Submission to HESSD
"Estimation of flood warning runoff thresholds in ungauged basins with asymmetric error functions", by Elena Toth

Author's Reply to Referee #2's Comments

I do warmly thank the Referee who carefully revised the paper pointing out some of its weaknesses (and identifying some wrong or confusing sentences too…) and providing very useful suggestions to improve it, that I will certainly follow in the revised version.

Detailed comments

Comment: It is indicated that the annual maximum flow records for some stations are available for as few as only 5 years. How was the quantile of interest in this work estimated and how meaningful is the estimate done using such a short data set?

The only quantile that is estimated in this work is the 2-years one and it was estimated as the median of the available historical records of flood maxima. Even if of course it would be preferable having longer time-series, five years should be sufficient for such a short return period, for example according to the classical guideline by Cunnane (1987), that suggests not to extrapolate statistical inference beyond a return period of 2 times the sample length (and for the shortest records in the present application, it is inferred a quantile with a return period that is less than half the sample length). In addition, the stations with less than 8 years of data are only 9, so that I believe the dataset, in terms of the length of the records, may be considered, overall, sufficiently meaningful for the purpose of estimating the 2-years return period flows.

Comment: Why were only three classes of catchment descriptors used to sample representative catchments from for the three groups of catchments? Are they not too few to enable a fair distribution of different ranges of the catchment characteristics evenly across all the three groups?

I fully agree that such choice is subjective and that a different number of classes could have been chosen; given the small number of features characterizing the catchments (the 3 first principal components); I believe that 3 classes should be sufficient for identifying training, cross-validation and test sets that are sufficiently similar. I report below the graphs showing the mean value (red dash) and the 90% and 10% percentiles of the resulting sets, for each of the three input variable (PC1, PC2 and PC3). The graphs seems to highlight a good degree of similarity in the distribution of the values over the three sets. Such graphs might be added in the revised manuscript at the end of Section 3.2.

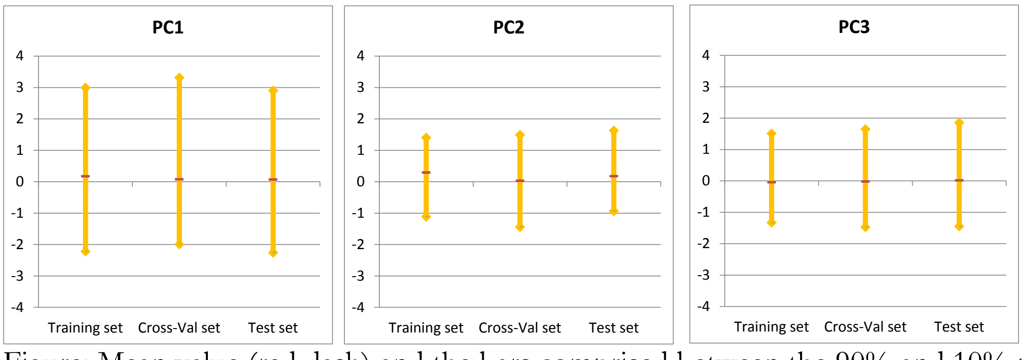

Figure: Mean value (red dash) and the bars comprised between the 90% and 10% percentiles of the resulting training, cross-validation and testing sets, for each of the three input variable (PC1, PC2 and PC3).

Comment: How were the output values standardized in the range between -1 and 1 (page 6023, line 12). Here I assume the output variables to be the 2-year flood values.
Related to my previous comment, are the error terms in Equations 3-5 estimated from the normalized 2-years flood values or from the actual values? If they are estimated from the actual values, as it looks is the case by looking at the values of MEA and RMSE in Table 1 and the errors in Figure 2, how was the scale inconsistency at the different stations handled? It is mentioned somewhere that these values range between 10 and 1000m3/s.

