Jul 13th, 10:50 AM - 11:10 AM

DATA-DRIVEN RAINFALL/RUNOFF MODELLING BASED ON A NEURO-FUZZY INFERENTIAL SYSTEM

N. Bartoletti
University of Florence, nicola.bartoletti@stud.unifi.it

F. Casagli
University of Florence, francesca.casagli@stud.unifi.it

S. Marsili-Libelli
University of Florence, stefano.marsililibelli@unifi.it

A. Nardi
University of Florence, arianna.nardi@stud.unifi.it

A. Oliva
University of Florence, alessandro.oliva@stud.unifi.it

See next page for additional authors

Presenter/Author Information

N. Bartoletti, F. Casagli, S. Marsili-Libelli, A. Nardi, A. Oliva, and L. Palandri

International Environmental Modelling and Software Society (iEMSs)
8th International Congress on Environmental Modelling and Software
Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2016-proceedings

DATA-DRIVEN RAINFALL/RUNOFF MODELLING BASED ON A NEURO-FUZZY INFERENTIAL SYSTEM

N. Bartoletti [a], F. Casagli [a], S. Marsili-Libelli [b], A. Nardi [a], A. Oliva [a], L. Palandri [a]

[a] *Graduate student, Department of Civil and Environmental Engineering, School of Engineering, University of Florence, Italy*
(nicola.bartoletti/francesca.casagli/arianna.nardi/alessandro.oliva/lorenzo.palandri@stud.unifi.it)
[b] *Department of Information Engineering, School of Engineering, University of Florence, Italy*
(stefano.marsililibelli@unifi.it)

Abstract: The development of rainfall/runoff models may involve extensive computation and differing platforms, including GIS. In this paper we present a simple data-driven approach which avoids the use of GIS, but is based on a combination of Principal Component Analysis (PCA) and an Adaptive Neuro Fuzzy Inference System (ANFIS) to produce a simple and effective output flow prediction based on previous rainfall/runoff data. Given the ANFIS internal complexity, the emphasis of the paper is on how to set-up the most representative and parsimonious data structure that produces an efficient output flow estimation. In the preliminary data reduction stage, the PCA approach is compared to the equivalent rainfall computed by the Thiessen polygons, involving GIS, and it is demonstrated that the former approach yields a better data set for the ANFIS processing in terms of algorithm complexity and output accuracy. The algorithm is applied to two differing medium-size catchments in Tuscany, central Italy, and provides an excellent approximation of the output discharge with reduced computational complexity.

Keywords: Rainfall/runoff models; Principal Component Analysis; Neuro-Fuzzy Networks; ANFIS.

1 INTRODUCTION

Rainfall-runoff models have reached an advanced stage of maturity and the diversity of the many differing mathematical techniques on which they are based is staggering, as shown by Chow et al., (1988), Voinov et al., (2004), Brunner, (2010), Wu et al., (2014), to name but a few. All these studies, though, concentrate more on the rainfall-runoff relationship than on the amount and possible redundancy of the input data. On the other hand, the deployment of rain gauges in small river catchment is not always optimal, with the result that the recorded data are often redundant and little informative. It is also often difficult to relate the output flow to any specific rain gauge in the catchment. The method described in this paper combines two popular data-driven algorithms to reduce the redundancy of the rainfall data and to model the relation between the input rainfall and the output flow using a neuro-fuzzy network. The reduction of the precipitation data is performed by either the Thiessen polygons (TP) or through principal component analysis (PCA). With the latter method an important data reduction can be obtained, while controlling the loss of information, before applying this equivalent input to the neuro-fuzzy model estimating the output discharge. An effective way of combining the fuzzy modelling flexibility with the learning capability of neural networks is the Adaptive Neuro Fuzzy Inference System (ANFIS) (Jang, 1993). The combined algorithmic pathway is shown in Figure 1. First the measured rainfall data $\boldsymbol{R}(t)$ are transformed either by the Thiessen polygons of through PCA, which performs both de-correlation and dimension reduction. In either way an equivalent set of rainfall data $\hat{\boldsymbol{R}}(t)$ is obtained. After partitioning these data into a training and a validation set, the ANFIS is trained with the equivalent precipitation data, organized as specified later, to return the estimated output discharge $\hat{Q}(t)$. In the PCA case the ANFIS output returns the transformed estimated discharge in the PC reference space, therefore the inverse PCA transform must be applied to obtained the estimated discharge.

