

Evaluating machine learning methods and satellite images to estimate combined climatic indices

Esmaeel Norouzi*, Saeed Behzadi**

ARTICLE INFO

Article history:

Received:

June 2019.

Revised:

July 2019.

Accepted:

August 2019.

Keywords:

Climatic phenomena

Remote sensing

Machine learning

Decision tree

Hybrid indicator

Abstract:

The reflections recorded on satellite images have been affected by various environmental factors. In these images, some of these factors are combined with other environmental factors that cannot be distinguished. Therefore, it seems wise to model these environmental phenomena in the form of hybrid indicators. In this regard, satellite imagery and machine learning methods can play a unique role in modeling and data mining of climatic phenomena as a result of their significant advantages, including their availability and analysis. Therefore, addressing the improvement and expansion of machine learning methods and modeling algorithms using remote sensing data is inevitable. In this study, 7 well-known machine learning algorithms are applied with different initial data to show that satellite images are able to estimate the combined indices more accurately. A new index (HT) is also defined by combining the quantities of relative humidity and temperature. Then, machine learning algorithms are trained for each of these three quantities. For each of the temperature and relative humidity quantities, four optimal bands were selected using the PCA method, then a combination of these optimal bands was determined for the HT index. Two criteria are used to validate the results: Root Mean Square Error (RMSE) statistic and comparing the map of the interpolation method with the result of this study. RMSE values show that satellite imagery could have a high ability to model and estimate composite indices. Classification-KNN-Coarse and Ensemble-Bagged Trees with accuracy of 79.8626 % and 84.9281% are identified as the best machine learning methods for temperature and relative humidity, while the best accuracy to estimate the HT index is 92.8792% for Matern 5/2 GPR. Therefore, it can be said that by changing the methods of database preparation, the modeling results can be changed effectively in order to train models.

1. Introduction

The reflection recorded on satellite images contains a substantial amount of information, because these reflections are affected by many environmental factors [1]. Therefore, it seems logical that the considered environmental factor can be discovered by establishing an appropriate mathematical relationship between several optimal bands of satellite images. For example, by eliminating the common environmental factors between the bands, the desired factor is achieved. On the other hand, some environmental features are incorporated in satellite imagery.

* MSc. Student in Geographic Information System (GIS), Shahid Rajaei Teacher Training University, Tehran, Iran.

** Corresponding Author: Assistant Professor in Survey Engineering, Faculty of Civil Engineering, Shahid Rajaei Teacher Training University, Tehran, Iran, Email: behzadi.saeed@gmail.com,

In other words, they should be extracted as the combined indicators of the satellite imagery. Since relative humidity and temperature indices are highly interrelated, their effect on satellite images may be indiscriminately recorded.

Today satellite imagery is widely available at very low cost or even free of charge. Therefore, using remote sensing data to model climate data is a convenient and economical way.

In recent years, the artificial intelligence approaches such as Coactive Neuro-Fuzzy Inference System (CANFIS), Adaptive Neuro-Fuzzy Inference System (ANFIS) which is hybrid of Artificial Neural Networks (ANN) and Fuzzy Inference System (FIS), Fuzzy-Logic (FL) [2], Radial Basis Neural Network (RBNN) which is a type of ANN, Support Vector Machines (SVM), Generalized Regression Neural Network (GRNN) which is a type of ANN, Genetic Algorithm (GA), Wavelet Transformation (WT) and Multi-

Layer Perceptron Neural Network (MLPNN) have been significantly utilized in diverse fields such as modelling climate indicators [3-15]. In order to model the target data, researchers used terrestrial station data as inputs, which have many limitations. Thus, many researchers tend to use satellite imagery [16-23]. In studying the model trees and neural network for modeling the rainfall, [12] it was concluded that tree models could be a suitable substitute for the neural network for precipitation modeling. MeikeKühnlein, in a study [16] about improving the accuracy of rainfall rates from optical satellite sensors with machine learning showed that, using machine learning methods can accurately improve the rates of precipitation, even up to hourly rates. Doña et al [20] used remote sensing to estimate the temporal variation of the flooded area, and their associated hydrological patterns related to the seasonality of precipitation and evapotranspiration. He applied several inverse modeling methods, such as two-band and multispectral indices, single-band threshold, classification methods, artificial neural network, support vector machine and genetic programming. The genetic programming approach yielded the best results, with a kappa value of 0.98, and a total error of omission-commission of 2%. Xu et al [19] upscaled evapotranspiration from eddy covariance flux tower sites to the regional scale with machine learning algorithms. Five machine learning algorithms were employed for evapotranspiration upscaling including artificial neural network; Cubist, deep belief network, random forest, and support vector machine. They demonstrated that the artificial neural network, Cubist, random forest, and support vector machine algorithms have almost identical performance in estimating evapotranspiration and have slightly lower Root Mean Square Error than deep belief network at the site scale.

