Reply to Prof. Dr. Solomatine

I wish to deeply thank Professor Solomatine for his review: his comments are particularly valuable given his deep knowledge of the subject of the paper. I will consider with particular care any additional suggestion that he may offer.

The Reviewer's Comments are in blue, boldface font, and the author's responses are in black, plain type.

Specific comments

1) Typically Kohonen network is profiled as a clustering method, and not a classification one. Classification results in a learned machine that is able to attribute new data to one of the existing classes. The training data for such machine should include the known (observed) output, since this is a supervised learning method. On the contrary, clustering methods belong to the group of non-supervised learning methods since the output is not known. (Often, the clusters found by a clustering method can be interpreted as classes, data is labeled accordingly, and then a classifier is trained – but this was not done in the paper.) The task solved by Kohonen network in the paper is a clustering task. It is suggested to consider using the term "clustering" or "grouping" instead of "classification". Interestingly, such use of terms can be found in other publications as well; for example, this reviewer has made exactly the same comment on the paper "Clustering of heterogeneous precipitation fields for the assessment and possible improvement of lumped neural network models for streamflow forecasts" by N. Lauzon, F. Anctil, and C. W. Baxter, Hydrol. Earth Syst. Sci., 10, 485–494, 2006 (which originally had "classification" in the discussion paper version.

2) The reader may ask, if SOM is a clustering method, how (page 910) it could be used to perform classification for the new data vectors to attribute them to a particular cluster/model. Was a separate classifier trained? What method was used for classification? The author is invited to clarify this.

The weigths of the nodes of the SOM are adjusted, through the learning process described in Section 5.1 (p. 906 L. 13 -> p. 907 L. 4), on the vectors of the calibration set. In the learning process all the calibration input vectors are processed through the SOM incrementally, one after the other, re-iteratively: for each sample input vector **x**, the weigths of the winner node and of the nodes in its neighborhood are changed closer to **x** (in the input data space). During the learning process, individual changes may be contradictory, but the net outcome in the process is that ordered values for the **W** emerge over the array.

At the end of the learning phase, the weigths have reached a final, tuned value and the SOM may be used (without changing the weigths any more) to classify the calibration vectors: for each vector the best-matching (that is the closest) unit in the SOM output layer is identified. In exactly the same way, the tuned SOM may used to associate any new vector to one of the units of the SOM output layer, thus attributing the new data to the clusters identified before.

In this way, maybe the SOM may be considered a classification method, even if unsupervised? Given the strong experience and deep knowledge of Dr. Solomatine on such themes, I will, of course, follow any suggestion he will provide on the denomination option.

The use of the SOM for the classification of the validation data (data that were not used during the tuning of the weights of the SOM nodes) is explained, probably too quickly, in Section 6, from l. 26 of p. 909 to l. 4 of p. 910, which is, indeed, too late: in the revised version it will be further detailed and anticipated in Section 5, where the classification is described.

3) It would have helped, if the ways the data is partitioned into the training, crossvalidation (if any) and validation sets had been presented. It is not specified what is the total size of the data set.

No cross-validation data set was used since, to mitigate overfitting and to improve generalization, a Bayesian regularization of the learning function (Foresee and Hagan, 1997; Anctil et al., 2004) is instead used. The third (cross-validation) data set is in fact generally used for an early stopping technique in the training phase, for generalisation purposes. The regularisation approach, instead, takes into account the goodness-of-fit as well as the network architecture parsimony, modifying the objective function with the addition of a term which is the sum of the squares of the network weights, offering an intrinsic regularization procedure providing a viable alternative to early stopping technique.

The partition of the data in the presented application is simply the chronological order of the subsequent hydrological years (beginning at the end of the dry season, that is at the end of August): the first three entire hydrological years were

used in calibration, the last part of the observation period, including the last entire hydrological year, 1994-1995, plus the last four months (from September to December 1995) were used in validation. This kind of subdivision is typical in hydrological studies, where, for allowing the plotting of a continuous hydrograph, all the examples are contiguous in time (Corzo and Solomatine, 2007) and such division corresponds, by the way, to the approach that would be used in any real-world application, where the calibration is carried out over all the data available up to the moment in which the model is built, and it is operatively used for forecasting in the following period.

As required also by Dr. Ehret, a brief description of the watershed hydro-meteorological regime and of the calibration and validation data, including their lengths (equal to 26280 hourly data in calibration and 11688 hourly data in validation) will be added in the revised version.

4) The author uses the following logic: build 9 models – analyse the multi-model performance on the test (validation) set – reduce number of modules from 9 to 4 – analyse the results again. This means that the test set is used in building the model (deciding on the number of its modules) which is methodologically questionable. Indeed, sometimes, in the situations of data shortage this is done, but in the considered case, as I understand, there is no data shortage, so a separate, cross-validation set could be used. The author is invited to present the justification for this approach.

Actually, the two modular approaches with 9 and 4 modules are two different models, to be compared, with equal dignity, to the global model and to the persistent one and should not be considered parameterisations of the same model. The validation phase is thoroughly independent because the validation data were not used for training the models.

