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

Reply Letter

I warmly thank the Editor, Dr. Stacey Archfield, for her kind encouragement and support and the two Referees for having so carefully read the paper, providing very useful and constructive insights, that prompted me to integrate and improve the work.

In the replies that follow, I explain how and where the required information were added and the suggested modifications implemented, for each comment.

For each remark, the original comment of the referee is copied in bold, the reply follows afterwards and the change made in the revised version are in red.

Please note that the insertion of the new figures and the new table added following the suggestions of the Referees have modified the overall numbering of the existing ones, and in particular:

Figure 2 of the original manuscript is Figure 4 in the revised version

Table 1 of the original manuscript is Table 2 in the revised version

Following the replies, the new manuscript is attached, where all the changes to the original version are highlighted in yellow.

I do hope the revisions adequately address the Referees' comments and I am at your disposal for any clarification that may be needed.

My best regards,

Elena Toth

Author's Reply to Referee #1's Comments

General Comments

Comment 1) • What are the advantages of using a biased estimate of $Q_{T=2yr}$? […]

I do thank the Referee for having pointed out the importance to better underline in the paper the limitations and advantages of the proposed approach, and in particular with respect to a probabilistic one.

It is certainly true that a probabilistic approach (it is possible to find examples also for neural network models, see for example Khosravi et al., 2011) may be able to add very valuable insights for a more complete evaluation of the prediction model, supplementing the information provided by point-value predictions, and in the future I intend to attempt to investigate the uncertainty of the issued predictions, (as I have added in the concluding session in the revised version) but I do not believe it would be possible performing such complex analysis here. It should in fact be considered that uncertainty assessment methods should take into account all error sources (see for example Montanari, 2007) and not only those related to the choice of their parameters (the majority of the uncertainty methods deal only with a single source of uncertainty, for instance, Monte Carlo-based methods analyze the propagation of uncertainty of parameters only) - and are subject, as well as the prediction model itself, to errors in their underlying assumptions and structure as well as in the determination of their parameters (Xiong et al., 2009) so that it is needed, even if it is far from easy, to correctly evaluate also their quality (in many methodologies it is hypothesised and not verified if the distribution of the forecasts is the real one). As a consequence, implementing a correct, fully comprehensive procedure for a consistent and reliable estimation of the global uncertainty is certainly not straightforward (nor it would be possible to describe it briefly) and this is why it may be subject of a separate, future work.

On the other hand, I do not believe it is necessary to add the implementation of a probabilistic framework here, since the presented methodology is a deterministic one, where an optimal point forecast is obtained by minimizing the conditional expectation of the future loss.

Such framework has not the pros of a probabilistic one in terms of quantification of the uncertainty, but its advantage is the operational value of the forecast in terms of an optimal decision that minimizes the cost; in fact asymmetric loss functions are more appropriate in many types of decision settings, as shown by recent forecasting literature analysing the statistical properties of optimal predictions under asymmetric loss (e.g. Christoffersen and Diebold, 1997, Granger and Pesaran, 2000, Patton and Timmermann, 2004; and in particular Zellner, 1986, 2004, showed that once the symmetric loss function is abandoned, optimal forecasts need not be unbiased) and showing that in many " practical applications, asymmetric loss functions can be critical to effective forecasting" (Elliott et al., 2006).

Minimising the asymmetric error function has the purpose of minimizing the cost, thus optimizing the threshold from an operational point of view. A probabilistic forecasting approach applied to the symmetric error function (provided that the methodology is able to include all sources of uncertainty and its quality may be objectively assessed/verified) would certainly provide awareness on the uncertainties associated with the point forecasts, but identifying the upper (e.g. 95%) uncertainty bound would not allow the decision-maker to choose the optimal value for the threshold in terms of costs/operational utility, since such value (upper bound) would be (if reliable) the one that identifies an assigned risk of underestimation (and, even if this is not the point here, it would, I expect, result in a very high value for a small assigned risk, given the large uncertainty of the approach, mainly due to the intrinsic limitations and shortcomings of the data set for such an heterogeneous area…) but it would not take into account in any way the overestimation costs resulting from high negative errors, nor it would consider the balance between the costs of positive and negative errors, as it may be done, instead, within an asymmetric loss approach.

In the revised version I have better specified the purpose of the proposed approach, along with considerations on the advantages/disadvantages in respect to a probabilistic framework, adding two new paragraphs in Section 2 and in the Conclusions.

Comment 2) • Regional Flood Frequency analysis is not regression. In a couple of locations in the text, page 6014 line 14–29 and page 6030 lines 10–18, there seems that there is the direct association between Regional Flood Frequency analysis to Regression with catchment attributes (regression or related techniques like ANN's). […]

In the revised version I will certainly rephrase ll. 24-27 p. 6014 and ll. 13-14 p. 6030, since I definitely did not mean to reduce Regional Flood Frequency Analysis to the application of regression techniques, but only to refer to that thematic area, because the runoff threshold literature generally does not include these issues. I fully agree (as highlighted, as the Referee underlines, also in the chapter on floods prediction that I co-authored of the 2013 book) that regression methodologies are only one of the possible methods (statistical and process-based) to predict floods in ungauged basins and in particular I should better specify in the text that their use is especially frequent only as far as the estimation of the index flood values is concerned.

In the revised version I have rephrased both paragraphs (in Section 1 and in the Conclusions)

Comment 3) • Relative error could be also very valuable. For assessing the performance of several variants of the proposed method, the measures MAE and RMSE are proposed, both functions of the error. Given the large range of discharges considered in the study, it could be also very valuable to report additionally boxplots of the relative errors. […]

I definitely agree that a more comprehensive description of the errors would be very helpful to interpret the results, especially given the large discharge range, as underlined by the Referee; I am not sure the relative errors would be the fairest way to analyse the results in the presented decision setting framework (see reply to first Comment), given that the costs are weighted in respect to the 'not-relative' errors in the loss functions, so I would prefer, if the Referees agrees, adding in the revised paper the scatterplots of observations/predictions, that I believe allow the most complete visualization of the results over the entire discharge range, showing every single prediction in respect to the corresponding observation. In addition, also OverH and UnderH are already defined in relative terms, since they represent the number of errors greater than 30%.

In the revised version I have added a new figure (Figure 3) showing the scatterplots of the results issued by the 5 models over the independent test set, showing every prediction in the respect to the corresponding observation.
Admittedly the scatterplot highlight that the errors are far from negligible for both the traditional and the asymmetric networks (as already underlined in the original paper); but the test is indeed an exacting one: a single regression model applied over a fully independent test set that includes extremely diverse catchments (from Alpine to Mediterranean) and based on a dataset (the only one currently available at national scale) that unfortunately does not include important influencing factors, the most important one being information on the rainfall extremes.

On the other hand, the objective of the work is to propose a new methodological approach, that may be applied to much better databases and the Italian case study is proposed for a comparative analysis, where all the applied models have the same limitations due to the dataset.

Detailed Comments

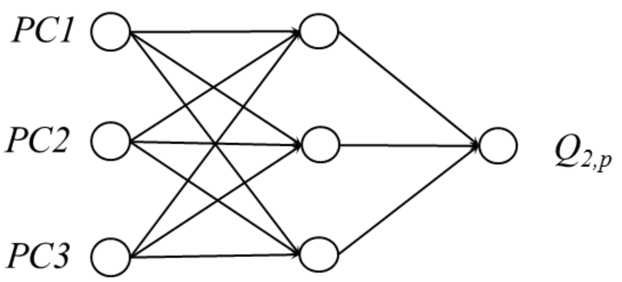
Page 6013, Line 27: is "real-world" here "real-time"?
Actually, I mean both: 'real-time' warning systems actually implemented by 'real-world' organization (that is not only in literature simulation studies): I will rephrase to make it clearer.

In the revised version I have replaced "real-world" with "operational real-time". (Introduction section)

Page 6017, Lines 17–19: The author defines here the error as the observed minus the predicted value. To my knowledge, in runoff prediction in ungauged basins, it is almost a consensus to define error as predicted minus observed. If the author wants to define it here inversely, a stronger warning to the reader should be given, in order to avoid confusions.
I used the notation by Elliot et al (2005), for consistency with their definition of the loss function in Eq 1.

In the revised version I have modified the definition of error as suggested by both Referees, since it indeed has generated confusion (both in the readers and in the author…). Accordingly, I have modified Fig 1, Equations 1 and 3, all the references to positive/negative errors throughout the paper and also the box-plot (ex-Figure 2).

Pages 6021–6023: Maybe adding an schematic figure with the structure of the selected ANN could help the reader.
I fully agree with your suggestion: I will add a figure showing the ANN architecture:



In the revised version I have added the above figure of the network architecture (Fig. 1c, section 4.2).

Page 6027, line 4: is here "scour" the Q2?
Page 6029, line 7: ... the errors <u>are</u> not negligible...
In the revised version I have amended the mistakes.

Page 6027, line 9: I think "prudence" is not the right word here. Maybe "tendency to over/underestimate"?
Page 6031, line 13: Again, "prudentially" is not the right word here.
In the revised version I have rephrased both sentences as suggested.

