Hydrology Research | 46.1 | 2015

Local vs. external training of neuro-fuzzy and neural networks models for estimating reference evapotranspiration assessed through *k*-fold testing

Jalal Shiri, Pau Marti, Amir Hossein Nazemi, Ali Ashraf Sadraddini, Ozgur Kisi, Gorka Landeras and Ahmad Fakheri Fard

ABSTRACT

The improvement of methods for estimating reference evapotranspiration (ET_0) requiring few climatic inputs is crucial, due to the partial or total lack of climatic inputs in many situations. The current paper compares the effect of local and external training procedures in neuro-fuzzy and neural network models for estimating ET_0 relying on two input combinations considering *k*-fold testing. Therefore, different data set configurations were defined based on temporal and spatial criteria allowing for a complete and suitable testing scan of the complete data set. The proposed methodology enabled the comparison in each station of models trained with local data series and models trained with the data series from the remaining stations. Results showed that the external training based on a suitable input choice and a representative pattern collection might be a valid alternative to the more common local training.

Key words | evapotranspiration, *k*-fold testing, modeling, neural networks, neuro-fuzzy

Jalal Shiri (corresponding author)
Amir Hossein Nazemi
Ali Ashraf Sadraddini
Ahmad Fakheri Fard
Water Engineering Department,
Faculty of Agriculture,
University of Tabriz, Tabriz,
Iran
E-mail: *j_shiri2005@yahoo.com*

Pau Marti
Departament d'Enginyeria Rural i Agroalimentària,
Universitat Politècnica de València,
c/Camí de Vera s/n 46022,
València,
Spain

Ozgur Kisi
Civil Engineering Department,
Architecture and Engineering Faculty,
Canik Basari University, Samsun,
Turkey

Gorka Landeras
NEIKER, AB,
Basque Country Research Institute for Agricultural Development,
Alava, Basque Country,
Spain

INTRODUCTION

Accurate estimation of evapotranspiration (*ET*) (the process of water loss to the atmosphere by the combined processes of evaporation and transpiration), is essential for the computation of crop water requirements, water resources management, water balance analysis, selecting the crop pattern of agricultural lands, modeling crop water production functions, and determination of the water budget, especially under arid conditions, where water resources are scarce and fresh water is a limited resource. *ET* can be quantified directly by relatively high cost aerodynamic as well as irradiative Bowen ratio methods or by utilization of lysimeters based on a water balance in a controlled crop area (Allen *et al.* 1998). Notable researches have been carried out so far for studying the physical laws governing the *ET* phenomenon on the analytical base leading to the evolution of some basic *ET* concepts (Katerji & Rana 2011).

The reference term *ET* (ET_0) was introduced by the Food and Agriculture Organization of the United Nations (FAO) as a methodology for computing crop evapotranspiration regardless of crop type, its stage of development and its management (Doorenbos & Pruitt 1977), because the interdependence of the factors affecting the *ET* makes the study of the evaporative demand of the atmosphere difficult. According to Allen *et al.* (1998), ET_0 is the evapotranspiration from a hypothetical crop having 0.12 m height, 0.23 albedo, and fixed canopy resistance of 69 s/m. In this way,

doi: 10.2166/nh.2013.112

the Penman–Monteith model has been adopted as a reference equation for estimating ET_0 and calibrating other ET_0 equations (Allen *et al.* 1998) in the absence of experimental ET_0 values.

According to Landeras *et al.* (2008), the adapted Penman–Monteith equation (which will be referred to as FAO56-PM in short) has two important advantages: (1) it can be applied in a great variety of environments and climate scenarios without local calibration; and (2) it has been validated using lysimeters under a wide range of climatic conditions. Nevertheless, the need for a large number of meteorological variables (e.g., air temperature, relative humidity, solar radiation, and wind speed) is a major disadvantage of the FAO56-PM equation.

During recent years, artificial intelligence (AI) approaches (e.g., artificial neural networks (ANN), adaptive neuro-fuzzy inference system (ANFIS), etc.) have been widely applied in water resources engineering issues. Broad evidence has shown that AI techniques can be successfully applied in modeling water resources engineering components (e.g., ASCE 2000; Cigizoglu & Kisi 2005, 2006; Kisi 2008; Shiri & Kisi 2011a, b; Kim *et al.* 2012; Kisi & Shiri 2012; Landeras *et al.* 2012; Shiri *et al.* 2012). Artificial neural networks (ANNs) have been widely applied for estimating evaporation as well as ET_0 values (e.g., Kim *et al.* 2009, 2012; Kumar *et al.* 2011).

Cigizoglu & Kisi (2005) predicted daily streamflows by three back-propagation techniques using *k*-fold partitioning of neural network training data and they showed that with a data period much shorter than the whole training duration similar flow prediction performance could be obtained. Cigizoglu & Kisi (2006) used *k*-fold partitioning in suspended sediment estimation and reported that partitioning of the training data set showed that similar or even superior sediment estimation performances can be obtained with quite limited data provided that the training data statistics of the subset are close to those of the testing data.

ANFIS is a combination of an adaptive neural network and a fuzzy inference system (FIS). An adaptive neural network is a superset of all kinds of feed-forward neural networks (Jang 1993). The parameters of the FIS are determined by the ANN learning algorithms. Since this system is based on the FIS, reflecting extensive knowledge, an important aspect is that the system should always be interpretable in terms of fuzzy IF-THEN rules. ANFIS is capable of approximating any real continuous function on a compact set (Jang *et al.* 1997).

Kisi (2006) investigated the ability of ANFIS technique to improve the accuracy of daily pan evaporation estimation. Kisi & Ozturk (2007) used the ANFIS computing technique for ET_0 estimation. Shiri *et al.* (2011) compared ANFIS to ANNs for modeling daily pan evaporation values. Kisi *et al.* (2012) developed and validated a generalized neuro-fuzzy based model for estimating daily pan evaporation values using weather data. Pour Ali Baba *et al.* (2013) applied ANFIS model for estimating ET_0 using available and estimated climatic variables. Kim *et al.* (2013) applied ANN and ANFIS techniques for modeling daily pan evaporation values using different lag-time patterns to analyze the effect of the data time series on modeling accuracy.

The assessment of AI models for ET_0 estimation is usually based on a single assignment of the data sets required for the application of the training algorithm and for testing. As stated by Martí *et al.* (2011a), this evaluation of the model performance, based on a single and limited test pattern collection, might be misleading or partially valid. In order to perform a suitable assessment of the AI model the leave-one-out (Stone 1974; Shao 1993; Mehrotra & Sharma 2005; Hrachowitz *et al.* 2010) based data set configurations offer a good solution. According to them, a complete testing scan of the data set can be fulfilled based on a previously defined minimum test set size or a maximum number of train-test stages. The most rigorous approach would be to leave a single pattern for testing (leave-one-out). However, this would involve too high computational costs; so, usually a larger test size is considered (*k*-fold testing). Through the *k*-fold test (where k refers to the number of necessary train-test stages) the data set is scanned according to the established test size or, conversely, according to the defined number of train–test stages. In addition to the single data set assignment, most AI approaches dealing with ET_0 estimation consider a local calibration of the models, i.e., models are trained and tested using data from the same stations. Only few studies have tackled the external performance of an AI model, i.e., when the test patterns belong to a station not considered for training (Kisi 2007; Martí & Gasque 2010; Martí *et al.* 2010, 2011a, b; Shiri *et al.* 2011, 2013a; Kisi *et al.* 2012; Pour Ali Baba *et al.* 2013).

However, these did consider neither leave-one-out procedures nor *k*-fold testing.

Recently, Shiri *et al.* (2013b) compared local and external training procedures of gene expression programming (GEP) models for estimating pan evaporation values in six stations in the USA using various input configurations based on physical/empirical pan evaporation estimation models. Results suggested that external training might be a valid alternative to local training if the models were fed with a suitable combination of inputs. This work aims at application of that data management scenario using ANFIS and ANN models for estimating ET_0 in a different climatic context, Iran. Therefore, two input configurations were defined and the data set was split up in several training and testing configurations according to temporal and spatial criteria.

