Computers and Electronics in Agriculture 144 (2018) 232–240

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Generalizability of gene expression programming and random forest methodologies in estimating cropland and grassland leaf area index

Sepideh Karimi[a], Ali Ashraf Sadraddini[a,*], Amir Hossein Nazemi[a], Tongren Xu[b], Ahmad Fakheri Fard[a]

[a] Water Engineering Department, Faculty of Agriculture, University of Tabriz, Iran
[b] State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China

ARTICLE INFO

Keywords:
Meteorological data
Leaf area index
Gene expression programming
Random forest

ABSTRACT

Leaf Area Index (LAI) is a very important structural attribute of ecosystems which affects the energy, water and carbon exchanges between the land surface and atmosphere. Direct measurement of LAI is costly and time consuming so indirect measurement approaches have been developed for determining its magnitude. The present paper aimed at modeling LAI in cropland and grassland sites using the available meteorological data through two heuristic data driven techniques, namely, gene expression programming (GEP) and random forest (RF). Different data set organizations were designed using local (temporal) and external (spatial) norms to provide a thoroughgoing data scanning strategy. The results showed that the external GEP and RF models (EGEP and ERF) might be suitable approaches for modeling LAI by average scatter index (*SI*) values of 0.275 and 0.270 (for cropland) and 0.273 and 0.279 (for grassland) when compared to the local GEP and RF models with average *SI* values of 0.207 and 0.204 (cropland), and 0.249 and 0.204 (grassland), respectively. The presented methodology allowed the evaluation in each site of models developed (trained) using local patterns and the models developed using the exogenous data (patterns from ancillary sites).

1. Introduction

Leaf area index (LAI) is a dimensionless variable defined as the total one-sided area of photosynthetic tissues per unit ground surface area (Watson, 1947). LAI is an important structural characteristic of ecosystem as it influences the exchanges of water, energy, and carbon between the land surface and atmosphere (Sellers et al., 1988; Wulder et al., 1998; Sonnentag et al., 2007). It determines the size of the plant–atmosphere interface, and therefore plays a key role in the energy and mass exchanges between the canopy and the atmosphere (Weiss et al., 2004). Xu et al. (2014) showed that the inclusion of LAI in their variational data assimilation model improved the simulation of surface water and energy fluxes. Li et al. (2009) indicated that using LAI in Xinanjiang hydrologic modeling improved rainfall-runoff modeling. Coopersmith et al. (2014) and Chen et al. (2015) obtained a better soil moisture prediction by incorporating LAI in Integrated Biosphere Simulator (IBS) and HYDRUS-1D models.

The ground-based measurement methods have been developed to measure LAI accurately (Asner et al., 2003; Jonckheere et al., 2004; Qu et al., 2014). However, those methods can only obtain LAI at point scale during limited time periods due to their high cost and time consumption. Therefore, different models have been developed to acquire LAI over large spatial scales based on remotely sensed data.

Currently, there are mainly three kinds of methods for retrieving LAI from remotely sensed data, i.e., the empirical relationships, radiative transfer models, and heuristic data driven models. The empirical methods are used to link LAI with remotely sensed vegetation index (i.e. NDVI) or with reflectance data with regression equations (Combal et al., 2003), which are relatively simple and accurate. However, these methods are sensor dependent and site specific, and have a major drawback of local calibration need. The physical laws are used in radiative transfer models to explicitly describe associations between the vegetation properties and canopy spectra, and produce reasonable LAI at regional scale (Meroni et al., 2004; He et al., 2013). However, the radiative transfer models are usually complicated and time consuming (Jacquemoud et al., 2000). Thus, heuristic data driven techniques are used to estimate LAI at larger scales (Xiao et al., 2014; Liang et al., 2014). The GLASS-LAI product is produced via the reflectance data in the visible and infrared bands based on heuristic data driven techniques (Liang et al., 2014; Xiao et al., 2014).

In recent years, heuristic data driven techniques (e.g., gene expression programming (GEP) and random forest (RF)) have been

[*] Corresponding author.
E-mail address: sadraddini@tabrizu.ac.ir (A.A. Sadraddini).

https://doi.org/10.1016/j.compag.2017.12.007
Received 10 August 2017; Received in revised form 4 December 2017; Accepted 7 December 2017
Available online 15 December 2017

utilized for modeling hydrological and eco-hydrological parameters. Genetic programming (GP), a generalization of genetic algorithm (GA) (Goldberg, 1989), was proposed by Koza (1992). It engages a "parse tree" structure for exploring the solutions. Gene expression programming (GEP) is equivalent to GP. The chromosomes in GEP collect multiple genes, each gene converting a smaller subprogram. Moreover, the systematic organization of the linear chromosomes provides the unrestrained behavior of important genetic operators such as mutation, transposition and recombination (Ferreira, 2006). Major dominances of GP (i.e., GEP) are that it can be applied to areas where (a) the interrelationships among the pertinent factors are less clarified, (b) finding the conclusive solution is difficult, (c) normal mathematical investigation cannot supply analytical solutions, (d) a rough solution is acceptable, (e) small improvements in the performance are routinely measured and highly valued, and (f) there is a large amount of data which require evaluation, classification, and integration (Banzhaf et al., 1998). One of the major advantages of GEP is that it can generate an explicit equation between input(s) and output of the underlying problem. Such an equation might be subjected to some interpretation to find the governing rules of the studied process.

A number of studies (e.g., Walthal et al., 2004; Dunea and Moise, 2008; Xiao et al., 2014) applied data driven neural networks models for obtaining LAI from remotely sensed data and filling the gaps between the recorded data. Everingham et al. (2009) applied heuristic techniques to forecast regional sugarcane crop production. Torres et al. (2011) applied support vector machine to estimate daily potential evapotranspiration with limited climatic data. Shiri et al. (2014a) used heuristic techniques to model dew point temperature. Shiri et al. (2014b) showed the generalizability of GEP in modeling daily evapotranspiration in local and regional scales. Karimi et al. (2017) used GEP for simulating daily evapotranspiration through a cross-station approach.

Commonly, lots of heuristic-based applications contemplate only a single data set assignment where models are developed and validated utilizing data of the same site. Apart from not executing a perfect performance evaluation of the local patterns, another important drawback of this data set assignment type is that the generalization ability of the achieved models is not evaluated outside the locations that have been used to train the models (Marti et al., 2013; Shiri et al., 2014b).

The present study aimed at assessing the performances of GEP and RF techniques in local and external cross-station scales for simulating LAI, using available meteorological and NDVI data. By relying only on meteorological data, LAI can be obtained in the long past time and future when no remotely sensed data exists. To the best of authors' knowledge, this is the first assessment of heuristic methods in estimating LAI in local and cross-station scales. The robust k-fold testing cross validation technique was used for assessing the applied methodologies in both local and external scales.

