Computers and Electronics in Agriculture 162 (2019) 70–81

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Data splitting strategies for improving data driven models for reference evapotranspiration estimation among similar stations

Jalal Shiri[a,*], Pau Marti[b], Sepideh Karimi[a], Gorka Landeras[c]

[a] Water Engineering Department, Faculty of Agriculture, University of Tabriz, Tabriz, Iran
[b] Departament de Biologia, Àrea d'Enginyeria Agroforestal. Universitat de les Illes Balears. Carretera de Valldemossa km 7.5, 07022 Palma, Spain
[c] NEIKER, AB. Basque Country Research Institute for Agricultural Development, Alava, Basque Country, Spain

ARTICLE INFO

Keywords:
Data driven models
Evapotranspiration
Ancillary inputs
Gene expression programming

ABSTRACT

In the last years, different heuristic data driven models have been proposed to estimate reference evapotranspiration (ET_o) with high performance accuracy as an alternative to empirical and physically-based approaches. However, these models, despite their complexity and soundness, rely on finite data series, like the empirical approaches, and their actual practical validity highly depend on the data management adopted in their development and assessment, in particular on the data splitting adopted. A major issue for ensuring a sound assessment of the heuristic model performance is the definition of a suitable criterion for splitting the data series in training and testing data. The present study evaluates new different data set splitting strategies based on the adoption of ancillary external inputs for enhancing the performance of the Gene Expression Programming- based models for estimating ET_o. All models are assessed using k-fold validation considering annual test sizes. The results show that it is preferable to incorporate the external target variable as input to feed the new model, rather than to incorporate the original external input variables of the model. Regarding the external performance of the models, it is crucial to select a suitable training station for each testing station for providing accurate enough estimations. This way, the applicability of such approaches is not limited to local emergency models, but it allows estimating ET_o elsewhere without the need of training previously a local model using local targets. Finally, it is important to select properly which station/s will provide external ancillary ET_o inputs to the training process, because otherwise they introduce noise to the model and decrease their generalizability.

1. Introduction

Reference evapotranspiration (ET_o) represents the evapotranspiration from a hypothesized reference crop [grass] (height 0.12 m, surface resistance 70 s m^{-1} and albedo 0.23) (Doorenbos and Pruitt, 1977; Allen et al., 1998). Accurate estimation of this parameter is crucial in irrigation and water resources engineering, assessing crop production functions, modeling ecosystems and determining the crop actual water requirement, as well as determining water budget especially in the regions with water scarcity.

The Penman-Monteith ET_o model revised by Food and Agriculture Organization of United Nations (FAO), i.e. the FAO56-PM model, is commonly used by agronomists, irrigation engineers and other scientists as the standard model for estimating ET_o, as well as for calibrating and validating other equations (Droogers and Allen, 2002). As the application of this model requires large number of meteorological parameters (e.g., air temperature, relative humidity, solar radiation and wind speed) which are often not available/reliable in many weather stations, especially in developing regions, other empirical and semi-empirical ET_o models have been developed by using fewer meteorological data, e.g. the well-known temperature-based Hargreaves-Samani (Hargreaves and Samani, 1985) [HS], and the radiation-based Priestley-Taylor (Priestley and Taylor, 1972) [PT] models. However, the need for local climatic data for calibrating these models is a major drawback of these empirical models.

As an alternative, in the last years different heuristic data driven models have shown high performance accuracy for estimating ET_o., (e.g. Trajkovic et al., 2004; Kisi and Yildirim, 2005; Kisi and Ozturk, 2007; Yassin et al., 2016). However, those models, despite their complexity and soundness, rely on finite data series, like the empirical approaches, and their actual practical validity highly depend on the data management adopted in their development and assessment, in particular on the data splitting adopted. A major issue for ensuring a sound assessment of the heuristic model performance is the definition

* Corresponding author.
E-mail address: j_shiri2005@yahoo.com (J. Shiri).

https://doi.org/10.1016/j.compag.2019.03.030
Received 1 December 2018; Received in revised form 28 March 2019; Accepted 30 March 2019

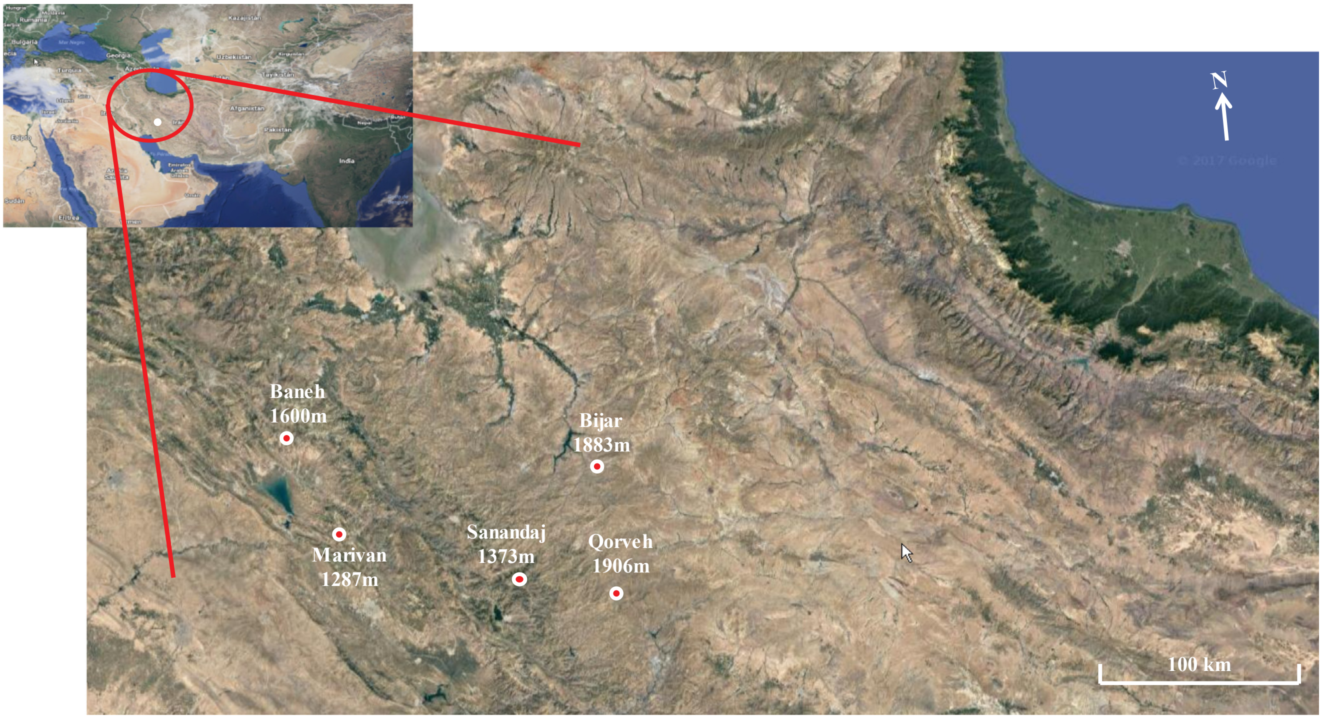


Fig. 1. Geographical positions of the studied locations.

of a suitable criterion for splitting the data series in training and testing data. When tackling the models based on FAO56-PM targets, it should be crucial to distinguish between local and external assessment, as well as between hold out and k-fold validation. In the literature a lot of studies present a simple local assessment. In the local assessment, the models are trained and tested using data from the same station separately, i.e. the models are not evaluated outside the training station. On the other hand, in the external assessment, which is unfortunately less frequent in the literature, the models are tested using data series from stations not considered in the training stage. So, if models are assessed just locally, even if the proposed heuristic models present higher performance accuracy than other approaches, their validity will be restricted to local emergency cases, where the standard method used for providing the targets, normally the FAO56-PM model, cannot be applied (Marti et al., 2015). Indeed, in those studies comparing locally trained models with calibrated or non-calibrated conventional empirical approaches, the heuristic models present the crucial advantage of relying on local patterns and/or on considering more training parameters than calibrated empirical models, where a simple linear calibration is commonly applied. On the other hand, the external assessment should be always considered, because it describes the actual estimation accuracy of a model in those conditions where it is intended to be useful, i.e. requiring no preliminary targets from the testing station for adjusting the model parameters (Shiri et al. 2014).

Further, the hold out validation, also very common in the literature, just considers a single data set assignment, i.e. a single splitting for defining the training and testing series, whereas the k-fold validation involves the definition of several splitting configurations, ensuring a complete scanning of the data set. Accordingly, a k-fold validation should be always desirable, because the conclusions drawn according to a hold-out assessment are only valid for the specific data sets used for testing and training, and might be different for the other patterns not considered. Moreover, as only one training and testing sets are defined, respectively, a minimum testing size of 30% is usually recommended, with the subsequent decrease of available training patterns. Hence, a k-fold validation might allow the definition of larger training set, allowing eventually for a more effective application of the training algorithm and pattern extraction.