MATERIALS AND METHODS

Data set and input combinations

In the present paper, daily weather data from five weather stations in Iran were used for modeling ET_0. The data sample consisted of daily maximum and minimum air temperatures (T_{max} and T_{min}, respectively), relative humidity (RH), wind speed (S_W) at 2 m above land surface, and solar radiation (R_S) covering a period of 6 years (from January 2003 to December 2008). The complete information of the studied weather stations as well as the average values of the considered meteorological parameters is given in Table 1. Also, Table 2 sums up the temporal variations of ET_0 values in the studied stations. From both tables the spatial and temporal variations of the ET_0 values (in terms of global average and standard deviation values) are observed. Due to the absence of experimental measurements, ET_0 values calculated according to FAO56-PM equation were considered as the targets for the ANFIS/ANN implementation and evaluation, which is an accepted and very common practice in this situation, in agreement with the FAO recommendation (Allen *et al.* 1998). The considered input combinations used to feed the ANFIS/ANN models in the present paper are

(1) T_{max}, T_{min}, T_{mean}, R_a (ANFIS1, ANN1)
(2) T_{mean}, R_S, RH (ANFIS2, ANN2)

where R_a is the extraterrestrial radiation (mm/day). These inputs are required for the application of the well-known Hargreaves–Samani (Hargreaves & Samani 1985) and Turc (Turc 1961) empirical models, respectively. Input combination (1) was selected because it only requires temperature records. Input combination (2) might allow assessment of the combined effect of RH and R_S on ET_0 in comparison to (1). It must be pointed out that the higher accuracy of AI methods over their corresponding empirical equations has been widely demonstrated (e.g., Landeras *et al.* 2008; Shiri *et al.* 2012, 2013a; Pour Ali Baba *et al.* 2013). Thus, the mentioned empirical equations will not be assessed in this study.

The obtained results are demonstrated for two time periods: (a) for the complete study period (including the

Table 1 | Summary of the studied weather stations (locations and meteorological parameters)

	UTM coordinate			Meteorological parameters					
Station	φ (°N)	τ (°E)	z (m)	T_{mean} (°C)	ΔT (°C)	RH (%)	S_W (m/s)	R_S (MJ/m²d)	ET_0 (mm/day)
Bojnurd	37.28	57.16	1,112	13.5	12.9	87.9	2.4	16.7	3.1
Quazvin	36.15	50.3	1,279.2	14.5	14.4	78.3	1.9	18.3	3.8
Shiraz	29.32	52.36	1,484	18.4	16	60.1	2.2	20.3	4.1
Tehran	35.41	51.19	1,190.8	18.6	10	73.8	3.2	15.3	3.4
Zanjan	36.41	48.29	1,665	11.5	14	63.1	3.2	16.6	3.1
Average				15.3	13.46	72.64	2.58	17.44	3.5
Standard deviation				3.11	2.23	11.34	0.59	1.92	0.44

φ: latitude; τ: longitude; z: altitude with respect to sea level; T_{mean}: daily mean air temperature; ΔT: difference between maximum and minimum air temperature; RH: relative humidity; S_W: wind speed at 2 m height above ground surface; R_S: solar radiation; ET_0: daily FAO56-PM evapotranspiration.

Table 2 | Temporal variations of ET_0 values per test year of each station

	Bojnurd	Quazvin	Shiraz	Tehran	Zanjan
Annual ET_0 (mm/year)					
2003	1,013.13	1,338.25	1,499.43	1,204.93	1,441.52
2004	1,134.51	1,341.50	1,526.98	1,220.14	1,437.47
2005	1,008.16	1,357.14	1,473.85	1,216.15	1,438.52
2006	1,181.49	1,409.29	1,487.29	1,271.38	1,418.93
2007	1,120.45	1,331.19	1,466.74	1,231.44	1,351.84
2008	1,130.24	1,450.56	1,535.58	1,302.59	1,460.82
Average	1,097.99	1,371.32	1,498.31	1,241.10	1,424.85
Standard deviation	70.88	47.98	28.03	37.84	38.16

annual results of the applied models); and (b) for the warmest period (including the time period between May and August). The time period (b) was selected since the study of the performance of the models during the warmest period of the year is crucial from irrigation and water allocation points of view.

Adaptive neuro-fuzzy inference system (ANFIS)

There are two approaches for FISs, namely the approach of Mamdani (Mamdani & Assilian 1975) and the approach of Sugeno (Takagi & Sugeno 1985). The neuro-fuzzy model used in this study implements Sugeno's fuzzy approach to obtain the values of the output variable from those of the input variables. In the implementation of fuzzy logic, several types of membership functions (MFs) (the curves that define how each pattern in the input variables is mapped to a degree of membership between 0 and 1) can be used. However, recent studies have shown that the type of MF does not affect the results fundamentally (Vernieuwe *et al.* 2005). In the present study, the triangular MFs were used, as they are commonly used for practical applications (Russel & Campbell 1996). The number of MFs was determined iteratively. Here, two or three triangular MFs were sufficient for establishing the models based on trial and error. A large number of MFs of input variables should be avoided for the sake of saving time and computational efforts (Keskin *et al.* 2004). The hybrid optimization method (the combination of least squares and back-propagation gradient descent methods) was used for training the membership functions parameters to emulate the calibration (train) data. The grid partitioning identification methods of the Sugeno FIS models are applied for mapping the nonlinear relationship among the input–output variables. Here the input variables of the ANFIS models are the weather data and the output layer corresponds to the ET_0 values. The grid partitioning method proposes independent partitions of each antecedent variable through defining the membership functions of all antecedent variables. The ANFIS approaches were implemented using MATLAB. Figure 1 illustrates a schematic representation of the ANFIS model for ET_0 estimation using weather data.

As a simple example a FIS with two inputs x_1 and x_2 and one output y is explained. Here, x_1 and x_2 might correspond, for instance, to mean air temperature T_{mean} and solar radiation R_S, while the output y would represent the reference evapotranspiration (ET_0). Suppose that the rule base contains two fuzzy IF-THEN rules

Rule 1: IF x_1 is A_1 and x_2 is B_1, THEN

$$y = p_1x_1 + q_1x_2 + r_1 \tag{1}$$

Rule 2: IF x_1 is A_2 and x_2 is B_2, THEN

$$y = p_2x_1 + q_2x_2 + r_2 \tag{2}$$

in which the IF (antecedent) part is fuzzy in nature, while the THEN (consequent) part is a crisp function of an antecedent variable (as a rule, a linear equation). Applied on the above example for mean air temperature (T_{mean}) and solar radiation (R_S), Equations (1) and (2) read as follows:

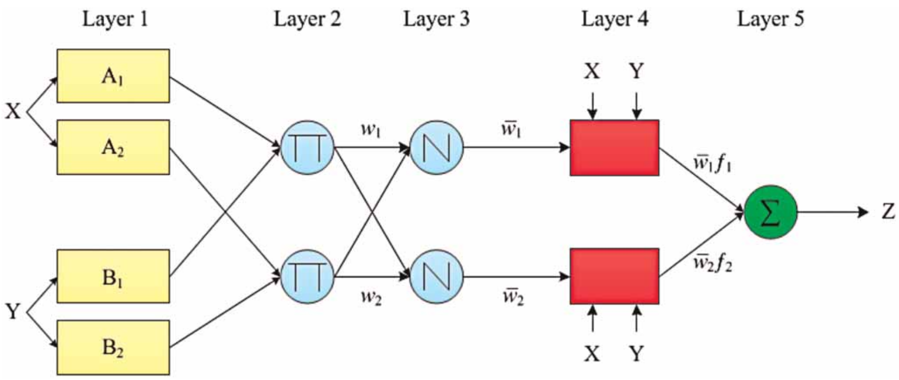

Figure 1 | Schematic ANFIS representation.

Rule 1: IF T_{mean} is LOW and R_S is LOW, THEN

$$ET_0 = p_1 T_{\text{mean}} + q_1 R_s + r_1$$

Rule 2: IF T_{mean} is HIGH and R_S is MEDIUM, THEN

$$ET_0 = p_2 T_{\text{mean}} + q_2 R_s + r_2$$

A common rule set may have n inputs and m IF-THEN rules and can be expressed as

$$y = k_i x_1 + l_i x_2 + \ldots\ldots p_i x_{n-1} + q_i x_n + r_i \quad (3)$$

where $k_i, l_i, \ldots p_i, q_i$ and r_i are parameters with $i = 1, 2, 3, \ldots, m$ corresponding to Rule 1, Rule 2, Rule 3, …, Rule m. The corresponding equivalent ANFIS architecture is represented in Figure 2. The node function in the same layer of the same function family is described as follows (Jang 1993).

Layer 1: Every node i in this layer is an adaptive node with node function given by

$$O_i^1 = \mu_A(T_{\text{mean}}) \quad (4)$$

where T_{mean} is the input to the ith node and μ is the membership function of A_i which is a linguistic label (such as HIGH, or LOW) associated with this node function. A similar equation as Equation (4) may be considered for the input R_S.