2. Materials and methods

2.1. Data

The GEP and RF-based models were trained and tested extensively over ten experimental sites (with five cropland and five grassland sites). The meteorological data of the ten experimental sites were obtained via Fluxnet website (http://www.fluxnet.ornl.gov/). The site locations and data temporal coverage were summarized in Table 1. Half-hourly or hourly micrometeorological data such as wind speed, air temperature and humidity, atmospheric pressure, solar radiation, and incoming longwave radiation were measured at the ten experimental sites. The 16- day MODIS NDVI/EVI (MOD13A2) (Huete et al., 2002) and the 8-day MODIS FPAR/LAI (MOD15A2) products (Myneni et al., 2002) with 1 km spatial resolution were also collected in this research. The daily NDVI and LAI values were temporally interpolated from the 16- day or 8- day averages using linear interpolation.

In this study, the meteorological data including maximum air temperature (T_{max}), minimum air temperature (T_{min}), mean air temperature (Ta) and mean relative humidity (RH), as well as the remotely sensed normalized difference vegetation index (NDVI) were used as input parameters of the applied models for simulating the LAI. With these parameters, the vegetation growth could be well controlled, and thus could provide useful dynamic information for the LAI estimation (Stockli et al., 2008; Yao et al., 2008; Qu et al., 2012). Table 2 presents the statistical characteristics of the applied LAI records.

2.2. Study flowchart

The performances of the applied GEP and RF-based models were assessed via a cross-validation k-fold test procedure. Accordingly, the available patterns were divided into k blocks and the train-test process was repeated k times till a complete data scanning was achieved (Marti et al., 2013; Roushangar et al., 2014a). Each time, a separate set of data was reserved for testing. In the present study, two different criteria were developed for defining the minimum test set size for executing the k-fold assessment, e.g. external (spatial) and temporal (local) criteria. First, train-test processes (5-fold spatial assessment) per land cover type were performed leaving each time the whole available patterns of one site for testing. This procedure allowed evaluating the external generalizability of the GEP and RF- based models (Shiri et al., 2015). Nonetheless, the local (temporal) performance was assessed per location through a k-fold temporary assessment. This case was carried out separately at each site, where one year was used for testing and the rest data were utilized for training. Finally, a global approach was developed where the applied models were trained using all data from cropland stations, and tested in grassland sites, and vice versa.

Four statistical criteria were used for assessing the applied models, namely, the coefficient of determination (r^2), the scatter index (SI), the mean absolute error (MAE), and the Nash Sutcliffe coefficient (NS), expressions for which are given below:

$$r^2 = \left[\frac{\sum_{i=1}^{n}(LAI_{io}-\overline{LAI_o})(LAI_{iM}-\overline{LAI_M})}{\sqrt{\sum_{i=1}^{n}(LAI_{io}-\overline{LAI_o})^2\sum_{i=1}^{n}(LAI_{iM}-\overline{LAI_M})^2}}\right]^2 \quad (1)$$

$$SI = \frac{RMSE}{\overline{LAI_o}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(LAI_{i0}-LAI_{iM})^2}}{\overline{LAI_o}} \quad (2)$$

$$MAE = \frac{1}{n}\left|\sum_{i=1}^{n}(LAI_{i0}-LAI_{iM})\right| \quad (3)$$

$$NS = 1-\frac{\sum_{i=1}^{n}(LAI_{i0}-LAI_{iM})^2}{\sum_{i=1}^{n}(LAI_{i0}-\overline{LAI_o})^2} \quad (4)$$

where LAI_{i0} denotes the target (observed) LAI value at the i^{th} time step and LAI_{iM} represents the corresponding simulated value. n stands for the number of time steps, $\overline{LAI_o}$ shows the mean of the observed values and $\overline{LAI_M}$ is the mean value of the simulations.

2.3. Gene expression programming (GEP)

GEP is similar to genetic programming (GP) where it additionally evolves computer programs with wide size domains and shapes encoded in linear chromosomes with fixed lengths. The chromosomes in GEP consist of multiple genes, where every gene is encoding a subprogram. In addition, the structural and functional organization of the linear chromosomes provide the unconstrained operation of important genetic operators, e.g. mutation, transposition and recombination. One strength of the GEP approach is that the creation of genetic diversity is extremely simplified as genetic operators work at the chromosomes level. Another strength of GEP is its uniqueness of multigenic nature

Table 1
Summary of the studied locations.

Station code	Site ID	Site	Latitude (°N)	Longitude (°W)	Altitude (m)	Available data
Cropland stations						
Crop1	US-ARM	Brasschaat (De Inslag Forest)	36.605	97.488	314	2003–2005
Crop2	US-Bo1	Bondville	40.006	88.290	219	1996–2007
Crop3	US-Ne1	NE - Mead - irrigated continuous maize site	41.165	96.476	361	2001–2005
Crop4	US-Ne2	NE - Mead - irrigated maize-soybean rotation site	41.164	96.470	362	2001–2005
Crop5	US-Ne3	NE - Mead - rain fed maize-soybean rotation site	41.179	96.439	363	2001–2005
Grassland stations						
Grass1	CA-Let	Lethbridge	49.709	112.940	960	1998–2005
Grass2	CA-Mer	Eastern Peatland- Mer Bleue	45.409	75.518	70	1998–2005
Grass3	FI-Kaa	Kaamanen Water Bodiesland	69.140	27.295	155	2000–2006
Grass4	US-FPe	MT - Fort Peck	48.307	105.101	634	2000–2006
Grass5	US-var	CA - Vaira Ranch- Ione	38.413	120.950	129	2001–2006

Table 2
Statistical indices of the LAI series in the studied locations.

	X_{max}	X_{min}	X_{mean}	SD	C_V	C_{SX}
Cropland sites						
Crop1	1.900	0.100	0.627	0.675	0.798	1.566
Crop2	3.300	0.100	0.740	0.835	1.128	1.444
Crop3	2.300	0.100	0.808	0.557	0.689	0.607
Crop4	2.600	0.100	0.838	0.685	0.817	0.920
Crop5	2.700	0.100	0.846	0.335	0.534	1.000
Grassland sites						
Grass1	1.700	0.100	0.503	0.350	0.694	1.143
Grass2	3.300	0.100	0.960	0.820	0.860	0.670
Grass3	1.800	0.000	0.578	0.408	0.705	0.825
Grass4	1.400	0.000	0.384	0.297	0.773	1.238
Grass5	3.500	0.100	1.033	0.653	0.632	1.334

which allows the evolution of more complex programs composed of several subprograms. As a result, GEP surpasses the old GP system in 100–10,000 times (Ferreira, 2001a, 2001b). The advantages of GEP are (Ferreira, 2006): (i) the chromosomes are simple entities: linear, compact, relatively small, and easy to manipulate genetically (replicate, mutate, recombine, etc.), (ii) the expression trees are exclusively the expression of their respective chromosomes; they are entities upon which selection acts, and according to fitness, they are selected to be reproduced with modification. However, there are also some problems regarding the GP (GEP) application. For instance, in some cases, the depth of parse tree starts growing which leads to produce nested functions (i.e., the Bloat Phenomena) (Shiri et al., 2014b). In such cases, penalization of complex models through e.g. Parsimony Pressure (Ploi and McPhee, 2008) should be established for producing parsimonious relations.