The algorithm is applied to two minor catchments in Tuscany, central Italy, for which abundant rainfall data are available from a large number of rain gauges. In both cases the combined PCA+ANFIS method produces an efficient flood representation with a much simpler computational structure compared to the Thiessen polygons approach.

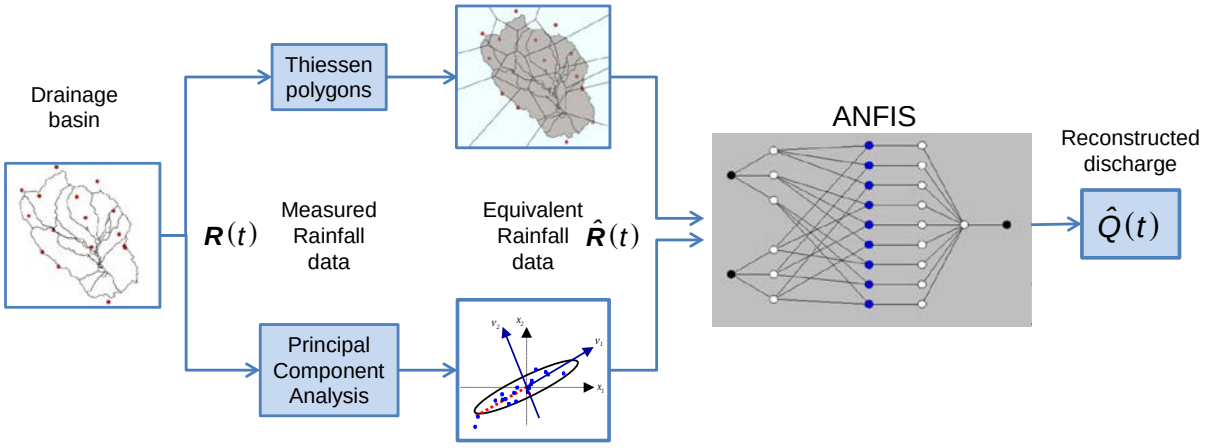

Figure 1. Data organization and discharge modelling. The measured rainfall data are reduced either by Thiessen polygons or through PCA. They are then used to train a neuro-fuzzy network (ANFIS) to produce an estimate of the output discharge.

2 ALGORITHM DESCRIPTION

The core of the algorithm is represented by the Adaptive Neuro Fuzzy Inference System (ANFIS) (Jang, 1993), which in the vast paradigm of artificial neural networks (Wu et al., 2014) has the merit of combining the learning features of a neural network with the approximating characteristic of a fuzzy inference system. The generic ANFIS inner structure is shown in Figure 2. It consists of a bank of antecedent membership functions which imply (in the fuzzy sense) an equivalent number of linear Sugeno consequents. Two feedback paths provide the adjusting mechanism to adapt the Neuro-Fuzzy network to the data in order to minimize a quadratic error functional extended to the training data set.

$$\Delta Q = \sum_t \left(\hat{Q}(t) - Q(t)\right)^2 . \tag{1}$$

The optimization algorithm can be either a back-propagation, a least-squares, or a combination of both.