Focusing on GEP applications, where the model presents the advantage of being translated into a relative simple expression, Parasuraman et al. (2007), Guven et al. (2008), Izadifar and Elshorbagy (2010), Guven and Kisi (2011), Traore and Guven (2013), and Yassin et al. (2016), among others, applied GEP for estimating ET_o. However, these authors used a holdout procedure. Shiri et al. (2012) compared two criteria for evaluating GEP models based, respectively, on local and pooled data for training-testing, and concluded that GEP is able to approximate ET_o through local, cross-station and pooled scenarios successfully. Also Shiri (2018&2019)) and Shiri et al. (2015) used a k-fold validation for assessing the performance accuracy of GEP-based ET_o models. However, these studies have considered only temporal or spatial data scanning procedures separately. Some studies have proposed to mix local and external patterns according to different criteria in order to improve the generalizability of the ANN models for estimating ET_o (e.g. Martí and Gasque, 2010, Martí et al., 2011a, b). However, these were still not applied using other heuristic models. Further, there are still pending different new criteria for handling the combined use of local and external patterns during the training and testing of the models.

Martí et al. (2011a, b) argued the use of ancillary data supply strategy for modeling ET_o magnitudes in target stations. The present study aims at assessing new approaches for splitting the dataset and incorporating external ancillary inputs for enhancing the performance of data driven models for estimating ET_o. GEP models were applied for assessing the performance accuracy differences derived from these splitting procedures. The accuracy of these models was compared with the accuracy of GEP handled with the conventional splitting commonly used for the application of data driven models. Further, two very common input combinations, namely the temperature-based combination of the Hargreaves-Samani model, and the radiation-based combination of the Priestley-Taylor model were used as local inputs. A comparison with further data driven modeling techniques was beyond the scope of this work.

Table 1
Summary of the studied locations.

Station	Latitude (°*N*)	Longitude (°*E*)	Elevation (m)	I_A	T_A (°C)	ΔT (°C)	R_S (MJ $m^{-2}day^{-1}$)	S_W (m s^{-1})	H_R (%)	ET_o (mm day^{-1})
Marivan (*dry sub-humid*)	35.31	46.12	1286.80	0.68	13.53	16.51	17.68	1.04	53.37	3.24
Baneh (*semi-arid*)	36.00	45.54	1600.00	0.42	14.05	10.26	17.45	2.53	45.75	4.20
Sanandaj (*semi-arid*)	35.20	47.00	1373.40	0.25	14.77	16.59	18.34	1.53	47.87	3.87
Bijar (*semi-arid*)	35.53	47.37	1883.40	0.24	12.00	11.19	17.48	2.92	46.33	4.17
Qorveh (*semi-arid*)	35.10	47.48	1906.00	0.23	12.64	11.73	18.37	2.33	47.68	4.00

Note- I_A: aridity index (ratio between the annual precipitation and the ET_o); T_A: mean air temperature; ΔT: average temperature difference (difference between maximum and minimum air temperature values); R_S: solar radiation; S_W; wind speed at 2 m above ground level; H_R: relative humidity; ET_o: reference evapotranspiration.

2. Materials and methods

2.1. Data set and targets used

Data from five locations in Northwestern Iran were utilized in the current study. The data set was limited to 5 stations due to the high number of study cases derived from the different data splitting methodologies adopted in this paper, which are explained in Section 2.1. A higher number of stations would have involved too high computational costs. Data samples included daily air temperature (T_A) [maximum (T_{max}), mean (T_{mean}) and minimum (T_{min}) air temperature], relative humidity (H_R), solar radiation (R_S) and wind speed (S_W) records, comprising a period of 6 years (1-January 2009 to 31-December 2014). Fig. 1 shows the geographical positions of the studied locations. The available data set was carefully analyzed and screened for any inconsistency and any possible missing values were filled using the common data filling processes as has been discussed by Yozgatligil et al. (2013). Table 1 presents a summary of the studied locations and used data. Stations are classified as semi-arid and sub-humid (Marivan) according to aridity index values (UNEP, 1997). Table 2 sums up the statistical descriptions of the applied parameters in the studied locations. For the use of ancillary inputs from other stations, it might be desirable that the ancillary stations considered were not climatically very different from the main training station. So, this reduces the number of stations that can be considered and data suppliers. Therefore this study considers a limited set of stations. Due to the absence of experimental ET_o values, they were calculated by the standard FAO56-PM model and were used as target values for training and testing the applied models (Allen et al., 1998). This is a very common and extended practice among the *ET* community according to the recommendation of the FAO56 paper:

$$ET_o = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T_A + 273} S_W (e_S - e_a)}{\Delta + \gamma(1 + 0.34 S_W)} \tag{1}$$

where, ET_o: reference evapotranspiration (mm day^{-1}), Δ: slope of the saturation vapor pressure function (kPa $°C^{-1}$), γ: psychometric constant (kPa $°C^{-1}$), R_n: net radiation (MJ $m^{-2}day^{-1}$), G: Soil heat flux density (MJ $m^{-2}day^{-1}$), T_A: mean air temperature (°C), S_W: average 24 h wind speed at 2 m height above ground level (m s^{-1}), e_S: saturation vapor pressure (kPa), e_a: actual vapor pressure (kPa), and λ: latent heat of evaporation (MJ Kg^{-1}).

2.2. Gene expression programming (GEP)

GEP evolves computer programs of various sizes and shapes encoded in linear chromosomes with fixed lengths. The chromosomes consisted of multiple genes, each gene encoding a smaller subprogram. Moreover, the systemic and functional arrangement of the linear chromosomes provides the uncompelled working of crucial genetic operators e.g. mutation, transposition and recombination. Genetic variety creation in GEP is very easy as genetic operators act at the chromosome level. Nonetheless, GEP has a unique, multigenic nature which provides the evolution of complex programs consisted of several subprograms (Ferreira, 2001; 2006). The ET_o modeling algorithm through using GEP is as follows:

- Selection of the fitness function: the Root Mean Squared Error (*RMSE*) was employed here according to Shiri et al. (2012).
- Choosing the terminal set and functions set to produce the chromosomes: the terminals set composed of meteorological variables, and the function set includes arithmetic operators (+, −, *, /) as well as some of the other basic mathematical functions ($\sqrt{}$, $\sqrt[3]{}$, ln(x), e^x, x^2, x^3) as has been advised by Shiri et al. (2012).
- Selection of the chromosomal architecture: Commonly used h = 8 (length of head), and three genes per chromosomes were employed here (Ferreira 2001).
- Choosing the linking function: addition linking function was used here, based on the results obtained by Shiri et al. (2012).
- Selecting the basic GEP operators.

A flexible modeling tool, i.e. GeneXpro program that is a user-friendly application of GEP, was employed here for ET_o modeling (www.gepsoft.com).

2.3. Study flowchart

Two modeling categories were used here to build the input configurations of the GEP models: temperature-based and radiation-based models:

- Temperature-based models, comprising T_{max}, T_{min}, T_{mean}, and R_a as inputs;
- Radiation-based models, comprising T_{max}, T_{min}, T_{mean}, R_a, and R_S as inputs.

Where R_a stands for the extraterrestrial radiation.

2.4. Data splitting and model evaluation

Here, k-fold testing was used for assessing the models through temporal and spatial data scanning. Accordingly, four different data splitting scenarios were adopted for modeling ET_o for both the temperature-based and radiation-based models:

Approach 1 (Ap.1): A normal local 6-fold validation was carried out per station. So, the models were trained each time using the daily data from 5 years, and then were tested using the remaining year. The procedure was repeated until all the patterns blocks (here, the years) were tested. A total of 30 train-test scenarios (6 years*5 stations) were evaluated per input combination (temperature-based and radiation-based). So, 60 train-test procedures were required in this approach.

Approach 1E (Ap.1E): This is the external testing of the Ap.1, where the complete local data set is used for training, then the developed model is tested using the complete external data set from the remaining stations.

Approach 2 (Ap.2): In each stage of Approach1, the corresponding patterns of the other 4 stations are added for training, in particular only

Table 2
Statistical summary of the meteorological data.