REFERENCES

Christoffersen P.F. and Diebold, F.X. (1997). *Optimal Prediction under Asymmetric Loss*, Econometric Theory, 13, 808-817.

Elliott, G., Granger C.W.J, and Timmerman A. (Editors) (2006), *Handbook of Economic Forecasting*, Elsevier, Amsterdam, The Netherlands, 1012 pp.

Khosravi, A., S. Nahavandi, D. Creighton, and A.F. Atiya (2011), *Comprehensive review of neural network-based prediction intervals and new advances*, IEEE Transactions on Neural Networks, 22 (9), pp. 1341–1356.

Montanari, A. (2007), *What do we mean by uncertainty? The need for a consistent wording about uncertainty assessment in hydrology*, Hydrol. Processes, 21, 841– 845, doi:10.1002/hyp.6623.

Xiong, L. H., Wan, M., Wei, X. J., and O'Connor, K. M. (2009), *Indices for assessing the prediction bounds of hydrological models and application by generalised likelihood uncertainty estimation*. Hydrolog. Sci. J., 54(5), 852-871.

Zellner, A . (1986), *Bayesian Estimation and Prediction Using Asymmetric Loss Functions*, Journal of the American Statistical Association, 81, 446-451.

Zellner, A. (2004), *Statistics, econometrics and forecasting*, Cambridge University press, Cambridge, New York, XVII, 163 pp.

Author's Reply to Referee #2's Comments

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.

In the revised version I have added such observation in Section 3.1.

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.

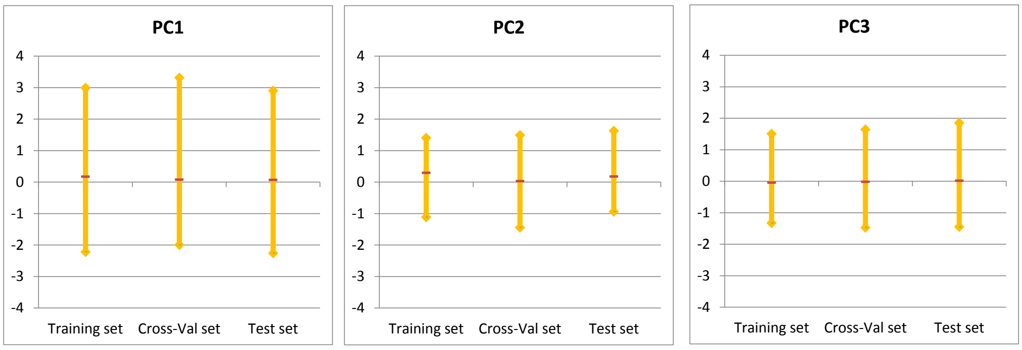


Figure: Mean value (red dash) and the bars comprised between the 90% and 10% percentiles of the resulting training, cross-validation and testing sets, for each of the three input variable (PC1, PC2 and PC3).

In the revised version I have added the above graphs in a new figure (Fig. 2) at the end of section 3.2.

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 standardized $Q_{2,p}$ over the independent test 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).

In the revised version I have better explained the standardization method in Section 4.2 and in Section 5.1 I have explained that the model results are compared after having de-standardized the network outputs.

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.

In the revised version, also due to the change in the definition of the errors (as suggested by the Referee in the comment below), I have amended the mistaken sentence and also rephrased the following part of page 6028 that was indeed not clear.

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 the revised version I have modified the definition of error as suggested by both Referees , since it indeed has generated confusion (both in the readers and in the author…). Accordingly, I have modified Fig 1, Equations 1 and 3, all the references to positive/negative errors throughout the paper and also the box-plot (ex-Figure 2).

Other comments

Comment: Define the variable M in Equation 3.

Thank you for pointing this out:
In the revised version I have added that M is the number of records in the set (either the early-stopping validation set or the test one).

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 new Table (new Table 1):

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.

On the other hand, the objective of the work is to propose a new methodological approach, that may be applied to much better databases and the Italian case study is proposed for a comparative analysis, where all the applied models have the same limitations due to the dataset.

In the revised version I have listed the descriptor in a new table (new table 1) and I have added in section 3.1 that "the CUBIST set is currently the only database available in the Italian hydrologists community at national scale".

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.

Estimation of flood warning runoff thresholds in ungauged basins with asymmetric error functions

E. Toth[1]

[1]{Department DICAM, School of Engineering, University of Bologna, Bologna, Italy}

Correspondence to: E. Toth (elena.toth@unibo.it)

Abstract

In many real-world flood forecasting systems, the runoff thresholds for activating warnings or mitigation measures correspond to the flow peaks with a given return period (often the 2-year one, that may be associated with the bankfull discharge). At locations where the historical streamflow records are absent or very limited, the threshold can be estimated with regionally-derived empirical relationships between catchment descriptors and the desired flood quantile. Whatever is the function form, such models are generally parameterised by minimising the mean square error, that assigns equal importance to overprediction or underprediction errors.

Considering that the consequences of an overestimated warning threshold (leading to the risk of missing alarms) generally have a much lower level of acceptance than those of an underestimated threshold (leading to the issuance of false alarms), the present work proposes to parameterise the regression model through an asymmetric error function, that penalises more the overpredictions.

The estimates by models (feedforward neural networks) with increasing degree of asymmetry are compared with those of a traditional, symmetrically-trained network, in a rigorous cross-validation experiment referred to a database of catchments covering the Italian country. The analysis shows that the use of the asymmetric error function can substantially reduce the number and extent of overestimation errors, if compared to the use of the traditional square errors. Of course such reduction is at the expense of increasing underestimation errors, but the overall accurateness is still acceptable and the results illustrate the potential value of choosing an asymmetric error function when the consequences of missed alarms are more severe than those of false alarms.

1 Introduction

In the operation of flood forecasting systems, it is necessary to determine the values of threshold runoff that trigger the issuance of flood watches and warnings. Such critical values might be used for threshold-based flood alert based on real-time data measurements along the rivers (WMO, 2011) or for identifying in advance, through a rainfall-runoff modelling chain, the rainfall quantities that will lead to surpass such streamflow levels, as in the Flash Flood Guidance Systems framework (Carpenter et al., 1999; Ntelekos et al., 2006; Reed et al., 2007; Norbiato et al., 2009).

A runoff threshold should correspond to a 'flooding flow', that is to a value that may produce flood damages, and it is very difficult to determine on a regional or national scale: it may be defined as a flow that just exceeds bankfull conditions, but in practice, both in gauged and in ungauged river sections, such conditions are arduous to quantify due to the lack of local information (Reed et al., 2007; Hapuarachchi et al., 2011).

In absence of more sophisticated physically-based approaches, based on detailed information of each specific cross-section that are rarely available due to limited field surveys, the literature suggests to estimate the bankfull flow as the flood having a 1.5 to 2 years return period (Carpenter et al., 1999; Reed et al., 2007; Harman et al., 2008; Wilkerson, 2008; Hapuarachchi et al. 2011; Cunha et al., 2012; Ward et al., 2013) and a flow that is slightly higher than bankfull may be identified with the 2-year return period flood (Carpenter et al., 1999; Reed et al., 2007).

Many operational systems all around the world adopt a statistically-based definition of the flooding flow and the flows associated with given return periods are used as threshold stages for activating flood warning procedures.

The 2-year recurrence is used by many River Forecast Services in the United States, as suggested by Carpenter et al. (1999), also due to the fact that "*the good national coverage of the 2-yr return period flows that the U.S. Geological Survey (USGS) maintains nationwide supports its use*" (Ntelekos et al., 2006), as well as in British Columbia (Canada).

However, the floods with different annual exceedance probabilities, associated with different levels of risk, are also frequently adopted in operational real-time flood warning systems: for example in the Czech Republic, flood watch usually corresponds to a 1- to 5-year flow return period (Daňhelka and Vlasák, 2013). In Italy, where a national directive issued in 2004 introduces a system articulated on at least two levels of flow thresholds, many Regions have identified the alert levels as flood quantiles with return periods of 2, 5 or 10 years (e.g. the Abruzzo, Lombardia, Puglia Regions). In the South of France, the AIGA flood warning system compares real-time peak discharge estimated along the river network (on the basis of rainfall field estimates and forecasts) to flood frequency estimates of given return periods (with three categories: yellow for values ranging from the 2-year to the 10-year flood, orange for between the 10 and the 50-year flood, and red for peaks exceeding the 50-year flood) in order to provide warnings to the national and regional flood forecasting offices (Javelle et al., 2014).

For river sections where the streamflow gauges are newly installed or where historical rating curves are not available, the observations of the annual maxima are absent or very limited and it is not possible obtaining a reliable estimate of flood quantiles on the basis of statistical analyses of series of observed flood peak discharges.