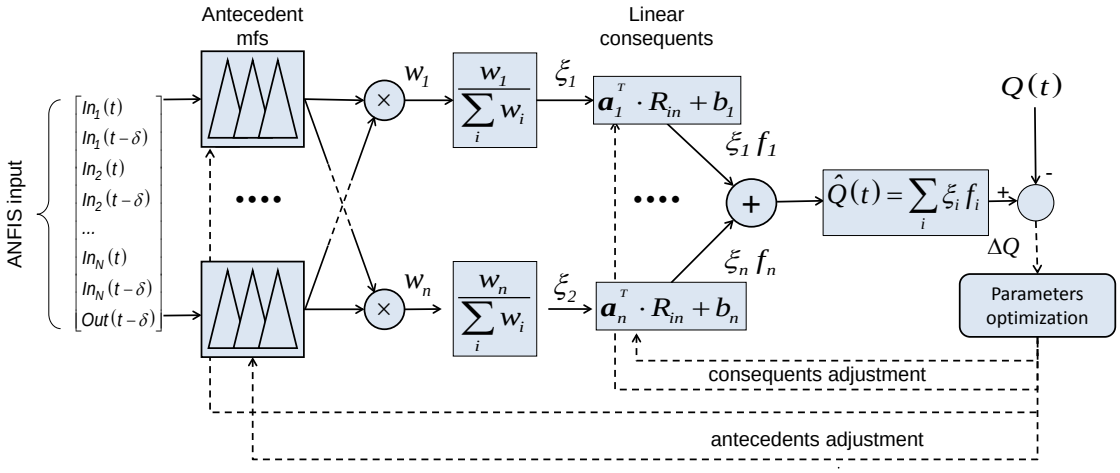

Figure 2. ANFIS structure adapted to the rainfall/runoff model, showing the adaptation feedback path to adjust both the antecedent membership functions and the coefficients of the linear consequents $[\boldsymbol{a}\ \boldsymbol{b}]$. The ANFIS input is composed of past samples of rainfall and output flow.

The ANFIS structure available in the Matlab® Fuzzy Toolbox requires that the data are structured as a matrix in which each column represents an input, while the last one contains the output data to be used for training or checking. Since ANFIS is not conceived to model dynamical systems, it does not possess any inner time-lapse capability. The time-dependence must therefore be introduced by the user in the data matrix by creating additional columns containing delayed copies of the input variables. Thus, with reference to Figure 2, the generic row of the input data matrix at time *t* would have the following form, as fully described in Marsili-Libelli, (2016), supposing that there are N rain gauges deployed in the catchment

$$Data(t) = \left[In_1(t) \quad In_1(t-\delta) \quad In_2(t) \quad In_2(t-\delta) \quad ... \quad In_N(t) \quad In_N(t-\delta) \quad Q(t-\delta) \quad Q(t) \right], \quad (2)$$

where δ is the observed input/output delay, the only restriction being that the training output must be placed in the last (rightmost) column of the *Data(t)* matrix. The exact composition of the data vector defined by eq. (2) depends on the method used to represent the equivalent rainfall data, either using the Thiessen polygons or through PCA, as will be exemplified later. The implications of using either data pre-processing method are now assessed.

2.1 Equivalent rainfall by Thiessen polygons

Thiessen Polygons are Voronoi Cells (de Berg et al., 2008; Aurenhammer et al., 2013) used to distribute the measured precipitation to the entire polygon assigned to each rain gauge. In this way, given a set of point measurements, the equivalent area-weighted precipitation can be inferred as,

$$\boldsymbol{R}_{th}(t) = \frac{\sum r_i(t) \cdot A_i}{\sum A_i}, \quad (3)$$

where the basin is partitioned into a number of Voronoi cells of surface A_i each containing one rain gauge, and $r_i(t)$ is the precipitation recorded by each rain gauge. The computation of eq. (3) is normally carried out within the GIS used to define the spatial characteristics of the catchment. The ANFIS input matrix is then composed of past rainfall samples obtained from eq. (3) plus the last two output flow samples, i.e.

$$\hat{\boldsymbol{R}}(t) = \left[R_{th}(t-\delta) \; R_{th}(t-2\delta) \; ... \; R_{th}(t-k\delta) \; Q(t-\delta) \; Q(t) \right], \quad (4)$$

where δ is the sampling interval and *k* is the maximum delay considered in the model. The actual structure of $\hat{\boldsymbol{R}}(t)$ obtained by eq. (4) depends on the specific features of the catchment.