The node function is the membership function of A_i and specifies the degree to which the given input T_{mean} (or R_S) satisfies the quantifier A_i. The membership function for A_i is usually described by bell-functions, such as

$$\mu_A(T_{\text{mean}}) = \frac{1}{1 + [(T_{\text{mean}} - C_i)/a_i]^{2b}} \quad (5)$$

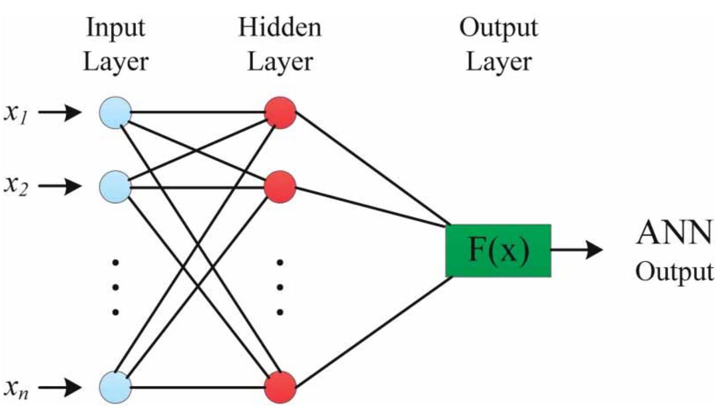

Figure 2 | Schematic ANN representation.

or

$$\mu_A(T_{\mathrm{mean}}) = \exp\left\{-\left(\frac{T_{\mathrm{mean}} - C_i}{a_i}\right)^2\right\} \tag{6}$$

where $\{a_i, b_i, c_i\}$ is the parameter set and μ is the membership function of A_i. As the values of these parameters change, the bell-shaped function varies accordingly, thus exhibiting various forms of membership functions depending on the linguistic label A_i. In fact, any continuous and piecewise differentiable functions, such as commonly used triangular or trapezoidal membership functions, are also qualified candidates for the node function in this layer. Parameters in this layer are referred to as premise parameters.

Layer 2: This layer consists of circle nodes labeled TT which multiply incoming signals and sending the product out. For instance:

$$O_i^2 = w_i = \mu_A(T_{\mathrm{mean}})\mu_B(R_S)\ i = 1, 2. \tag{7}$$

Each node output represents the firing strength of a rule.

Layer 3: In this layer, the circle nodes labeled N, calculate the ratio of the ith rule firing strength to the sum of all rule firing strengths

$$O_i^3 = \bar{w}_i = \frac{w_i}{w_1 + w_2}, \quad \text{for } i = 1, 2 \tag{8}$$

The outputs of this layer are referred to as normalized firing strengths.

Layer 4: All of the nodes in this layer are adaptive with a node function

$$O_i^4 = \bar{w}_i y_i = w_i(P_i T_{\mathrm{mean}} + q_i R_S + r_i) \tag{9}$$

where $\bar{w}_i$ is the output of layer 3, and $\{p_i, q_i, r_i\}$ is the parameter set. Parameters in this layer are called consequence parameters.

Layer 5: The single circle node of this layer, labeled Σ, computes the overall outputs as the summation of all incoming signals:

$$O_i^5 = \sum_{i=1}^{i} \bar{w}_i y_i = \frac{\sum_{i=1}^{i} w_i y_i}{\sum_{i=1}^{i} w_i} \tag{10}$$

Thus, an adaptive network which is functionally equivalent to a Type 3 FIS has been constructed. Detailed information about ANFIS may be found, for example, in Jang (1993).

Artificial neural networks

ANNs are basically parallel information-processing systems. The internal architecture of ANNs is similar to the structure of a biological brain with a number of layers of fully interconnected nodes or neurons. Each neuron is connected to other neurons by means of direct communication links, each with an associated weight. The neural network usually has two or more layers of neurons in order to process nonlinear signals. The input layer admits the incoming information, which is processed by the hidden layer(s), and the output layer presents the network result. During the learning process, the weights of the interconnections and the neural biases are adjusted in trial and error procedures, to minimize the errors. Figure 2 illustrates a schematic ANN structure. Multilayer feed forward networks with radial basis functions were used for modeling ET_0. The basic details and concepts of the working of an ANN can be found in Bishop (1995) or Haykin (1999).

K-fold testing

To perform a suitable assessment of the model performance, the data set was scanned in several successive training–test stages, ensuring that all the patterns were tested. The assignment of the required train and test set configurations was based on the k-fold testing approach, defining previously a minimum assumable test set size. Additionally, the test set was defined according to two different criteria, a spatial/external (S) and a temporal/local (T). Accordingly, the S-test was performed as follows. In each stage, the data set of one station was reserved for testing, whereas the data sets of the remaining four stations were used for the application of the training algorithm. Thus, the S-testing involved five train–test stages (five-fold testing), including one per test station. In this case, 7,702 patterns were available for training, while 2,192 patterns were used for testing. On the other hand, a T-test was performed per station. Namely, defining a minimum test size of 1 year, in each stage of the T-test, the annual data of one station were reserved for testing, while the other 5 years of that station were used for training. Accordingly, six train–test stages (six-fold test), one per year, were necessary to fulfill the temporal leave-one-out per station. In each stage, 1,825 patterns were available for training, while 365 patterns were used for testing. Thus, 30 (5×6) train–test stages were required to perform the T-test approach.

Statistical indices

Four statistical evaluation parameters were used to assess the models' performances

(1) The coefficient of determination (r^2)

$$r^2\left(\frac{\sum_{i=1}^{n}(ET_{i0}-ET_{0\text{mean}})(ET_{iM}-ET_{M\text{mean}})}{\sqrt{\sum_{i=1}^{n}(ET_{i0}-ET_{0\text{mean}})^2\sum_{i=1}^{n}(ET_{i0}-ET_{M\text{mean}})^2}}\right)^2 \tag{11}$$

(2) The scatter index (SI)

$$SI=\frac{\text{RMSE}}{ET_{0\text{mean}}}=\frac{\sqrt{(1/n)\sum_{i=1}^{n}(ET_{iM}-ET_{i0})^2}}{ET_{0\text{mean}}} \tag{12}$$

(3) The Nash–Sutcliffe coefficient (NS)

$$NS=1-\frac{\sum_{i=1}^{n}(E_{iM}-E_{i0})^2}{\sum_{i=1}^{n}(E_{i0}-ET_{0\text{mean}})^2} \tag{13}$$

(4) The mean absolute error (MAE)

$$MAE=\frac{\sum_{i=1}^{n}|ET_{i0}-ET_{iM}|}{n} \tag{14}$$

where ET_{iM} and ET_{io} denote the values generated by different models and FAO56-PM equation at the ith time step, respectively. Also, $ET_{0\text{mean}}$ and $ET_{M\text{mean}}$ represent the mean ET values of FAO56-PM and other applied models, respectively. n is the number of time steps (here, total number of days in each case). The perfect value of r^2 index is 1, representing the best fit of observed versus simulated values, but it cannot be alone used as the unique performance evaluation index since it is sensitive to outliers (Legates & McCabe 1999). Nevertheless, r^2 does not consider biases in mean and standard deviation that may be affecting the result. Therefore, it is better to apply other indices along with r^2, e.g., weighted root mean square error (RMSE) (expresses RMSE as a percentage of mean observed value, here called SI), MAE, or NS. The combined use of these indices can give a good insight about the performance of the applied models.

RESULTS AND DISCUSSION

Table 3 presents the global average statistical parameters (for the annual and seasonal periods) obtained for the two neuro-fuzzy and neural network approaches using local (T) and external (S) training, respectively. For instance, the parameters of the first row correspond to the average of the six-fold temporal test (one per year) in the five stations (i.e., 30 models). Similarly, the second row presents the parameters of the five-fold spatial test (one per station), i.e., the predicted vectors are put together and the statistical indexes are calculated for the global estimation vector. Subsequently, the first row of the second subsection (average seasonal results) corresponds to the average of the six-fold temporal test (one per year) in the five stations, during the warmest period of each year.

