Answers to Referee #3 comments

First of all the authors would like to express their gratitude to Referee #3 because his remarks were really appropriate and allowed the authors to improve the original version of the manuscript.

C: *I found the development of the methodology more than adequate except for the section dealing with TNDs for more than one forecasting model. For the multi-model TND, the authors basically assign a threshold for the best performing model thus restricting the TND to one model. It is unclear to me why this would properly take into account heteroscedasticity particularly in cases where the competing models have almost the same predictive skills. Does this methodology only work for the published case study or is it more robust?*

REPLY: The authors thank the reviewer because he pointed out a question which, on one hand was not sufficiently clarified in the manuscript and on the other hand was not exactly what it was meant to be.

Before answering to this question, the authors would like to make an introductory note. The use of the Truncated Normal Distributions (TNDs) is needed to take into account the heteroschedasticity of the data. In particular, the nature of the hydrological data and the application of the Normal Quantile Transform (NQT), tend to enhance the heteroschedasticity of the distribution of low and high flows. The application of the NQT tends to enlarge the variability of the low flow values while reducing that of high values in the Normal Space. Given that the number of high flows is also much smaller than the corresponding number of low flows, the latter have an overwhelming weight in predictive uncertainty assessment, which often leads to overestimate the variance and at the same time to underestimate the mean of the predictive uncertainty relevant to high flow values. This behavior is not specific of the presented cases, but it is common in most hydrological application, because it is mainly due to the application of the NQT to hydrological data, which are characterized by a limited number of extreme values, particularly in the upper tail. Figs. 1 and 2 show that the threshold values, for both TETIS and TOPKAPI models, correspond to a change in the slope of the cumulated distribution of the original forecasted variable. Hence, analyzing the NQT function it is rather simple to understand why the error variance is reduced above the threshold and increases below it.

On the contrary, when the homoschedastic conditions are met, the introduction of a truncation threshold and the use of the TNDs do not affect the joint distribution representation, and the obtained solution is practically identical to what one would obtain using a Normal Joint Distribution. This is the case of the ANN on the Baron Fork River and the same happens when applying the MCP to the Po River, which is a second case study that will be added in the revised manuscript.