2.2 Equivalent rainfall by principal component transform

An alternative method for constructing the equivalent precipitation is to aggregate the rainfall measured from all the rain gauges by Principal Component Analysis (PCA) (Dunteman, 1989; Jolliffe, 2002) and retain only the most significant ones, according to the scree diagram, assessing the contribution of each PC to the total data variability. The data reduction is essential to limit the ANFIS complexity and is limited to the first few PCs which conserve a minimum of about 85% of the original information. In the following examples it was observed that very few PCs are required to obtain an excellent percentage of the total variability. Since the PCA-transformed data bear no resemblance to the original ones, to retrieve the estimated output discharge the neuro-fuzzy network output must undergo the inverse transformation process. Assuming that there are *N* rain gauges in the catchment and that the network will consider the previous output flow $Q(t-\delta)$ together with the rainfall from each rain gauge at the two previous sampling instants $R_i(t-\delta)$ and $R_i(t-2\delta)$, with $i = 1,...,N$, the ANFIS input data matrix has $q = N + 2$ columns, so that its generic row has the following structure

$$\hat{\boldsymbol{R}}(t) = \left[R_1(t-\delta) \; R_1(t-2\delta) \; ... \; R_N(t-\delta) \; R_N(t-2\delta) \; ... \; Q(t-\delta) \; Q(t) \right], \quad (5)$$

Such a high-dimensional input would represent a formidable challenge for the ANFIS structure in terms of number of parameters and rules. However, transforming the matrix $\hat{\boldsymbol{R}}(t)$ through PCA yields

$$\boldsymbol{Z} = \hat{\boldsymbol{R}} \cdot \boldsymbol{P} \quad \rightarrow \quad \boldsymbol{P} = \left[\boldsymbol{P}_a \mid \boldsymbol{P}_{q-a} \right] \quad \rightarrow \quad \boldsymbol{Z}_a = \hat{\boldsymbol{R}} \cdot \boldsymbol{P}_a \tag{6}$$

where $\boldsymbol{P}_a$ represents the first a principal components, which are retained in the PCA transform. Selecting $a = 3$ provides a sufficient approximation of the data variability, so that the reduced input matrix $\boldsymbol{Z}_a$ can be used as the ANFIS input instead of $\hat{\boldsymbol{R}}$, with a considerable saving in the network complexity. Since the ANFIS output will be provided in the PCA transformed space, the original output discharge will be recovered as the last (rightmost) column of the inverse PCA transform, i.e.

$$\hat{Q} = \boldsymbol{Z}_a \left(\textit{last column} \right) \cdot \boldsymbol{P}_a^T \ . \tag{7}$$

While the specific form of the data input matrix of eq. (2) will be defined later in Sect. 3, it is important to point out that PCA data pre-processing produces a considerable simplification in the ANFIS structure and enhances its generalization capability. The PCA-ANFIS combination is shown in Figure 3.

