Reply to Anonymous Referee #2

I wish to deeply thank the Referee for his/her time and consideration and for his/her valid and pertinent comments that I examined with great care and that will certainly lead to an improvement in the paper. I hope that the reviewer finds that his/her concerns have been adequately addressed in this response.

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

General comments

Related work is cited sufficiently, however, the new contribution of the work could be pointed out clearer. Please also make clearer that the reported improvement is due to the new approach (see also specific comment 1).

I will make clearer this aspect in the last paragraph of section 4.

The principal difference with all the cited works is that the present approach issues multi-step ahead forecasts: to do so it makes use, in the classification phase, not only of past streamflow data but also of past rainfall data; this is not done in many of the cited modular applications, as written in Section 4, but it is instead crucial for multi-step forecasts given the importance of past precipitation in the evolution of future further flows.

Also in the work by Jain and Srinavasulu (2006), a single step ahead forecast is issued and the alternative methods they propose for decomposing the calibration data are always based on a separation of the hydrograph in different rising and falling limbs, that is they are based on past streamflow data alone. I therefore believe that, even if from the text of their paper I understood that the rainfall values were included in the SOM mapping, the resulting classification is mainly seen and interpreted as a classification of the streamflow magnitude, as the authors state:

"It may be noted that the SOM classifications correspond to different magnitude flows e.g. low, medium, and high based on the descriptive statistic in terms of mean."

This is not what is obtained in the present work, where instants with similar flow values may belong to different classes if they represent a different hydrometeorological condition. The reason may be due to the fact that the number of nodes of the SOM are small (3 and 4 respectively), but it is always difficult to interpret the data and analyses developed by other authors. It follows that in their work the potential of the SOM method is not thoroughly explored nor exploited and they obtained better results with the other modular approaches, based on the decomposition of the hydrograph. This is not in disagreement with what obtained here, since for one-step ahead forecasts the role of past precipitation is certainly less important than for longer forecast horizons and the proposed decomposition methods might be more efficient than the SOM for interpreting the behaviour of the flow progress alone. On the contrary, for issuing forecasts over longer lead-times, the role of past precipitation is crucial and the algorithms based on the decomposition of the hydrograph cannot take it into account.

Specific comments

1. Increasing the complexity of the model often improves prediction results. From the study it is not entirely clear whether improved performance is due to the increased number of ANN (4 compared to 1 in the reference case) or whether the use of SOMs for data pre-clustering is the reason for the improved performance. Could you address the question whether you expect comparable results to be achieved using some other partitioning method in the discussion of your results?

With data-driven models, it may happen that increasing the complexity of the model when considering the same calibration set actually deteriorates the performances in validation, because the model becomes lacking in generalisation capability: the training performance improves as the complexity increases, whereas the performance on an external validation set tends to deteriorate.

The same happens in the present study with the first modular approach which, in spite of being formed by an even larger number of ANNs (9 compared to 4), issues validation forecasts less reliable than those obtained with the modular approach with 4 classes. The small classes obtained dividing the calibration data in nine classes leads to an information content that is insufficient for the adequate calibration of the corresponding R/R module and therefore to a generalisation failure in validation.

The SOM is only one of the possible classification methods, with the advantages exposed in the paper (mainly the ability to preserve the topological structure of the data, thus allowing also an evaluation of the affinity between the clusters, so that wider classes may be formed, that are always hydrologically homogeneous) Honestly, I believe that also a different method for the partitioning of the input data, provided that it is able to take in due account both the past streamflow and rainfall observations and it is able to adequately identify the different hydro-meteorological conditions, would provide analogous results.

2. Please point out more clearly how the method presented in manuscript differs from other rainfall-runoff modelling studies using SOMs (as reviewed by Kalteh (2008)). This will help to better understand the innovation resulting from the presented work.

See Reply to General Comment above.

3. p900L22ff Please also mention the purpose of the model for the Sieve River. If the goal is high flow prediction, then the presented study does not substantially improve predictions compared to the global model presented later on. The Nash Sutcliffe coefficient of efficiency is mainly influenced by high flow periods, and only minor improvements were achieved in terms of this measure.