For these ungauged or poorly gaged basins, the peak flow of given frequency to be associated with the watch/warning threshold can be estimated transferring information from data-rich sites to data-poor ones, as it is done in the corpus of methodologies applied in RFFA (Regional Flood Frequency Analysis) at ungauged sites, that have always received considerable attention in the hydrologic literature (Bloeschl et al., 2013). Among the possible approaches (statistical and process-based) to predict floods in ungauged basins, many researchers have traditionally applied regression-like regionalisation methods for i) the estimation of the index flood (Darlymple, 1960), usually defined as either the mean or the median (that is the 2-year return period quantile) of the annual maximum flood series, or for ii) the direct estimate of other quantiles of annual maxima in ungauged basins

(Stedinger and Lu, 1995; Salinas et al., 2013). Such methods are based on the assumption that there is a relationship between catchment properties and the flood frequency statistics and are implemented through a regression-type model that relates the flood quantile or the index flood to a number of relevant morpho-climatic indexes. Linear or power (often linearized through a log-transformation) forms, with either a multiplicative or additive error term, are the most commonly used functions (see e.g. Stedinger and Tasker, 1985; GREHYS, 1996; Pandey and Nguyen, 1999; Brath et al., 2001; Kjeldsen et al., 2001, 2014; Bocchiola et al., 2003; Merz and Bloeschl, 2005; Griffis and Stedinger, 2007; Archfield et al., 2013; Smith et al., 2015).

In order to allow more flexibility to the model structure (whose 'true' form is of course not known), the international literature has recently proposed methods based on the use of artificial neural networks (ANN), providing a non-linear relationship between the input and output variables without having to define its functional form a priori. Successful applications of ANN for the estimation of index floods or flood quantiles at ungauged sites are reported in Muttiah et al., 1997; Hall et al., 2002; Dawson et al., 2006; Shu and Burn, 2004; Shu and Ouarda, 2008; Singh et al., 2010; Simor at al., 2012; Aziz et al., 2013.

Both the traditional power form or linear regression methods and the neural networks models are generally parameterized by minimizing the mean or root mean of the squared errors, that is a symmetric function assigning the same importance to overestimation and underestimation errors.

Nevertheless, the consequences of under or overestimating the runoff threshold when used for early warning are extremely different.

Adopting a watch threshold that is higher than the runoff/stage that actually produces flooding damages would in fact lead to missing such events, failing to issue an alarm. Underestimating the runoff threshold may instead determine the issue of false alarms.

False alarms may certainly lead to money losses and also "undermine the credibility of the warning organisation but are generally much less costly than an unwarned event." (UCAR, 2010): in fact the costs of failing to issue an alarm grow rapidly in a real emergency, since a totally missed event has strongly adverse effects on preparedness. Not only the costs of false warnings are commonly much smaller than the avoidable losses of a flood, but they cannot match up to indirect and/or intangible flood damages such as loss of lives or serious injuries (Pappenberger et al., 2008; Verkade and Werner, 2011).

Furthermore, regarding the effects of false alarms, "in opposition to 'cry wolf' effect, for some they may provide an opportunity to check procedures and raise awareness, much like a fire practice drill." (Sene, 2013)

Overall, false alarms have usually a higher level of acceptance than misses and this entails that the estimate of flood warning thresholds should be cautionary, so as to reduce, conservatively, the number of missed alarms.

For the development of watches and warnings it is therefore important to obtain estimates as accurate as possible, minimising both positive and negative errors, but, considering that an error will always be present, it is better underpredicting rather than overpredicting the threshold estimate, for safety reasons.

To obtain a conservative estimate of the thresholds, penalising more the predictions that exceed the "observed" values (in the present case represented by the quantile estimate based on the statistical analysis of observed flow peaks) than those that underestimate them, in the present work it is proposed, for the first time to the Author's knowledge, a parameterisation algorithm that weights asymmetrically the positive or negative errors, in order to decrease the consistency of overestimation and therefore the risk of missing a flooding occurrence.

It is important to underline that the proposed asymmetric error function is here applied for optimising a neural network model for predicting the 2-year return period flood (due to its association with the bankfull conditions) but it might be used to improve any other kind of methodology for the estimate of flood warning thresholds associated to any return period.

Section 2 presents the asymmetric error functions; the next one describes the information available in a database covering the entire Italian country and the identification of the subsets to be used for a rigorous cross-validation approach. Section 4 presents the implementation of the models for estimating the 2-year return period flood in ungauged catchments, consisting in artificial neural networks calibrated using respectively the symmetric square error and the asymmetric error functions. The results are presented and then discussed in section 5 and section 6 concludes.

2 The asymmetric error function

The scientific literature on forecasting applications, in any scientific area, adopts almost exclusively an objective function based on the sum or mean of the squared discrepancies, that is a symmetric quadratic function, due to the well-established good statistical properties of the minimum mean square error estimator.

On the other hand, in economics as well as in engineering and other many fields, there are cases where the forecasting problem is inherently non-symmetric and, in the financial forecasting literature, the use of mean squared error, even if still widely applied, is nowadays not always accepted.

Error (or loss) functions devised to keep into account an asymmetric behaviour have been proposed, such as the linear-exponential, the double linear and the double quadratic (Christoffersen and Diebold 1996; Diebold and Lopez 1996; Granger 1999; Granger and Pesaran 2000; Elliot et al. 2005; Patton and Timmerman, 2006). In particular, Elliot et al. (2005) recently presented a family of parsimoniously parameterized error functions that nests mean squared error loss as a special case (Patton and Timmerman, 2006).

Such function, adapted from Elliot et al. (2005) and defining the error ε as the prediction minus the observed value (that is, a negative error corresponds to underestimation, a positive one to overestimation), reads:

$$L(p,\alpha) = 2 \cdot \left[\alpha + (1-2\alpha) \cdot \mathbf{1}\{\varepsilon > 0\}\right] \cdot |\varepsilon|^{p}, \qquad (1)$$

where **1(·)** is a unit indicator, equal to one when $\varepsilon > 0$ and zero otherwise; p is a positive integer that amplifies the larger errors (corresponding to a quadratic error when equal to 2) and $\alpha \in (0,1)$ is a parameter representing the degree of asymmetry.

For $\alpha < 0.5$ the function penalises more the overestimation errors ($\varepsilon > 0$), while for $\alpha > 0.5$ more weight is given to negative forecast errors (under-predictions); for $\alpha = 0.5$ the loss weights symmetrically positive and negative errors.

When $p = 2$ and $\alpha \neq 0.5$, the error becomes the asymmetric double quadratic (Quad-Quad) loss function (see Christoffersen and Diebold 1996), that is used in the present work for a fair comparison with the traditional mean square error estimator. When $p = 2$ and $\alpha = 0.5$, Eq. (1) corresponds in fact to the 'traditional', symmetric, square error:

$$L(2,0.5) = \varepsilon^2 \tag{2}$$

Figure 1 shows the asymmetric Quad-Quad loss function (with α varying from 0.1 to 0.9) compared with the squared error (SE).

In the water engineering field, the asymmetric Elliot error function with quadratic amplification (p = 2) has been recently applied to parameterise a model for estimating the expected maximum scour at bridge piers, in order to obtain safer design predictions through the reduction of underestimation errors by Toth (2015).

It should be noted that the proposed methodology is a deterministic one, where an optimal point forecast is obtained by minimizing the conditional expectation of the future loss; such framework has not the pros of a probabilistic one in terms of quantification of the uncertainties of the prediction, but it aims at identifying the optimal value for the threshold in terms of operational utility.

In Section 4, the asymmetric quadratic error function is proposed for optimizing the parameters of an input-output model, based on artificial neural networks, between the input variables summarising a set of catchment descriptors (obtainable also for ungauged river sections) and the 2-year return period flood, thus warranting that overestimation errors, that would increase the risk of missing flood warnings, are weighted more than underestimation ones.

3 Available information: the national data set of Italian catchments

The case study refers to a database of almost 300 catchments scattered all over the Italian peninsula, compiled within the national research project "CUBIST – Characterisation of Ungauged Basins by Integrated uSe of hydrological Techniques" (Claps et al., 2008).

3.1 Input and output variables

The 12 geomorphological and climatic descriptors are listed in Table 2. The dataset unfortunately lacks information on other hydrological properties (e.g. on soils, land-cover, vegetation) and the climatic characterisation is very limited (for example information on extreme rainfall would be extremely important), but the CUBIST set is currently the only database available in the Italian hydrologists community at national scale.

The dataset is described in Di Prinzio et al. (2011), where, following a catchment classification procedure based on multivariate techniques, the descriptors were used to infer regional predictions of mean annual runoff, mean maximum annual flood and flood quantiles through a linear multiregression model.

As described in such work, in order to reduce the high-dimensionality of the geomorphological and climatic descriptors set, a Principal Components (PC) analysis was applied, obtaining a set of derived uncorrelated variables. The PC variables are as many as the original variables, but they are ordered in such a way that the first component has the greatest variability, the second accounts for the second largest amount of variance in the data and is uncorrelated with the first and so forth. In the present data set, the first three principal components explain more than three quarters of the total variance (see Di Prinzio et al., 2011) and such three first PCs are here chosen as input variables to the models described in the following, assuming that they may adequately represent, in a parsimonious manner, the main features of the study catchments.