		T_{min} (°C)	T_{max} (°C)	T_{mean} (°C)	R_S (MJ m^{-2}day^{-1})	S_W (m s^{-1})	H_R (%)	ET_o (mm day^{-1})
Baneh	X_{mean}	8.92	19.19	14.06	17.46	2.53	45.75	4.20
	X_{max}	26.80	187.00	98.40	29.71	8.13	99.00	63.47
	X_{min}	−13.80	−6.90	−8.60	4.00	0.00	10.00	0.37
	X_{SD}	8.34	11.29	9.59	7.59	1.13	21.16	2.93
	C_V	0.93	0.59	0.68	0.43	0.45	0.46	0.70
	X_{skew}	−0.13	1.44	0.24	−0.11	1.13	0.37	4.00
	$X_{Kurtosis}$	−0.96	20.77	1.44	−1.29	2.37	−0.97	74.90
Bijar	X_{mean}	6.41	17.60	12.01	17.48	2.92	46.33	4.17
	X_{max}	24.20	38.20	30.10	29.72	10.94	96.50	12.41
	X_{min}	−19.60	−10.40	−14.40	4.02	0.00	7.00	0.36
	X_{SD}	8.46	11.36	9.80	7.58	1.65	20.19	2.64
	C_V	1.32	0.65	0.82	0.43	0.57	0.44	0.63
	X_{skew}	−0.30	−0.13	−0.19	−0.11	0.84	0.32	0.34
	$X_{Kurtosis}$	−0.60	−1.09	−0.94	−1.29	1.10	−0.80	−0.96
Marivan	X_{mean}	5.28	21.78	13.53	17.69	1.05	53.38	3.25
	X_{max}	26.00	41.20	32.00	29.61	9.63	98.00	10.04
	X_{min}	−20.80	−2.20	−10.60	4.08	0.00	14.50	0.40
	X_{SD}	7.35	10.75	8.66	7.53	0.92	17.62	2.06
	C_V	1.39	0.49	0.64	0.43	0.88	0.33	0.63
	X_{skew}	−0.38	−0.11	−0.16	−0.10	2.24	0.18	0.35
	$X_{Kurtosis}$	0.14	−1.20	−0.90	−1.26	9.38	−0.96	−0.95
Qorveh	X_{mean}	6.78	18.51	12.64	18.38	2.34	47.69	4.00
	X_{max}	24.30	37.90	30.30	30.63	9.63	97.00	10.16
	X_{min}	−21.40	−11.40	−16.00	4.13	0.19	10.50	0.32
	X_{SD}	8.59	11.07	9.69	7.22	1.18	19.96	2.42
	C_V	1.27	0.60	0.77	0.39	0.51	0.42	0.61
	X_{skew}	−0.34	−0.15	−0.21	−0.18	1.42	0.37	0.20
	$X_{Kurtosis}$	−0.46	−1.01	−0.84	−1.15	3.34	−0.78	−1.25
Sanandaj	X_{mean}	6.48	23.07	14.77	18.34	1.54	47.88	3.87
	X_{max}	26.30	42.70	33.00	30.62	7.91	94.50	9.99
	X_{min}	−18.60	−5.60	−11.20	4.11	0.00	15.50	0.48
	X_{SD}	8.11	11.24	9.37	7.23	0.94	17.87	2.39
	C_V	1.25	0.49	0.63	0.39	0.61	0.37	0.62
	X_{skew}	−0.04	−0.08	−0.04	−0.18	1.41	0.19	0.28
	$X_{Kurtosis}$	−0.61	−1.19	−1.05	−1.15	3.50	−1.02	−1.17

Note: X_{mean}: mean value; X_{max}: maximum value; X_{min}: minimum value; X_{SD}: standard deviation; C_V: coefficient of variation; X_{skew}: Skewness coefficient; $X_{Kurtosis}$: Kurtosis coefficient

those belonging to the same training period. Accordingly, for each training period, the data from the 5 stations are pooled together, and the model is tested with the remaining year in the 5 stations. So, only one model is trained in each stage for the 5 stations. The test sets for each year are then assessed per station, individually. A total of 12 models (6 years*2 input combinations) are assessed here. A similar approach was used in the past, namely the external assessment, where for the current stations, 4 stations would be used for training considering all years, and the models where tested in the remaining station, i.e. any pattern of the testing stations was considered for training. Ap. 2 could be observed as an external assessment, but using data from previous years in the testing station and omitting data from the testing year in the ancillary stations. These local data from previous years in the testing station could contribute to improve the performance of the external models.

Approach 3 (Ap.3): In each stage of Approach1, the corresponding patterns of the other 4 stations are added for training, but only those corresponding to the considered test year. The test sets for each year are then assessed per station. A total of 60 models (6 years*5 stations*2input combinations) were assessed here. This approach intends to improve Ap. 1 taking advantage of eventual pattern changes in the testing year in comparison to the training years. Therefore the testing years of the remaining stations are incorporated to the training series.

Approach 4 (Ap.4): This approach is based on the use of external ancillary inputs from other stations proposed by Martí and Gasque (2010) and Martí et al. (2011a, b), but using GEP instead of artificial neural networks. This approach involved considerable accuracy improvements when similar external stations were considered. Ap. 4 is similar to Ap. 1, but additional external ancillary inputs are added from the remaining 4 stations, namely the external ET_o patterns from the other stations are considered as further inputs of Ap.1. So, the input combination 1 is extended here and adds 4 further inputs, namely ET_o from the remaining 4 stations in the training years. More details can be found in Martí and Gasque (2010). A total of 60 models (6 years*5 stations*2 input combinations) were assessed here.

Approach 4E (Ap.4E). An external test of approach Ap4 was carried out as follows: The model was trained in one station using local inputs (e.g. temperature data) and ET_o external ancillary inputs. In the test stage, the trained model in the same station was tested in another station, i.e. using as local inputs from the test station and ET_o data from remaining stations (but testing period only). The procedure was repeated for all studied stations.

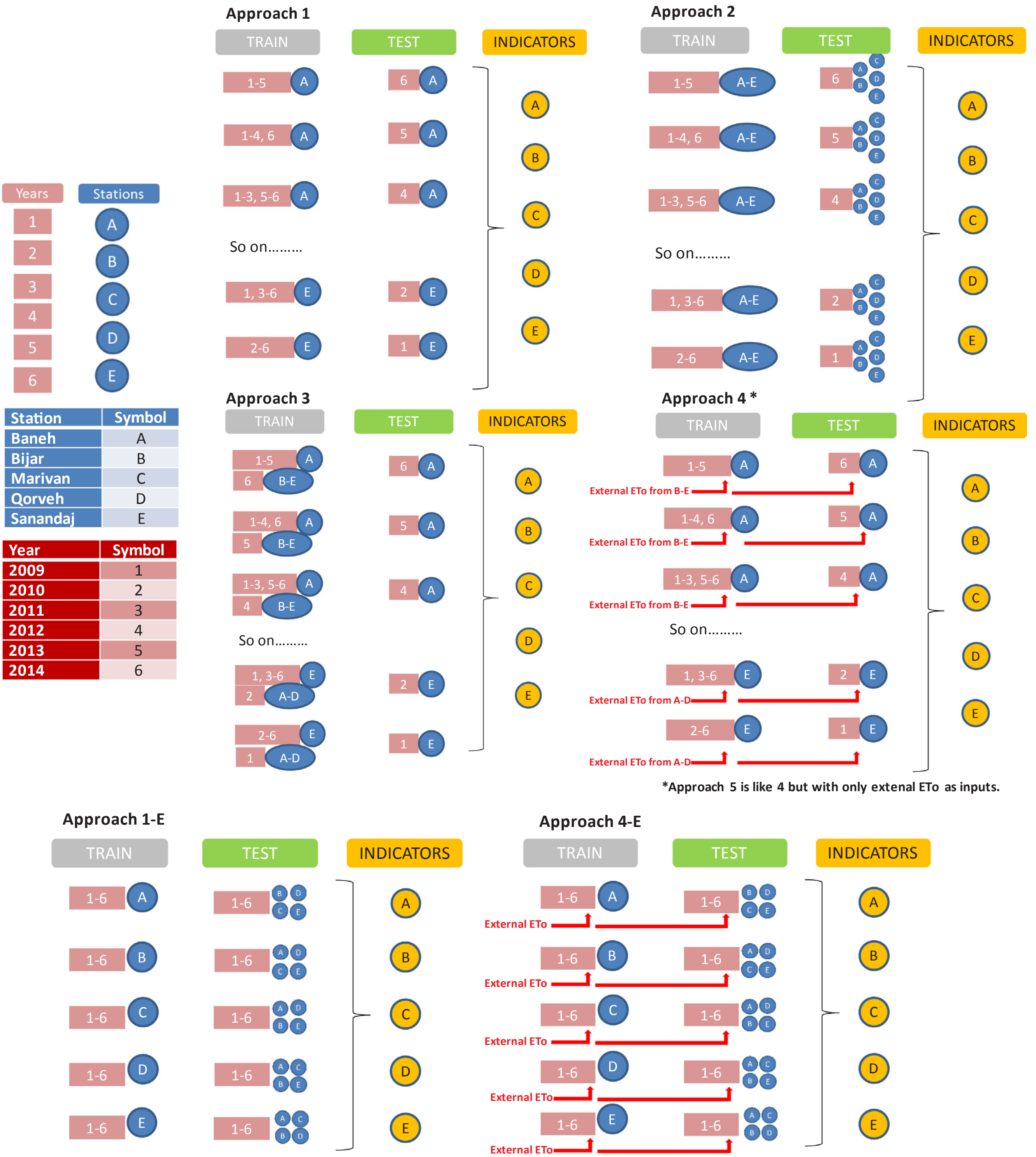
Approach 5 (Ap.5): This approach is similar to Ap. 4, but considers just external ET_o as inputs. It can be considered like a kind of interpolation.

A schematic representation of the study flowchart is illustrated in Fig. 2.