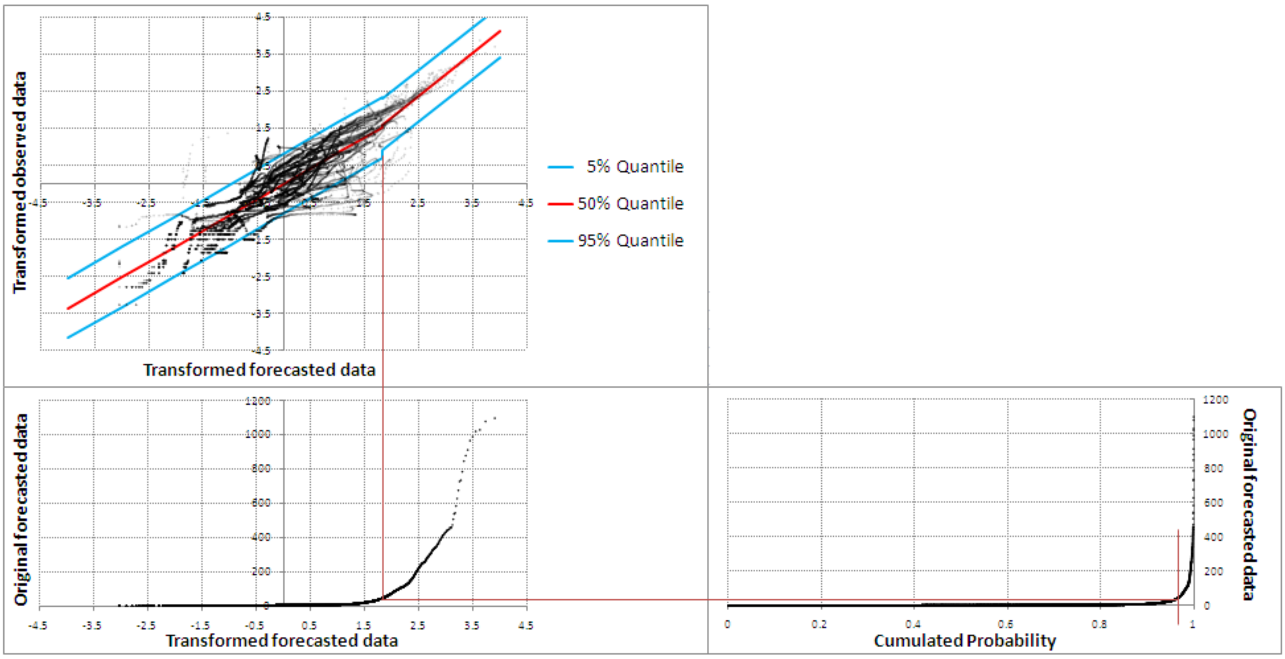

Figure 1. Representation of the Normal Space obtained using the MCP with the TOPKAPI forecasts. The relation between the threshold and the change in the slope of the cumulated distribution function of the original data is shown.

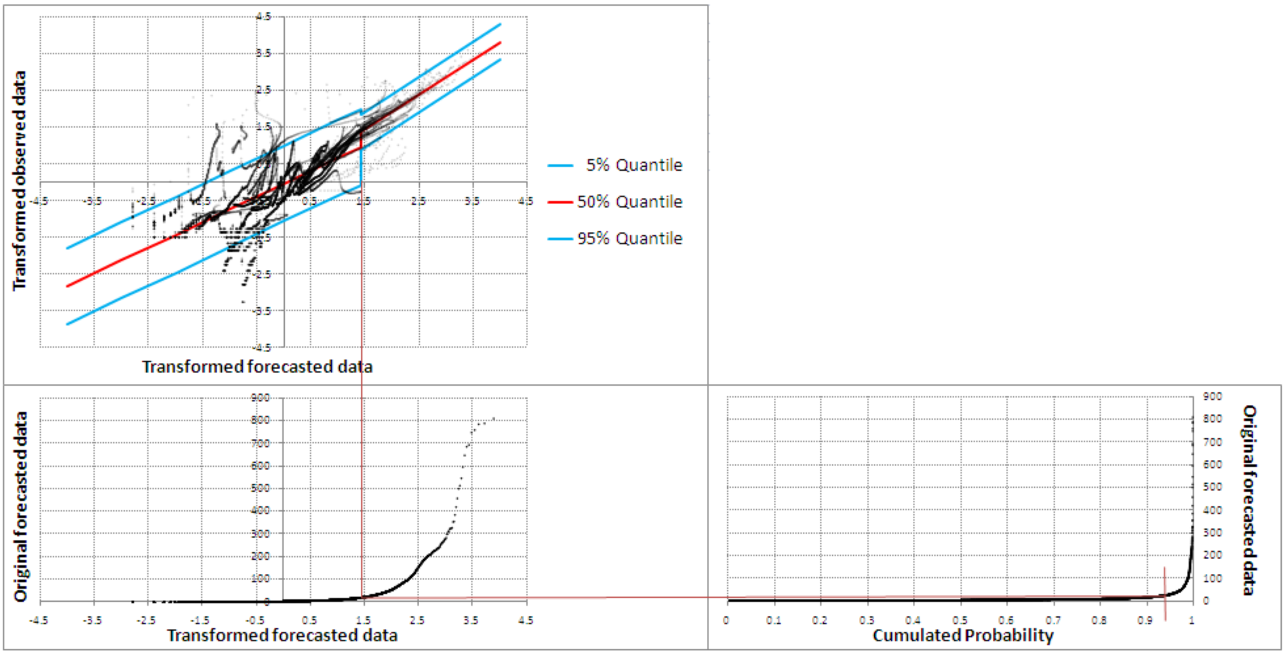

Figure 2. Representation of the Normal Space obtained using the MCP with the TETIS forecasts. The relation between the threshold and the change in the slope of the cumulated distribution function of the original data is shown.

The answer to the question is now easier. Concerning the multi-model approach the previous observations are still true, but the problem is that when *N* model forecasts are combined together, the joint distribution becomes an (N+1)-variate distribution, which would require 2^N TNDs if each model space is truncated by a threshold. In many cases, it would involve samples statistically not significant. As presented in the manuscript (and correctly understood by the Reviewer #3), the selected truncation threshold, based on the model which better represents the high flows, is used to separate the space in two portions. In practice, if in the basic approach the bi-dimensional joint space of η and $\hat{\eta}$ is divided by the straight line $\hat{\eta} = a_{\hat{\eta}}$, in the multi-model approach the N+1-

variate joint space of η and $\hat{\eta}_{k,k=1..N}$ is dived by the hyperplane $\hat{\eta}_k = a_{\hat{\eta}_k}$, where k represents the model that better reproduces the high flows.
We fully agree with the reviewer comments concerning the not full correctness of taking into account just one model to split the joint space, hence we decided to adopt a new hyperplane form. The truncation in two portions is now based on the hyperplane that cuts all the model axes at the same threshold value, this value is again identified as the one which minimizes the predictive variance of the high flows. The new hyperplane is $\sum_{i=1}^{N} \hat{\eta}_i = N \cdot a$. In any case, although more appropriate, the new hyperplane adoption has led to marginal differences from the previous results. However, the authors certainly agree with Referee #3 that the methodology should be tested on more cases and this will be made in future research.