The first step with GEP development is to select a fitness function. In this study, different fitness functions were evaluated and it was found that the mean absolute error (*MAE*) produce the lowest error and thus was chosen as the best fitness function. The second step consists of choosing the set of terminals and functions to create the chromosomes. Here, the terminal set includes T_{min}, T_{max}, T_a, R_H, and *NDVI*. The selection of an appropriate function set was carried out by analyzing different function sets and the following set was found as the optimum one: $\{+, -, \times, \div, \sqrt[3]{}, \sqrt{}, \ln, e^x, x^2, x^3\}$. The third step is choosing the chromosomal architecture. Following Ferreira (2001a) and Ferreira (2006), length of head was set to 8 ($h = 8$), and three genes per chromosomes were employed. The fourth step is to choose the linking function. The linking function must be "addition" or "multiplication" for algebraic sub-trees (Ferreira, 2001a). Results showed that the "addition" linking function produced the most accurate results. The final step is to choose the genetic operators. Table 3 shows the GEP operators used in this study. These values are the default values of GeneXpro and are usually used in the literature (e.g. Shiri and Kisi, 2012).

Table 3
Genetic operators used in the GEP models.

Number of chromosomes	30	One point recombination rate	0.3
Head size	8	Two point recombination rate	0.3
Number of genes	3	Gene recombination rate	0.1
Linking function	Addition	Gene transposition rate	0.1
Fitness function Error type	*MAE*	Insertion sequence transposition rate	0.1
Mutation rate	0.044	Root insertion sequence transposition	0.1
Inversion rate	0.1	Penalizing tool	Parsimony Pressure
Mutation	*Allows the evolution of good solutions for the studied models to virtually all problems*		
Inversion	*Inversion is restricted to the heads of genes*		
One-point recombination	*The parent chromosomes are paired and split up at exactly the same point*		
Two-point recombination	*Two parent chromosomes are paired and two points are randomly chosen as crossover points*		
Gene recombination	*Entire genes are exchanged between two parent chromosomes, forming two daughter chromosomes containing genes from both parents*		
Gene transposition	*An entire gene works as a transposon and transposes itself to the beginning of the chromosome*		
IS transposition	*Short fragments of the genome with a function or terminal in the first position that transpose to the heads of gene except the root*		
RIS transposition	*Short fragments with a function in the first position that transpose to the start position of genes*		

2.4. Random forest (RF)

Random forests (RF) is an assembling learning algorithm that supervises high-dimensional regression problems. RF is a tree-based assembling attitude, where all trees are dependent on a group of random variables, and the forest is grown from many regression trees put together and from a group (Breiman, 2001). The eventual decision is achieved by averaging the outputs, after fixing individual trees in entity (bagging procedure). The bias of the bagged trees is the same as the individual trees', though the variance is decreased by reducing the correlation values between the existing trees (Hastie et al., 2009). Different tree numbers were evaluated and the best tree numbers were selected when increasing the tree numbers makes negligible variations in the average squared error values of the simulation. Here, 15 cycles were found to be the optimum cycle number of the mean error calculation, by a trial-error process. Similarly, the percentage of decrease in training error was observed as 5%, minimum child node size to stop (for controlling the smallest permissible number in a child node, for a split to be applied) as 5, and the maximum number of levels (the depth of the tree as measured from the root node) as 10.

Table 4
Global statistical indices of the applied models.

Approach	Cropland sites				Grassland sites			
	LGEP	EGEP	LRF	ERF	LGEP	EGEP	LRF	ERF
r^2	0.960	0.934	0.963	0.913	0.935	0.927	0.942	0.915
SI	0.207	0.275	0.204	0.270	0.249	0.273	0.240	0.279
MAE	0.113	0.155	0.113	0.150	0.160	0.193	0.164	0.207
NS	0.922	0.861	0.933	0.910	0.880	0.851	0.889	0.869

3. Results and discussions

3.1. Global assessment of the models

Table 4 sums up the global statistical indicators of the applied models for both the land cover types. Expectedly, the local GEP and RF models (LGEP and LRF, respectively) presented the most accurate simulations because they relied on the local patterns, so they were trained and tested using the meteorological patterns of the same locations. This would make a great limitation to the applicability of these locally trained models, so that they could not be applied using data outside the trained location. Consequently, the external GEP (EGEP) and RF (ERF) models were built and assessed through the same procedure. From Table 4, although the local models have better performance accuracy in general, the difference between the performance of the local and external models (especially for grassland sites) is small. This is very important, because the external models utilize the exogenous data from ancillary sites for simulating the LAI magnitudes in the target location, so local patterns won't be necessary using these models. However, it should be noted that the external models cannot be validated if the test location presents different input-output trends than those sites utilized for training the model.

Since the GEP models could be expressed in relatively simple equations, further models were trained using the complete data set (patterns) of each location to include a more representative pattern collection. The corresponding GEP formulations are given in Table 5. Although the performance accuracy of the global GEP models (fed with all patterns of the studied locations) cannot be evaluated, the listed expressions presented higher performance accuracy (not presented here) than the mentioned external GEP models, since their training set consisted of the patterns of the sites that have been considered for testing the models in external scenario. Analyzing the equations shows that the minimum and average air temperature (T_{min} and T_a, respectively) parameters have not been picked by GEP in modeling LAI in both the land cover types. For the cropland model, *NDVI*, R_H and T_{max} have been used as inputs, while the grassland model does not pick the T_{max} records. Meanwhile, *NDVI* have had the highest weight (predictor importance) among the inputs for both the cases. On the other hand, the RF model utilizes all the introduced parameters for the same cases, giving them different weights. Similarly, *NDVI* has had the highest weight while T_{min} has had the minimum weight (importance) for modeling LAI. Apart from their similarity in selecting *NDVI*, T_{max} and humidity records, differences in selecting the rest of parameters might be explained by basic assumptions of the applied methodologies. Deschaine (2014) argues that GEP develops the general organization and constant values of the equations concomitantly, so the degree of similarity between the constant values of two specified sites would affect the formulation transferability in different locations. In contrast, RF consists of lots of decision forests run by building lots of decision trees at training time and outputting the average estimation of the separate trees. Strong regression-based relations of LAI with NDVI have been approved by the previous studies, too (e.g. Fan et al., 2009).

Table 5
Mathematical expressions of the GEP-based models.