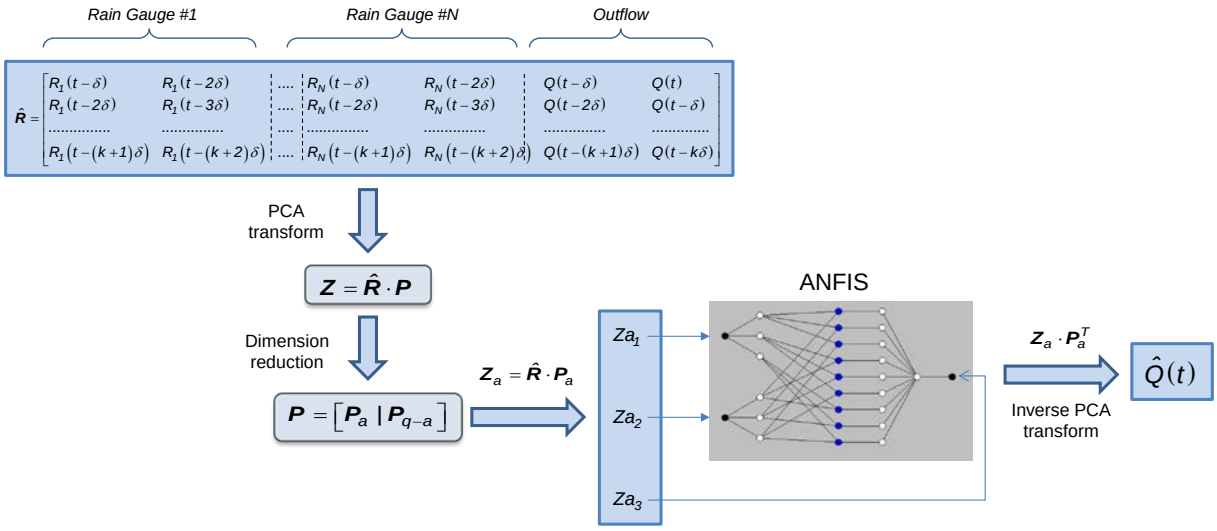

Figure 3. Data flow in the PCA-ANFIS combination. The data reduction is performed by retaining only the first three PCs obtained by PCA transform. The estimated outflow $\hat{Q}(t)$ is eventually retrieved by inverse PCA transform.

3 APPLICATION TO SMALL CATCHMENTS

Two small catchments of about the same surface area in the Tuscany region were selected for the study. Their location and extent is shown in Figure 4. In both cases there are a few rain gauges deployed over the catchment and one single hydrometer at the basin outlet. As explained in Sect. 2, the rainfall data were condensed into a single equivalent rainfall time-series $\hat{\boldsymbol{R}}(t)$, either through the Thiessen polygons or by PCA reduction. These two approaches, described in Sect. 2, are now adapted to the these case studies.

3.1 Application to the Ombrone Pistoiese catchment

This river is a right-bank tributary to the Arno river and flows through a steep hilly terrain. It collects water from a large number of tributaries covering a drainage basin of nearly 490 km^2. The hydrographic mesh is shown in Figure 4 A, in which the location of sixteen rain gauges is indicated together with the output flow hydrometer

3.1.1 Equivalent rainfall by Thiessen polygons

The corresponding Thiessen polygons were computed by Quantum GIS (www.qgis.org) (Shekar and Xiong, 2008; Petrasova et al., 2015) as shown in Figure 5. All the data were originally sampled every 15 minutes. After spline smoothing, the equivalent rainfall $\hat{\boldsymbol{R}}(t)$ over the entire catchment is computed according to eq. (3).

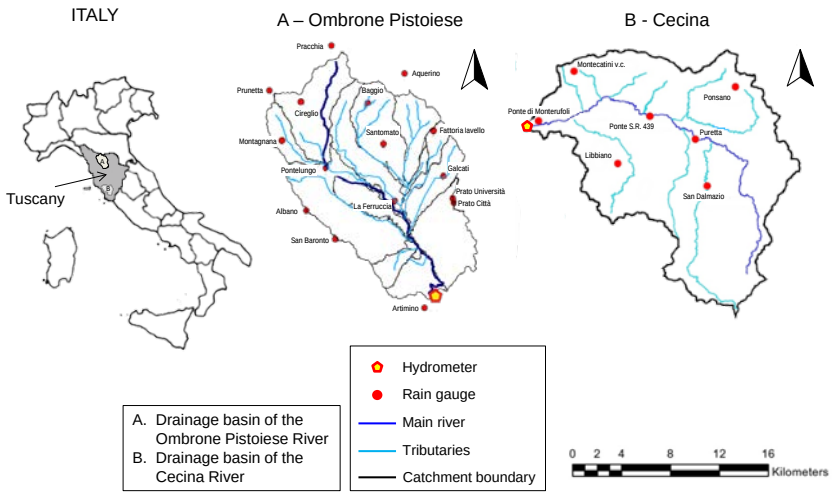

Figure 4. Location of the Ombrone Pistoiese and Cecina river catchments in Tuscany, showing the position of the rain gauges and of the outlet hydrometers.