The data base, in addition to the morpho-pluviometric information, includes the annual maxima flow records for periods ranging from 5 to 63 years, whose median values, corresponding to the 2-year return period, represent the output variable to be simulated by the models. Even the shortest records (and actually only 9 of the locations have less than 8 years of data) 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.

The data set covers a great diverseness of hydrological, physiographic and climatic properties and in order to partially reduce such heterogeneity, it was decided to limit the analysis to catchments having a 2-year flood included in the range 10-1000 m^3/s, that is 267 over the original 296 basins.

3.2 Identification of balanced cross-validation subsets with SOM clustering of input data

As will be detailed in Section 4, the database is to be divided in three disjoint subsets (called training, cross-validation and test sets) in order to allow a rigorous independent validation and also to increase the generalization abilities of the model when encountering records different from those used in the calibration (or 'training') phase, following an 'early stopping' parameterisation procedure.

The way in which the data are divided may have a strong influence on the performance of the model and it is important that each one of the three sets contains all representative patterns that are included in the dataset. As proposed in the recent literature (Kocjancic and Zupan, 2001; Bowden et al., 2002; Shahin et al., 2004) a self-organising map (SOM) may be applied to this aim. The SOM is a data-driven classification method based on unsupervised artificial neural networks that may be applied for several clustering purposes (for hydrological applications see, for example, Minns and Hall, 2005; Kalteh et al, 2008).

In the recent years, SOMs were also successfully applied for catchments classification either based on geo-morpho-climatic descriptors (Hall and Minns, 1999; Hall et al., 2002; Srinivas et al., 2008; Di Prinzio et al., 2011) or based on hydrological signatures (Chang et al., 2008; Ley et al., 2011; Toth, 2013); however, it is important to underline that the clustering is not carried out here in order to identify a pooling group of similar catchments for developing a region-specific model, but for the optimal division of the available data for the parameterization and independent testing of a single model to be applied over the entire study area.

The SOM is in fact used to cluster similar data records together: an equal number of data records is then sampled from each cluster, ensuring that records from each class (that is catchments with different features) are represented in the training, validation and test sets, that, as a result, have similar statistical properties (Bowden et al., 2002; Shahin et al., 2004).

A SOM (Kohonen, 1997) organizes input data through non-linear techniques depending on their similarity. It is formed by two layers: the input layer contains one node (neuron) for each variable in the data set. The output-layer nodes are connected to every input through adjustable weights, whose values are identified with an iterative training procedure. The relation is of the competitive type, matching each input vector with only one neuron in the output layer, through the comparison of the presented input pattern with each of the SOM neuron weight vectors, on the basis of a distance measure (here the Euclidean one). In the trained (calibrated) SOM, all input vectors that activate the same output node belong to the same class.

In the present application, the dimension of the input layer is equal to three (that is, the first three principal components of the catchments descriptors); as far as the output layer is concerned, there is not a predefined number of classes and, given the small dimension of the input variables, it was here chosen a parsimonious output layer formed by three nodes in a row, each one corresponding to a class.

The three resulting clusters are formed respectively by 121, 70 and 76 catchments; each cluster is then divided into three parts, and one third is assigned to the training, validation and test sets respectively. Overall, the training, validation and test sets are therefore equally numerous (91, 88 and 88 records respectively) and formed by the same proportion of catchments belonging to each of the clusters, having eventually a similar information content, as shown by the similar statistics of the three variables in the three sets represented in Figure 2.

4 Development of symmetric and asymmetric artificial neural networks models for estimating the 2-year return period flows at ungauged sites

4.1 Feedforward Artificial Neural Networks

Artificial neural networks are massively parallel and distributed information processing systems, composed by nodes, arranged in layers, that are able to infer a non-linear input-output relationship. ANN, and in particular feedforward networks have been widely used in many hydrological applications (see for example the recent review papers by Maier et al., 2010 and by Abrahart et al., 2012) and the readers may refer to the abundant literature for details on their characteristics and implementation.

Three different layer types can be distinguished: input layer, connecting the input information, one or more hidden layers, for intermediate computations, and an output layer, producing the final output; adjacent layers are connected through multiplicative weights and, in each node, the sum of weighted inputs and a threshold (called bias) is passed through a non-linear function known as an activation.

The models here applied are networks formed by one hidden layer, with tan-sigmoid activation functions, and a single output node (corresponding to the estimated flood with 2-year return period), with a linear activation function.

The identification of the network's weights and biases (called training procedure) is carried out with a non-linear optimization, searching the minimum of an error (or learning) function measuring the discrepancy between predicted and observed values, and feedforward networks are generally trained with a learning algorithm known as BackPropagation (Rumelhart et al. 1994) based on steepest descent or on more efficient quasi-newton methods.

In order to avoid overfitting, that degrades the generalisation ability of the model, the Early Stopping or Optimal Stopping procedure was applied (see, for example, Coulibaly et al., 2000). For applying Early Stopping, the available data have been divided into three disjoint subsets with a similar information content, as described in Section 3.2: a training set, an early-stopping validation set and a test set. While the network is parameterised minimising the error function on the training set, the error function on the early-stopping validation set is also monitored; if the error function on such second set increases continuously for a specific number of iterations, this is a sign of overfitting of the training set: the training is then stopped and network parameters at the lowest validation error are returned. The third set (test set) is not used in any way during the parameterization phase, but it is used for out-of-sample, independent evaluation of the resulting models.

4.2 Implementation of the symmetric model

Neural networks, including those applied in the recent hydrological literature for the estimation of index floods or flood quantiles at ungauged sites, are traditionally trained minimizing the square error function, which is symmetrical about the y-axes and negative or positive discrepancies of the same magnitude result in the same function value.

In the present work, the results obtained by a network trained with a 'conventional' square error function are compared with those obtained when parameterising the network through the minimisation of an asymmetric loss function, that takes into account both over and underestimation discrepancies but penalizes more the overprediction errors, since the consequences of missing alarms are more severe than those of false alarms.

For both type of models, the output values (2-year flood values) are rescaled as a function of the overall minimum and maximum values to the [-0.95,+0.95] range, to facilitate the optimization algorithms and also avoid saturation problems by accommodating possible extreme values occurring outside the range of available data (Dawson and Wilby, 2001). Each implemented architecture is randomly initialized for ten times to help avoiding local optima: the parameter set that results in the minimum error function on the early stopping validation data (second set) is chosen as the trained network.

The first implemented model is obtained through the minimization of the traditional, symmetric mean squared error, applying the quasi-Newton Levenberg-Marquardt BackPropagation algorithm (Hagan and Menhaj 1994), widely applied and regarded as one of the most efficient neural network training algorithms.

The input variables are the first three principal components of the catchment descriptors, so the input layer is formed by three nodes; the output node corresponds to the estimated flood with 2-year return period; as far as the dimension of the hidden layer is concerned, there is, unfortunately, no definitive established methodology for its determination, because the optimal network architecture is highly problem-dependent: different architectures with a number of hidden nodes varying from 2 to 6 were set up and the mean squared error of the estimates issued for the third, independent set resulted the lowest with the hidden layer formed by 3 nodes.

The architecture with three input nodes, three hidden nodes and 1 output node, represented in Figure 3, is therefore the network finally chosen; the network parameterized minimising the symmetric mean square error function will be denoted as ANN-Symm, and its results will be in Section 5 compared with those of the asymmetric models having the same architecture but a different error function.

4.3 Implementation of asymmetric models with varying degree of asymmetry

The Quad-Quad loss function described in Section 2 is here applied for calibrating the network parameters of the asymmetric models. The learning function to be minimized is therefore the average value of the double quadratic errors (Mean Quad-Quad Error, *MQQE*), obtainable averaging the *M* (number of records in the set) errors given by Eq. (1) when p=2:

$$MQQE = \frac{2}{M}\sum_{j=1}^{M}\left[\alpha - (1-2\alpha)\cdot \mathbf{1}\{\varepsilon > 0\}\right]\cdot \left|\varepsilon_j\right|^2 \qquad (3)$$

The value of α, corresponding to the degree of asymmetry of the loss function, cannot be fixed a priori, since such choice should be based on a location-specific cost-benefit analysis, keeping into account the avoidable losses (that is the direct and indirect losses, provided they may be quantifiable, that may be prevented by mitigation actions following an alarm issue) and the cost of the mitigation measures themselves. Such analysis is acknowledged to be extremely difficult, especially since it involves also intangible costs such as life losses, but also warning credibility issues; furthermore, the costs may change over time and are also dependent on the warning lead-time (see e.g. Martina et al., 2006; Verkade and Werner, 2011, Montesarchio et al., 2011/2014).

For this reason, in the present application, different asymmetric networks, with α varying from 0.4 to 0.1, are implemented, in order to compare the results obtainable with a different asymmetry degree, that is a different extent of importance given to over/underestimation errors. Such asymmetrically trained network are in the following denoted as “Asymm- 0.4”, “Asymm- 0.3”, “Asymm- 0.2”,“Asymm- 0.1”.