2.5. Statistical performance evaluation

Two statistical parameters were used for assessing the models' performance, namely, the scatter index (*SI*), and the Nash-Sutcliffe (*NS*) index, defined as follows:

$$SI = \frac{RMSE}{\bar{ET_o}} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(ET_{iM} - ET_{io})^2}}{\bar{ET_o}} \qquad (2)$$



Station	Symbol
Baneh	A
Bijar	B
Marivan	C
Qorveh	D
Sanandaj	E

Year	Symbol
2009	1
2010	2
2011	3
2012	4
2013	5
2014	6

Fig. 2. Schematic representation of the study flowchart.

$$NS = 1 - \frac{\sum_{i=1}^{N}(ET_{io} - ET_{iM})^2}{\sum_{i=1}^{N}(ET_{io} - \bar{ET}_o)^2} \quad (3)$$

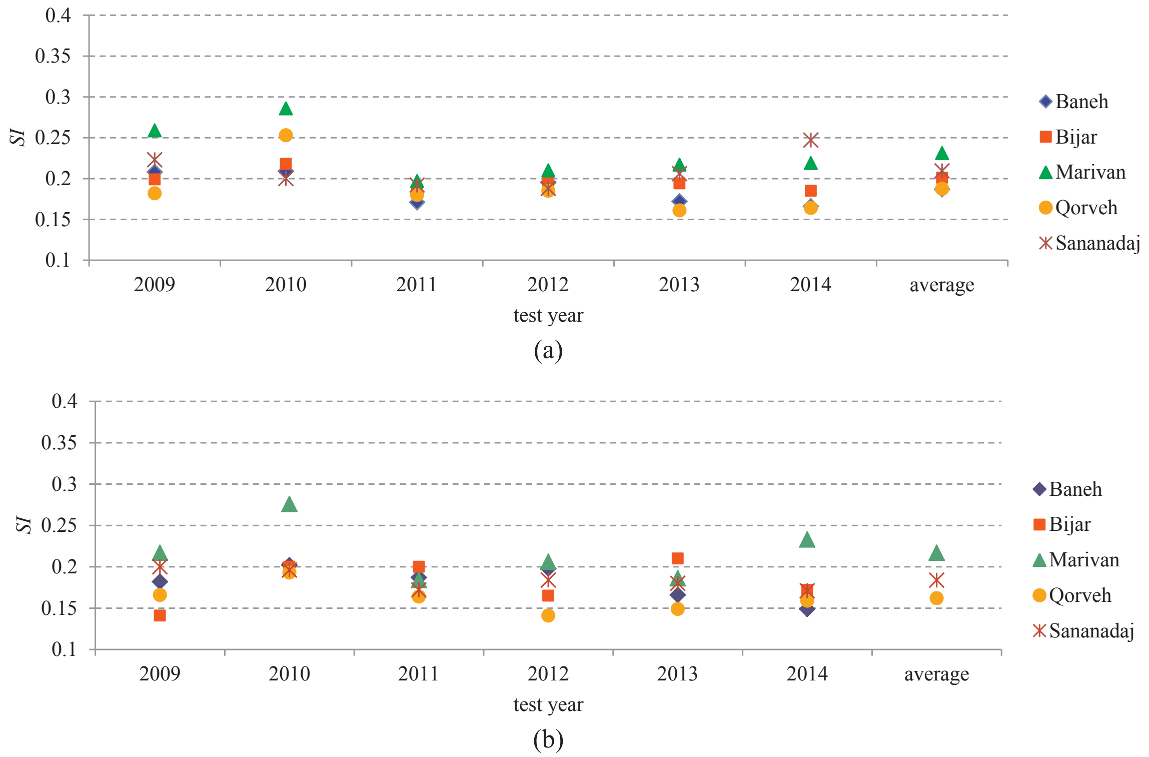
where, ET_M and ET_o denote the estimated and reference values, at the i^{th} time step, respectively. $\bar{ET}_M$ and $\bar{ET}_o$ represent the corresponding mean *ET* values, respectively. *N* is number of time steps. *RMSE* (that shows the average error values via assigning more weights to large errors) varies from 0 (perfect fit) to ∞ (the worst fit). The dimensionless *RMSE* index (*SI*) can also provide suitable information for assessing the performance accuracy of the employed models. Nonetheless, the Nash-Sutcliffe coefficient (*NS*) coefficient can be applied for the relative assessment of the models' performances. The higher the *NS* value the better the model performance, and vice-versa. A perfect match between the estimated and target ET_o magnitudes would yield $NS = 1.0$ (Legates and McCabe, 1999).

Table 3
SI values of the GEP models for the evaluated approaches.

	Ap.1	Ap1. E average	Ap1. E-optimum	Ap.2	Ap.3	Ap.4	App4.E average	App4.E-optimum	App.5
	Temperature based								*ET_o based*
Baneh	0.187	0.306	0.240	0.206	0.187	0.171	0.234	0.132	0.186
Bijar	0.201	0.273	0.233	0.206	0.203	0.177	0.286	0.233	0.181
Marivan	0.231	0.374	0.244	0.246	0.231	0.183	0.378	0.276	0.183
Qorveh	0.188	0.225	0.198	0.198	0.181	0.132	0.280	0.173	0.149
Sanandaj	0.209	0.286	0.221	0.216	0.201	0.153	0.261	0.214	0.178
AVERAGE	0.203	0.293	0.227	0.214	0.201	0.163	0.283	0.231	0.175
	Radiation based								
Baneh	0.181	0.368	0.347	0.171	0.162	0.163	0.221	0.104	–
Bijar	0.181	0.275	0.234	0.194	0.18	0.164	0.264	0.192	–
Marivan	0.217	0.490	0.236	0.206	0.21	0.167	0.456	0.368	–
Qorveh	0.162	0.238	0.151	0.171	0.163	0.121	0.242	0.201	–
Sanandaj	0.184	0.367	0.174	0.189	0.18	0.144	0.303	0.204	–
AVERAGE	0.185	0.347	0.228	0.186	0.179	0.152	0.297	0.213	–

Table 4
NS values of the GEP models for the evaluated approaches.

	Ap.1	Ap1. E- average	Ap1. E-optimum	Ap.2	Ap.3	Ap.4	App4.E- average	App4.E-optimum	App.5
	Temperature-based								ET_o based
Baneh	0.911	0.800	0.892	0.889	0.912	0.926	0.797	0.867	0.914
Bijar	0.896	0.806	0.894	0.925	0.910	0.913	0.789	0.862	0.917
Marivan	0.861	0.626	0.851	0.819	0.850	0.914	0.700	0.805	0.925
Qorveh	0.898	0.858	0.892	0.89	0.908	0.95	0.775	0.919	0.936
Sanandaj	0.882	0.775	0.894	0.874	0.893	0.936	0.812	0.878	0.924
AVERAGE	0.890	0.773	0.884	0.879	0.895	0.928	0.775	0.866	0.923
	Radiation-based								
Baneh	0.927	0.746	0.852	0.923	0.931	0.931	0.835	0.911	–
Bijar	0.899	0.805	0.862	0.901	0.9	0.947	0.813	0.906	–
Marivan	0.878	0.718	0.861	0.837	0.884	0.927	0.731	0.798	–
Qorveh	0.926	0.833	0.937	0.917	0.934	0.956	0.837	0.892	–
Sanandaj	0.909	0.749	0.919	0.901	0.904	0.942	0.824	0.888	–
AVERAGE	0.908	0.770	0.886	0.896	0.911	0.941	0.808	0.879	–



(a)

(b)

Fig. 3. *SI* values of the GEP models for (a) temperature-based and (b) radiation-based input configurations using the Approach1 split up per test year.

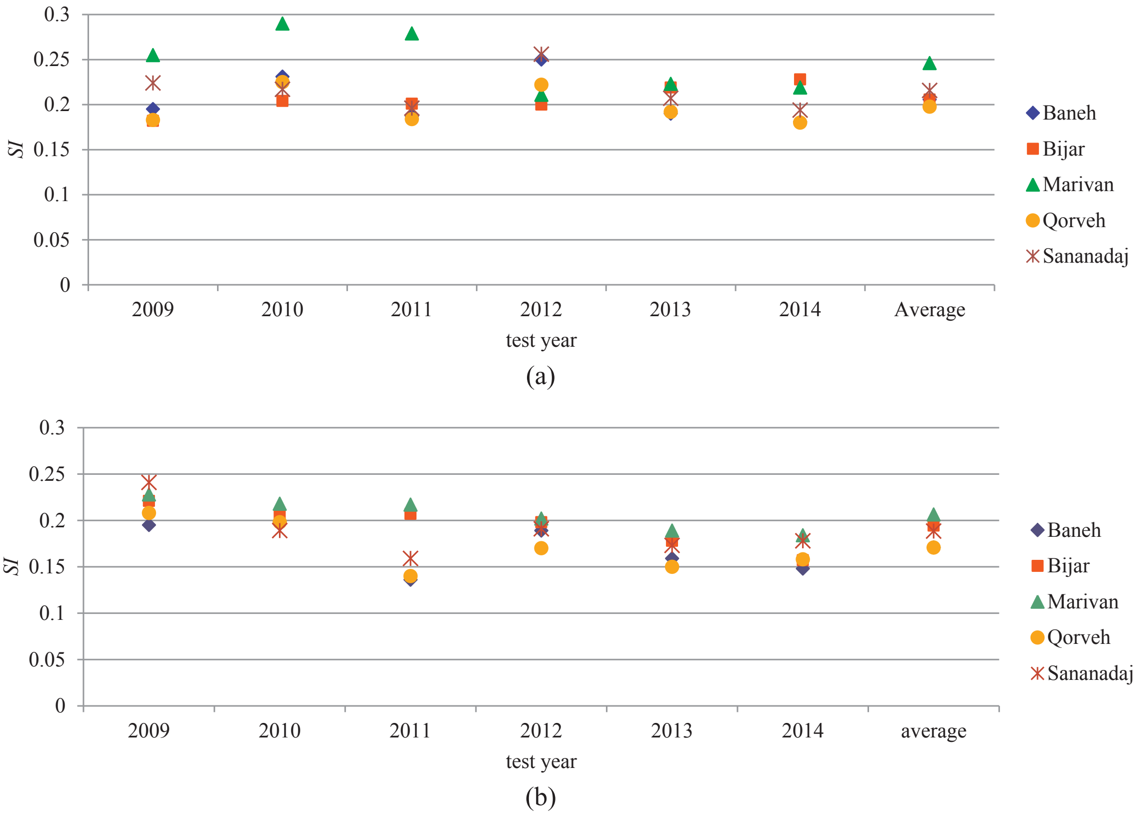


Fig. 4. *SI* values of the GEP models for (a) temperature-based and (b) radiation-based input configurations using the Approach2 per test year.