As could be expected, the most accurate estimations correspond to locally trained ANFIS2 and ANN2, followed by those of locally trained ANFIS1 and ANN1, externally trained ANFIS2 and ANN2, and externally trained ANFIS1 and ANN1, respectively. Thus, according to the SI values, the consideration of RH and R_S improves the global average accuracy in 0.02 in the locally trained models, and in 0.03 in the externally trained models. The combined effect of RH and R_S on ET_0 will be analyzed per station in the next paragraph. The accuracy decrease in SI terms derived from considering external training instead of local training is 0.06 for ANFIS1, 0.04 for ANN1, 0.05 for ANFIS2, and 0.03 for ANN2, respectively. Thus, ANFIS2 and ANN2 models (R_S–RH-based models) might allow for a higher spatial generalizability, due to a more suitable input–output mapping. However, the most interesting results can be stated when comparing external

Table 3 | Global average performance parameters

Model	Inputs	Training	*MAE* (mm/day)	*SI*	*NS*	r^2
Average annual results						
Input combination (i)						
ANFIS1	$T_{\mathrm{mean}}, T_{\mathrm{max}}, T_{\mathrm{min}}, R_a$	Local	0.35	0.12	0.945	0.947
		External	0.47	0.18	0.920	0.945
ANN1	$T_{\mathrm{mean}}, T_{\mathrm{max}}, T_{\mathrm{min}}, R_a$	Local	0.35	0.13	0.942	0.954
		External	0.45	0.17	0.917	0.945
Input combination (ii)						
ANFIS2	$T_{\mathrm{mean}}, R_s, RH$	Local	0.30	0.10	0.959	0.957
		External	0.39	0.15	0.940	0.950
ANN2	$T_{\mathrm{mean}}, T_{\mathrm{max}}, T_{\mathrm{min}}, R_a$	Local	0.32	0.11	0.958	0.954
		External	0.39	0.14	0.831	0.951
Average seasonal results						
Input combination (i)						
ANFIS1	$T_{\mathrm{mean}}, T_{\mathrm{max}}, T_{\mathrm{min}}, R_a$	Local	0.43	0.09	0.943	0.743
		External	0.66	0.14	0.892	0.710
ANN1	$T_{\mathrm{mean}}, T_{\mathrm{max}}, T_{\mathrm{min}}, R_a$	Local	0.45	0.07	0.956	0.744
		External	0.55	0.12	0.915	0.692
Input combination (ii)						
ANFIS2	$T_{\mathrm{mean}}, R_s, RH$	Local	0.37	0.08	0.958	0.748
		External	0.57	0.12	0.911	0.745
ANN2	$T_{\mathrm{mean}}, T_{\mathrm{max}}, T_{\mathrm{min}}, R_a$	Local	0.38	0.06	0.960	0.736
		External	0.41	0.10	0.948	0.745

ANFIS2/ANN2 and local ANFIS1/ANN1 estimations (0.15 vs. 0.12 of SI for ANFIS and 0.14 vs. 0.14 of SI for ANN, respectively). This result reveals that it might be preferable to train external ANFIS2/ANN2 models, rather than local ANFIS1/ANN1 models, because the accuracy of the prediction might be only slightly worse. On the other hand, externally trained models exempt us from the need of data availability in the test station to train a local model, which might be a decisive advantage. The accuracy differences between locally and externally trained models seem to decrease with a suitable input selection allowing for an optimum input–output mapping. The same conclusions can be drawn on the basis of the other performance parameters.

The results of the models during the warmest period (May–August) present a similar trend, where the locally trained ANFIS2/ANN2 models are the most accurate models followed by the locally trained ANFIS1/ANN1 models. The externally trained ANFIS2/ANN2 and ANFIS1/ANN1models are ranked as the third and fourth models (based on their accuracy), respectively. In similarity to the global average annual results, the SI differences between the externally trained ANFIS21/ANN2 models with locally trained AFIS1/ANN1 models are 0.03 and the same conclusions might be drawn as for the annual results.

Figures 3 and 4 present the performance parameters of the ANFIS and ANN models split up per test station for the total and warmest periods, respectively. In each station, the results of the local training (T-parameters) correspond to the global six-fold temporal testing, whereas the S-parameters correspond to each individual stage of the five-fold spatial testing. This analysis allows evaluation of whether locally calibrated models are always more accurate than externally calibrated ones or if this depends on the specific climatic conditions of each station. Attending to temperature-based models (ANFIS1/ANN1), local calibration is always more accurate, except in stations 2 and

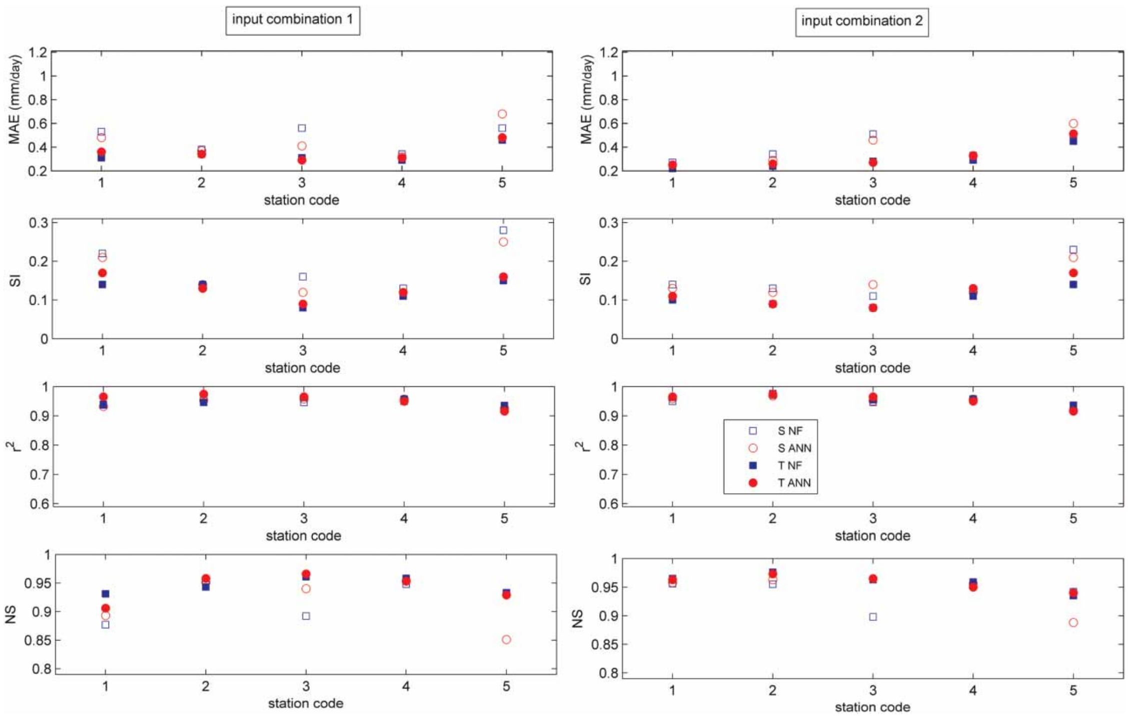

Figure 3 | External vs. local performance split up per station for total data period (station codes 1–5 corresponding to Bojnurd, Quazvin, Shiraz, Tehran, and Zanjan stations, respectively).

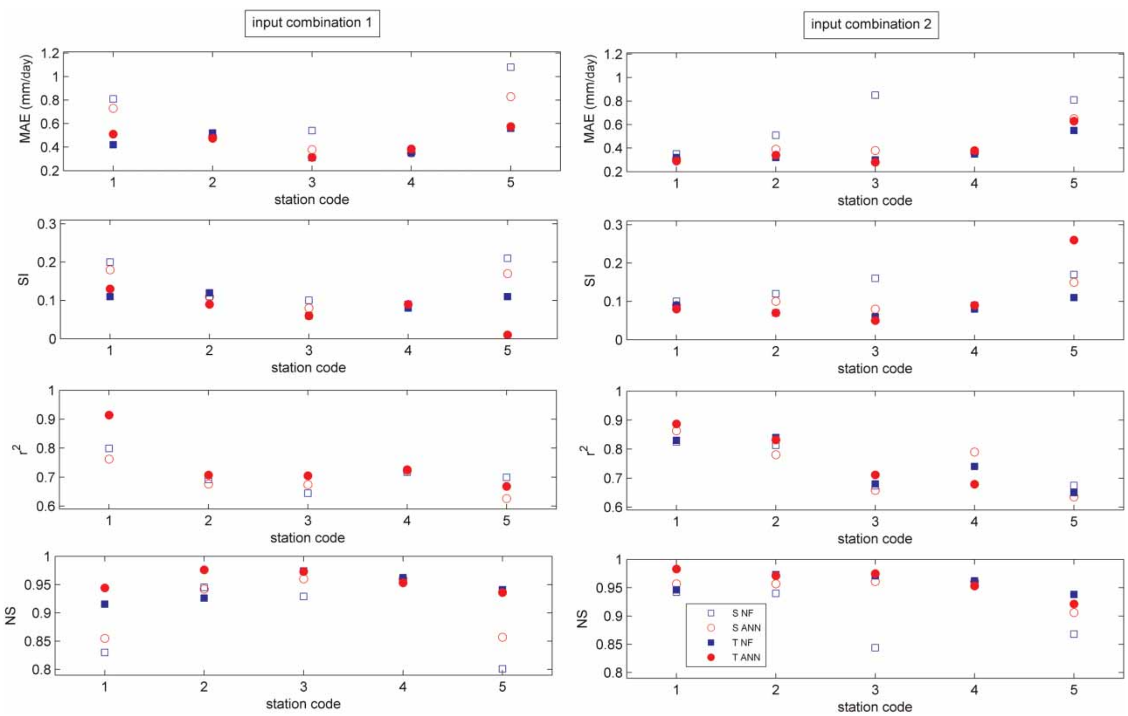

Figure 4 | External vs. local performance split up per station for the warmest period (station codes 1–5 corresponding to Bojnurd, Quazvin, Shiraz, Tehran, and Zanjan stations, respectively).