The model is meant to simulate all the flow regimes and seosons: the distinction of the different hydro-meteoreological conditions has exactly the aim to identify and simulate better the different regimes. I will make it clearer at the beginning of the paper.

The efficiency coefficient is widely used for assessing the performances of the reproduction of not only high but of all flows: its main advantage is its easy interpretation, since the meaningful value of 0.0 provides a convenient reference point to compare the model with the predictive abilities of the observed mean; its use is opportune also because of its frequent use (it is probably the most widely used goodness-of-fit criterion for hydrological models). On the other hand, as underlined in the paper, it tends to inflate the higher errors and therefore the errors on the higher flows. Also for this reason (the second reason being to present also an error measure in the same units of the simulated variable, as suggested also by Legates and McCabe, 1999), I added another performance index, that is the mean absolute error, MAE, that does not inflate the highest errors but it gives the same weigth to all errors. Also the comparison with a persistent model is useful for assessing the performances of all flow regimes, since it is not biased towards high flows.

4. The introduction reads as if the term "system theoretic" is interchangable with data driven? Wheater et al. (1993) seems to be the relevant reference, which unfortunately is not easily available to me. I do not agree with this interchangability of terms. It is my understanding, that system theory starts from a system understanding, defining system boundaries, system components and their interactions. Maybe you could clarify on your view of system theory.

The first phrase of the Introduction is probably misleading: actually, Wheater et al use the term “metric” rather than “systemic” or “system-theoretic”, that is instead used, for example, in Sivapalan et al (2003) or in Nayak et al (2005), who explain the meaning that I, too, give to the term:

“In the physical approach, the primary motivation is the study of physical phenomena and their understanding, while in the system theoretic approach the concern is with the system operation, not the nature of the system by itself or the physical laws governing its operation.” (Nayak et al., 2005)

“Whereas the mechanistic or reductionist approach emphasizes the individual components or processes that make up the whole, the systems approach emphasizes the whole ([…]). Also, recognizing that the properties of the individual processes are not intrinsic properties of the whole system, any study of the properties of the individual processes or components is carried out only from the point of view of understanding the whole system, and the focus is much more on the interactions, feedbacks, and functional relationships between the various parts of the whole. In other words, the systems approach focuses more on networks or patterns or linkages between the various components of the whole rather than on the individual components themselves.” And “In his ‘Linear theory of hydrologic systems’, Dooge (1973: 5–6) states: <<In systems analysis, we are concerned only with the way in which the system converts input to output. If we can describe this system operation, we are not concerned in any way with the nature of the system—with the components of that system their connection with one another, or with the physical laws which are involved>>” (Sivapalan et al., 2003).

In particular, as far as ANNs are concerned, several papers in the literature classify them as “systemic” or “system-theoretic”, for example: Rajurkar et al. (2002); Jain and Prasad Indurthy (2003); Xiong et al. (2006), Nayak et al. (2005).

It is also true that in hydrology the two terms often describe the same models, as stated by Sivapalan et al (2003):

" In hydrology, the systems approach attracted a lot of interest in the 1960s and 1970s ([…]) but, interestingly, tended to focus on deriving input–output relationships from data in a most efficient manner, often based on formal optimization procedures […]".

I agree that the two terms do not mean the same thing and I will rephrase line 23, specifying that the two things do not coincide but that systemic models are ALSO data-driven.

5. p902L9 Use of equations could make the following easier to understand: "In correspondence of a computational node, each one of the entering values is multiplied by a connection weight. Such products are then all summed with a neuronspecific parameter, called bias, used to scale the sum of products into a useful range. The computational node finally applies an activation function to the above sum producing the node output."