After testing several differing parametrizations, the best ANFIS structure proved to be the one including the output flow one hour previously and the rainfall sampled at 5 and 10 hours earlier. After smoothing the original rainfall and flow data and resampling them at 1 h intervals, the ANFIS input data matrix of eq. (2) now takes the form of Figure 6

$$\hat{\boldsymbol{R}}(t) = \left[R_{th}(t-10) \quad R_{th}(t-5) \quad Q(t-1) \quad Q(t) \right] . \tag{8}$$

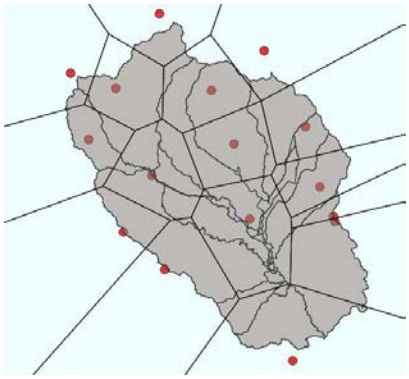

Figure 5. Thiessen polygons computed for the Ombrone Pistoiese catchment by Quantum GIS.

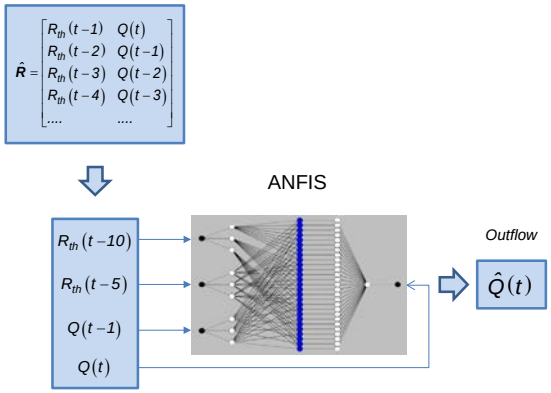

Figure 6. ANFIS structure used to estimate the output flow on the basis of two previous rainfall and one flow samples, scaled by multiples of the sampling interval $\delta = 1\ h$.

3.1.2 Equivalent rainfall by PCA transform

The alternate method, illustrated in Sect. 2.2, consists of transforming the rainfall and flow data through PCA and to retain only the first three PCs, thus the ANFIS input data matrix of eq. (2) now takes the simple form

$$\left[Z_{a1}(t)\, Z_{a_2}(t)\, Z_{a3}(t) \right], \tag{9}$$

where the transformed data $\boldsymbol{Z}_a$ were obtained from eq. (6) and the input-output delays are already accounted for by inheriting the structure of the original data matrix $\hat{\boldsymbol{R}}(t)$ of eq. (5).

3.1.3 ANFIS rainfall/runoff reconstruction

The ANFIS network of Figure 6 fed by the data structured as in eq. (8) was used to estimate the output flow, which is compared to the observed data in Figure 7, where the first part of the rain event was used for training, and the remaining data from the same event were used for validating the network. The ANFIS structure was generated from the data using a FCM (Bezdek, 1981) clustering, using the `genfis3` method which extracts a set of rules that models the data behaviour. By comparing the two plots, it can be seen that the performance of the two algorithms is comparable, but the PCA case performs better on terms of output approximation and has the additional merit of a better generalization. In fact it can accurately predict the peak in the validation data (Figure 7 B), which is higher than any peak appearing in the training phase.