Land cover type	GEP expression
Cropland	$LAI = 2.383[NDVI]^9 + 2[NDVI]^2 + 0.007[2R_H + T_{max}]$
Grassland	$LAI = NDVI^4[3.9601{*}NDVI{-}R_H] + NDVI[\sqrt[3]{R_H}{-}R_H] + NDVI$

3.2. Cropland models

The local (temporal) and external (spatial) performances per station of the applied GEP and RF models are shown in Figs. 1 and 2 for the cropland and grassland stations, respectively. In each location, the presented statistics of the local GEP and RF models belong to the global k-fold temporal testing, while the statistics of the external GEP and RF models show the indicators of the 5-fold external testing. From the figure, all r^2, *SI*, *RMSE* and *MAE* statistics present high variability in all locations. In case of the cropland sites, the global *SI* values of the local GEP and RF models, respectively range between 0.167 and 0.175 for the US-Ne3 site (code: crop5) and between 0.260 and 0.267 for the US-Bo1 site (code: Crop2). This may be attributed to the statistical characteristics of the LAI series at each location (Table 2), where the crop5 station presents the lowest values of the standard deviation and coefficient of variations for LAI, while the Crop2 has the highest amount of these statistics among others. Higher values of these statistics could make it difficult to get more accurately outcomes for LAI simulation in a specified location. The other performance indicators present similar variations for the stations. Consequently, it is evident that Crop2 and Crop3 sites are of lowest performance accuracy when applying GEP and RF models. Such variability might be attributed to the relationships/differences between the training and testing patterns, too.

Regarding the external performance of the cropland GEP and RF models, Crop4 site presents the lowest error magnitudes in terms of *SI*. On the other hand, Crop1 and Crop4 sites present similar performance of the external models in terms of *MAE* and the lowest *NS* is observed in Crop3. Overall, Crop5 shows the poorer performance than the other stations while its r^2 values is greater for the external model. This could be linked to the linear inherent of this index, which can picture only the linear dependency between the observed and simulated variables. Nevertheless, as discussed by Legates and McCabe (1999), higher sensitivity to the outliers is another weakness of this index, which totally makes its application limited, so that other weighted (dimensionless) indices e.g. *SI* should be applied together with this index for a better judgment on the models performance accuracy. The weak performance of the external models in Crop5 might be due to the differences between the training and testing patterns applied to feed the models. As can be seen from Table 1, Crop5 shows the lowest variations (in terms of C_V) for LAI data among the studied locations, which can make it difficult to extrapolate its LAI values using the more scattered data from other locations. However, the values of *SI* difference (ΔSI) between the local and external models are about 0.166 (GEP) and 0.109 (RF), which show the external models ability in simulating LAI using exogenous data. Comparing the performances of the local and external GEP and RF models in Fig. 1 reveals that the accuracies of the models are not similar for every station, where local models may give better performance in a specified station than the external models and vice versa. Although the better performance accuracy of the local models could be anticipated (since they are relying on the local patterns for the training and testing phases), the better performance of external models show the high capability of the generalized GEP and RF models in simulating LAI values using data outside the test points. These models could be of great utility in locations with partial or total absence of local input data.

3.3. Grassland models

Fig. 2 displays the statistical indices of the applied models for the grassland sites. Similar to the cropland sites, there are obvious variations for the statistical indicators among the studied sites for the local

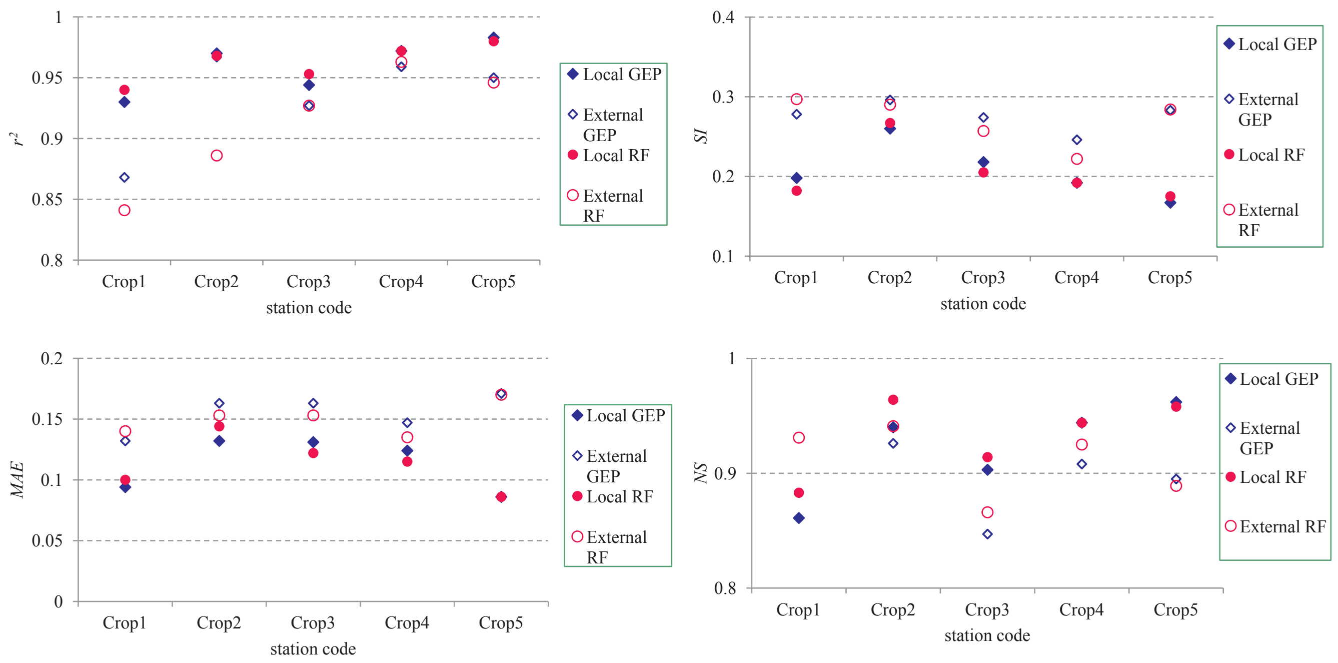

Fig. 1. Statistical criteria for cropland LAI simulations split up per test station.