3. Results and discussions

3.1. Overall comparison of the 5 approaches.

Table 3 sums up the overall statistical indices of the evaluated approaches (Ap.1-Ap.5) at the studied stations for both input combinations. As it could be expected, the solar radiation-based models surpass the temperature-based models in all studied locations. Comparing the performance accuracy of the models for the five applied approaches (the averages) revealed that the fourth approach, i.e. using the external ET_o values of the ancillary stations as input parameters for simulating ET_o at target stations, produced the most accurate results among the applied approaches, followed by Ap.5, just slightly worse. On the other hand, the second approach, i.e. using the training patterns of the ancillary stations as inputs along with the local patterns of each station, provided the worst outcomes. Approach 3, where the inputs corresponding to the testing period in the remaining stations are used as training patterns, gives slightly better results than the first and second approaches. The superior performance of the first approach towards the second approach might be due to using exclusively local patterns to feed the models, which allows for a more suitable input-output mapping, while the second approach includes the training data from different locations, which might decrease the model accuracy, due to the difficulty for approximating local ET_o from external inputs. In the third approach, however, the exogenous training patterns (corresponding in this case to the period used for testing) might reflect the similar climatic conditions of the testing period and increase the modeling accuracy, but very slightly. Finally, for the last approach, since the ET_o magnitudes (which comprise the effect of all influential meteorological parameters) are introduced as input vectors, the modeling accuracy is increased and GEP-based models can produce the outcomes with relatively higher performance accuracy than the previous models.

The performance pattern presented above for the statistical averages is almost the same in the case of the statistical indices split per location. The performances (*SI* values) at Marivan varies between 0.246 (Ap. 2) and 0.183 (Ap. 4) for temperature based GEPs; 0.217 (Ap. 1) and 0.167 (Ap. 4) for solar radiation based GEPs. Marivan is the station with the poorest performances for both the temperature based and solar radiation based models. This could be related with the climatic characteristics of Marivan, which are wetter than at the rest of the locations, with higher values of H_R and lower values of ET_o.

Approach 4 clearly increases the performance of temperature and solar radiation GEPs at all locations. Qorveh presents the highest performance values and the highest increment of performance associated with approach 4 with regard to the classical k-fold approach 1 (*SI* = 0.188 (Ap. 1) vs *SI* = 0.132 (Ap. 4) for temperature based GEPs; and *SI* = 0.162 (Ap. 1) vs *SI* = 0.121 (Ap. 4) for solar radiation based GEPs. So, the use of data from neighbor locations in the training set is an interesting alternative, in agreement with Martí and Gasque (2010).

The presented approaches 1 to 5 can be valid alternatives to conventional approaches for estimating ET_o. However, they require local targets for performing the training. Therefore, this might lack their applicability and limit it to local emergency models. In practice, the models used for providing the targets would be used for estimating ET_o. Therefore, aiming at assessing the actual usefulness of these models, 2 of them, Ap.1 and Ap.4 were assessed outside the training station, i.e. testing them in the remaining stations (Ap.1E and Ap.4E). Table 3 presents two scenarios for this application, the average performance and the optimum performance. The average performance corresponds to the mean performance of the 4 training stations (models) per testing station, while the optimum performance corresponds to the training station (model) with the optimum performance per testing station. As can be observed, attending to the average indices, the external performance is considerably worse than the local performance (*SI* of 0.203 vs. 0.293 for temperature-based Ap1 and Ap.1E, respectively, *SI* of 0.185 vs. 0.347 radiation-based Ap1 and Ap.1E, *SI* of 0.163 vs 0.283 for temperature-based Ap.4 and Ap.4E, respectively, and *SI* of 0.152 vs. 0.297 for radiation-based Ap.4 and Ap.4E, respectively). However, a

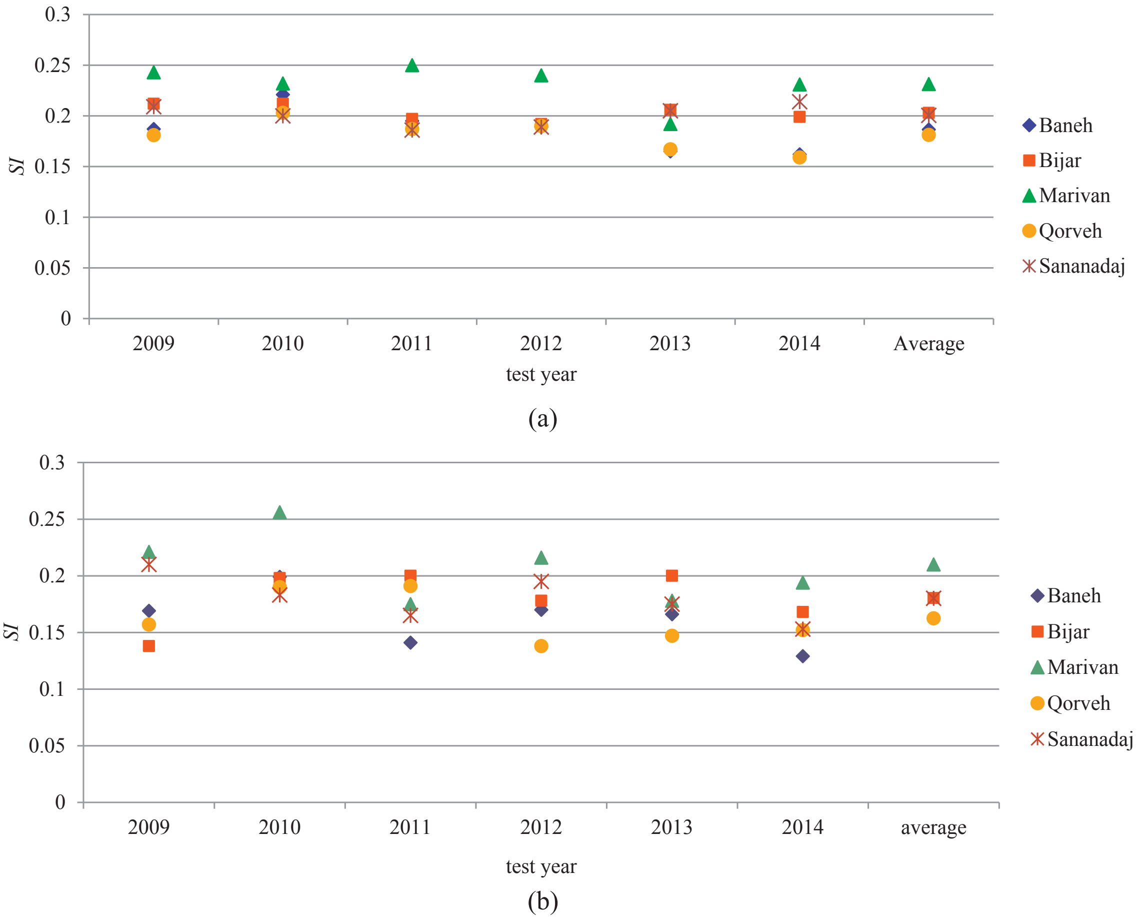


Fig. 5. *SI* values of the GEP models for (a) temperature-based and (b) radiation-based input configurations using the Approach3 per test year.

preliminary selection of the most suitable training station/s per testing station seems reasonable, i.e. the selected training station/s should present a similar input-output mapping patterns and similar patterns variation ranges than the testing station. Accordingly, if the optimum performance of Ap.1E and Ap.4E is considered it can be observed that the performance improvement is noteworthy (*SI* basis: 0.227 vs. 0.293 in temperature-based Ap.1E, 0.228 vs. 0.347 in radiation-based Ap.1E, 0.231 vs. 0.283 in temperature-based Ap.4E, and 0.213 vs. 0.297 in radiation-based Ap.4E), despite slightly worse than the local performance, which could be expected. Similar conclusion can be drawn from the *NS* basis, Table 4. Thus, the application of these approaches would be justified, because they would provide accurate estimations in the testing stations, while not requiring local patterns for preliminary training of the models. This point will be analyzed with more detail in subsections 3.7.