I will modify the text and add the equation according to this suggestion:

"In correspondence of a computational node *J* (see Figure 1), each one of the N_j entering values (IJ_i) is multiplied by a connection weight (w_{ij}). Such products are then all summed with a neuron-specific parameter, called *bias* (b_j), used to scale the sum of products into a useful range. The computational node finally applies an *activation function (f)* to the above sum producing the node output (*OJ*):

$$OJ = f\left(\sum_{i=1}^{N_j} w_{ij} IJ + b_j\right)$$"

6. p903L9 Please give a short description why the use of independent ANN with a single output node for each lead time is of advantage compared to having a single ANN with multiple output nodes, one for each lead time.

In the past I have often used a direct multi-output approach for the prediction of different hydro-meteorological variables (for example in Toth et al., 2000 and Toth and Brath, 2002): such approach sets the number of the output nodes equal to the number of the lead-times of the prediction: each output node represents one time step to be forecasted, so that the forecasts for all the lead-times are issued simultaneously. The multi-output approach is certainly more efficient than the recursive approach, sometimes used in the first ANNs studies, where an only output node represents the forecast for one step ahead, so that, for issuing the forecasts over longer lead-times, past forecasts are provided in input in substitution of the future observations that are not, yet, available.

On the other hand, in the ANN model, there is no need to issue a simultaneous forecasting over all the lead-times, because the future occurrences are not interdependent (like they would be, for example, in a recursive approach). Each one of the future streamflow occurrences may be considered as a separate variable to be predicted and therefore a separate model may be tailored to fit the streamflow evolution for each specific lead-time, starting from the same hydro-meteorological conditions, that is giving in input to the network the same input variables available in the forecast instant.

In recent rainfall-runoff modelling experiments, I obtained that the performances in validation of the method here used (separate networks) were better than those obtained with a single ANN with multiple nodes. This is not surprising, since the parameter are tuned to fit the future flow on a specific time horizon, rather than trying to simultaneously fit all the 6 forecasts, resulting in a trade-off that may penalise one or more of the forecasts. The system, even if more complex to describe, is in reality more parsimonious and less susceptible to the risk of overfitting: in fact, in comparison to the multioutput network, each single-output network has less parameters to be calibrated, since there are less weigths (in the present case the spared weigths are 5 times the number of the hidden nodes), and less biases (5), while the amount of input data available for the training is always the same.

If the Referee agrees, I would rather to not add these considerations in the revised paper, since this aspect is not the main focus of the paper.

7. I was not able to find information about the units for Q and P. Please clarify this in a revised version.

The units are m^3/s and mm for the Q and for the P respectively: this is indicated only in Table 1 (and in Table 2, through the unit of the MAE) in the original manuscript and I will add it in Section 2, when describing the case study, in the revised version.

8. P907L19 How does the result depend on the network geometry/number of nodes.

As far as the geometry is concerned, I do not believe that the geometry plays an important role on the obtained results; in order to guarantee a symmetry (isotropy) in the lateral interaction between neighbouring output nodes, I chose an hexagonal grid (where that diagonal neighbours have the same distance as horizontal and vertical neighbours), as suggested by Kohonen himself (Kohonen, 2001) and by several works on SOM maps in different areas (e.g. Van der Voort et al., 1996; Hsu and Halgamuge, 2003; Shirazi and Menhaj, 2005).

As far as the number of nodes is concerned, 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.

In addition, following the Reviewer's following remark, in the reply to Comment 12 I describe the results obtained by a SOM with 4 nodes, that was analysed in the course of the research progress. The slightly inferior performances obtained with this data partitioning are probably due to the fact that the classes result less well-identified (being the paper already pretty long, I did not include such description in the original paper).