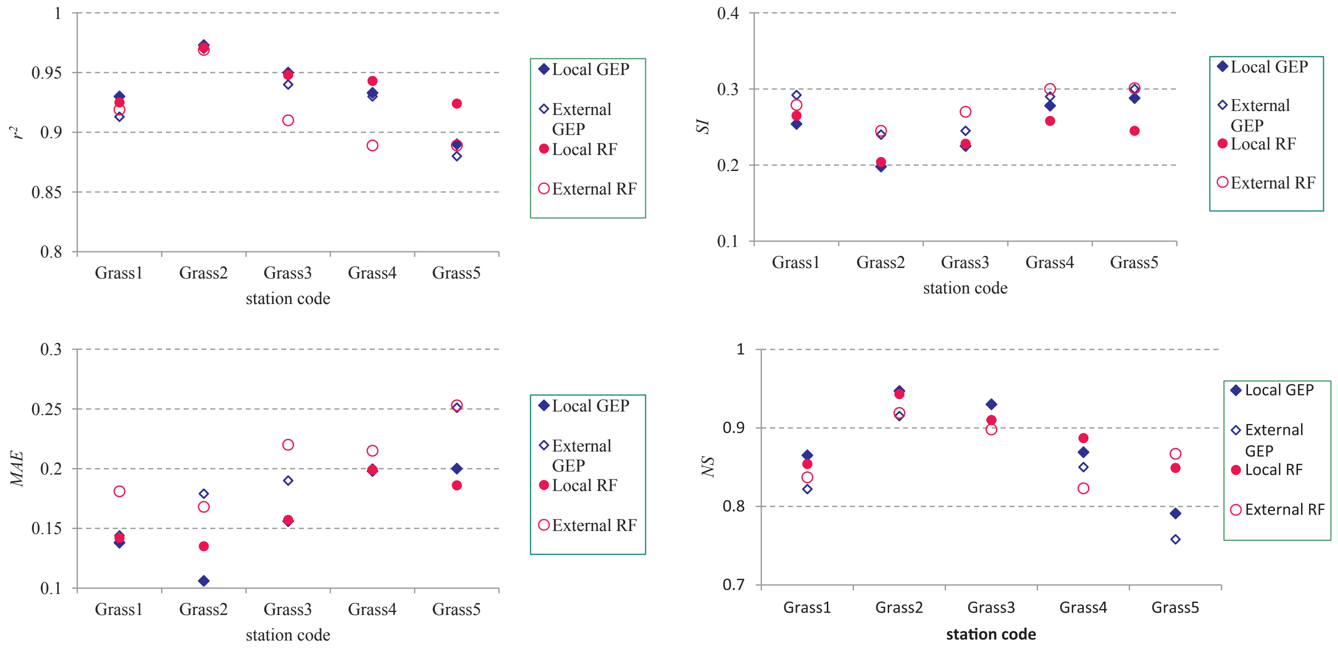

Fig. 2. Statistical criteria for grassland LAI simulations split up per test station.

and external models. Regarding the local GEP and RF models, Grass2 presents the more accurate results with the highest r^2 and *NS* and the lowest *SI* and *MAE* magnitudes, and the rest of the locations (except Grass3), showed similar performance accuracy. The best performance of the models in Grass2 might be linked to the lowest skewness values (the lowest deviation from the normal distribution) of the LAI values in this station that makes the LAI simulation easy using the local patterns. Nonetheless, analyzing the correlations between the input variables and LAI (not presented here) showed that the highest correlation values belonged to Grass2 that could be a possible reason for better performance of the models in this location. For all studied locations, the relative humidity (R_H) showed a small negative correlation with LAI, while the highest correlation corresponded to NDVI, followed by temperature components, except Grass5 where only NDVI presented high correlation with LAI and the rest of inputs had small correlation magnitudes. Given that the land cover is similar for all these stations (grassland) such difference for LAI responses to the applied inputs may be due to the higher differences between the observed LAI values (which ranges between 0.1 and 3.5) in this site as well as the geographical position of the site (the lowest latitude). In case of the external grassland models, again the Grass2 site presents the most accurate results.

3.4. General statements

The results for both the land cover types revealed that in spite of slight differences in the local and external models' performance accuracy; it could be more practical to utilize the external GEP and RF

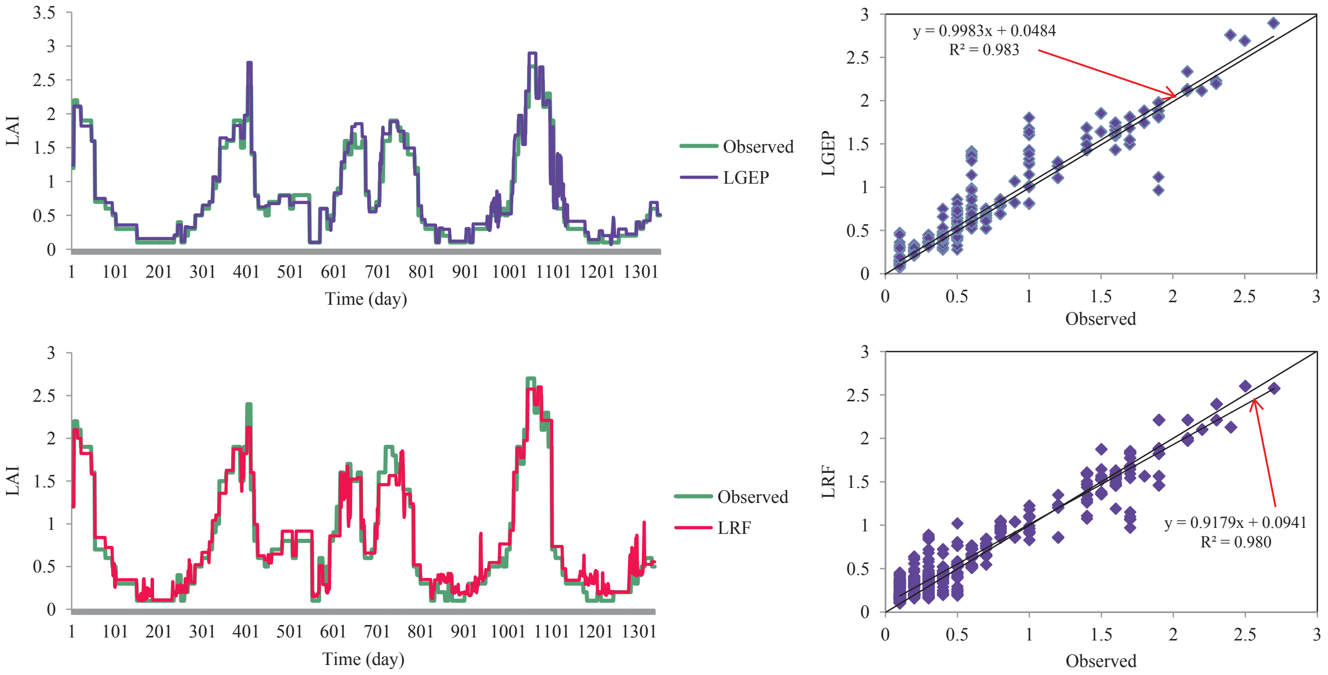

Fig. 3. Observed vs. simulated values of local models in Crop5.

models, as it would not be necessary to train a specific model at each site. So, no local patterns would be necessary to previously train/calibrate the local models, as discussed by Shiri et al. (2014c). The performance fluctuations among the studied sites dictate the necessity of applying a complete data set scanning for evaluating the applied models.