3.2. Approach 1

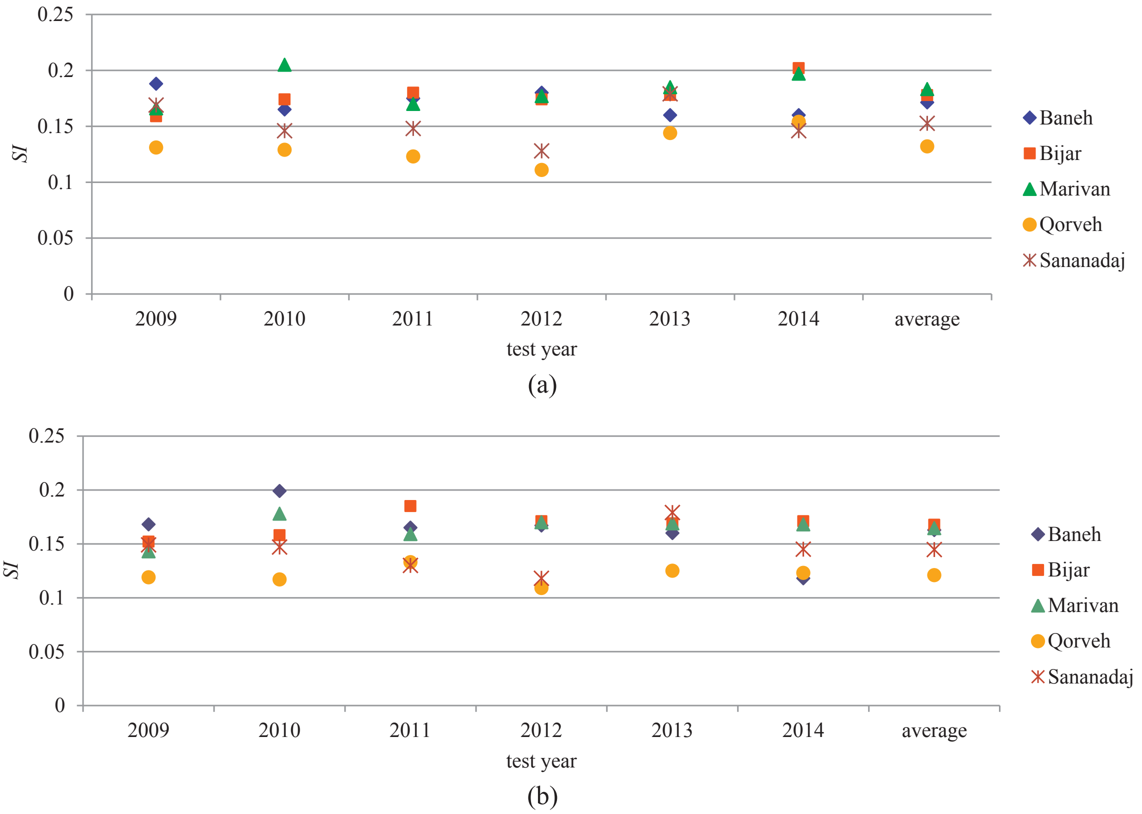
Fig. 3 presents the statistical indices of the GEP models for the temperature-based and radiation-based input combinations split up per test year. As could be expected, the models relying on input combination 2 (i.e. radiation-based models) are more accurate. It can be stated that Ap.1 produces the most accurate results in the Baneh, Bijar and Qorveh stations for both models. Further, the performance accuracy of the models fluctuates between the test years. The performance (*SI*) of the temperature based GEPs varies between 0.29 (2010) and 0.19 (2011) at Marivan; 0.25 (2014) and 0.18 (2012) at Sanandaj; 0.25 (2010) and 0.16 (2013) at Qorveh; and 0.22 (2010) and 0.17 at Baneh and Bijar. Comparing the performance accuracy of the GEP models for both input combinations it can be seen that their performance is the similar for some test years (e.g. Marivan in 2010 & 2012; Baneh in 2013, etc), while some notable difference are observed among them.

So, in the case of temperature based GEPs, the variability of performance at Marivan and Qorveh is higher than at the rest of the locations. At Marivan and Qorveh depending on the testing/training years there are quite significant differences. Marivan, which is the location with the highest variability, is a dry-subhumid location with higher H_R values and lower wind speed values than the rest of the locations (which are semiarid locations). So, in the case of Marivan, H_R component might have a higher influence on ET_o estimation than at the rest of the locations.

As it was mentioned above, the performance of the solar radiation based GEPs is more accurate than the temperature based GEPs, but the relative differences of performance among locations and years are similar than in the case of temperature based GEPs. Again Marivan presents the highest variability of performance (*SI* values between 0.27 (2010) and 0.18 (2013)) due to its climatic characteristics which are slighter different to the rest of the locations. Either way, a complete testing scan of the whole data set should be desirable, because the results might differ between years. Therefore, the practice of using a single data set assignment, i.e. a hold-out validation, very common between the ET_o modeling community, should be avoided if possible.

3.3. Approach 2

The *SI* values per test year of the temperature-based and radiation-based models corresponding to the second application (Ap.2) are illustrated in Fig. 4. Alike to the first application, the models relying on the radiation-based models present more accurate results than the temperature-based models. The general trend of the models' accuracy is



(a)

(b)

Fig. 6. *SI* values of the GEP models for (a) temperature-based and (b) radiation-based input configurations using the Approach4 (local testing) per test year.

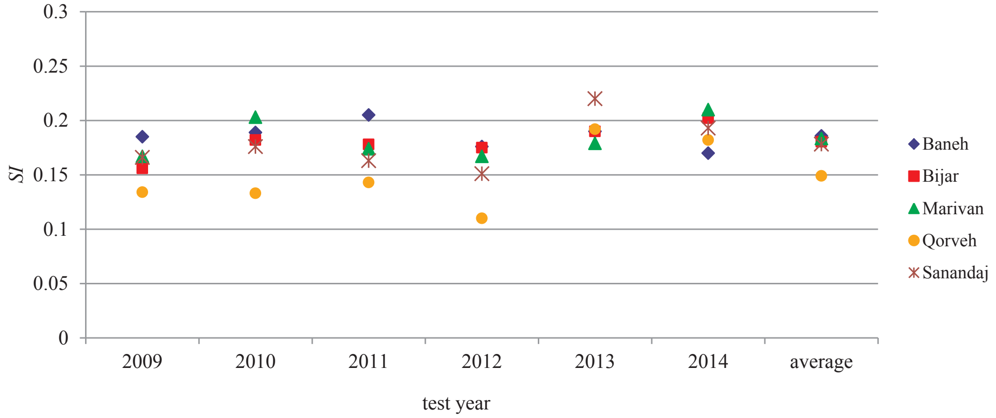


Fig. 7. *SI* values of the GEP models using the Approach5 per test year.

similar to the first application, where the models present the most accurate results in Baneh, Bijar and Qorveh stations. In general, the accuracy of the models using this approach is slightly lower than the accuracy of models belonging to the first approach in the case of temperature based GEPs. This fact might be attributed to the inclusion of training data from the other studied stations, which might difficult the mapping between the inputs and outputs of the model. Nevertheless, the differences between approach 1 and 2 are very low. And in the case of solar radiation based GEPs, there are not significant differences of performance between both approaches. So, the inclusion of data from neighbor stations in the training set does not increase the performance of ET_o estimation in the conditions of this study.

The performance accuracy of the models varies between the stations and the test year, while there are some years with similar statistical indices. The performance (*SI*) of the temperature based GEPs varies between 0.29 (2010) and 0.21 (2011) at Marivan; 0.25 (2012) and 0.19 (2014) at Sanandaj; and 0.25 (2012) and 0.18 (2009) at Qorveh, Baneh and Bijar. Comparing the *SI* values per year of temperature based GEPs of approaches 1 and 2 it can be stated that the performance at Marivan clearly decreases (higher SI values) introducing data from neighbor stations (approach 2), especially for year 2011. 2011 is the year with the lowest ETo values at Marivan due to low temperature values. As a result of this, and due to the differences in precipitation/relative humidity patterns between Marivan (dry subhumid) and the rest of the locations, the inclusion of data from other locations in the training set reduces the performance at Marivan in 2011.

The performance of the solar radiation based GEPs for approach 2 varies between 0.23 (2009) and 0.18 (2014) at Marivan; 0.24 (2009) and 0.16 (2011) at Sanandaj; 0.23 (2009) and 0.17 (2013) at Bijar; and 0.21 (2009) and 0.14 (2011) at Qorveh, and Baneh. It is worth noting that in the case of Marivan and comparing approaches 1 and 2, the performance clearly increases with the introduction of data from other