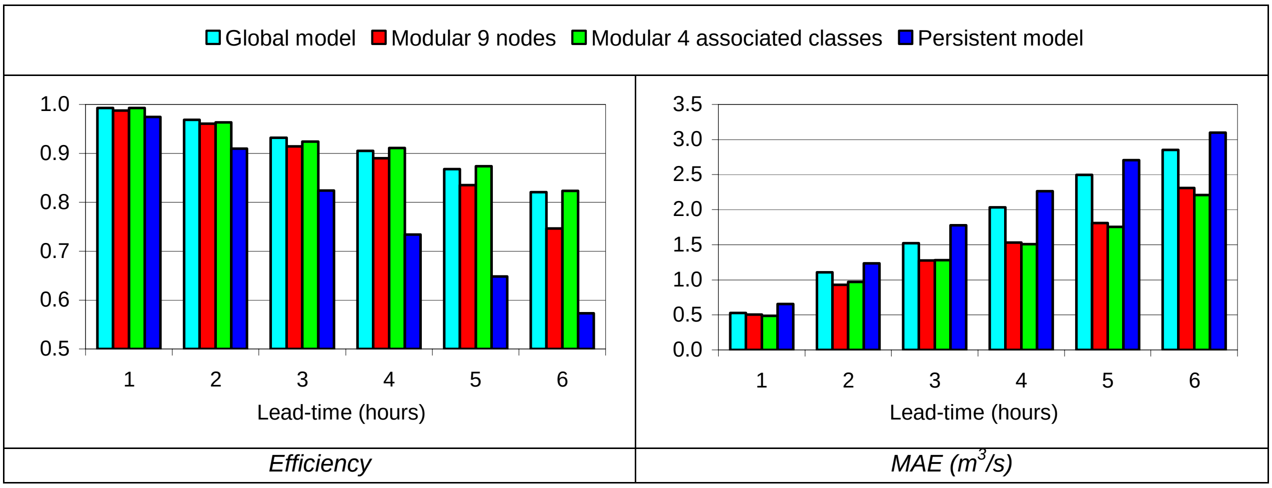
9. P908L17-25 Fig1 may not be representative for the entire simulation period. For example, it seems that nodes 4,5 and 9 form the class of the late recession period - whereas the text states that these nodes are less easily identifiable. Maybe a longer time series could be shown using a different representation of the class assignment?

It is certainly true that the part of the hydrograph represented in Figure 1 is not representative of all the conditions. Indeed, it is on the base of the visual inspection of several zooms of the hydrograph that I stated that classes 4, 5 and 9 are nor clearly identifiable. In fact the corresponding forecast instants are placed on different positions of the hydrograph, always in correspondence of average to low flows, but not always on recession limbs, since in some cases they are placed where the flow has a slight increase.

I confess that since each forecast instant may correspond to a different class, I do not have in mind a different graphical representation that may show the class assignment and a longer hydrograph would be unreadable (in figure 1 the markers are already a bit too squeezed together).

10. Please give a graphic representation of Table 2 as this is much easier to read.

I may show the following bar plots:



I have to thank twice the Referee for this comment, since, preparing the plots, I found out a mistake in Table 2, as far as the performances for LT=5 hours of the global model are concerned: the MAE is 2.50 m^3/s and the efficiency is 0.868.

11. P910 L6ff. I like the use of multiple performance measures and the use of a benchmark model as suggested by Schaefli and Gupta (2007). To make your work more transparent, maybe you could highlight how you decided on which performance measures to use?

See reply to comment 3), of which, following the Reviewer suggestion, I can add a part in the revised text, as suggested.

12. P911 L16-29 Please clarify why you don't calculate a new SOM with fewer nodes instead of combining nodes in a somewhat subjective manner? I suggest to recalculate the SOM with only 4 nodes.

Actually, I did implement also SOM networks with a different number of nodes, and in particular with a 4-nodes output layer, whose results are shown below. I did not add these results in the paper since I feared it would make it excessively long, but the Reviewer's comment (extremely pertinent) suggests me to cite this experiment in the revised version.

This 4-nodes modular approach would represent in the paper the third modular approach, in addition to the first one, based on 9 classes corresponding to the 9 nodes of the original SOM, and to the second one, based on the reasoned association of the original 9 classes in 4 homogeneous broader classes. In the third approach, 4 classes are formed automatically by a SOM with 4 nodes in its output layer.

