Dear Editor,

please find attached the new version of the manuscript “Estimation of flood warning runoff thresholds in ungauged basins with asymmetric error functions”, which I am submitting following the last comments received from the Editor and the Referee.

I would like to warmly thank the Reviewer and the Editor, for taking time in reading again and suggesting improvements to the paper.

I report below the Editor’s and Referee’s comments, in Italic and blue font, followed by my replies.

In the revised manuscript I highlighted in green the changes with respect to the previously revised version.

Thank you very much for your kind consideration of this revised version of our manuscript.
Sincerely yours,
Elena Toth

Editor’s Comments

The paper has now gone out for re-review after the first round of comments and revisions. The reviewer found that the manuscript is suitable for publication subject to minor revision. I agree with this assessment and I would now request that the author revise or respond to the minor revisions noted by the review.

I would also ask the author add labels to the axes in the figures and explain any abbreviations in the table and figure captions. In particular, figures 4 and 5 are missing labels on the figures.

Thank you for having highlighted that I uploaded the scatterplots without the labels identifying the models and also missing the axis titles: I have added them in the revised Figure 5.

In Figure 4 I have added the title of the y-axis.

The abbreviations are reported in the text, but of course I may repeat them in the captions.

Referee’s Comment

I thank the author for taking into consideration the comments I made on the manuscript and coming up with an updated version. Although I may not fully agree with the replies given to some of the issues I raised, I feel that the manuscript is just fine given that its objective is more on demonstrating an approach for estimating a given flood level. Nevertheless, I would like to make the following comments and suggestions:

I do thank the Referee: indeed the data set is certainly not the best one and also part of the analysis is based on subjective choices, but indeed, as the Referee highlights, the main objective is mainly to demonstrate the feasibility of the proposed approach and much more information would be needed on the study catchments in order to provide reliable estimates of the actual thresholds.

• In my view, the suggestion in Cunnane (1987) not to extrapolate statistical inference beyond a return period of 2 times the sample length does not mean that one can use very short records for inference as long as one keeps the return period within 2 times the sample length. I would remove that part of the text.
At p. 8, lines 7-10, I have changed:

“Even the shortest records (and actually only 9 of the locations have less than 8 years 4 of data) should be sufficient for such a short return period, for example according to the 5 classical guideline by Cunnane (1987), that suggests not to extrapolate statistical inference 6 beyond a return period of 2 times the sample length.”

in

“Even if of course it would be preferable having longer time-series, only 9 of the locations have actually less than 8 years of data and the records are deemed to be sufficient for such a short return period.”

• It is fine to use three classes of catchment descriptors to sample representative catchments from for the three groups. However, I find the reasoning given for the choice of three classes less than convincing. Is it the dimension of the input variables or the variability that should be considered for deciding the number of classes?

The choice is indeed totally subjective; I first chose a parsimonious model and then verified that the resulting sets were not too dissimilar in terms of information content (a different choice could certainly have been made, but, again, the main objective of the paper is to show the possible approach and not estimating actual flood values).

I have removed the reference to input dimension in the paper at p. 9, ll. 21-22.

• I find the MAE and RMSE as defined in the manuscript (Equations 4 and 5, respectively) not so robust as goodness-of-fit measures for the particular study presented in the manuscript. The q2 values vary between 10 – 1000m3/s (page 8, line 10). Given this wide range of the magnitudes at the stations, the measures can be misleading unless the errors are brought to a similar scale.

I do agree that, in principle, the use of relative errors might be a more proper way to both set up (training) and evaluating the performances of this kind of models, but, since in the present application the costs are weighted in respect to the ‘not-relative’ errors in the loss functions, I believe that showing the total error, plus the scatterplot may suffice. In a future study, it would be interesting to use a (asymmetric) cost function based on relative errors and then, consistently, analyse the results in the same way.
Estimation of flood warning runoff thresholds in ungauged basins with asymmetric error functions

E. Toth

[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 et 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 $\varepsilon$ as the prediction minus the observed value (that is, a negative error corresponds to underestimation, a positive one to overestimation), reads:

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

(1)

where $\mathbf{1}(\cdot)$ 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:

$$L(2,0.5) = \varepsilon^2$$

(2)

Figure 1 shows the asymmetric Quad-Quad loss function (with $\alpha$ 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 1a. 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 it was not possible to obtain homogeneous information on these characteristics nationwide and 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 database, 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 if of course it would be preferable having longer time-series, only 9 of the locations have actually less than 8 years of data and the records are deemed to be sufficient for such a short return period.

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 m³/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; 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.
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 \( p=2 \):

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

The value of \( \alpha \), 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 \( \alpha \) 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 therefore evaluated through a set of indexes that describe the prediction error, $\varepsilon$, that is the difference between 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)

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\varepsilon(i)|,$$

RMSE (root mean square error)

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\varepsilon(i))^2}.$$

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

Such metric shows the general tendency of the model to overestimate (or to underestimate, since 100- \(Over\%\) represents, conversely, the proportion of underpredictions), 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:

\[
OverH\% (\text{percentage of high overprediction errors})
\]

\[
OverH\% = \frac{\{i=1, \ldots, N | Q_{2,p}(i) > 1.3Q_{2,o}(i)\}}{N}
\] (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.

\[
UnderH\% (\text{percentage of high underprediction errors})
\]

\[
UnderH\% = \frac{\{i=1, \ldots, N | Q_{2,p}(i) < 0.7Q_{2,o}(i)\}}{N}
\] (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, shown in Figure 5 that shows 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) In Fig.4, 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, again 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 $\alpha$ 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 positive vs
negative 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 \( \alpha \), with OverH% reaching a value that is much lower
than that of the ANN-Symm model when \( \alpha \) 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 the boxes become less compact
and, as expected, their position shifts downwards with increasing asymmetry. The length of
the upper whiskers substantially decrease with \( \alpha \) 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 networks, 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-04, 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


Tables

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

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

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.

<table>
<thead>
<tr>
<th>Model\Index</th>
<th>MAE ($m^3/s$)</th>
<th>RMSE ($m^3/s$)</th>
<th>Over%</th>
<th>OverH%</th>
<th>UnderH%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symm</td>
<td>98</td>
<td>133</td>
<td>46%</td>
<td>34%</td>
<td>32%</td>
</tr>
<tr>
<td>Asymm-04</td>
<td>104</td>
<td>147</td>
<td>42%</td>
<td>32%</td>
<td>35%</td>
</tr>
<tr>
<td>Asymm-03</td>
<td>105</td>
<td>152</td>
<td>41%</td>
<td>30%</td>
<td>37%</td>
</tr>
<tr>
<td>Asymm-02</td>
<td>108</td>
<td>162</td>
<td>36%</td>
<td>27%</td>
<td>41%</td>
</tr>
<tr>
<td>Asymm-01</td>
<td>115</td>
<td>178</td>
<td>31%</td>
<td>18%</td>
<td>47%</td>
</tr>
</tbody>
</table>
Figure Captions

Figure 1. Asymmetric Quad-Quad loss function (with $\alpha$ 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 ($\varepsilon$=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 $\alpha$ 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 ($\varepsilon = Q_{2,o} - Q_{2,p}$) of the 2-year floods estimates obtained by symmetric and asymmetric networks on the independent test set of catchments.

(added title of y-axis)
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.

(added title of x- and y-axis)