller variance than the one produced by HUP. If needed an AR1 (or even more complex ARMA) model can be used as one of the alternative models. The point is that MCP benefits from taking into account the cross correlation among the alternative model forecasts (which is not irrelevant since they all aim at being highly correlated to the predictand), whereas in HUP the model forecast errors are assumed to be independent from the AR1 errors, which is not always guaranteed.

5) **C:** *Much of Section 2.4 outlines the new work in this paper, the use of truncated normal distributions. As such putting it a section titled 'existing approaches' is confusing. I suggest that work should be in a separate section.*

A: The authors completely agree with Referee #1 about the erroneous position of the section describing the use of the Truncated Distributions, therefore it will be moved, also reflecting the comments of Referee #2.

6) **C:** *At the start of Section 2.4 the authors highlight some of the difficulties encountered due to transforming the data before describing the joint distribution. In the processors outlined earlier in the work the transformation had clear advantages, i.e. the marginal distributions were Gaussian and a multivariate Gaussian distribution was used. What is the purpose of maintaining the transformation with the truncated normal joint distribution?*

A: The advantages of using the Normal Quantile Transform (NQT), correctly pointed out by Referee #1, are still valid when dealing with the Truncated Normal Distributions (TNDs). In fact, the assumption of TNDs as joint distribution is more realistic when we are working in a Normal Space. Moreover, the use of NQT also makes more acceptable the hypothesis that marginal distributions are obtained by the two truncated Normal Distributions resulting from the division of the Normal Space.

7) **C:** *Section 2.4.3 should be written to improve the clarity of the exposition. In particular*

- *It is often unclear which of the two truncated normal distributions was being discussed;*
- *The estimation of the parameters of the truncated normal distributions is reported in the appendix, yet this is entirely unmentioned in the text;*
- *Some notation is reused for different variables resulting in repetition e.g. Eq 25 and 26;*
- *Eq 21 and the purpose of $\hat{\eta}^*$ is opaque. If $\hat{\eta}^*$ is a realisation of the random variable $\hat{\eta}$ do you mean $f(\eta|\hat{\eta} = \hat{\eta}^*, \hat{\eta}^* > a)$?*
- *No mention is made of how to select a, or how to select the number of truncated normal distributions. These selections are crucial to applying the technique;*

- *The NQT transform used results in η and $\hat{\eta}$ having Gaussian marginal distributions. Does the truncated multivariate normal representation maintain this property? If not what does this imply about the validity of using this transform? For example what is the predictive if there is no evaluation of think of forecasting without an evaluation $\hat{\eta}$? Also if one is not interested in maintaining this marginal property why not consider alternative representations (e.g. mixtures of normal distributions) and different transforms?*

A: • All the points mentioned by Referee #1 It will be better clarified in the manuscript which of the two truncated normal distributions is being discussed.

• Appendix A is mentioned after Equation 24 and after Equation 34.

• Equations 25 and 26 have the same notation because they have the same meaning even if they refer to the upper Truncated Distribution in the case of eq. 25 and to the lower one in the case of eq. 26. The authors think that additional notations to identify the upper or lower TND can generate further unclearness since the same Equation is actually taken into account, whilst only the data used to compute the distribution moments change. This repetition can be better explained and made more understandable making clearer which TND the Equations refers to, as mentioned above and correctly noted by Referee #1.

• The authors agree with Referee #1 to substitute $f(\eta|\hat{\eta} > a, \hat{\eta}^*)$ with the suggested notation$f(\eta|\hat{\eta} = \hat{\eta}^*, \hat{\eta}^* > a)$.

• As described in chapter 2.4.4, the TNDs are identified on the basis of the predictive variance of the upper sample and the threshold *'a'* represents a sort of limit between low and high flows. The same is made when only one model is being considered; in fact, in this work, the threshold is identified as the value that minimizes the predictive variance of the upper sample. Moreover, the threshold search is lower and upper limited in order to count with significant samples for computing the moments of the truncated distribution. Referee #1 correctly noted that in the paper is not described how to select the threshold and it will be added. Anyway, there are relatively ample margins for choosing the truncation point. One must realize that the splitting point is not that critical. What is important is the presence of a double model to cope with the non stationarity of the data.