4, where both local and external approaches present similar accuracy. In these stations, although being locally trained, there might be considerable fluctuations in the variables' ranges throughout the studied years. Or conversely, the input–output mapping encountered for the external training stations is also suitable for the testing station patterns. In stations 1, 3, and 5 the testing patterns might be too different from the training patterns in externally/spatially calibrated models, and the encountered input–output mapping might allow no generalization in those testing stations for that input combination. The SI differences between the T- and S-approach in these stations are 0.08, 0.08, and 0.13 (for ANFIS1) and 0.04, 0.03, and 0.09 (for ANN1), respectively. Attending to the ANFIS2/ANN2 models, the performance parameters present a qualitatively similar trend, i.e., T-approaches are more accurate. However, the performance of the S-models is more accurate and less fluctuating than for ANFIS1/ANN1 (average SI of 0.125 vs. 0.16 for ANFIS, and 0.12 vs. 0.14 for ANN, respectively), excluding station 5, which presents, comparatively, a very high error (SI of 0.23 vs. 0.28, and 0.21 vs. 0.28 for ANFIS and ANN models, respectively). According to these results, the consideration of RH and R_S in addition to temperature allows for an improvement of the generalizability in stations 1, 3, 4, and 5, due to a more suitable input–output mapping. Further, station 5 might present different climatic patterns or variables' ranges than the other stations because of presenting a clearly higher altitude and, therefore, sufficient accurate estimations are not achieved. The risk that the input–output mapping defined from the training patterns is not valid for the testing patterns increases the lower the number (and/or the level of significance) of considered inputs. Nevertheless, the consideration of a wider and more representative training set might be enough to deal with this limitation. Attending to the local approaches, the consideration of RH and R_S in addition to T_{mean} only involves an improvement in stations 1 and 2 (0.04 and 0.05 of SI lower for ANFIS, and 0.06 and 0.04 of SI lower for ANN, respectively). Thus, the local performance of the models in stations 3, 4, and 5 is unaffected by RH and R_S, or, conversely, temperature-based model allows for a suitable mapping between the training and testing patterns.

Comparing the performance of T-ANFIS1/T-ANN1 and S-ANFIS2/S-ANN2 estimations per station, Figures 3 and 4 show that the locally trained models are more accurate in stations 3 (by 0.03 of SI), 4 (by 0.01 of SI), and 5 (by 0.08 of SI), whereas externally trained models are more accurate in station 2 (by 0.01 of SI). In station 1 (Bojnurd), both approaches present the same SI. Neglecting station 5 (Zanjan), where the inaccuracy of the estimation might be dealt with using a more representative training set, these results suggest that it might be preferable to train externally a model relying on a suitable input combination than a local model. The estimation accuracy might be similar or only slightly worse, and we would not require local data series in the test stations for training a local model. Finally, considering stations where RH and R_S might be more significant in the estimation of ET_0, the accuracy decrease of S-approaches in comparison to T-approaches might be even lower.

Tables 4–7 present the performance parameters of the locally trained ANFIS and ANN models split up per station and test year. Attending to the SI values of Bojnurd in the tables, a high variability can be stated throughout the considered period. ANFIS1 performance ranges between 0.05 (2007 and 2008) and 0.24 (2003), ANFIS2 performance ranges between 0.03 (2008) and 0.17 (2003), ANN1 performance ranges between 0.05 (2008) and 0.32 (2004), and ANN2 performance ranges between 0.05 (2008) and 0.15 (2003). Similarly, the SI performance of ANFIS1 in Quazvin ranges between 0.12 (2005) and 0.2 (2003). This variability can be linked to the relationships/differences between the training and testing patterns. The key point in this regard is that the consideration of a single data set assignment for testing and for the application of the training algorithm would have not allowed a suitable assessment of the model performance, leading to partially valid conclusions. Therefore, the application of this single data set assignment procedure, which is a very common practice, should be questioned. Leaving out one procedure, or at least k-fold testing if the computational costs of the former one are not assumable, can be a good choice to face this limitation. Regarding the performance in Shiraz, Tehran, and Zanjan, throughout the considered period a lower variability can be stated. As mentioned, this might be caused by a lower climatic variability within the considered years.

The average maximum SI difference is smaller for ANFIS2 than for ANFIS1 (0.048 vs. 0.066, respectively).

Table 4 | Temporal six-fold testing parameters of ANFIS 1 models

	Bojnurd	Quazvin	Shiraz	Tehran	Zanjan
MAE (mm/day)					
2003	0.47	0.39	0.40	0.31	0.50
2004	0.39	0.40	0.33	0.30	0.47
2005	0.27	0.33	0.28	0.31	0.45
2006	0.48	0.36	0.25	0.29	0.44
2007	0.13	0.33	0.27	0.26	0.49
2008	0.13	0.37	0.30	0.29	0.46
SI					
2003	0.24	0.20	0.10	0.12	0.16
2004	0.19	0.16	0.10	0.12	0.15
2005	0.11	0.12	0.08	0.12	0.15
2006	0.22	0.13	0.08	0.11	0.14
2007	0.05	0.13	0.08	0.10	0.16
2008	0.05	0.13	0.09	0.11	0.15
NS					
2003	0.869	0.894	0.943	0.956	0.921
2004	0.907	0.938	0.951	0.952	0.936
2005	0.967	0.960	0.968	0.954	0.934
2006	0.864	0.950	0.972	0.960	0.945
2007	0.990	0.957	0.969	0.964	0.924
2008	0.991	0.956	0.965	0.959	0.935
r^2					
2003	0.869	0.898	0.946	0.956	0.924
2004	0.916	0.938	0.953	0.951	0.938
2005	0.990	0.962	0.970	0.956	0.934
2006	0.870	0.961	0.972	0.960	0.946
2007	0.990	0.959	0.970	0.964	0.928
2008	0.992	0.960	0.965	0.960	0.935

Table 5 | Temporal six-fold testing parameters of ANFIS 2 models

	Bojnurd	Quazvin	Shiraz	Tehran	Zanjan
MAE (mm/day)					
2003	0.32	0.26	0.33	0.31	0.48
2004	0.28	0.24	0.32	0.31	0.46
2005	0.20	0.22	0.25	0.30	0.45
2006	0.32	0.24	0.26	0.29	0.42
2007	0.10	0.23	0.26	0.28	0.47
2008	0.10	0.25	0.30	0.27	0.46
SI					
2003	0.17	0.10	0.10	0.12	0.15
2004	0.13	0.09	0.10	0.12	0.15
2005	0.11	0.08	0.07	0.12	0.15
2006	0.15	0.10	0.08	0.11	0.13
2007	0.04	0.09	0.08	0.10	0.16
2008	0.03	0.09	0.09	0.10	0.15
NS					
2003	0.935	0.972	0.950	0.957	0.930
2004	0.957	0.976	0.952	0.950	0.936
2005	0.970	0.979	0.968	0.956	0.932
2006	0.940	0.975	0.973	0.960	0.949
2007	0.994	0.976	0.969	0.963	0.925
2008	0.994	0.978	0.967	0.965	0.936
r^2					
2003	0.935	0.972	0.909	0.956	0.933
2004	0.958	0.976	0.944	0.950	0.938
2005	0.956	0.979	0.971	0.956	0.933
2006	0.942	0.976	0.973	0.960	0.950
2007	0.994	0.977	0.969	0.963	0.929
2008	0.994	0.978	0.965	0.966	0.936

Thus, the consideration of RH and R_S in addition to T_{mean} might allow for a better input–output mapping, and as a result, a lower variability within the annual performances. However, regarding the performance of both input combinations in Shiraz, Tehran, and Zanjan, it seems that RH and R_S might not be so decisive in these stations when performing a local training. It can be observed that both input combinations provide the same accuracy except in 2005 (Shiraz), 2008 (Tehran), and 2003, 2006 (Zanjan), where ANFIS2 is slightly more accurate. These results confirm the conclusions of Shiri *et al.* (2013b) on modeling pan evaporation using GEP. Hence, based on a suitable input choice, the external training might be a valid alternative to local training. Although this procedure might provide slightly less accurate estimations, it presents the decisive advantage of not requiring data series in the test stations to train a local model. So, only climatic measurements for the testing points would be required.