There is a saying that apples shouldn't be compared with oranges or in other words, do not compare two items or group of items that are practically incomparable. However, the lack of comparability can be overcome if the two items or groups are somehow standardized or brought on the same scale. In a similar way, normalizing the RMSE facilitates the comparison between datasets or models with different scales.

The Heat Index (HI) is an index that combines air temperature and relative humidity in shaded areas, to posit a human-perceived equivalent temperature as how hot it would feel if the humidity were a different value in the shade. Like the wind chill index, the HI contains assumptions about the human body mass and height, clothing, amount of physical activity, individual heat tolerance, sunlight and ultraviolet radiation exposure, and the wind speed. Significant deviations from these will result in HI values which do not accurately reflect the perceived temperature [24-26].

New experimental techniques, as well as modern variants on known methods, have recently been employed to investigate the fundamental reactions underlying the oxidation of bio-char. The purpose of study executed by Bastistella and et.al [27] was to experimentally and statistically study how the relative humidity of air, mass, and particle size of four biochars influenced the absorption of water and increase in temperature. A random factorial design was employed using the intuitive statistical software Xlstat. A simple linear regression model and an analysis of variance with a pairwise comparison were performed. The experimental study was carried out on the wood of *Quercus pubescens*, *Cyclobalanopsis glauca*, *Trigonostemon huangmosun*, and *Bambusa vulgaris*, and involved five relative humidity conditions (22, 43, 75, 84, and 90%), two mass samples (0.1 and 1 g), and two particle sizes (powder and piece). Two response variables including water adsorption and temperature increase were analyzed and discussed. The temperature did not increase linearly with the adsorption of water. Temperature was modeled by nine explanatory variables, while water adsorption was modeled by eight. Five variables, including factors and their interactions, were found to be common to the two models. Sample mass and relative humidity influenced the two qualitative variables, while particle size and biochar type only influenced the temperature [27].

In this research, the relationship between Landsat 8 satellite imagery and terrestrial station data is investigated in estimating combined quantities using different machine learning methods. These include Artificial Neural Network, Neuro-Fuzzy (ANFIS), Classification-KNN, Regression-Robust Linear, Gaussian SVM, Matern 5/2 GPR, and Ensemble-Bagged Trees. Then the different methods are compared to find the optimal method. This study is conducted with the following objectives: (i) to select appropriate combination for input variables in the models using two ways, one of which is Principal Component Analysis (PCA) and the other one is applying standard deviation and Correlation as an indicator; (ii) to calibrate and validate the heuristic models with selected input variables; and (iii) to compare the results of the listed models with those of the interpolation based models, IDW.

2. Material and methods

2.1. Study area and data acquisition

The study area includes three provinces of Tehran, Alborz and Qazvin in Iran which is situated between 48° 43' 38.83" E to 53° 09' 11.70" E longitude and 34° 50' 14.29" N to 36° 47' 1.33" N latitude at an altitude of 1495.9m above MSL (Mean Sea Level) (Fig.1). The mean annual rainfall is about 1500mm at the study area. The area has six climatic sub regions, including dry, semi-arid, Mediterranean, semi-

humid, humid and very humid. In Fig.1, the study area and the position of stations applied in this study are shown on the region map. Daily relative humidity and temperature data are collected from Meteorological Organization and Water Resources Management Organization of Iran.