The number of TNDs to be used is at user discretion, depending on the data structure of the specific case. In particular, in order to avoid the processor to excessively adapt to the calibration data with the risk of losing generalization ability, the authors think that is advisable to use as less TNDs as possible. In all the cases studied by the authors, it was decided that two TNDs were enough to well represent the Normal Space, also avoiding unnecessary waste of computational time.

• When dealing with TNDs, the Normal Space is divided in two parts over $\hat{\eta}$ and this assumption does not have any consequence on the marginal distribution of η, which is still a Normal Standard Distribution. On the other hand, the marginal distribution of $\hat{\eta}$ is assumed to be composed by two Truncated Normal Uni-variate Distributions which moments depend on the threshold according to which the Normal Space is split. The assumption of normality (or truncated normality) for marginal and joint distributions is more acceptable if it is made in a Normal Space, for this reason the use of the NQT is still necessary.

8) **C:** *Section 2.4.4 should be altered to reflect the changes in Section 2.4.3.*

A: The same changes described in the answer to comment 7 will be done also in Section 2.4.4, where the use of the TNDs with more than one forecasting model is described.

9) **C:** *The assumption around line 10 on pg 9238 in section 2.4.4 implies a single split and two truncated multivariate normal distributions. But if each model had a single a values there could be* 2^M *truncated normal distributions. The computational advantages of the assumption are clear, but the loss of accuracy in the representation of the predictive uncertainty compared to considering more complex splits need exploring.*

A: The authors think that the truncation over just one model is not a strong approximation because it may be considered one of the possible Joint Distributions that can be used to represent the Normal Space. Moreover, as mentioned in the answer to the comment #7, the threshold *a* represents a sort of criteria to define low and high flows and it should assume similar values for each model since they are representing the same process. For this reason, the cases when some models' forecasts are lower than its threshold and others are higher (hereafter called mixed TNDs) are not very frequent and they just occur in few situations during the rising or recession limbs of the main events. Moreover, due to the few occurrences of these cases, the mixed TNDs should be identified using very small data sets, which are not usually sufficient to compute significant sample moments. Hence, the authors think that the introduction of one threshold for each model could lead to worse approximation than using only one model to divide the Normal Space.
Anyway, situations with many data also in mixed TNDs can happen and also these TNDs can be taken into account. As mentioned in the answer to the comment #7, this work aims at introduce an alternative way to represent the joint distribution used to compute the Predictive Uncertainty, hence many attempts can be done to analyze the use of different kinds of TNDs and how it can improve the Predictive Uncertainty assessment. This may be the objective of future researches.

10) **C:** *The results of example application show that the PU technique outlined does not appear to fully capture the timing errors on the rising limb (Fig 10 & 11). Please comment as to why this is the case.*

A: In the presented conceptualization it is not possible to identify timing errors, because the inputs to the processor are the models' forecasts at a specific time horizon (i.e.: 6 hours in the Baron Fork application). This means that if the forecasts have a systematic delay, during the rising limb the model will subestimate the observed value while it will overestimate it during the recession limbs. The processor is not able to recognize if the predicted value belongs to a rising or a recession limb. In order to identify timing errors a different conceptualization, where the input are forecasts at different horizon times, must be used (concerning this issue, please, refer to the last point of answer to comment #4).

11) **C:** *Section 4 is very difficult to follow, though the explanation of the results in comparison to Figures 1 & 2 is worth reporting. The interpretation of Figure 14 is very unclear, the labelling of the y axis as 'Observed occurrences' makes little sense when related to the text and figure label.*

A: The analysis of the correctness of the probability to exceed an alert level is necessary to validate the Predictive Uncertainty assessment and to verify the robustness of the processor results in view of operational decisions. The authors agree with Referee #1 concerning the unclearness of Figure 14, hence it will be better explained.

12) **C:** *The conclusions introduce a number of summary results that would be better reported and discussed in Sections 3 and 4. The results on line 6 Pg 9247 indicate some misrepresentation of the predictive uncertainty that may require further discussion.*

A: The misrepresentation of Predictive Uncertainty, which Referee #1 refers to, can be caused by the two main assumptions done in the presented methodology: the assumptions concerning the Joint Distribution shape and the marginal distributions' tails. The authors think that the percentages of observed values falling outside the 90% uncertainty band presented in the conclusions can be considered a good result, especially considering that these values have been computed over the entire validation period, which includes a wide number of data.