Figures 5 and 6 display the scatterplots of the externally trained ANFIS and ANN models. Comparison of the fit line equations and r^2 values in the scatterplots indicates that the ANFIS2/ANN2 models perform better than the ANFIS1/ANN1 models for the Bojnurd, Quazvin, Tehran, and Zanjan stations. For the Shiraz station, however,

Table 6 | Temporal six-fold testing parameters of ANN 1 models

	Bojnurd	Quazvin	Shiraz	Tehran	Zanjan
MAE (mm/day)					
2003	0.46	0.34	0.31	0.29	0.48
2004	0.71	0.36	0.31	0.39	0.50
2005	0.27	0.31	0.27	0.29	0.42
2006	0.48	0.34	0.24	0.28	0.41
2007	0.13	0.33	0.27	0.25	0.47
2008	0.13	0.35	0.35	0.34	0.62
SI					
2003	0.24	0.13	0.09	0.11	0.15
2004	0.32	0.14	0.10	0.15	0.16
2005	0.12	0.12	0.08	0.11	0.14
2006	0.22	0.13	0.07	0.11	0.14
2007	0.06	0.12	0.08	0.10	0.16
2008	0.05	0.12	0.10	0.13	0.20
NS					
2003	0.877	0.960	0.968	0.961	0.934
2004	0.749	0.949	0.957	0.918	0.929
2005	0.967	0.963	0.970	0.961	0.942
2006	0.864	0.959	0.976	0.962	0.952
2007	0.990	0.961	0.972	0.967	0.933
2008	0.992	0.959	0.953	0.949	0.888
r^2					
2003	0.877	0.958	0.961	0.961	0.935
2004	0.747	0.949	0.959	0.919	0.933
2005	0.991	0.964	0.971	0.962	0.942
2006	0.870	0.960	0.976	0.962	0.953
2007	0.991	0.960	0.972	0.967	0.933
2008	0.992	0.960	0.953	0.948	0.889

Table 7 | Temporal six-fold testing parameters of ANN2 models

	Bojnurd	Quazvin	Shiraz	Tehran	Zanjan
MAE (mm/day)					
2003	0.29	0.27	0.06	0.31	0.47
2004	0.28	0.35	0.34	0.40	0.65
2005	0.30	0.22	0.29	0.31	0.43
2006	0.28	0.25	0.26	0.29	0.42
2007	0.23	0.23	0.26	0.27	0.50
2008	0.11	0.24	0.38	0.36	0.61
SI					
2003	0.15	0.10	0.02	0.12	0.15
2004	0.13	0.12	0.11	0.16	0.22
2005	0.15	0.08	0.09	0.12	0.14
2006	0.13	0.11	0.08	0.11	0.13
2007	0.09	0.08	0.08	0.10	0.18
2008	0.04	0.08	0.12	0.13	0.20
NS					
2003	0.950	0.972	0.997	0.958	0.936
2004	0.960	0.960	0.944	0.915	0.855
2005	0.947	0.979	0.965	0.956	0.940
2006	0.954	0.971	0.973	0.967	0.951
2007	0.976	0.976	0.972	0.965	0.915
2008	0.994	0.981	0.941	0.944	0.890
r^2					
2003	0.952	0.972	0.997	0.959	0.937
2004	0.961	0.960	0.945	0.915	0.856
2005	0.951	0.979	0.965	0.956	0.941
2006	0.955	0.975	0.973	0.961	0.953
2007	0.979	0.977	0.973	0.965	0.918
2008	0.994	0.981	0.941	0.945	0.891

externally trained temperature-based ANFIS1/ANN1 models comprising only inputs T_{max}, T_{min}, T_{mean}, R_a seem to be more accurate than the ANFIS2/ANN2 models. Different climatic characteristics (high ΔT and ET_0) of this station may be the reason for this. The underestimations of the ANFIS and ANN models are clearly seen for Zanjan station in Figures 5 and 6. The reason for this may be the fact that this station has higher ET_0 data than the other stations and the externally trained models encounter extrapolation difficulties for the testing patterns. However, adding RH and R_S inputs to the ANFIS and ANN models improves the models accuracies and decreases the underestimations.

The time series plot of the ANFIS and ANN models for the warmest period of year 2003 in Bojnurd station is given in Figure 7, where the overestimations of the externally trained temperature-based models can be clearly stated, especially for the ANN1 model. Locally trained ANN1 and ANFIS1 models significantly underestimate ET_0 of the July period. It is clear from Figure 7 that the externally trained R_S–RH based ANN2 and ANFIS2 models are more accurate than the locally trained temperature based ANN1

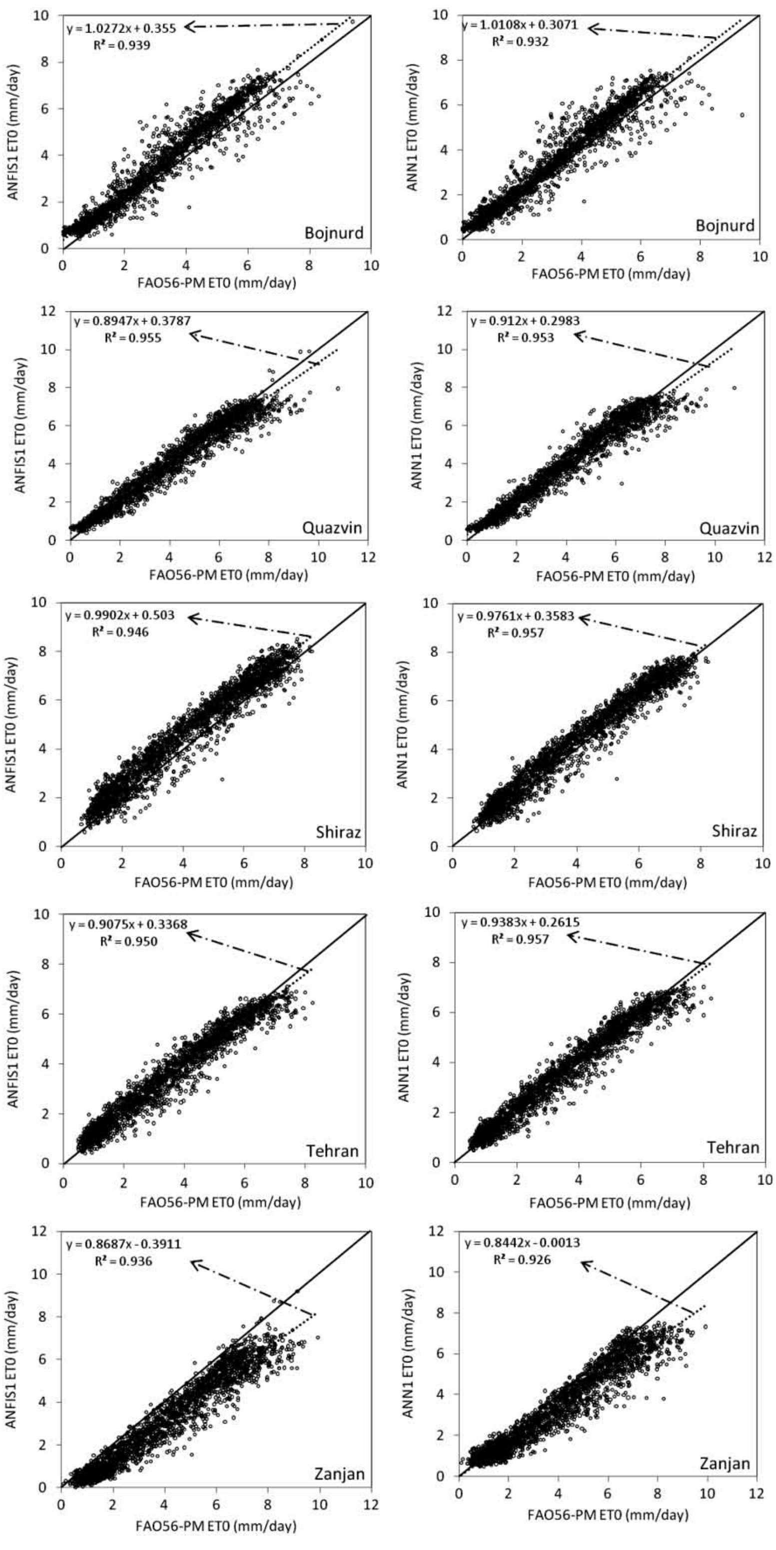

Figure 5 | Scatterplots of the externally trained ANFIS1 and ANN1 models at the test stations.