Figs. 3 and 4 show the observed and simulated LAI values of the local GEP and RF models for the best models (as stated before). As observed, both LGEP and LRF models show considerable scatters around identity line, although the regression coefficients (a and b, respectively) of the fitted lines (in terms of $y = ax + b$) are respectively closer to 1 and 0 which shows their tendency to the identity function ($y = x$). On the other hand, as the LAI values are generally low (less than 3.5 in general), the discrepancy between the observed and simulated values for the studied models (which show accurate results) will be low and the observed scatters will have lower magnitudes as can be seen from Figs. 3 and 4. In case of the cropland sites, LGEP presents more scatter (in terms of both underestimation and overestimation) for values greater than 0.5 while LRF shows such trend for all ranges of LAI amount. In Grass2 site, however, the scatters between the observed and

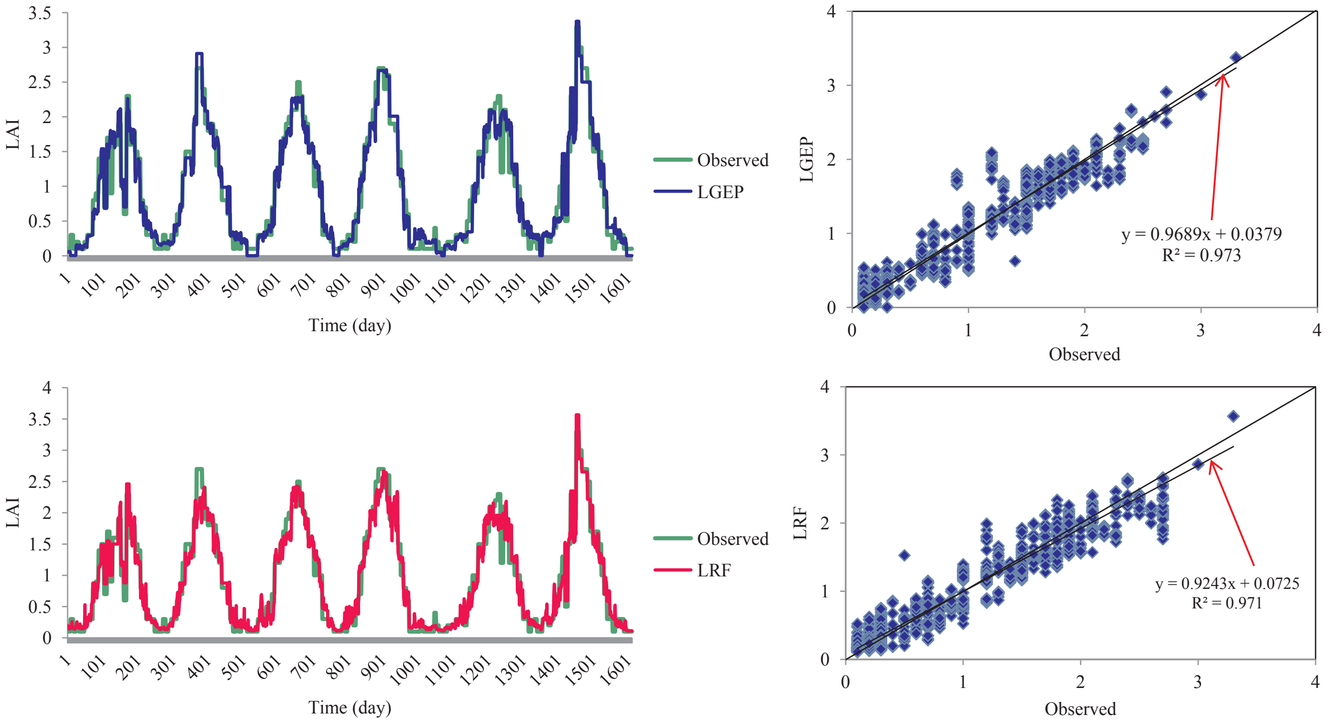

Fig. 4. Observed vs. simulated values of local models in Grass2.

simulated LAI amounts are distributed for all LAI magnitudes. Some unusual scatters in the plots might be explained by taking into consideration the quality of the applied data as well as sharp variations of LAI among different days/months which has made it difficult to get a more accurate simulation using the same input parameters with generally mild variations. Nevertheless, any possible variations of the input parameters which can affect LAI magnitudes and its simulation should be taken into account. The similar statements would stand for the external models.

For both the cropland and grassland sites, it seems difficult to assess the model performance in the climatic context of each location, as the present study considers only 5 sites for each land cover type (due to the high computational costs involved in the k-fold testing procedures as well as limitation of necessary records for modeling issue), and because these present similar climatic parameters (not presented here). In cropland sites, Crop5 has higher temperature magnitudes and slightly lower LAI records than the other locations. Further, Crop3 has slightly higher relative humidity records and the highest LAI. Similar variations of the meteorological variables are observed in grassland sites, where Grass5 presents the highest air temperature and LAI records while Grass4 shows the lowest NDVI and LAI values. Nevertheless, much couldn't be inferred here since the number of considered sites was low.

The performances of the GEP and RF models were also assessed temporally (per test year) at each site. For instance, the *SI* variations of the local GEP and RF models split per test year have been presented in Fig. 5 for sample stations. The figure clearly shows the considerable fluctuations of the *SI* values within test years. In case of Crop5, *SI* varies between 0.102 (test year 2005) and 0.231 (test year 2004) for GEP and between 0.115 and 0.24 (the same test years) for RF. Similarly, in Grass2, the *SI* oscillations is between 0.098 (test year 2004) and 0.29 (test year 2002) for GEP and between 0.1 and 0.3 (the same test years) for RF models. Such variability may be due to the relationships between the train/test patterns considered at each stage of modeling phase, where a part of patterns (here one year) was reserved for testing, then the model was trained using the rest of patters. In this way, any inconsistency or outlier values within the test/train patterns would crucially affect the performance accuracy of the trained models. The outcomes of this parts confirms the requirement of the complete data scanning procedure for evaluating the applied methodologies, otherwise, the obtained results would be partially valid as discussed by Marti et al. (2011).

Finally, the global cross-station assessment was achieved where all patterns of each land cover type were pooled and the GEP and RF models were trained using the pooled data from all sites of cropland sites, then tested using all patterns of the grassland sites. The same procedure was repeated by training the models using grassland data and testing using cropland data. Based on the *SI* values, GEP and RF models have *SI* values of 0.295 and 0.300 (for the first case) and 0.301 and 0.315 (for the second case), respectively.

Although the number of sites is limited due to the high computational costs as well as the availability of data (as stated before), and the chronological period of the applied patterns is generally short (limited study period), the applied models (i.e. GEP and RF) have lower degree of sensitivity to the amount of utilized patterns, so the applied techniques can be utilized with any chronological period (number of patterns; here the years), provided that the information content in the data set is sufficient to evolve a generalized formulation, as discussed by Deschaine (2014) and Shiri (2017).