The training of the four asymmetric networks, based on the minimisation of the Mean Quad-Quad Error, is carried out through the generalization of the backpropagation algorithm proposed by Crone (2002) and applied by Silva et al. (2010), that may be used for parameterising artificial neural networks with any differentiable (analytically or numerically) error function.

5 Results and discussion

5.1 Goodness-of-fit measures and plots

As described in section 4.2, the neural networks are trained over the standardized (rescaled) output values of the training and cross-validation sets and they are successively used for predicting the output over the independent test set: such ANN output values are then scaled back, obtaining the predictions $Q_{2,p}$.

The performances of the models are evaluated through a set of indexes that describe the prediction error, ε, that is the difference between the de-standardised predictions, $Q_{2,p}$ issued by the models (as a function of morpho-climatic attributes only) and the 'observed' 2-year flood values (the median of historical annual maxima), $Q_{2,o}$, on the third set (test set), formed by N=91 catchments distributed all over the country, whose data have not been used in any capacity in the models' development.

The following error statistics have been computed:

MAE (mean absolute error)

$$MAE=\frac{\sum_{i=1}^{N}\left|\varepsilon(i)\right|}{N} \tag{4}$$

RMSE (root mean square error)

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}\left(\varepsilon(i)\right)^2}{N}} \tag{5}$$

MAE and RMSE both represent a symmetric accuracy, corresponding to the distance of the predictions from the observations independently of the error sign (and the RMSE, being quadratic, emphasizes more the larger errors).

In order to keep into account the differences in sign of the errors, representing the extent of overpredictions as compared to underpredictions, the overall percentage of positive errors (Over%), is computed:

Over% (percentage of overestimates)

$$Over\% = \frac{\{i = 1, \ldots, N | Q_{2,p}(i) > Q_{2,o}(i)\}}{N} \quad (6)$$

Such metric shows the general tendency of the model to overestimate (or to underestimate, since *100- Over%* represents, conversely, the proportion of underpreditions), but these indexes do not distinguish among errors of different magnitude, since they count also predictions that may be only barely above (or below) the targets, that is very good predictions, with minimum errors.

It is therefore computed also the number of the 'high' overestimation errors, keeping into account only the more relevant, and therefore potentially more dangerous, overpredictions. It was here considered as 'high overprediction' an estimate that is more than 30% higher than the corresponding target value:

OverH% (percentage of high overprediction errors)

$$OverH\% = \frac{\{i = 1, \ldots, N | Q_{2,p}(i) > 1.3 \cdot Q_{2,o}(i)\}}{N} \quad (7)$$

The more conservative is the threshold estimate, the lower is the value of OverH%.

On the other hand, even if - as discussed – generally less crucial in terms of consequences, also the number of high underestimation errors should be monitored, since excessively low values imply the tendency of the model to establish thresholds leading to the issuance of too many false alarms.

UnderH% (percentage of high underprediction errors):

$$UnderH\% = \frac{\{i = 1, \ldots, N | Q_{2,p}(i) < 0.7 \cdot Q_{2,o}(i)\}}{N} \quad (8)$$

In addition to the goodness-of-fit measures (reported in Table 2), the boxplot of the errors (predicted minus observed quantiles) is shown in Figure 4: the bottoms and tops of the rectangular boxes are respectively the lower and the upper quartiles, the horizontal segment inside the box is the median and the whiskers represent the 5th and 95th percentiles.

The results may be evaluated also through the scatterplots of predicted (y-axis) vs observed (x-axis) quantiles, presented in Figure 5 that show every prediction $Q_{2,p}$ in respect to the corresponding 'observation' $Q_{2,o}$.

5.2 Discussion of the results

The boxplot (Fig. 4) allows to visually assess both the accuracy and the tendency to over/underestimate of the models: the boxes should be compact and close to the dotted line representing zero error but at the same time it is better if the data lie below such line, thus indicating that the method do not tend to overpredict the thresholds and the warning system is therefore less subject to miss a potentially dangerous flood.

It may be seen that for the network that was trained minimising the traditional Square Error (ANN-Symm) the box and whiskers are centred on the zero-error line and the quantiles (top/bottom of the box, top/bottom whiskers) are at a similar distance from such line, showing that the errors are equally distributed among overestimation and underestimations. The box is compact, demonstrating the good accurateness of the method for a substantial part of the test set, but, due to the symmetric disposition of the errors, many overestimation errors, also remarkably high, are issued, as shown by the position of the upper whisker.

Analysing Table 2, the relatively good accuracy of the ANN-Symm model is demonstrated by the values of the MAE and RMSE, that are the lowest among the implemented models. The symmetric distribution of the overall errors is shown by an Over% close to 50% and the similar values of the OverH% (34%) and UnderH% (32%) confirm that also the high relative errors are equally split among over and underestimates.

Such results were expected since the training is based on a symmetric loss function, but the consequence is that the ANN-Symm model issues a remarkable number of significant overprediction errors, in fact for about one third of the test catchments the estimates are more than 30% higher than the observations.

The analysis of Table 2 shows that the asymmetrically trained networks tend, for decreasing α values, to reduce the number of overestimations (positive errors). For the overall errors this is shown by the different proportion of over/underestimations, that moves from a value that corresponds, approximately, to a balance, to a much more skewed distribution of negative vs positive errors, with Over% decreasing up to 31%.

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

Conversely, as expected, the more asymmetric is the network, the higher are the underprediction errors, as shown by the values of UnderH%: the number of significant negative errors gradually increases from one third up to 47% of the total.

Also the accuracy (given by the total amount of the discrepancies independently of their sign) deteriorates when the asymmetry is more pronounced, but the drop is moderate and the RMSE and MAE values are not so far from those of the ANN-Symm network.

Looking at the parallel boxplots (Fig. 4), it may be seen that with increasing asymmetry the boxes become less compact and, as expected, their position shifts downwards. The length of the upper whiskers substantially decrease with α but the length of the lower whiskers does not increase at the same rate, thus compensating for the fact that the boxes are taller for the more asymmetric models. It follows that the global distances from the 5% to the 95% percentiles (given by the distance between the ends of the top and bottom whiskers) are very close for the symmetric (ANN-Symm) and for the two most asymmetric, thus showing that the variability of the errors for the vast majority (middle 90%) of the data is similar. On the other hand, overall, the errors are moving towards the underestimation side for increasing asymmetry (as confirmed also by the corresponding median values) and for Asymm-01, the upper part of the box indicates that only about one quarter of the errors are overestimations.

It may be noted, in particular from the scatterplots (Fig. 5), that, for both symmetric and asymmetric models, the errors are not negligible: this is due to the shortcomings of the available data set but mainly to the intrinsic limitations of a regional approach applied to the extreme variability of the study area. As already underlined in Section 3.1, the national data set lacks important information that may help to characterise the hydrological behaviour and the phenomena governing formation of extreme flows. In addition to the unavoidable risk of erroneous data, the absence in the database of additional influences certainly further hampers the possibility to obtain a reliable relationship with the flood quantiles. Most importantly, the data set covers the entire Italian peninsula, characterised by extremely different hydro-climatic settings (from Alpine to Mediterranean ones) and this high heterogeneity is certainly an additional reason that limits the performance.

Notwithstanding the limitations of the dataset, that affect equally all the proposed models, the results demonstrate that the use of the double quadratic error function, even if at the expense of more substantial underestimation errors, can substantially decrease the number and extent of overestimation errors, if compared to the use of the traditional square errors.

In the application to a specific cross section, the degree of asymmetry might be identified as proportional to the “risk averseness” of the situation: the more the impact of false alarms is, comparatively, small, the more the decision-makers are reluctant to the consequences (economic and social) of a flood and, rather than risking a missed alarm, can accept many cases of false alarm with the associated costs.

6 Conclusions

A crucial issue in the operation of flood forecasting/warning systems at ungauged locations is how to assess the possible impacts of the forecasted flows, that is the identification of streamflow values that may actually cause flooding, to be associated to thresholds that trigger the issuance of flood watches and warnings. The values that may produce damaging conditions (or “flooding flows”), when in absence of detailed local information on each cross-section, are in many parts of the world estimated as the peak floods having a certain return period, often the 2-year one, that is generally associated with the bankfull discharge.

For locations where the gauges are new or where historical rating curves are not available, the series of past annual flow maxima are absent or very limited, and the peak flow of given frequency to be associated with the watch/warning threshold can be estimated with regionally-derived empirical relationships, such as those that may be applied for the estimation of the index flood at ungauged sites. Such regression-like methods consist in a relation between a set of catchment descriptors that may be obtained also for ungauged sites and the desired flood quantile; linear or power forms are the most commonly used functions, but recent studies have successfully applied artificial neural network models, due to their flexibility, to flood quantile and index flood estimation.

Whatever is the function form, such models are generally parameterised by minimising the mean square error, that assigns equal weight to overprediction or underprediction errors, whereas, instead, the consequences of such errors are extremely different when the estimates are to be used as warning threshold. In fact, false alarms (due to an underprediction of the warning threshold) generally have a much higher level of acceptance than misses (that would derive from an overestimated threshold).