C: *As already noted, the reliability diagram in Figure 14 represents the most important test of the MCP performance. Because of the importance of this figure more detail about its construction should be provided. In addition, how many separate flood events were used, how many data points were used in estimating exceedance frequencies and what was the range of thresholds. […]*
Figure 14 shows that most of the points fall below the 1:1 line suggesting MCP is consistently overestimating the probability of exceeding a threshold. A more critical discussion of this shortcoming would be appropriate.

REPLY: The authors agree with Referee #3 that Fig. 14 represents the most important test of the MCP performance. Also according to the comments of Referee #1, the following details about the construction of Fig. 14 will be added to the manuscript.
Figure 14 has been obtained considering the entire validation period for the MCP and just one threshold value (75 m^3s^{-1}). The verification has not been made on successive events, but considering all the continuous data record (35 months of hourly prediction with a 6 hours time horizon). The observations of Referee #3 suggested testing the MCP performance considering not just one threshold, but rather the PU quantiles with a 5% interval. This allows to explore all the range of possible threshold values and to have the maximum number of available data falling inside each interval. The manuscript will be updated with the new results.
Concerning this new verification, the results highlighted a systematic overestimation of the predictive variance. This overestimation justifies the positive bias for low probabilities values and, although less evident, the negative bias for high probability values observed in Fig. 14. Further analysis allowed identifying the cause of this overestimation, which is due to the failure of the normality hypothesis in the tails. The residuals analysis showed that they are almost normally distributed in the central part, but their empirical distribution has fatter tails due to few data with large residuals.
This important issue, resulting from the analysis suggested by Referee #3, has been addressed and will be discussed in an appropriate section in the new manuscript version.

C: *The reliability diagram must be constructed using events that are independent of the events used in calibration. It appears that Figure 14 was constructed from events in the validation period for MCP as shown in Figure 6. However, this period overlaps with the calibration period for Topkapi and the verification period for the ANN.*

REPLY: Referee #3 correctly pointed out that there is an overlap between the MCP validation period, the TOPKAPI calibration and the ANN verification ones. The authors agree with Referee #3 that the reliability diagram must be constructed using events that are independent from the events used in calibration. In particular validation data, or better operational forecasts data should

be used for (a) calibrating and (b) validating MCP. In the revised version, this will be shown on the basis of a second example where the Po river operational forecast data will be used.
A fair way of using the data would have been the one depicted in the following figure, where both calibration and validation of MCP is only performed using data in the validation period of all the models.