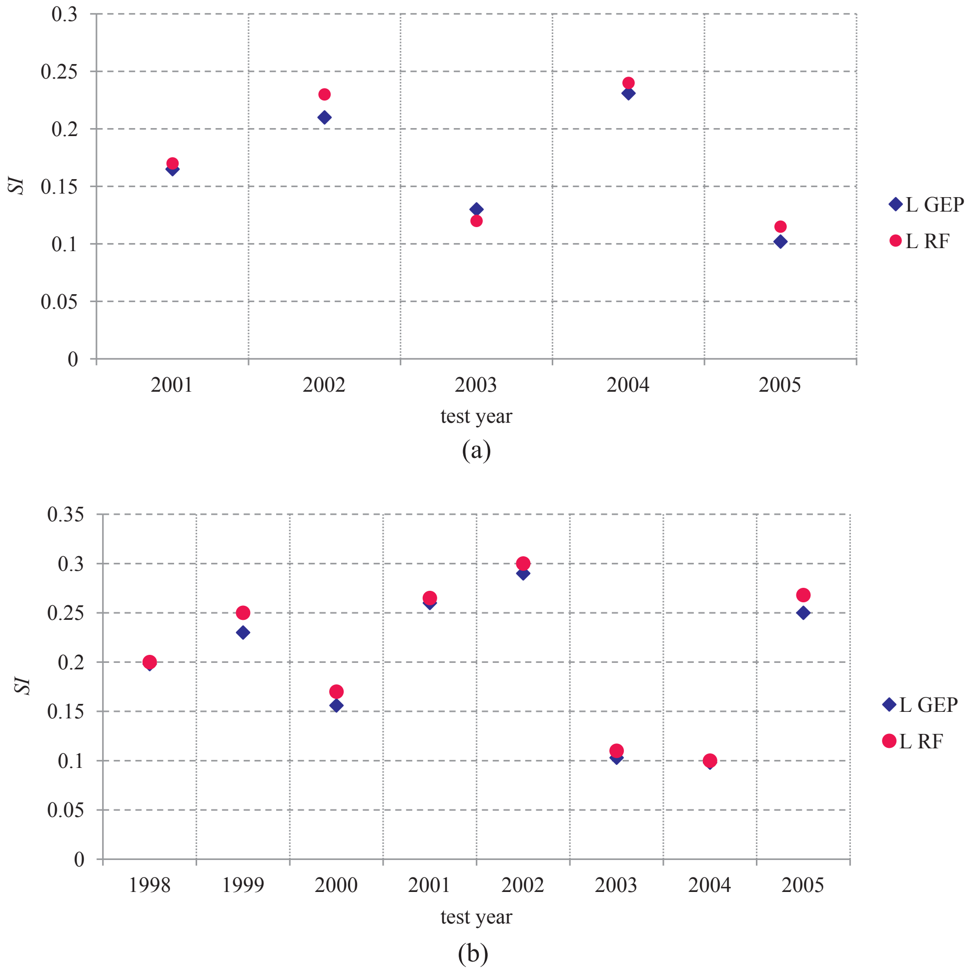

Fig. 5. Temporal variations of the *SI* values for sample sites:a) cropland (Crop5); b) grassland (Grass2).

Summarizing, the results obtained in the present study revealed that a complete cross validation data scanning procedure (i.e. k-fold testing) is very necessary to better assessment of the models' performance which confirms the results obtained by previously published literature (e.g. Marti et al., 2015; Roushangar et al., 2014b). The presented k-fold validation is infrequently applied in practical issues for assessing heursitic models in case of LAI simulation. Further, the presented external validation of data driven approaches outside the training station might be a valid alternative to conventional locally trained models. Further researches would be necessary using data from other land cover types and using other sites/techniques for confirming the outcomes of the present study. Nonetheless, in this study, LAI is predicted with remotely sensed LAI as the target. However, these estimates may not accurate when the remotely sensed LAI have large uncertainties. Therefore, the ground-based LAI should be collected broadly and used as target to enhance the models' prediction abilities.

4. Conclusions

The present study reports a new application of heuristic data driven models for simulating cropland and grassland LAI using meteorological variables. Meteorological and remotely sensed data from the mentioned land cover types were utilized to estimate the LAI through gene expression programming (GEP) and random forest (RF) techniques. A most robust cross validation approach, i.e. k-fold testing was adopted here for both local (temporal) and external (spatial) assessment of the applied models. Accordingly, the test patterns were selected either chronologically or geographically independent of the training patterns, so a complete data scanning procedure was fulfilled. In both land cover types, temporal (local) models gave the most accurate results since they were relying on the local inputs which were used for the train and test stages. However, the performance accuracy of the external models was comparable to the local ones, even though they were trained without considering data set of the test sites. This was a big step forward in case of the LAI modeling because the training of the local models would not be required if sufficient necessary data were available at other locations. The obtained results showed that externally trained GEP and RF models could be valid alternatives to locally trained models. This study was the first attempt for applying the local and external heuristic models (e.g. GEP and RF) for simulating LAI in literature. The presented application of local and external k-fold testing processes provided a sound evaluation of the GEP and RF performances.

References

Asner, G.P., Scurlock, J.M.O., Hicke, J.F., 2003. Global synthesis of leaf area index observation: implications for ecological and remote sensing studies. Glob. Ecol. Biogeogr. 12, 191–205.

Banzhaf, W., Nordin, P., Keller, P.E., Francone, F.D., 1998. Genetic Programming. Morgan Kaufmann, San Francisco, CA, pp. 512.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Chen, M., Willgoose, G.R., Saco, P.M., 2015. Investigating the impact of leaf area index temporal variability on soil moisture predictions using remote sensing vegetation data. J. Hydrol. 522, 274–284.

Combal, B., Bret, F., Trubuil, A., Mace, D., Pragnere, A., Mynenei, R., Knyazikhinc, Y., Wang, L., 2003. Retrieval of canopy biophysical variables from bidirectional reflectance—using prior information to solve the ill-posed inverse problem. Remote Sens. Environ. 48 (1), 1–15.

Coopersmith, E.J., Cosh, M.H., Daughtry, C.S.T., 2014. Field-scale moisture estimates using COSMOS sensors: a validation study with temporary networks and leaf-area-indices. J. Hydrol. 519 (Part A), 637–643.

Deschaine, L.M., 2014. Decision Support for Complex Planning Challenges: Combining Expert Systems, Engineering-Oriented Modeling, Machine Learning, Information Theory, and Optimization Technology. Chalmers University of Technology, Sweden 233 P.

Dunea, D., Moise, V., 2008. Artificial neural networks as a support for leaf are index modeling in crop canopies. In: 12th WSEAS International Conference on COMPUTERS, Heraklion, Greece, July 23–25, 440–445.

Everingham, Y.L., Smyth, C.W., Inman-Bamber, N.G., 2009. Ensemble data mining approaches to forecast regional sugarcane crop production. Agric. For. Meteorol. 149 (3–4), 689–696.

Fan, L., Gao, Y., Bruk, H., Bernhofer, C.H., 2009. Investigating the relationship between NDVI and LAI in semiarid grassland in Inner Mongolia using in-situ measurements. Theor. Appl. Climatol. 95, 151–156.

Ferreira, C., 2001a. Gene expression programming in problem solving. In: 6th Online World Conference on Soft Computing in Industrial Applications (invited tutorial).

Ferreira, C., 2001b. 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, pp. 478.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison –Wesley, Reading MA 432 pp.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learnin. Springer, New York.

He, B.B., Quan, X.W., Xing, M.F., 2013. Retrieval of leaf area index in alpine wetlands using a two-layer canopy reflectance model. Int. J. Appl. Earth Observ. Geoinform. 21, 78–91.

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83, 195–213.