For this reason, in the present work, the regression model (a feed-forward neural network) is parameterised minimising an asymmetric error function (of the double quadratic type), that penalizes more the overestimation than the underestimation discrepancies. The predictions of models with increasing degree of asymmetry are compared with those of a traditional (trained on the symmetric mean of square errors) neural network, in a rigorous cross-validation experiment referred to a database of catchments covering all the Italian country.

The results confirm, as expected, that the more asymmetric is the network, the more numerous and higher are the underprediction errors, and the less numerous and less severe are the overestimation errors. As also expectable, the symmetric accuracy decreases when the asymmetry is more pronounced, but the drop is moderate and the RMSE and MAE values are not so far from those of the traditionally trained network.

Undoubtedly, the nature of the regional approach, as well as the shortcomings of the dataset and the extreme heterogeneity of the study area, generate errors much greater than those obtainable with detailed local studies. On the other hand, where no alternatives exist, the proposed methodology may provide a preliminary estimate of the threshold runoff that do not, ~~prudentially~~, overestimate the actual flooding flow.

Notwithstanding the acknowledged limitations of the dataset, that affect equally all the proposed models, the analysis shows that the use of the asymmetric error function substantially reduces the number and extent of overestimation errors, if compared to the use of the traditional square errors. Of course such reduction is at the expense of increasing underestimation errors, but the overall precision is still acceptable and the study highlights the potential benefit of choosing an asymmetric error function when the consequences of missed alarms are more severe than those of false alarms.

Minimising the asymmetric error function has the purpose of optimizing the threshold from an operational point of view, in a deterministic framework: future analyses may be devoted to investigate the uncertainty of the issued predictions, since a probabilistic approach (provided that the methodology is able to include all sources of uncertainty and its quality may be objectively assessed) may provide very valuable insights for a more complete evaluation of the model, supplementing the information provided by point-value predictions.

It is important to highlight that the asymmetric error function is used, in this study, to parameterise a neural network, but of course it might be used to optimize any other model or equation, when aiming at obtaining conservative estimates, for safety reasons.

The appropriate degree of asymmetry might be identified depending on the risk-averseness of the specific flood-prone context. The quantification of risk aversion is extremely difficult and case-specific: it should keep into account that the perception of society may be very different from a technical appraisal of the involved costs and it should include also indirect, intangible and long-term impacts. More research on the societal perception in different contexts would greatly improve the process of risk-based decision-making (Merz et al., 2009), including the choices concerning flood-warning thresholds. Hopefully, in the next years, a more direct collaboration between the hydrologic and socio-economic research communities, as auspicated in the new Panta Rhei science initiative (Montanari et al., 2013; Javelle et al., 2014), will provide a progress in this direction.

Acknowledgements

The author thanks Monica Di Prinzio and Attilio Castellarin for the elaboration of the dataset carried out in 2011 within the Italian National Programme CUBIST.

The present work was developed within the framework of the Panta Rhei Research Initiative of the International Association of Hydrological Sciences (IAHS), Working Group on “Data-driven hydrology”.

References

Abrahart, R.J., Anctil, F., Coulibaly, P., Dawson, C.W., Mount, N.J., See, L.M., Shamseldin, A.Y., Solomatine, D.P., Toth, E., and Wilby, R.L.: Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting, Progress in Physical Geography, 36(4), 480–513, doi:10.1177/0309133312444943, 2012.

Archfield, S.A., Pugliese, A., Castellarin, A., Skøien, J.O., and Kiang, J.E.: Topological and canonical kriging for design flood prediction in ungauged catchments: an improvement over a traditional regional regression approach?, Hydrol. Earth Syst. Sci., 17, 1575-1588, 2013.

Aziz, K., Rahman, A., Fang, G., and Shreshtha, S.: Application of Artificial Neural Networks in Regional Flood Frequency Analysis: A Case Study for Australia, Stochastic Environment Research & Risk Assessment. 28, 541-554, DOI 10.1007/s00477-013-0771-5, 2013.

Bloeschl, G., Sivapalan M., Wagener, T, Viglione A. and Savenije, H. (Eds): Runoff prediction in ungauged basins: Synthesis across processes, places and scales, Cambridge University Press, New York, USA, 2013.

Bocchiola, D., De Michele, C., and Rosso, R.: Review of recent advances in index flood estimation, Hydrol. Earth Syst. Sci., 7, 83–296, doi:10.5194/hess-7-283-2003, 2003.

Bowden, G. J., Maier, H. R., and Dandy, G. C.: Optimal division of data for neural network models in water resources applications, Water Resour. Res., 38(2), 1010, 10.1029/2001WR000266, 2002.

Brath, A., Castellarin, A., Franchini, M., and Galeati, G.: Estimating the index flood using indirect methods, Hydrological Sciences Journal, 46(3), 399-418, 2001.

Carpenter, T. M., Sperfslage J. A., Georgakakos K. P., Sweeney T., and Fread D. L.: National threshold runoff estimation utilizing GIS in support of operational flash flood warning systems, J. Hydrol., 224, 21–44, 1999.

Chang, F. J., Tsai, M. J., Tsai, W. P., and Herricks, E. E.: Assessing the Ecological Hydrology of Natural Flow Conditions in Taiwan, J. Hydrol., 354, 75–89, 2008.

Christoffersen P.F., and Diebold F.X.: Further results on forecasting and model selection under asymmetric loss, J App Econom, 11, 561–571, 1996.

Claps and the CUBIST Team, Development of an Information System of the Italian basins for the CUBIST project, Geophysical Research Abstracts, 10, EGU2008-A-12048, 2008.

Coulibaly, P., Anctil, F., and Bobee, B.: Daily reservoir inflow forecasting using artificial neural networks with stopped Training Approach, J. Hydrol., 230, 244–257, 2000.

Crone, S.F.: Training Artificial Neural Networks using Asymmetric Cost Functions, in: Proceedings of the 9th International Conference on Neural Infomation Processing (ICONIP'OZ) , 18-22 november 2002, Singapore, Vol. 5, IEEE, Singapore, 2374- 2380, 2002.

Cunha, L. K., Krajewski W.F., and Mantilla R.: A framework for flood risk assessment under nonstationary conditions or in the absence of historical data, J. Flood Risk Manage., 4(1), 3–22, 2011.

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

Dalrymple, T.: Flood frequency analyses., Water Supply Paper 1543-A. U.S. Geological Survey, Reston, Virginia, USA, 1960.

Daňhelka, J., and Vlasák, T: Evaluation of Real-time Flood Forecasts in the Czech Republic, 2002-2012, Czech Hydrometeorological Institute Report, last access date: 17 June 2015, 2013. (http://www.chmi.cz/files/portal/docs/poboc/CB/pruvodce/vyhodnoceni_en.html)

Dawson, C.W., Abrahart, R.J., Shamseldin, A.Y., and Wilby R.L.: Flood estimation at ungauged sites using artificial neural networks, J Hydrol., 319, 391–409, 2006.

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

Di Prinzio, M., Castellarin,A., and E. Toth E.: Data-driven catchment classification: application to the PUB problem, Hydrology and Earth System Sciences, 15, 1921-1935, 2011.

Diebold, F. X., and Lopez, J. A.: Forecast Evaluation and Combination, in G. S. Maddala and C. R. Rao (Eds.), Handbook of Statistics, Vol. 14, Amsterdam, 241-268, 1996.

Elliott, G., Komunjer, I., and Timmermann, A.: Estimation and Testing of Forecast Rationality under Flexible Loss, Review of Economic Studies, 72(4), 1107-1125, 2005.

Granger, C. W. J.: Outline of Forecast Theory Using Generalized Cost Functions, Spanish Economic Review, 1, 161-173, 1999.

Granger, C. W. J., and Pesaran, M. H.: A Decision Theoretic Approach to Forecast Evaluation, in W. S. Chan, W. K. Li and H. Tong (Eds.) Statistics and Finance: An Interface, Imperial College Press, London, 261-278, 2000.

GREHYS (Groupe de recherche en hydrologie statistique): Presentation and review of some methods for regional flood frequency analysis, J. Hydrol., 186, 63–84, 1996.

Griffis, V. W., and Stedinger, J. R.: The use of GLS regression in regional hydrologic analyses, J. Hydrol., 344(1–2), 82–95, 2007.

Hagan, M. T., and Menhaj, M. : Training feedforward networks with the Marquardt algorithm, IEEE T.Neural Networ. 5(6), 989-993, 1994.

Hall, M.J., and Minns A.W.: The classification of hydrologically homogeneous regions, Hydrol. Sci. J. 44(5), 693–704, 1999.

Hall, M.J., Minns, A.W., and Ashrafuzzaman, A.K.M.: The application of data mining techniques for the regionalization of hydrological variables, Hydrology and Earth System Sciences, 6(4), 685–694, 2002.