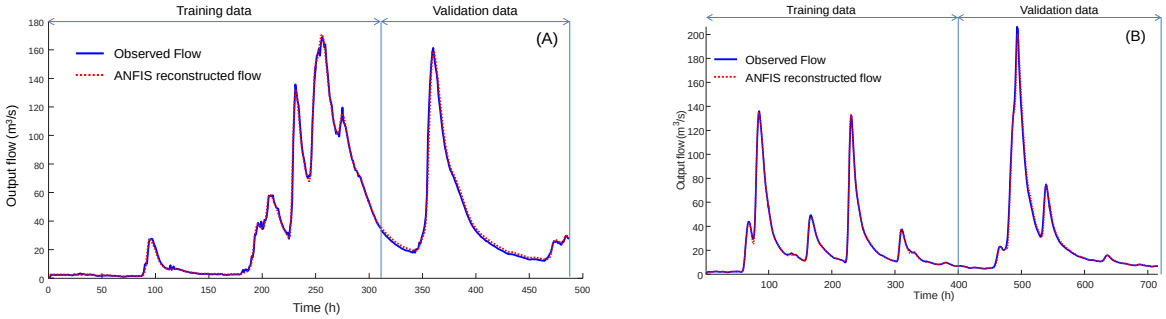

Figure 7. Comparison between observed and ANFIS-reconstructed output flow for the Ombrone Pistoiese catchment using a FCM clustering initialization in two separate events. In (A) data pre-processing by Thiessen polygons are shown, while (B) depicts the result obtained with PCA data pre-processing.

3.2 Application to the Cecina catchment

The portion of the Cecina drainage basin considered in this study is shown in Figure 4 B. Located in southern Tuscany, it has a surface area of about 633 km^2 and a perimeter of over 150 km. There are seven rain gauges in the basin and their data show a considerable correlation, given the relatively short distance among them. Therefore a PCA data reduction technique was directly considered, to lower their dimensionality and eliminate the inherent correlation. The rainfall and flow data used in this exercise are the daily averages recorded in during the 2013 – 2014 period. Given the less mountainous morphology of the catchment, its response is considerably slower than in the previous case. Also the available data represented daily means. So it was decided to adopt a daily sampling interval.

3.2.1 Equivalent rainfall by PCA transform

According to Figure 4 there are seven rain gauges in the Cecina basin providing daily averages. If the ANFIS data matrix were to include the rainfall from the seven rain gauges at any time *t* and at the previous sampling time $t-\delta$, with $\delta = 1\ d$, plus the current output flow $Q(t)$, each row of the input data matrix would include sixteen data

$$\hat{\boldsymbol{R}}(t) = \left[R_1(t-\delta)\ R_1(t-2\delta)...R_7(t-\delta)\ R_7(t-2\delta)\ Q(t-\delta)\ Q(t) \right] \tag{10}$$

and the network complexity would be overwhelming. Instead the reduced ANFIS input data matrix takes the form of eq. (9) and the response of the combined PCA+ANFIS algorithm is shown in Figure 8.

4 CONCLUSION

In data-driven rainfall/runoff models the emphasis has mainly been focussed on the flow routing aspect rather than on the choice of the input data. In this paper both aspects have been put into context by challenging the use of GIS in sorting the often over-abundant wealth of rainfall data. The approach pursued here is based on a two-step procedure to select a significant input data set via PCA reduction, and to use those data to train an adaptive neuro-fuzzy inference system (ANFIS). This approach compared favourably to a conventional equivalent rainfall method based on the Thiessen polygons and it turned out that the equivalent rain gauge obtained by PCA reduction provides much

more efficient data for training the ANFIS network. Its complexity is considerably lower than in the Thiessen case, for which an *ad hoc* structure is required, to be defined on a case-by-case basis. Conversely, the PCA transform not only makes the use of GIS unnecessary, but also drastically reduces the data dimension, eliminates their cross-correlation, and results in a much simpler network with only two inputs and one output. Further, the resulting network has better generalization capabilities. Thus it represents an efficient data representation for the training of the ANFIS network. Of course the (minor) price to pay is the inverse PCA transform to reconstruct the estimated output flow.