Data standardization is generally used in neural network in order to ensure that the data receive equal attention during the training process (Maier and Dandy, 2000) and it is also important for the efficiency of training algorithms (Dawson and Wilby, 2001).
In the present case, the output data are rescaled as a function of the minimum and maximum values to the [-1 1] range (actually to the [-0.95 +0.95] range, to avoid the problem of saturation – I had not explained this issue in order not to further complicate the explanation of the neural networks working). Such rescaled values are those that are simulated by the ANN model: those corresponding to the training and cross-validation sets are used, as 'target', for training the neural networks; when the model is successively used for predicting the $Q_{2,p}$ over the independent validation set, the ANN output values are then transferred back reversing the function, restoring the values - in a proportional manner - to the original ranges.
Finally, the error statistics presented in section 5 (and in Table 1 and Figure 2) are calculated on the errors between the actual observed values and the re-transformations of the values issued by the neural network.
It is certainly necessary to better explain the procedure in the revised manuscript, both in Section 4.2 and 5.1 (and in the latter it is indeed important to clarify that the prediction $Q_{2,p}$ is not directly the value issued by the model but its 'de-standardised' value, since the present wording is, as highlighted by the Referee, confusing).

Comment: I find the whole text on page 6028 messy. Most of the discussion on results is presented on this page, but it is very confusing. The author mentions that negative errors mean overestimation and a couple of lines later a contradictory statement is made (statements on line 6 and 10). Similarly, it is mentioned somewhere that the overestimation error reduces with increasing alpha value and the opposite is mentioned elsewhere. There is even little consistency between what is discussed here and the referred Table 1 and Figure 2.

Thank you indeed for having identified two mistakes in the same sentence (I really have to apologise: I had in my last version reworded the sentence changing the focus from under to overestimation errors and I have inadvertently maintained a part of the original sentence…): lines 10-12 should read:

"At the same time, and more importantly, the number of negative (~~under~~ overestimation) errors larger than 30% substantially decreases with *a*, with OverH% reaching a value that is much lower than that of the ANN-Symm model when *a* arrives at ~~0.4~~ 0.1 (18% vs 34%)"

I do hope that amending the wrong sentence , the text will become more clear and it should be consistent with the results shown in Table 1 and Figure 2.

Comment: Why did the author choose to define the error term as the observed minus the simulated values? Defining it in a more conventional way would have helped to avoid such inconsistency.

In the manuscript I used the notation by Elliot et al (2005), for consistency with their definition of the loss function in Eq 1.

In order to better warn the hydrologists readers, I may add the equation:

$$\varepsilon = O - P$$

instead of defining it only in words, or I might (if you do not deem such addition to be sufficient) reverse the definition of errors throughout the whole paper.

Other comments

Comment: Define the variable M in Equation 3.

Thank you for pointing this out: I have to add that M is the number of records in the set.

Comment: I suggest that the catchment descriptors be listed in a table. I am a bit astonished to read that data on soils and land cover are missing when there are open data sources on both that are often used in modeling.

I will list, as suggested, the descriptors in a Table:

1	*Long* - UTM longitude of catchment centroid
2	*Lat* - UTM latitude of catchment centroid
3	*A* - Catchment Drainage area
4	*P* - Catchment perimeter
5	z_{max} - Maximum altitude of the catchment area
6	z_{min} - Elevation of the catchment outlet
7	z_{mean} - Mean altitude of the catchment area
8	*L* - Length of the Maximum Drainage Path
9	S_L - Average slope along the Maximum Drainage Path
10	S_A - Catchment average slope
11	Φ - Catchment orientation
12	*MAP* - Mean Annual Precipitation

And I definitely take the Referee's point that it would be extremely helpful to extend the database content, and working on a consistent, comprehensive database of Italian catchments with validated and reliable information on other important features of the catchments' areas.
However, such compilation of an extended database for the Italian country was not the object of the present analysis, that presents a comparison of methodologies applied utilizing the same dataset, and I based the analysis on the data made available by the CUBIST project (the most recent National project of characterization of the Italian basins) and already used in Di Prinzio et al. (2011).
I do hope, in the (hopefully near) future, that the colleagues who prepared the CUBIST database (and who have already developed the analyses for the delineation of the catchment boundaries) will find the time (and I may certainly offer my help, too) to include additional descriptor to the national database.

REFERENCES

Cunnane, C. Review of statistical models for flood frequency estimation, in Hydrologic frequency modeling (Ed. V. P. Singh), pp. 49-95, 1987.

Dawson C. and Wilby, R. Hydrological modelling using artificial neural networks, Progress in Physical Geography 25,1, 80–108, 2001.

Maier, H.R. and Dandy G.C. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications, Environmental Modelling & Software 15, 101 – 123, 2000.