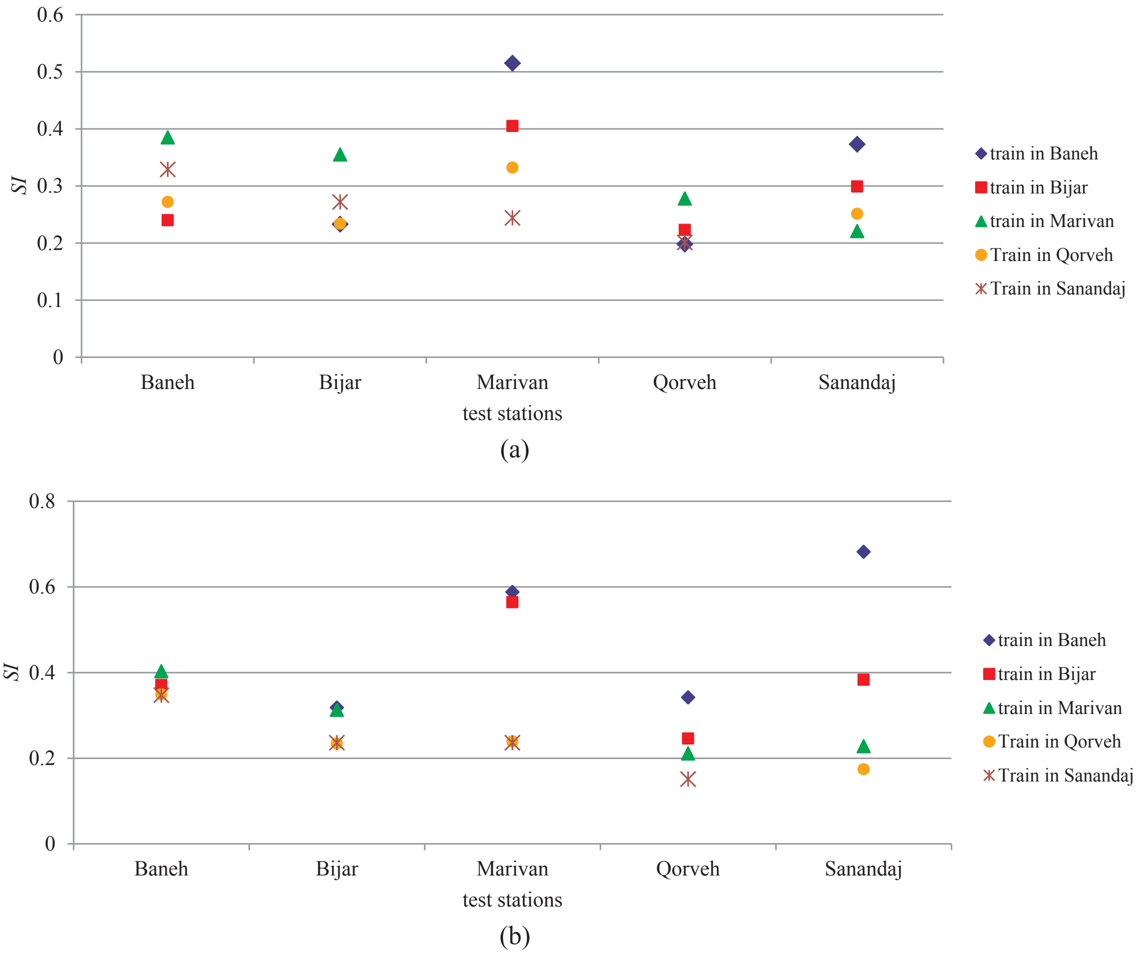


Fig. 8. *SI* values of the GEP models for (a) temperature-based and (b) radiation-based input configurations using the Approach1E (AP.1E) per test station.

locations in the training set in 2010 and 2014. These years (2010 and 2014) present the highest solar radiation and temperature values at Marivan. As Marivan is in general terms the wettest location, the inclusion of patterns from the rest of the locations (locations in which days with high temperature and solar radiation values are common) increases the performance at Marivan in 2010 and 2014, because these years are characterized by higher temperature and solar radiation values than the rest of the years.

3.4. Approach 3

Fig. 5 shows the *SI* values of the models for the third application (Ap.3) split up per test year. In general terms approach 3 and approach 2 present similar *SI* values. The difference between these two approaches (2 and 3) is that in the case of approach 2 the training patterns data are composed of data from the rest of the locations for the same training period, and approach 3 includes data from the other stations in the training set, but these data belong to the test year of the target station. However there are some differences of performance between approaches 2 and 3. In the case of year 2012 at Quorveh and Sanandaj (attending to the *SI* values) taking data from the testing year (2012) from the rest of the locations as training data, increases the performance of temperature based GEPs with regard to taking data of the non-testing years from the rest of the locations as training patterns. The opposite occurs in the case of Marivan.

Regarding solar radiation based GEPs it is interesting to analyze year 2009. The adoption of approach 3 instead of approach 2 clearly increases the performance at Qorveh, Bijar and Baneh. 2009 is the year with the lowest solar radiation values. As Bijar and Baneh are locations very similar from a climatic point of view, using data from 2009 from the rest of the stations in the training set, instead of using data of 2010–2014 period, increases the performance of 2009 testing year.

3.5. Approach 4

The *SI* values per year corresponding to Ap.4 are presented in Fig. 6. It can be stated that the performance accuracy of the GEP models fluctuates considerably among the stations and test years. Generally, the models relying on temperature and radiation records produce more accurate results than those obtained using the temperature records. The lowest *SI* values (average) corresponded to the Qorveh (with the highest altitude and the lowest latitude and aridity index values) for both the temperature-based (0.132) and radiation-based (0.121) categories, while the highest *SI* values for these categories are, respectively as 0.183 and 0.167 for Marivan (with the lowest altitude), which might be linked to its climatic context (dry-sub humid when compared to the rest of the stations which are semi-arid stations).

As it was mentioned previously, Ap.4 is based on the inclusion of external ET_o values from ancillary stations as inputs of the model. Comparing this Ap.4 with approaches 1, 2 and 3 it is possible to see a clear and significant accuracy improvement in the performance of temperature based GEPs. The breakdown of this increase per year is quite homogeneous (a reduction of about 0.5 in *SI* values). However the breakdown of this performance increase per location is not as homogeneous as the breakdown per year. The adoption of Ap.3 is especially positive (in the case of both temperature and solar radiation based GEPs) at Qorveh and Sanandaj. This fact could be explained taking into account that from a climatic point of view Sanandaj and Qorveh are

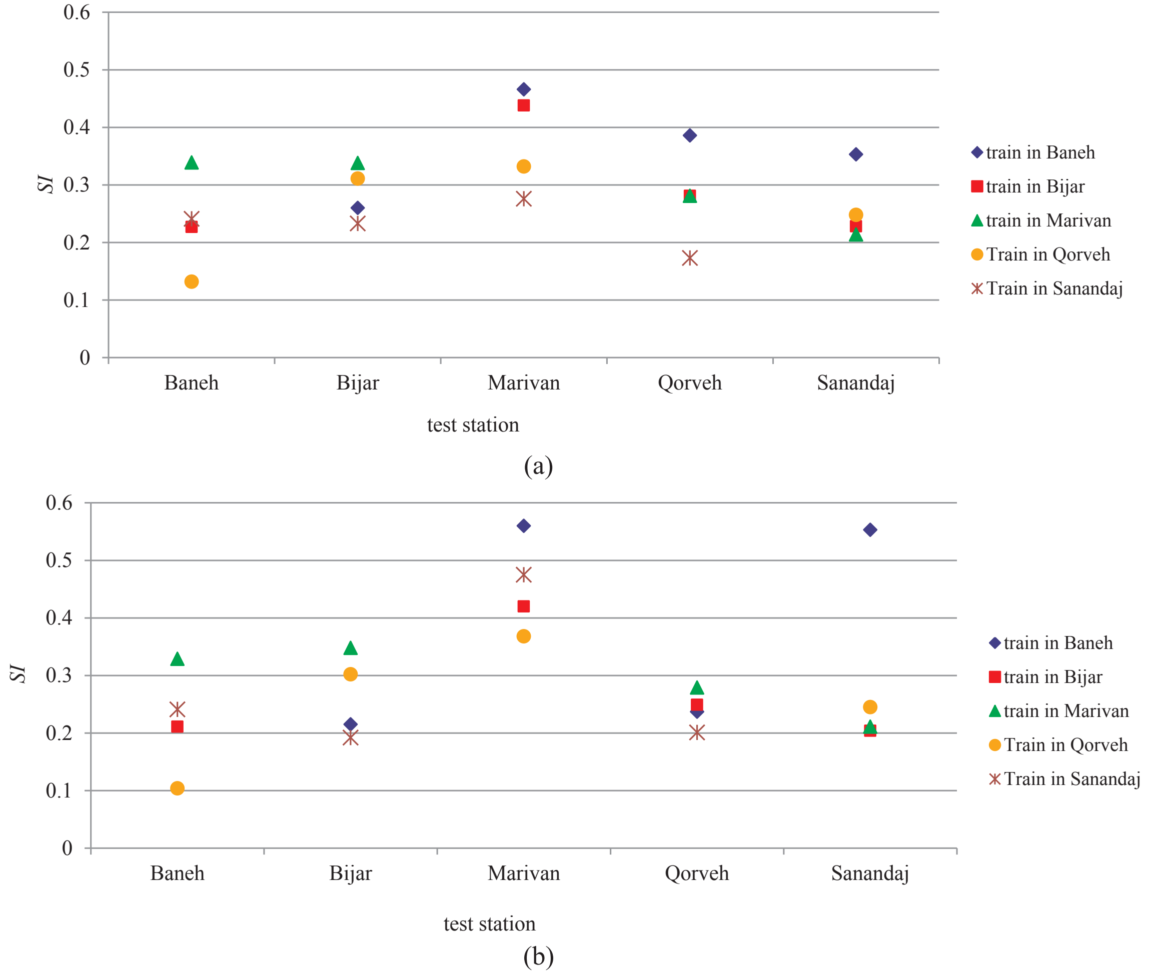


Fig. 9. *SI* values of the GEP models for (a) temperature-based and (b) radiation-based input configurations using the Approach4E (AP.4E) per test year.

midway between the Marivańs climatic characteristics (dry sub-humid) and Baneh/Bijaŕs. Due to the high number of cases studied and the subsequent calculation costs, Ap.4 just considers the case where all 4 exogenous ET_o inputs from the remaining stations are included among the inputs of the models. Nevertheless, a deeper analysis might have assessed the performance of this approach considering different number of exogenous inputs attending to the similarities between the ancillary station and the training station providing the targets. According to this, the inclusion of external ET_o records as inputs from a station which might have different climatic trends to the training station might introduce noise to the model. So, further research might be required about the criterion used to incorporate ancillary inputs. Martí and Gasque (2010) assessed the effect on the estimation performance of an increasing number of ancillary ET_o inputs according to the differences in the continentality index between ancillary and training stations.