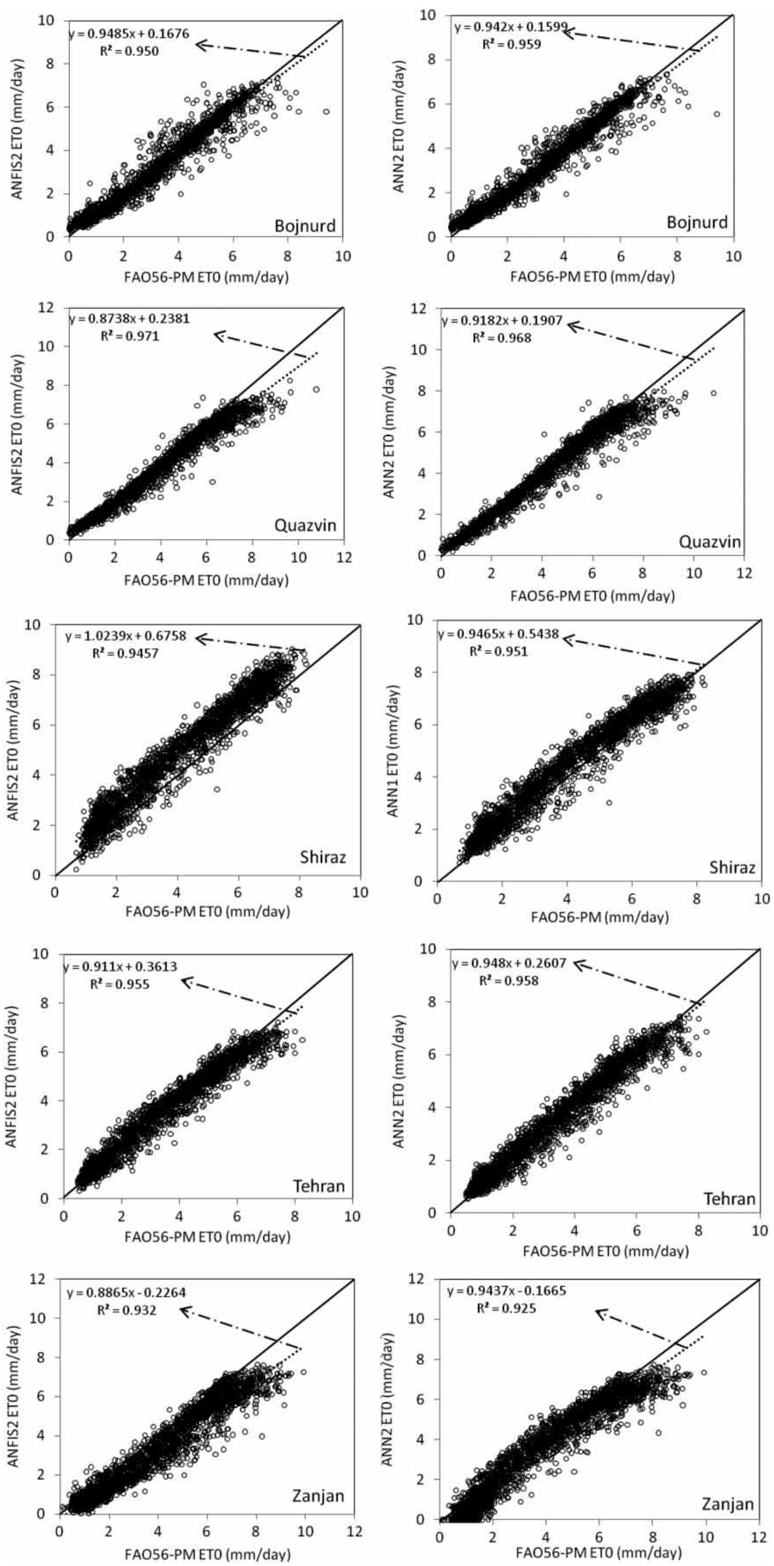

Figure 6 | Scatterplots of the externally trained ANFIS2 and ANN2 models at the test stations.

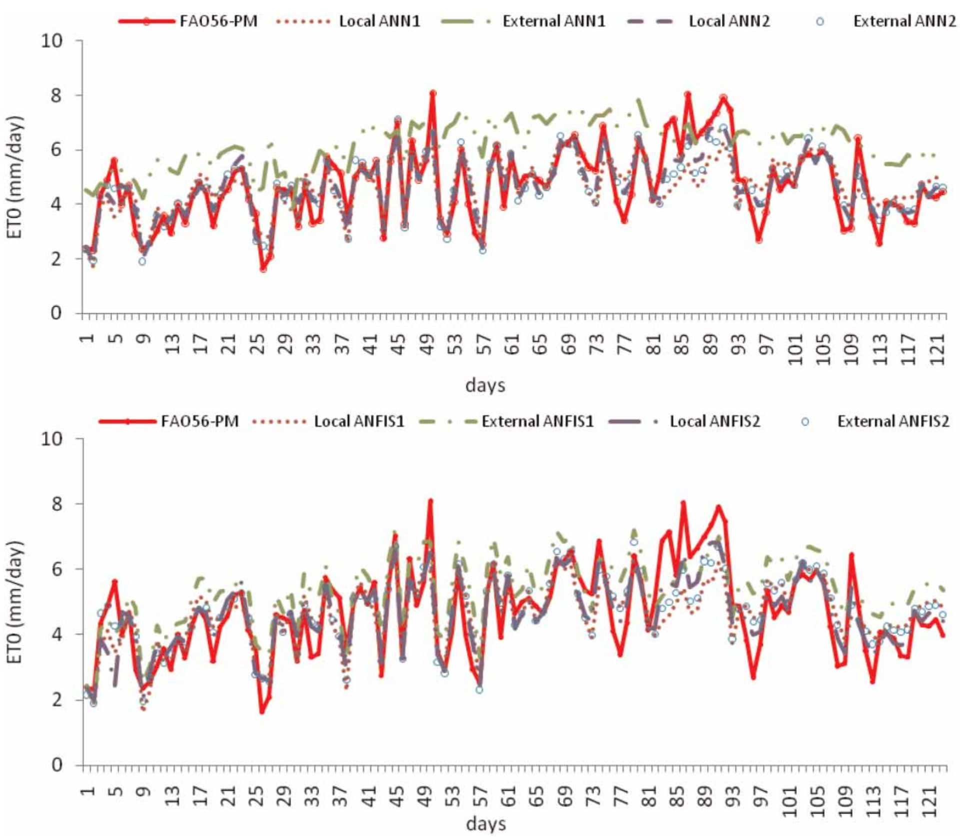

Figure 7 | Time series plot of the applied ET_0 models during the warmest period of the test year 2003 for Bojnurd station.

and ANFIS1 models. As stated before, it might be preferable to train externally a model relying on a suitable input combination than a local model.

The annual variability stated in the models' performance confirms the need to apply data set scanning procedures (LOO, *k*-fold testing) to properly evaluate it, as suggested by Martí *et al.* (2011a, b). Further research might tackle the effect of each climatic input on the target variable and the associated improvement in the generalizability of the model considering a wider number of stations and input combinations based on a data set scanning assessment.

CONCLUSIONS

This paper reports the application of neuro-fuzzy and neural network approaches for estimating reference evapotranspiration considering a *k*-fold testing assessment. Different data set configurations were defined based on temporal and spatial criteria allowing for a complete testing scan of the data set. The proposed methodology enabled the comparison in each station of models trained with local data series and models trained with data series from the remaining stations, both fed by two different input combinations. Results show that the external training (i.e., using data series not belonging to the test station) based on a suitable input choice and a representative pattern collection might be a valid alternative to the more common local training. Although, comparatively, a slight accuracy decrease might be expected, this procedure presents the decisive advantage of not requiring data series in the test station for the application of the training algorithm. Only climatic measurements for the testing inputs would be required. Results also show that AI applications based on a single data assignment of the training and test sets might not be advisable for a proper evaluation of the model performance, as they might lead to only partially valid conclusions, i.e., for the single specific training and test sets defined, which might only cover a part of the complete patterns range spectrum of the considered data set.

ACKNOWLEDGEMENT

This work has been prepared as a part of Jalal Shiri's PhD thesis.

REFERENCES

Allen, R. G., Pereira, L. S., Raes, D. & Smith, M. 1998 Crop evapotranspiration. Guide lines for computing crop evapotranspiration. FAO Irrigation and Drainage Paper No. 56, Rome, Italy.