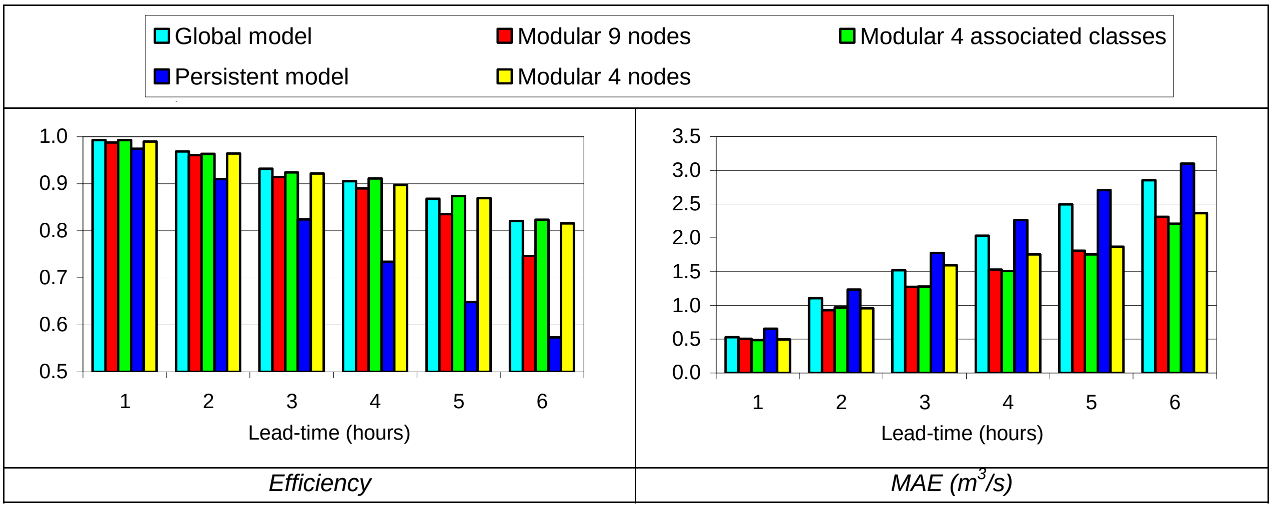
The differences between the four classes identified automatically by this 4-nodes SOM and the classes obtained by the reasoned combination (here named with letters) of the nodes of the original 9-nodes SOM are presented in the table below, where the number of elements and the mean of the input variables of each class are shown.

Class	Members number	Q_t (m^3/s)	Q_{t-1} (m^3/s)	Q_{t-2} (m^3/s)	Q_{t-3} (m^3/s)	P_t (mm/h)	P_{t-1} (mm/h)	P_{t-2} (mm/h)
2nd modular approach (reasoned association of the original 9-classes)								
A (7+8)	*17628*	4.29	4.30	4.32	4.33	0.00	0.00	0.00
B (4+5+9)	*5640*	19.09	19.14	19.22	19.33	0.11	0.09	0.11
C (1+2)	*1470*	17.46	15.85	14.94	14.43	1.69	1.75	1.62
D (3+6)	*1542*	117.43	118.61	119.01	118.87	0.33	0.35	0.39
3rd modular approach (4-nodes SOM)								
1	*17120*	3.8	3.8	3.8	3.8	0.01	0.01	0.01
2	*4850*	20.2	20.3	20.5	20.6	0.02	0.02	0.02
3	*1463*	8.5	8.3	8.2	8.1	0.54	0.46	0.49
4	*2847*	75.5	75.3	75.1	74.8	0.93	0.98	0.95

The classes do not coincide: in particular, the 4-nodes SOM does not seem able to clearly identify the group of the rising limbs, characterised by the highest rainfall (classes 1 and 2 of the original 9-nodes SOM, joined in one class, named class C, in the reasoned associated classes) and to distinguish it from the data that are around the peak and at the beginning of falling limb (classes 3 and 6 of the original 9-nodes SOM, joined in one class, D, in the reasoned associated classes).