Jacquemoud, S., acour, C., Poilve, H., Frangi, J.P., 2000. Comparison of four radiative transfer models to simulate plant canopies reflectance: direct and inverse mode. Remote Sens. Environ. 74 (3), 471–481.

Jonckheere, I., Fleck, S., Nackaerts, K., Muysa, B., Coppin, P., Weiss, M., Baret, F., 2004. Review of methods for in situ leaf area index determination Part I. Theories, sensors and hemispherical photography. Agric. For. Meteorol. 121, 19–35.

Karimi, S., Kisi, O., Kim, S., Nazemi, A.H., Shiri, J., 2017. Modelling daily reference evapotranspiration in humid locations of South Korea using local and cross-station data management scenarios. Int. J. Climatol. 37 (7), 3238–3246.

Koza, J.R., 1992. Genetic Programming: on the Programming of Computers by Means of Natural Selection. The MIT Press, Cambridge, MA, pp. 840.

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

Li, H., Zhang, Y., Chiew, F.S.H., Xu, S., 2009. Predicting runoff in ungauged catchments by using Xinanjiang model with MODIS leaf area index. J. Hydrol. 370 (1–4), 155–162.

Liang, S., Zhang, X., Xiao, Z., Cheng, J., Liu, Q., Zhao, X., 2014. Leaf Area Index. Global LAnd Surface Satellite (GLASS) Products. Springer Briefs in Earth Sciences 3–31.

Marti, P., Manzano, J., Royuela, A., 2011. Assessment of 4-input artificial neural network for ET_0 estimation through data set scanning procedures. Irrig. Sci. 29, 181–195.

Marti, P., Shiri, J., Duran-Ros, M., Arbat, G., Cartagena, F.R., Puig-Bargues, J., 2013. Artificial neural networks vs. gene expressions programming for estimating outlet dissolved oxygen in micro irrigation sand filters fed with effluents. Comput. Electron. Agric. 99, 176–185.

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 Manage. 149, 81–90.

Meroni, M., Colombo, R., Panigada, C., 2004. Inversion of a radiative transfer model with hyperspectral observations for LAImapping in poplar plantations. Remote Sens. Environ. 92 (2), 195–206.

Myneni, R., et al., 2002. Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. Remote Sens. Environ. 83 (1–2), 214–231.

Ploi, R., McPhee, N.F., 2008. Covariant parsimony pressure for genetic programming. Technical report CES-480, ISSN: 1744–8050.

Qu, Y., Zhang, Y., Wang, J., 2012. A dynamic Bayesian network data fusion algorithm for estimating leaf area index using time-series data from in situ measurement to remote sensing observations. Int. J. Remote Sens. 33, 1106–1125.

Qu, Y., Han, W., Fu, L., Li, C., 2014. LAINet – a wireless sensor network for coniferous forest leaf area index measurement: design, algorithm and validation. Comput. Electron. Agric. 108, 200–208.

Roushangar, K., Akhgar, S., Salmasi, F., Shiri, J., 2014a. Modeling energy dissipation over stepped spillways using machine learning approaches. J. Hydrol. 508, 254–265.

Roushangar, K., Mouaze, D., Shiri, J., 2014b. Evaluation of genetic programming-based models for simulating friction factor in alluvial channels. J. Hydrol. 517, 1154–1161.

Sellers, P.J., Hall, F.G., Asrar, G., Strebel, D.E., Murphy, R.E., 1988. The first ISLSCP field experiment (FIFE). Bull. Am. Meteorol. Soc. 69, 22–27.

Shiri, J., 2017. Evaluation of FAO56-PM, empirical, semi-empirical and gene expression programming approaches for estimating daily reference evapotranspiration in hyper-arid regions of Iran. Agric. Water Manage. 188, 101–114.

Shiri, J., Kisi, O., 2012. Estimation of daily suspended sediment load by using wavelet conjunction models. J. Hydrol. Eng. 17 (9), 986–1000.

Shiri, J., Kim, S., Kisi, O., 2014a. Estimation of daily dew point temperature using genetic programming and neural networks approaches. Hydrol. Res. 45 (2), 165–181.

Shiri, J., Sadraddini, A.A., Nazemi, A.H., Kisi, O., Landeras, G., Fakheri Fard, A., Marti, P., 2014b. Generalizability of gene expression programming-based approaches for estimating daily reference evapotranspiration in coastal stations of Iran. J. Hydrol. 508, 1–11.

Shiri, J., Marti, P., Singh, V.P., 2014c. Evaluation of gene expression programming approaches for estimating daily evaporation through spatial and temporal data scanning. Hydrol. Process. 28 (3), 1215–1225.

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.

Sonnentag, O., Talbot, J., Chen, J.M., Roulet, N.T., 2007. Using direct and indirect measurements of leaf area index to characterize the shrub canopy in an ombrot rophic peatland. Agric. For. Meteorol. 144, 200–212.

Stockli, R., Rutishauser, T., Dragoni, D., O'Keefe, J., Thornton, P.E., Jolly, M., Lu, L., Denning, A.S., 2008. Remote sensing data assimilation for a prognostic phenology model. J. Geophys. Res. 113.

Torres, A.F., Walker, W.R., McKee, M., 2011. Forecasting daily potential evapotranspiration using machine learning and limited climatic data. Agric. Water Manage 98 (4), 553–562.

Walthall, C., Dulaney, W., Anderson, M., Norman, J., Fang, H., Liang, S., 2004. A comparison of empirical and neural network approaches for estimating corn and soybean leaf area index from Landsat ETM+ imagery. Remote Sens. Environ. 92, 465–474.

Watson, D.J., 1947. Comparative physiological studies in the growth of field crops. I: variation in net assimilation rate and leaf area between species and varieties, and within and between years. Ann. Bot. 11, 41–76.

Weiss, M., Baret, F., Smith, G.J., Jonckheere, I., Coppin, P., 2004. Review of methods for in situ leaf area index (LAI) determination. Part II. Estimation of LAI, errors and sampling. Agric. For. Meteorol. 121, 37–53.

Wulder, M.A., LeDrew, E.F., Franklin, S.E., Lavigne, M.B., 1998. Aerial image texture information in the estimation of northern deciduous and mixed wood forest leaf area index (LAI). Remote Sens. Environ. 64 (1), 64–76.

Xiao, Z.Q., Liang, S., Wang, J.D., Chen, P., Yin, X.J., Zhang, L.Q., Song, J.L., 2014. Use of general regression neural networks for generating the GLASS leaf area index product from time-series MODIS surface reflectance. IEEE T. Geosci. Remote 52, 209–223.

Xu, T., Bateni, S.M., Liang, S., Entekhabi, D., Mao, K., 2014. Estimation of surface turbulent heat fluxes via variational assimilation of sequences of land surface temperatures from Geostationary Operational Environmental Satellites. J. Geophys. Res.-Atmos. 119, 10780–10798.

Yao, Y., Zhang, X., Dduan, Y., 2008. Impacts of climate change on pasture growth in Subalpine Meadows. Res. Sci. 30.