Other Remarks

- *I would consider the errors given for the observation of water level (Pg 9223 Line 24) to be optimistic (especially at sites without gauging structures)*

A: We concur with the Referee #1 comment. In fact the measurement errors referred to were the instrument errors. In practice, as pointed out in Dottori et al. (2009) water stage measurement errors, which also include water level oscillations, add up to some $\sqrt{5}$ cm, which results from taking into account an instrument with a standard error of 2 cm + a water level oscillation of 3 cm, which leads to something around 2.24 cm, which is still small. The authors will modify the text accordingly

- *The reference on Pg 9223 Line 23 is inadequate*
- *More evidence then one conference paper is required to suggest that a paradigm has been radically changed (Pg 9924 Line 11)*
- *Pg 9224 first paragraph. From personal experience I don't think I have met an operational flood forecaster who believes that a deterministic output from their model is correct. While they do use the model forecasts in issuing warnings to suggest an approach as naive as your text does is perhaps doing many a disservice.*
- *Pg 9224 Line 16: 'in' not 'into'. This statement need evidence*

A: Point taken. The authors will modify the manuscript accordingly.

- *Pg 9226 Line 15: To claim the HUP was the first 'correct' formulation of the prediction uncertainty needs further support. Even if excluding other inference methodologies other Bayesian formulations (with differing assumptions) had been given before (see references above)*
- *Pg 9226 Line 15: Why is the HUP 'hardly extendable' to multiple models?*

A: The authors acknowledge the fact that the statement that the HUP was the first correct formulation is misleading. The authors wanted to underline the fact that Krzysztofowicz (1999) with the HUP was the first in hydrological forecasting who clarified the concept of predictive uncertainty. The text will be rephrased.

The reason why the authors claim that the HUP is hardly extendible to multiple models is because its formulation would require deriving the joint multiple likelihood of all the models given the observations, which is quite hard to get or to perform the upgrading one model at a time, but the analytical formulation for this latter approach is not available and the authors doubt that can be analytically derived as in the case of the single model.

- *Pg 9230 Line 1: Please enlarge on the extensions required to accommodate the tails of the distribution or provide a reference.*

A: A description of the tail models will be added.

- *Pg 9231: As the authors point out on page 9232 this page outlines computing the conditional distribution from a multivariate normal distribution. Since this is well known it could be shortened.*

A: The Equations will be substituted by references.

- *Pg 9237 Line 9: Missing symbols?*

A: If Referee #1 refers to the acronym MTNDs, it stands for Multivariate Truncated Normal Distributions, an explication of the acronym will be added after the first .

- *Section 3 onwards: The standard of the English in these later section is not a high as earlier in the text and requires further proof reading. For example the word 'pick' is repeatedly used instead of 'peak'.*

A: Point taken. The authors will modify the manuscript accordingly.

- *pg 9243 Line 25: What 'early stopping procedure'?*

A: The 'early stopping' is a procedure that stops the Neural Network calibration as soon the evaluation indexes computed on the verification data set starts to decrease. It is used to avoid the overfitting of the calibration data. This explication will be inserted in the text.

- *Figures 3 & 4 appear quite blurred. It is not clear what Figure 5 adds to the presentation.*

A: Figure 5 can be removed, quality of Figures 3 and 4 will be improved.

Cited Literature

Di Baldassarre, G. and Montanari, A. 2009. Uncertainty in river discharge observations: a quantitative analysis, Hydrol. Earth Syst. Sci., 13, 913–921, doi:10.5194/hess-13-913-2009.

Dottori, F., Martina, M.V.L., Todini, E., 2009. A dynamic rating curve approach to indirect discharge measurement Hydrol. Earth Syst. Sci. Discuss., 6, 859–896.

Kalman R.E. 1960. A new Approach to linear filtering and prediction problems. J. Basic Eng. Trans. ASME, 82 D, 35-45.

Kalman R.E. and Bucy R.S. 1961. New results in linear filtering and prediction theory. J. Basic Eng. Trans. ASME, 83 D, 95-108.

Krzysztofowicz R. 1999. Bayesian theory of probabilistic forecasting via deterministic hydrologic model. Water Resour. Res., 35 (9), 2739–2750.

Todini, E., 2008. A model conditional processor to assess predictive uncertainty in flood forecasting. Intl. J. River Basin Management. Vol. 6 (2), 123-137