Hapuarachchi, H. A. P.,Wang, Q. J., and Pagano, T. C.: A review of advances in flash flood forecasting, Hydrol. Process., 25, 2771–2784, 2011.

Harman C., Stewardson M., and DeRose R.: Variability and uncertainty in reach bankfull hydraulic geometry, J Hydrol., 351, (1–2), 13–25, 2008.

Javelle, P., Demargne, J., Defrance, D., Pansu, J., and Arnaud, P.: Evaluating flash flood warnings at ungauged locations using post-event surveys: a case study with the AIGA warning system. Hydrological Sciences Journal, 59 (7), 1390–1402, 2014.

Kalteh, A.M., Hjorth, P., and R. Berndtsson, R.: Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application, Environmental Modelling & Software, 23, 835-845, 2008.

Kjeldsen, T. R., Smithers, J. C., and Schulze, R. E.: Flood frequency analysis at ungauged sites in the KwaZulu-Natal Province, South Africa, Water S.A., 27(3), 315-324, 2001.

Kjeldsen, T. R., Jones, D. A., and Morris, D. G.: Using multiple donor sites for enhanced flood estimation in ungauged catchments, Water Resources Research, 50 (8), 6646–6657, 2014.

Kocjancic, R., and Zupan, J.: Modeling of the river flowrate: the influence of the training set selection, Chemom. Intell. Lab. Syst., 54, 21–34, 2000.

Kohonen, T.: Self-Organizing Maps, 2nd ed., Springer, Berlin, ISBN 3-540-62017-6, 1997.

Ley, R., Casper, M. C., Hellebrand, H., and Merz, R.: Catchment classification by runoff behaviour with self-organizing maps (SOM), Hydrol. Earth Syst. Sci., 15, 2947–2962, doi:10.5194/hess-15-2947-2011, 2011.

Maier, H.R., Jain A., Dandy, G.C., and Sudheer, K.P.: Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions, Environmental Modelling & Software, 25(8), 891-909, DOI: 10.1016/j.envsoft.2010.02.003, 2010.

Martina M.L.V., Todini E., and Libralon A.: A Bayesian decision approach to rainfall thresholds based flood warning, Hydrology and Earth System Sciences, 10, 1-14, 2006

Merz, B., Elmer, F., and Thieken, A. H.: Significance of “high probability/low damage” versus “low probability/high damage” flood events, Nat. Hazards Earth Syst. Sci., 9, 1033–1046, doi:10.5194/nhess-9-1033-2009, 2009.

Merz, R., and Bloschl, G.: Flood frequency regionalisation—Spatial proximity vs. catchment attributes, J. Hydrol., 302, 283–306, 2005.

Minns A.W., and Hall M.J.: Artificial neural network concepts in hydrology. In Encyclopedia of Hydrological Sciences, Anderson MG, McDonnell JJ (eds). John Wiley and Sons: Chichester, UK, pp. 307-320, 2005.

Montanari, A., Young, G., Savenije, H.H.G., Hughes, D., Wagener, T., Ren, L.L., Koutsoyiannis, D., Cudennec, C., Toth, E., Grimaldi, S., Bloschl, G., Sivapalan, M., Beven, K., Gupta, H., Hipsey, M., Schaefli, B., Arheimer, B., Boegh, E., Schymanski, S.J., Di Baldassarre, G., Yu, B., Hubert, P., Huang, Y., Schumann, A., Post, D.A., Srinivasan, V., Harman, C., Thompson, S., Rogger, M., Viglione, A., McMillan, H., Characklis, G., Pang, Z., and Belyaev, V: Panta Rhei-Everything Flows: Change in hydrology and society-The IAHS Scientific Decade 2013-2022, Hydrological Sciences Journal, 58(6), 1256-1275, 2013.

Montesarchio, V., Ridolfi, E., Russo, F., and Napolitano, F.: Rainfall threshold definition using an entropy decision approach and radar data, Nat. Hazards Earth Syst. Sci., 11, 2061-2074, doi:10.5194/nhess-11-2061-2011, 2011.

Muttiah R.S., Srinivasan R., and Allen, P.M.: Prediction of two year peak stream discharges using neural networks, J Am Water Resour Assoc, 33(3), 625–630, 1997.

Norbiato, D., Borga, M., and Dinale, R.: Flash flood warning in ungauged basins by use of the flash flood guidance and model-based runoff thresholds, Meteorol. Appl., 16, 65–75, doi:10.1002/met.126, 2009.

Ntelekos, A. A., Georgakakos, K.P., and Krajewski, W.F.: On the uncertainties of flash flood guidance: Towards probabilistic forecasting of flash floods, J. Hydrometeorol., 7(5), 896–915, doi:10.1175/JHM529.1, 2006.

Pandey, G.R., and Nguyen, V-T-V.: A comparative study of regression based methods in regional flood frequency analysis, Journal of Hydrology, 225, 92-101, 1999.

Pappenberger, F., Bartholmes, J., Thielen, J., Cloke, H., Buizza, R., and de Roo, A.: New dimensions in early flood warning across the globe using grand-ensemble weather predictions, Geophys. Res. Lett., 35, L10404, doi:10.1029/2008GL033837, 2008.

Patton, A.J., and Timmermann. A.: Properties of Optimal Forecasts under Asymmetric Loss and Nonlinearity, Journal of Econometrics, 140(2), 884–918, 2007.

Reed, S., Schaake, J., and Zhang, Z.: A distributed hydrologic model and threshold frequency based method for flash flood forecasting at ungauged locations, J. Hydrol., 337(3–4), 402–420, 2007.

Rumelhart, D.E., Widrow, B, ans Lehr, M.A.: The basic ideas in neural networks. Communications of the ACM, 37(3), 87-92, 1994.

Salinas, J.LO., Laaha, G.; Rogger, M., Parajka, A., Viglione, A., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins – Part 2. Flood and low flow studies, Hydrol. Earth Syst. Sci., 17, 2637–2652, 2013.

Sene, K.: Flash floods: forecasting and warning, Springer, Dordrecht, 385 p. ISBN: 9789400751637, 2013.

Shahin, M., Maier, H., and Jaksa, M. : Data Division for Developing Neural Networks Applied to Geotechnical Engineering, J. Comput. Civ. Eng., 18(2), 105–114, 2004.

Shu C., and Burn, D.H.:Artificial neural network ensembles and their application in pooled flood frequency analysis. Water Resour. Res 40(9):W09301. doi:10.1029/2003WR002816, 2004.

Shu C., and Ouarda T.B.M.J.: Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system, J Hydrol, 349, 31–43, 2008.

Silva, D.G.E., Jino, M., and de Abreu, B.T.: Machine learning methods and asymmetric cost function to estimate execution effort of software testing, in: Proc. Third International Conference on Software Testing, Verification and Validation (ICST), Paris, 7-9 April 2010, IEEE, New York, 275-284, 2010.

Simor V., Hlavcova K., Silvia Kohnova S., and Szolgay J.: Application of Artificial Neural Networks for estimating index floods, Contributions to Geophysics and Geodesy, 42/4, 295–311, 2012.

Singh,·K.K., Pal, M. and Singh V.P.: Estimation of Mean Annual Flood in Indian Catchments Using Backpropagation Neural Network and M5 Model Tree, Water Resour Manage, 24, 2007–2019, DOI 10.1007/s11269-009-9535-x, 2010.

Smith, A., Sampson, C., and P. Bates, P.: Regional flood frequency analysis at the global scale, Water Resour. Res., 51, 539–553, doi:10.1002/2014WR015814, 2015.

Srinivas, V.V., Tripathi, S., Rao, A.R., and Govindaraju R.S.: Regional flood frequency analysis by combining self-organizing feature maps and fuzzy clustering, J. Hydrol., 348(1–2), 148–166, 2008.

Stedinger, J. R., and Lu, L.: Appraisal of regional and index flood quantile estimators, Stochastic Hydrol. Hydraul., 9(1), 49–75, 1995.

Stedinger, J. R., and Tasker, G. D.: Regional hydrologic analysis 1. Ordinary, weighted, and generalized least squares compared, Water Resour. Res., 21(9), 1421–1432, 1985.

Toth, E.: Catchment classification based on characterisation of streamflow and precipitation time series, Hydrology and Earth System Sciences, 17, 1149-1159, doi:10.5194/hess-17-1149-2013, 2013.

Toth, E.: Asymmetric Error Functions for Reducing the Underestimation of Local Scour around Bridge Piers: Application to Neural Networks Models., J. Hydraul. Eng., 141(7), 04015011, doi: 10.1061/(ASCE)HY.1943-7900.0000981, 2015.

UCAR (University Corporation for Atmospheric Research): Flash Flood Early Warning System Reference Guide 2010, ISBN 978-0-615-37421-5, last access date: 17 June 2015, 2010 (http://www.meted.ucar.edu/communities/hazwarnsys/ffewsrg/FF_EWS.pdf).

Verkade, J. S., and Werner, M. G. F.: Estimating the benefits of single value and probability forecasting for flood warning, Hydrol. Earth Syst. Sci., 15, 3751–3765, doi:10.5194/hess-15-3751-2011, 2011.