ASCE Task Committee 2000 Artificial neural networks in hydrology. II: hydrological applications. *ASCE J. Hydrolog. Eng.* **5** (2), 124–137.

Bishop, C. M. 1995 *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 504 pp.

Cigizoglu, H. K. & Kisi, O. 2005 Flow prediction by three back propagation techniques using k-fold partitioning of neural network training data. *Nordic Hydrol.* **36** (1), 49–64.

Cigizoglu, H. K. & Kisi, O. 2006 Methods to improve the neural network performance in suspended sediment estimation. *J. Hydrol.* **317**, 221–238.

Doorenbos, J. & Pruitt, W. O. 1977 Crop water requirements. FAO Irrigation and Drainage Paper No. 24, Rome, Italy.

Hargreaves, G. H. & Samani, Z. A. 1985 Reference crop evapotranspiration from temperature. *Appl. Eng. Agric.* **1** (2), 96–99.

Haykin, S. 1999 *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, Upper Saddle River, NJ, 842 pp.

Hrachowitz, M., Soulsby, C., Imholt, C., Malcolm, I. A. & Tetzlaff, D. 2010 Thermal regimes in a large upland salmon river: a simple model to identify the influence of landscape controls and climate change on maximum temperatures. *Hydrol. Process.* **24**, 3374–4339.

Jang, J. S. R. 1993 ANFIS: adaptive-network-based fuzzy inference system. *IEEE Trans. Syst. Man Cyb.* **23** (3), 665–685.

Jang, J. S. R., Sun, C. T. & Mizutani, E. 1997 *Neurofuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. Prentice-Hall, Upper Saddle River, NJ.

Katerji, N. & Rana, G. 2011 Crop reference evapotranspiration: a discussion of the concept, analysis of the process and validation. *Water Resour. Manage.* **25**, 1581–1600.

Keskin, M. E., Terzi, O. & Taylan, D. 2004 Fuzzy logic model approaches to daily pan evaporation estimation in western Turkey. *Hydrol. Sci. J.* **49** (6), 1001–1010.

Kim, S., Kim, J. H. & Park, K. B. 2009 Statistical learning theory for the disaggregation of the climatic data. *Proc. 33rd IAHR Congress 2009*, IAHR, Vancouver, Canada, pp. 1154–1162.

Kim, S., Shiri, J. & Kisi, O. 2012 Pan evaporation modeling using neural computing approach for different climatic zones. *Water Resour. Manage.* **26** (11), 3231–3249.

Kim, S., Shiri, J., Kisi, O. & Singh, V. P. 2013 Estimating daily pan evaporation using different data driven methods and lag-time patterns. *Water Resour. Manage.* **27**, 2267–2286.

Kisi, O. 2006 Daily pan evaporation modeling using a neuro-fuzzy computing technique. *J. Hydrol.* **329**, 636–646.

Kisi, O. 2007 Evapotranspiration modelling from climatic data using a neural network computing technique. *Hydrol. Process.* **21**, 1925–1934.

Kisi, O. 2008 The potential of different ANN techniques in evapotranspiration modeling. *Hydrol. Process.* **22**, 2449–2460.

Kisi, O. & Ozturk, O. 2007 Adaptive neuro-fuzzy computing technique for evapotranspiration estimation. *J. Irrig. Drain. Eng.* **133** (4), 368–379.

Kisi, O. & Shiri, J. 2012 River suspended sediment estimation by climatic variables implication: comparative study among soft computing techniques. *Comput. Geosci.* **43**, 73–82.

Kisi, O., Pour Ali Baba, A. & Shiri, J. 2012 Generalized neuro-fuzzy models for estimating daily pan evaporation values from weather data. *J. Irrig. Drain. Eng.* **138** (4), 349–362.

Kumar, M., Raghuwanshi, N. S. & Singh, R. 2011 Artificial neural networks in evapotranspiration modelling: a review. *Irrig. Sci.* **1**, 11–25.

Landeras, G., Ortiz-Barredo, A. & Lopez, J. J. 2008 Comparison of artificial neural network models and empirical and semi-empirical equations for daily reference evapotranspiration estimation in the Basque Country (Northern Spain). *Agri. Water Manage.* **95**, 553–565.

Landeras, G., Lopez, J. J., Kisi, O. & Shiri, J. 2012 Comparison of gene expression programming with neuro-fuzzy and neural network computing techniques in estimating daily incoming solar radiation in the Basque Country (Northern Spain). *Energy Conver. Manage.* **62**, 1–13.

Legates, D. R. & McCabe, G. J. 1999 Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic validation. *Water Resour. Res.* **35** (1), 233–241.

Mamdani, E. H. & Assilian, S. 1975 An experiment in linguistic synthesis with a fuzzy logic controller. *Int. J. Man Machine Stud.* **7** (1), 1–13.

Martí, P. & Gasque, M. 2010 Ancillary data supply strategies for improvement of temperature-based ET_0 ANN models. *Agri. Water Manage.* **97** (7), 939–955.

Martí, P., Royuela, A., Manzano, J. & Palau, G. 2010 Generalization of ET_0 ANN models through data supplanting. *J. Irrig. Drain. Eng.* **136** (3), 161–174.

Martí, P., González-Altozano, P. & Gasque, M. 2011a Reference evapotranspiration estimation without local climatic data. *Irrig. Sci.* **29** (6), 469–495.

Martí, P., Manzano, J. & Royuela, A. 2011b Assessment of a 4-input artificial neural network for reference evapotranspiration estimation through data set scanning procedures. *Irrig. Sci.* **29** (3), 181–195.

Mehrotra, R. & Sharma, A. 2005 A nonparametric nonhomogeneous hidden Markov model for downscaling of multisite daily rainfall occurrences. *J. Geophys. Res.* **110**, D16108.

Pour Ali Baba, A., Shiri, J., Kisi, O., Fakheri Fard, A., Kim, S. & Amini, R. 2013 Estimating daily reference evapotranspiration using available and estimated climatic data by adaptive neuro-fuzzy inference system and artificial neural networks. *Hydrol. Res.* **44** (1), 131–146.

Russel, S. O. & Campbell, P. F. 1996 Reservoir operating rules with fuzzy programming. *J. Water Resour. Plann. Manage.* **122** (3), 165–170.

Shao, J. 1993 Linear model selection by cross-validation. *J. Am. Statist. Assoc.* **88** (422), 486–494.

Shiri, J. & Kisi, O. 2011a Application of artificial intelligence to estimate daily pan evaporation using available and estimated climatic data in the Khozestan Province (Southwestern Iran). *J. Irrig. Drain. Eng.* **137** (7), 412–425.

Shiri, J. & Kisi, O. 2011b Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. *Comput. Geosci.* **37** (10), 1692–1701.

Shiri, J., Dierickx, W., Pour-Ali Baba, A., Nemati, S. & Ghorbani, M. A. 2011 Estimating daily pan evaporation from climatic data of the state of Illinois, USA using adaptive neuro-fuzzy inference system and artificial neural network. *Hydrol. Res.* **42** (6), 491–502.

Shiri, J., Kisi, O., Landeras, G., Lopez, J. J., Nazemi, A. H. & Stuyt, L. C. P. M. 2012 Daily reference evapotranspiration modeling by using genetic programming approach in the Basque Country (Northern Spain). *J. Hydrol.* **414–415**, 302–316.

Shiri, J., Nazemi, A. H., Sadraddini, A. A., Landeras, G., Kisi, O., Fakheri Fard, A. & Marti, P. 2013a Global cross station assessment of neuro-fuzzy models for estimating daily reference evapotranspiration. *J. Hydrol.* **480**, 46–57.

Shiri, J., Marti, P. & Singh, V. P. 2013b Evaluation of gene expression programming approaches for estimating daily pan evaporation through spatial and temporal data scanning. *Hydrol. Process.* Doi:10.1002/hyp.9669.

Stone, M. 1974 Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. B* **36**, 111–147.

Takagi, T. & Sugeno, M. 1985 Fuzzy identification of systems and its application to modeling and control. *IEEE Trans. Syst. Man Cyber.* **15** (1), 116–132.

Turc, L. 1961 É valuation des besoins en eau d'irrigation, évapotranspiration potentielle. *Ann. Agron.* **12** (1), 13–49.

Vernieuwe, H., Georgieva, O., De Baets, B., Pauwels, V. R. N., Verhoest, N. E. C. & De Troch, F. P. 2005 Comparison of data-driven Takagi-Sugeno models of rainfall-discharge dynamics. *J. Hydrol.* **302** (1–4), 173–186.

First received 17 June 2013; accepted in revised form 20 November 2013. Available online 20 December 2013