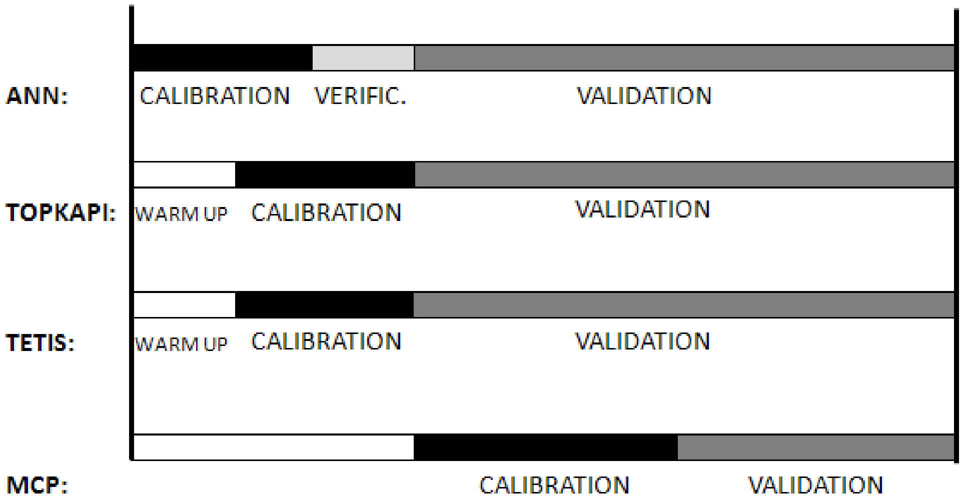

Figure 4 – A correct way of using and testing the model combination

Unfortunately this was not possible for several reasons. The main reason was the relatively short length of the record in which only few major events were available. A second reason was that the TETIS model results were not generated by the authors, but received from a second participant to the DIMIP 2 project.
This is why it was decided to consider the hydrological model forecasts regardless to their calibration and verification phase. In other words, apart from the ANN model, which validation period data were correctly used (verification data were not used to calibrate but only to stop calibration in order to prevent over-fitting and were also assumed as validation data), the output of the two hydrological models was used as an operational output, without identifying the calibration and the verification periods.

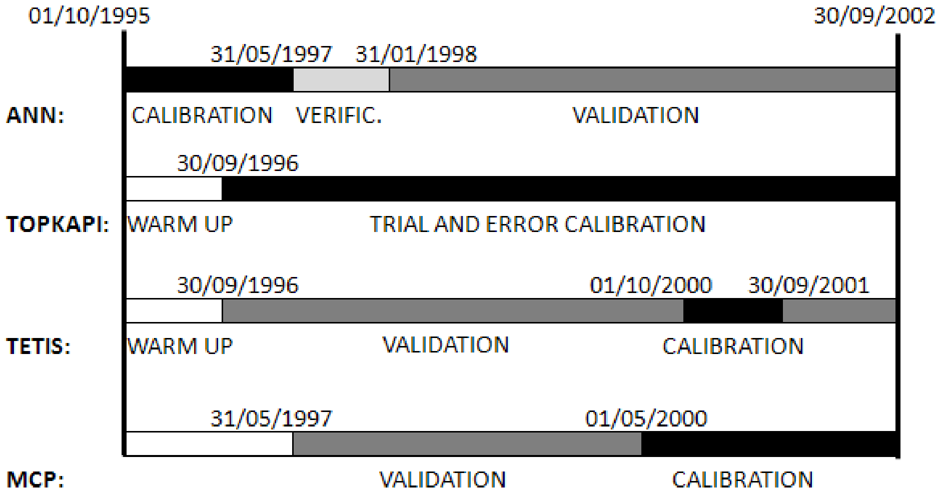

Figure 5 – The data set as it was used in the experiment

This procedure, although not fully correct, does not affect substantially the results of the model merging in terms of MCP, if one assumes that all the data used in the uncertainty assessment can be assimilated to “operational forecasts” over a period of time, since they were all produced with

the same parameter values, as it happens in operational real time forecasting. The MCP approach can still show the improvement that the combination of more models has with respect to the use of the single model.
As mentioned earlier, in the second example, the Po river example, the data used do not relate to any calibration period: they correspond to a series of operational flood forecasts.

C: *The final discussion and conclusions need to be more forthcoming about the shortcomings/limitations of the methodology. While the use of multiple models has improved forecast skill, there appears to remain a bias in forecast uncertainty. Moreover, while the TND appears to address much of the non-Gaussian joint probability behavior in the transformed variables in the Baron Fork case study, its applicability to other catchments with different response times and data quality remains untested.*

REPLY: The authors agree with the reviewer. The section named *Conclusions* will be moved inside the previous sections and more critical conclusions will be added, focusing on 3 main points.

1) Concerning the use of the NQT, this approach has some disadvantages. First of all, it implies to identify additional models to adjust the quantiles outside the range of the historical available data. The proposed technique is quite sensitive to the shape and to the parameters of these models and some precautions in the choice of the subset of observations used for calibrating the tails data must be taken (as it was mentioned when discussing the overestimation of the probability of low quantiles). They must contain a large variety of cases, as required by any Bayesian approach, and in order to reduce the uncertainty on the marginal distribution tails the calibration data must include the highest number of extreme cases.
2) Concerning the TNDs, this technique can be easily developed and applied obtaining good results such as for the study cases where it has been used. The results shown in Fig. 1 demonstrate that the joint distribution is well represented with this technique, even if some unavoidable approximations are still present. Nevertheless, the methodology should be tested considering other catchments with different features and for each specific application the correctness of the joint distribution representation must be verified. However, it must be noted that the use of the TNDs does not affect those cases when the data are homoscedastic.
3) Further discussions concerning the normality hypothesis of the joint distribution will be added.