3.6. Approach 5

Fig. 7 presents the *SI* values corresponding to Ap5, i.e. models relying only on external ET_o inputs. The performance pattern is similar, slightly worse, to Ap.4, because Ap.4 present the advantage of considering local temperature and radiation records, respectively, which might improve slightly the input-output mapping. However, Ap.5 presents the advantage of providing similar accuracy requiring fewer inputs. Ap. 5 is also more accurate than Ap. 2 and Ap. 3. Similarly to Ap.4, further research might be needed to find out a suitable criterion to select the ancillary stations providing the ancillary exogenous ET_o inputs.

3.7. External validation. Ap.1E and Ap.4E

Figs. 8 and 9 present the indices corresponding to the external validation of approaches 1 and 4, i.e. testing those approaches outside the training station. These indices, as could be expected, involve lower estimation accuracy in comparison to the indices corresponding to the local performance, which can be also clearly stated from Tables 3 and 4. However, a high fluctuation is stated in each testing station between the estimation accuracy of the different training stations. For instance, attending to the *SI* values of radiation-based models in Sanandaj (Fig. 9) it can be seen that they range between 0.2 (trained in Bijar) and 0.55 (trained in Baneh). In agreement with this, the highest performance in Bijar is obtained with the model trained in Sanandaj. This might involve that Bijar and Sanandaj present more similarities than Sanandaj and Baneh, for instance. Thus, it is crucial to select a suitable training station for each testing station, in agreement with Martí and Gasque (2010). Comparing the upper and lower sides of these figures, it can be observed that the mentioned similarities between stations depend on the inputs considered. For example, Baneh provides the less accurate estimation relying on Rs at Qorveh (Fig. 8), whereas it provides the most accurate estimation relying on temperature records at the same station. So, it might be more suitable to speak in terms of suitable input-output mapping instead of in terms of similarities between stations. In most cases the qualitative trend (order of accuracy) is the same for both input combinations, e.g. performance at Sanandaj and Marivan. A suitable selection of the training station might provide estimations with similar accuracy than in the local performance scenario, i.e. around 0.19–0.23 of *SI* in temperature-based models, and around 0.15–0.23 in radiation-based models, excluding Baneh, with an abnormal high *SI* value. This is crucial, because it points out that, using the correct training station, it is possible to provide accurate enough estimations at any station without the need of training previously a local model using local targets. Similar trends can be stated attending to Fig. 8, corresponding to Ap.4E, i.e. incorporating external ET_o inputs to the model. In this case, a higher variability between stations is stated for the optimum *SI* value, i.e. the *SI* ranges from around 0.13 in Baneh (training station Qorveh) to close to 0.3 in Marivan (training station Sanandaj) for temperature-based models, and from 0.1 in Baneh (training station Qorveh) to close to 0.4 in Marivan (training station Qorveh). In this case, it might be crucial not only the training station selected, but also the number and origin of external ancillary ETo inputs used. Accordingly, if the temperature-based model trained in Marivan presents low performance accuracy when tested in Baneh, it might seem reasonable not to include external ET_o inputs from Marivan when training in Baneh, and not to include external ET_o inputs from Baneh when training in Marivan. i.e. external ancillary ET_o inputs from stations with different climatic pattern might introduce noise to the model and decrease their generalizability. So, it might be desirable to select properly which station will provide external ancillary ET_o inputs to the training process.

Summarizing, it should be noted that the external performance might be deeply related to the climatic similarities between training and testing stations. However, in this regard, we should not evaluate these similarities based on the long term average values of different climatic parameters, but rather attending to the variation range of the climatic patterns considered. So, it is crucial for developing a good model to consider stations where climatic parameters present a wide spectrum of variation. Accordingly, the external performance of a model rather depends of the relationships in the variation ranges of the input and target variables in the training and testing stations.

4. Conclusions

This paper presents 5 different approaches for splitting the data set aiming at improving the performance of data driven methods for reference evapotranspiration estimation relying on two common input combinations. Therefore external ancillary inputs from secondary stations are used for training the models according to different criteria. Further, all models are assessed using k-fold validation considering annual test sizes. A comparison with further data driven modeling techniques was beyond the scope of this work.

It seems preferable to incorporate the external target variable, ET_o, as input to feed the new model, rather than to incorporate the original external input variables of the model, i.e. the same variables, temperature and/or radiation, used from the main local station. In these latter approaches, it seems slightly better to consider patterns from the testing period rather than from the training period of the main station.

Regarding the external performance of the models, i.e. outside the training station, it is crucial to select a suitable training station for each testing station for providing accurate enough estimations. This way, the applicability of such approaches is not limited to local emergency models, but it allows estimating ET_o elsewhere without the need of training previously a local model using local targets.

Regarding those models considering external ancillary ET_o inputs, it might be desirable to select properly which station/s will provide those external ancillary ET_o inputs to the training process. External ancillary ET_o inputs from stations with different climatic pattern might introduce noise to the model and decrease their generalizability.

Acknowledgment

The present work was partially supported by University of Tabriz.

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.

Doorenbos, J., Pruitt, W.O., 1977. Crop water requirements. FAO Irrigation and Drainage paper no. 24, Rome, Italy.

Droogers, P., Allen, R.G., 2002. Estimating reference evapotranspiration under inaccurate data conditions. Irrigation Drainage Systems 16 (1), 33–45.

Ferreira, C., 2001. Gene expression programming: a new adaptive algorithm for solving problems. Complex Syst. 13 (2), 87–129.

Ferreira, C., 2006. Gene expression programming: mathematical Modeling by an artificial intelligence. Springer, Berling, Heidelberg New York, NY, USA, pp. 478.

Guven, A., Aytek, A., Yuce, M.I., Aksoy, H., 2008. Genetic Programming-based empirical model for daily reference evapotranspiration estimation. Clean: Soil, Air, Water 36 (10–11), 905–912.

Guven, A., Kisi, O., 2011. Daily pan evaporation modeling using linear genetic programming technique. Irrig. Sci. 29 (2), 135–145.

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

Izadifar, Z., Elshorbagy, A., 2010. Prediction of hourly actual evapotranspiration using neural networks, genetic programming, and statistical models. Hydrol. Process 24 (23), 3413–3425.

Kisi, O., Ozturk, O., 2007. Adaptive neurofuzzy computing technique for evapotranspiration estimation. J. Irrig. Drain. Eng. 133 (4), 368–379.

Kisi, O., Yildirim, G., 2005. Discussion of "forecasting of reference evapotranspiration by artificial neural networks" by slavisa trajkovic, branimir todorovic, and miomir stankovic. J. Irrig. Drain. Eng. 131 (4), 390. https://doi.org/10.1061/(ASCE)0733-9437(2005) 131:4(390).

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.

Marti, P., Gasque, M., 2010. Ancillary data supply strategies for improvement of temperature-based ETo ANN models. Agric. Water Manag. 97, 939–955.

Marti, P., Manzano, J., Royuela, A., 2011a. Assessment of a 4-input artificial neural network for ETo estimation through data set scanning procedures. Irrig Sci 29, 181–195.

Marti, P., Gonzalez-Altozano, P., Gasque, M., 2011b. Reference evapotranspiration estimation without local climatic data. Irrig Sci 29, 475–495.

Marti, P., González-Altozano, P., López-Urrea, R., Mancha, L.A., Shiri, J., 2015. Modeling reference evapotranspiration with calculated targets: assessment and implications. Agric. Water Manag. 149, 81–90.

Parasuraman, K., Elshorbagy, A., Carey, S.K., 2007. Modelling the dynamics of the evapotranspiration process using genetic programming. Hydrol. Sci. J. 52 (3), 563–578.

Priestley, C.H.B., Taylor, R.J., 1972. On the assessment of surface heat flux and evaporation using large-scale parameters. Monthly Weather Rev. 100 (2), 81–92.

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., 2014. Comparison of heuristic and empirical approaches for estimating reference evapotranspiration from limited inputs in Iran. Comput. Electron. Agric. 108, 230–241.

Shiri, J., Sadraddini, A.A., Nazemi, A.H., Marti, P., Fakheri Fard, A., Kisi, O., Landeras, G., 2015. Independent testing for assessing the calibration of the Hargreaves-Samani equation: new heuristic alternatives for Iran. Comput. Electron. Agric. 117, 70–80.

Shiri, J., 2018. Improving the performance of the mass transfer-based reference evapotranspiration estimation approaches through a coupled wavelet-random forest methodology. J. Hydrol. 561, 737–750.

Shiri, J., 2019. Modeling reference evapotranspiration in island environments: assessing the practical implications. J. Hydrol. 570, 265–280.

Trajkovic, S., Stankovic, M., Todorovic, B., 2000. Estimation of FAO blaney-criddle b factor by RBF network. J. Irrig. Drain. Eng. 126 (4), 268–271.

Traore, S., Guven, A., 2013. New algebraic formulations of evapotranspiration extracted from gene-expression programming in the tropical seasonally dry regions of West Africa. Irrig. Sci. 31 (1), 1–10.

UNEP (United Nations Environmental Programme). 1997. World Atlas of Desertification. Editorial commentary by N. Middleton and D.S.G. Thomas. London: Edward Arnold.

Yassin, M.A., Alazba, A.A., Mattar, M.A., 2016. Artificial neural networks versus gene expression programming for estimating reference evapotranspiration in arid climate. Agric. Water Manag. 163, 110–124.

Yozgatligil, C., Aslan, S., Iyigun, C., Batmaz, I., 2013. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. Theor. Appl. Climatol. 112, 143–167.