The third forecasting modular approach was successively implemented, on the basis of the 4-nodes SOM classes, formed by 4 differently parameterised R-R ANN networks. Its performances on the validation data are shown in the table below and, for a comparison with the other models, in yellow in the bar plot below

	LT=1h	LT=2h	LT=3h	LT=4h	LT=5h	LT=6h
MAE (m^3/s)	0.499	0.957	1.593	1.753	1.871	2.367
Efficiency	0.990	0.964	0.922	0.898	0.869	0.816



The performances obtained with this last data partitioning are always worse, even if not excessively, than those obtained with the second partition (reasoned association of the 9 classes). This is probably due to the fact that the classes result to be not sufficiently well-identified, in comparison to the association of groups of nodes that represented more "sharply" the various hydro-meteo conditions, because of the more detailed original classification.

Lastly, I will certainly take into account the valuable suggestions on English revision (listed below) in the final version of the manuscript.

13. multi-network and multinetwork are both used - please be consistent. I would suggest to use multi-network

14. Have a native speaker check for English. Examples of sentences, that are hard to understand are:

p899 L11 where it is explicit the intention of modelling;

p899 L21 In order to keep openly into account the fact;

p911 L4-15 is hard to understand;

p901 L18-21 is hard to understand

REFERENCES NOT CITED IN THE PAPER

Legates, D. R., and G. J. McCabe (1999), Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation, *Water Resour. Res., 35*, 233-241.

Nayak, P.C., K.P. Sudheer, D.M. Rangan and K.S. Ramasastri, Short-term flood forecasting with a neurofuzzy model, Water Resour Res 41 (4) (2005) doi:1029/2004WR003562.

Sivapalan M, Blöschl G, Zhang L, Vertessy R. 2003. Downward approach to hydrologic prediction. Hydrological Processes 17: 2101-2111

Jain, A. and S. K. V. Prasad Indurthy, Comparative Analysis of Event-based Rainfall-runoff Modeling Techniques---Deterministic, Statistical, and Artificial Neural Networks, J. Hydrologic Engrg. 8, 93-98 (2003), DOI:10.1061/(ASCE)1084-0699(2003)8:2(93)

Rajurkar, M. P., Kothyari, U. C. & Chaube, U. C. (2002) Artificial neural networks for daily rainfall–runoff modelling. Hydrol. Sci. J.47(6), 865–877.

Xiong L., Guo S., O'Connor K.M. Smoothing the seasonal means of rainfall and runoff in the linear perturbation model (LPM) using the kernel estimator (2006) Journal of Hydrology, 324 (1-4), pp. 266-282.

Nayak, P.C., K.P. Sudheer, D.M. Rangan and K.S. Ramasastri, Short-term flood forecasting with a neurofuzzy model, Water Resour Res 41 (4) (2005) doi:1029/2004WR003562.

Toth E. e Brath A., Flood Forecasting Using Artificial Neural Networks in Black-Box and Conceptual Rainfall-Runoff Modelling, Proc. International Environmental Modelling and Software Society Conference 2002, 24-27 June 2002, Lugano (CH), Ed. A.E. Rizzoli and A.J. Jakeman, iEMSs Publ., Como, Vol. 2, pp. 166-171, 2002.

Toth, E., Brath, A. e A. Montanari, A. – Comparison of short-term rainfall prediction models for real-time flood forecasting, Journal of Hydrology, 239 (1-4), pp. 132-147, 2000.

Van der Voort, M., Dougherty, M., Watson, S., 1996. Combining Kohonen maps with ARIMA time series models to forecast traffic flow. Transportation Research C 4 (5), 307–318

Hsu A.L. and S.K. Halgamuge. An unsupervised hierarchical dynamic selforganizing approach to cancer class discovery and marker gene identification in microarray data (supplementary information on “Dynamic SOM with hexagonal structure for data mining”). Bioinformatics, Vol. 19(16), pp. 2131-2140, 2003.

Shirazi J. and Menhaj, M.B., A SOM Based 2500 - Isolated - Farsi – Word Speech Recognizer, in W. Duch et al. (Eds.): ICANN 2005, LNCS 3696, pp. 589 – 595, Springer-Verlag Berlin Heidelberg 2005.