The number of the modules is not a parameter of the model: a full and exhaustive analysis of the performances obtainable with all the possible numbers of modules may be interesting (even if not easy to carry out) but it is not performed here, as it is not performed in the great majority of the works on multi-networks modelling presented in the literature (and cited in the paper), which consider only one partition (e.g. 2 clusters for the K-means clustering in Corzo and Solomatine, 2007, 64 clusters in Abrahart and See, 2002) or, like in Jain and Srinivasulu (2007), two possible partitions (3 and 4 clusters respectively).

I have here presented two different modular models, based on two partitions, but the second one is in reality a reasoned interpretation of the first one: the choice of 9 nodes in the original classification is certainly subjective, since no known number of different hydro-meteorological situations may be fixed (but a number had to be picked up…). I believe such number to be a good trade-off between parsimony and a sufficiently wide range of different conditions, therefore allowing a satisfactory identification of the hydro-meteorological situations, and this seems confirmed by the analysis of the obtained classes in respect to the hydro-meteorological conditions as described in Section 5.

[By the way, also a SOM with 4 nodes was analysed in the course of the research progress (as described in the reply to the Anon. Ref. #2), but the performances in validation were slightly worse than those of the second modular approach, because the classes resulted not sufficiently well-identified) and, being the paper already pretty long, I did not include such description in the original manuscript].

5) In data-driven rainfall-runoff (RR) modelling input selection can be performed by using "physical" approach, for example determining the lags through studying the travel time through the catchment, or by studying the correlations and mutual information between lagged rainfalls and flows. Unfortunately, it is not clear how the lags for rainfall (3) and flows (4) were chosen for the case study considered? (page 905) It is recommended at least to mention such possibilities.

I will certainly cite the possibility to use "physical" indications (based on travel times and/or correlations) for input determination in the revised paper, along with the relative references.

The choice of the input variables and also the choice of the number of hidden nodes (see comment 8), is here based on the results obtained in previous works on the same watershed (as cited at ll. 1-5 at p. 903) where an extensive analysis of the results obtained with different input and hidden nodes was carried out. I did not report the (long and tedious) tables of the obtained results because the selection of the optimal ANN architecture was not the direct objective of the present paper, that focuses instead on the differences between the global and modular networks, considering always the same architecture.

6) In the model on page 905, out of 7 inputs, 4 inputs represent a very strong autocorrelation component (flow). The problem with such models is of course that the forecasted flow mainly depends on the flow of the previous day and much less on the precipitation. However, the most important use of such model is forecasting the increase of flow due to the past precipitation, and they may be tuned to react to a strong (but physically

uninteresting) signal of the past flow(s). This issue is not discussed, and it would be advisable to mention this problem.

Past streamflow data are provided in order to exploit the precious information coming from the real-time measurement of the actual discharge preceding the forecast instant since they represent the only available information on the conditions of the basin before the current storm, and therefore on the capability of the system to respond to rainfall perturbation. As observed since the very first works on ANN rainfall-runoff forecasting presented in the literature (e.g. Minns and Hall, 1996), ANN fed by precipitation data only do not allow satisfactory performances, especially when reproducing runoff at fine temporal scales (hourly or daily), because the state of the basin plays an important role in determining the streamflow formation process.

On the other hand, past precipitation is a crucial information, especially for forecasts for lead-times greater than one, as those issued in the present applications: the proposed models, being tuned for adequately forecasting directly the future streamflow values also for long time-horizons and due to the importance of precipitation data for such high lead-times, cannot but take into account adequately this kind of information, as proved by the satisfactory forecasting performances.

7) Clustering is performed in the same 7-variable space, which is characterized by the high (auto)correlation of flow components. It would be interesting to see if similar results can be achieved by clustering in a much simpler space – for example with the linear combinations (moving averages) of Q(t-L), L=0,: : :3.

On behalf of the Reviewer's suggestion I tested also this option, but the average of the streamflow over the past hours can not allow an adequate identification of the hydro-meteorological conditions and of their future evolution, since 1) a variable given by streamflow data only provides no information at all on the precipitation data, that are extremely important for longer lead-times, and 2) a moving average does not even allow to identify a falling limb from a rising one but it simply separates the forecast instants in non overlapping classes of different flow ranges, as in the results I obtained testing a SOM clustering based on the moving average on the last three streamflow values.

Class	Minimum value	Maximum value
1	10.51	15.88
2	15.90	23.53
3	46.27	71.65
4	7.47	10.48
5	23.56	33.78
6	71.95	712.97
7	0	5.01
8	5.01	7.46
9	33.87	46.20

The forecasts obtained in validation (and reported below) from the rainfall-runoff models calibrated over such classes are less good than those obtained by the other modular approaches because such division leads to classes that are not always homogeneous as far as the future behaviour of the catchment is concerned.

	LT= 1 hour	LT= 2 hours	LT= 3 hours	LT= 4 hours	LT= 5 hours	LT= 6 hours
Efficiency	0.988	0.933	0.860	0.860	0.765	0.662
MAE	0.53	1.08	1.54	1.73	2.17	2.53

Performance indexes in validation for a modular approach based on a SOM clustering of the moving average on the last three streamflow values

8) How the ANN topology was optimized (number of hidden nodes). Was crossvalidation set used?

See the reply to comments 3 and 5.

Technical corrections/suggestions

I will certainly take into account the valuable suggestions on English revision in the final version of the manuscript.