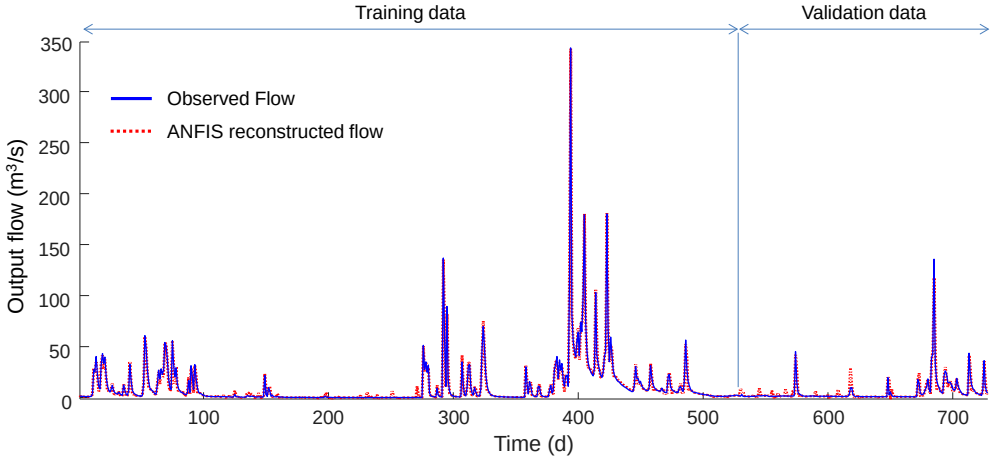

Figure 8. ANFIS performance in reconstructing the output discharge for the Cecina catchment of Figure 4 B, both in the training and in the validation phase, using a daily sampling interval.

The combined PCA+ANFIS method has been applied to two minor river basins in Tuscany and compared to the “conventional“ approach based on Thiessen polygons, which appeared to perform slightly worse, was more complex, had less numerical stability and generalization capability. It also required the preliminary use of a GIS to produce the Thiessen polygons partition and computation of the equivalent rainfall. Though the comparison between the two approaches will be fully discussed in the final paper in terms of their performance, some preliminary conclusions can already be drawn in terms of reduced computational complexity of the PCA approach, which can avoid the lengthy computation of the Thiessen polygons, for which a GIS is required. The input dataset obtained via PCA is much more efficient from the information viewpoint and therefore a simpler network structure can be used.

5 REFERENCES

Aurenhammer, F., Klein, R., and Lee, D. T., 2013. *Voronoi Diagrams and Delaunay Triangulations*, World Scientific Publ. Co., Singapore, pp. 337.

de Berg, M., Cheong, O., van Kreveld, M., and Overmars, M., 2008. *Computational Geometry: Algorithms and Applications*, Springer-Verlag, Berlin, pp. 386.

Bezdek, J. C., 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, pp. 256.

Brunner, G., 2010. *HEC-RAS river analysis system, Hydraulic reference manual, Version 4.1*, US Army Corps of Engineers Hydrologic Engineering Center, Davis CA, USA, pp. 790.

Chow, V. T., Maidment, D. R., and Mays, L. W., 1988. *Applied Hydrology*, McGraw-Hill, New York, USA, pp. 572.

Dunteman, G. H., 1989. *Principal Component Analysis*, SAGE Publ. Inc., Newbury Park, Calif., pp. 487.

Jang, J. R., 1993. ANFIS : Adaptive-Network-Based Fuzzy Inference System. *IEEE Trans. on Systems, Man, and Cibernetics*, 23, 665–685.

Jolliffe, I. T., 2002. *Principal Component Analysis, Second Edition*, Springer Series in Statistics, Berlin, pp. 487.

Marsili-Libelli, S., 2016. *Environmental Systems Analysis with MATLAB*, CRC Press, Boca Raton, FL., USA, pp. 546.

Petrasova, A., Harmon, B., Petras, V., and Mitasova, H., 2015. *Tangible Modeling with Open Source GIS*, Springer Int. Publ., Berlin, pp. 135.

Shekar, S. and Xiong, H., 2008. *Encyclopedia of GIS*, Springer Science & Business Media, New York, USA, pp. 1370.

Voinov, A., Fitz, C., Boumans, R., and Costanza, R., 2004. Modular ecosystem modeling. *Environmental Modelling & Software*, 19, 285–304.

Wu, W., Dandy, G. C., and Maier, H. R., 2014. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. *Environmental Modelling and Software*, 54, 108–127.