Ward, P.J., Jongman, B., Sperna Weiland, F.C., Bouwman, A., van Beek, R., Bierkens, M.F.P., Ligtvoet., W., and Winsemius, H.C.: Assessing flood risk at the global scale: Model setup, results, and sensitivity, Environmental Research Letters 8, 44019, doi:10.1088/1748-9326/8/4/044019, 2013.

Wilkerson G.V.: Improved bankfull discharge prediction using 2-year recurrence-period discharge, J Am Water Resour Assoc., 44(1), 243–258, DOI: 10.1111/j.1752-1688.2007.00151.x, 2008.

WMO, Manual on flood forecasting and warning, WMO Series No. 1072, 142 pp, ISBN: 978-92-631-1072-5, last access date: 17 June 2015, 2011. (http://www.wmo.int/pages/prog/hwrp/publications.php)

Tables

Table 1. Geomorphological and climatic descriptors of the CUBIST database of Italian catchments

1	Long - UTM longitude of catchment centroid
2	Lat - UTM latitude of catchment centroid
3	A - Catchment drainage area
4	P - Catchment perimeter
5	zmax - Maximum elevation of the catchment area
6	zmin - Elevation of the catchment outlet
7	zmean - Mean elevation of the catchment area
8	L - Length of the Maximum Drainage Path
9	SL - Average slope along the Maximum Drainage Path
10	SA - Catchment average slope
11	Φ - Catchment orientation
12	MAP - Mean Annual Precipitation

Table 2. Goodness-of-fit criteria of the 2-year floods estimates obtained by the symmetric and asymmetric networks on the independent test set of catchments.

Model\Index	*MAE (m^3/s)*	*RMSE(m^3/s)*	*Over%*	*OverH%*	*UnderH%*
Symm	98	133	46%	34%	32%
Asymm-04	104	147	42%	32%	35%
Asymm-03	105	152	41%	30%	37%
Asymm-02	108	162	36%	27%	41%
Asymm-01	115	178	31%	18%	47%

Figure Captions

Figure 1. Asymmetric Quad-Quad loss function (with α varying from 0.1 to 0.9) compared with the Squared Error (SE).

Figure 2: 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).

Figure 3. Architecture of the chosen network, with three input nodes, three hidden nodes and 1 output node.

Figure 4. Parallel box-plots of the errors (ε=Q2,o- Q2,p) of the 2-year floods estimates obtained by symmetric and asymmetric networks on the independent test set of catchments.

Figure 5. Scatterplots of the predicted (y-axis) vs observed (x-axis) 2-year floods estimates on the independent test set of catchments, for the symmetric and asymmetric models.

Figure 1. Asymmetric Quad-Quad loss function (with α varying from 0.1 to 0.9) compared with the Squared Error (SE).

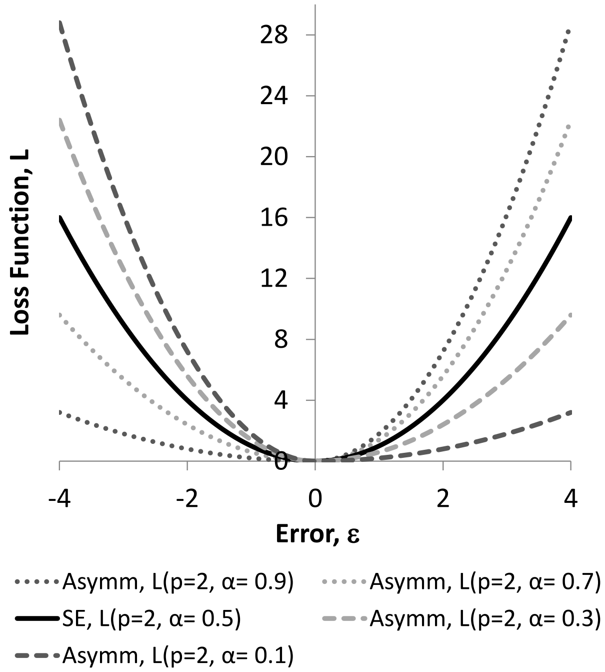


Figure 2: Mean value (red dash) and the bars comprised between the 90% and 10% percentiles of the resulting training, cross-validation and testing sets, for each of the three input variable (PC1, PC2 and PC3).

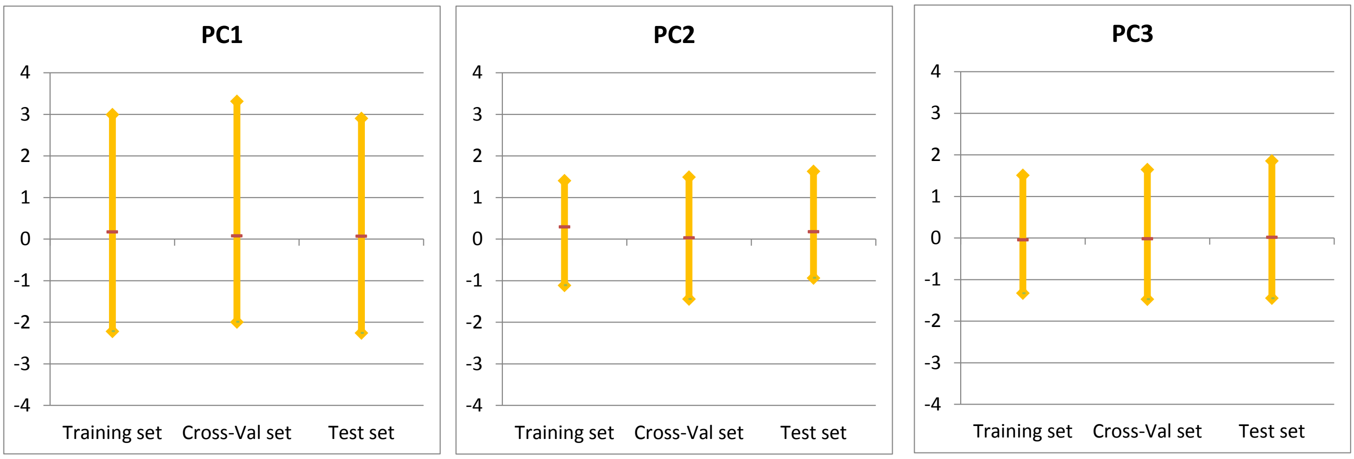


Figure 3. Architecture of the chosen network, with three input nodes, three hidden nodes and 1 output node.

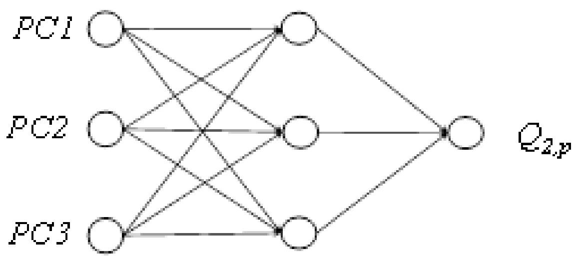


Figure 4. Parallel box-plots of the errors (ε=Q2,o- Q2,p) of the 2-year floods estimates obtained by symmetric and asymmetric networks on the independent test set of catchments.

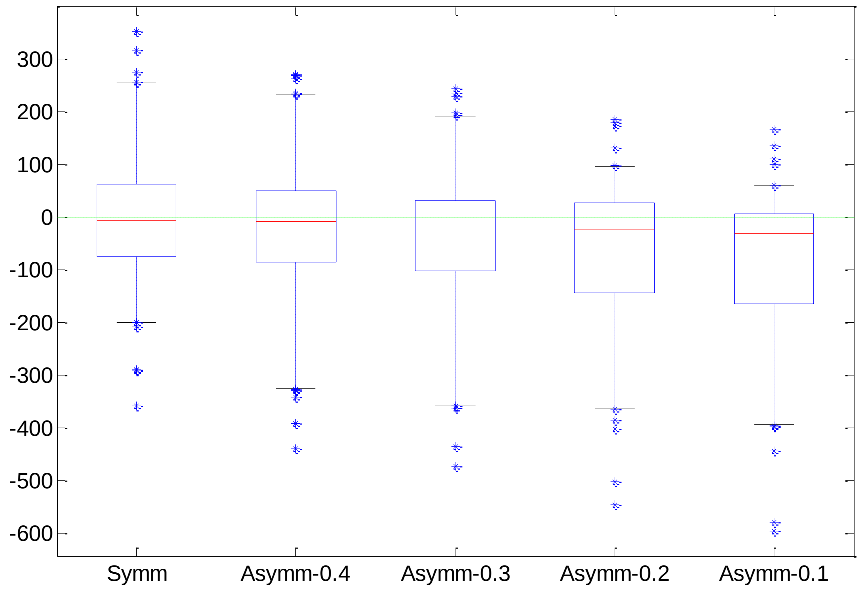


Figure 5. Scatterplots of the predicted (y-axis) vs observed (x-axis) 2-year floods estimates on the independent test set of catchments, for the symmetric and asymmetric models.