The primary data source was a series of Landsat-8 satellite images. The spatial resolution of the band 8 for the satellite is 15 m; other bands (1,2,3,4,5,6,7,9,10,11) have a spatial resolution of 30 m. The selected Landsat-8 scenes are path 164 / row 35 and path 165/ row 35, which covers the extent of the area of interest. Scenes that were mostly cloud-free from 2013 to 2016 are obtained from the United States Geological Survey Earth Explorer website [28]. All images were Level 1T products, which have been precision and terrain corrected in the GeoTIFF format and are in the UTM Zone 39S projection and WGS datum [29]. The resulting dataset comprised 34 full scenes.

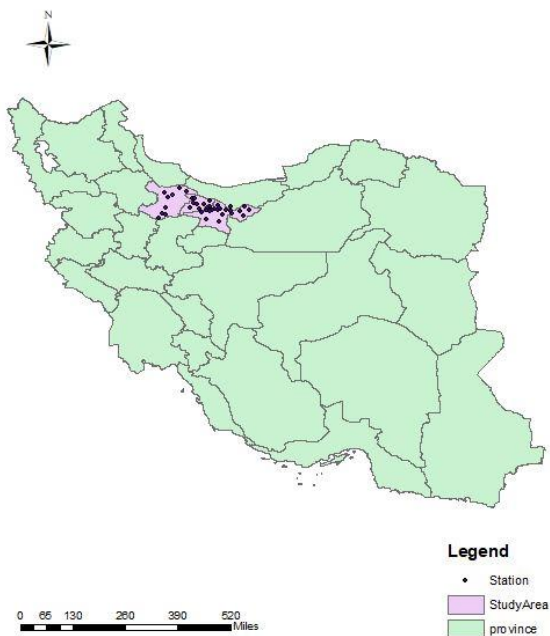


Fig. 1: Location map of the study area and Distribution of terrestrial stations

2.2. Preparing the initial datasets for Machine Learning

Machine Learning (ML) depends heavily on data, which is the most crucial aspect that makes algorithm training possible and explains why machine learning has become so popular in recent years. Regardless of actual terabytes of information and data science expertise, if data records are not prepared and organized, a machine will be nearly useless or perhaps even harmful. Without exception, all datasets needed correction. That is why data preparation is such an important step in the machine learning process. In a nutshell, data preparation is a set of procedures that helps make the dataset more suitable for machine learning. In broader terms, the data preparation also includes establishing the right data

collection mechanism. These procedures consume the utmost time spent on machine learning. Sometimes it takes months before the first algorithm is built [30, 31].

Knowing what must be modeled or estimated helps to decide which data may be more valuable to collect. When formulating the problem, data exploration must be conducted, and reasoned in the categories of classification, clustering, regression, and ranking. For instance, when an algorithm needs to answer binary yes-or-no questions, classification is the best method, or when it comes to finding the rules of classification and the number of classes, clustering is a suitable choice.

Generally since the surface evaporation dataset is formed by numerical values, regression algorithm is more beneficial to this case [30, 31]; however, (i) it must be considered that surface evaporation depends on numerous factors, which makes it too complicated to be formulated accurately. (ii) On the other hand, it is tempting to include as much data as possible, because of its reliability and quantity. Since the target attribute (the value-needed to be modeled) is known, common sense guides the rest. It can be assumed which values are critical and which ones are going to add more dimensions and complexity to the dataset without any useful contribution. This approach is called attribute sampling. (iii) Since missing values can tangibly reduce prediction accuracy, this issue must be addressed as a priority. In terms of machine learning, assumed or approximated values are “more right” for an algorithm than just missing ones. Hence in this study, all the well-known methods of machine learning are mostly applied with two statistical data preparation techniques, one of which is PCA and the other which uses standard deviation and correlation, in order to find the best algorithm for the purpose of the study.

2.2.1. Defining a combined index (HT)

In order to clarify the capabilities of satellite imagery and machine learning algorithms in the estimation and modeling of combined indices, daily temperature and relative humidity are used to define a new index Humidity-Temperature (HT) seen in formula (1). Then the result of this study is calculated and analyzed for each of these three parameters (temperature, relative humidity and HT index):

$$HT = \frac{H - T}{H + T} \quad (1)$$

Where, HT is the new combined index, H is the relative humidity and T is the temperature. A total of 1291 HT sample records (40 stations in 34 pairs of images during the 3 years) are collected. The data records are split into two sub-datasets: 80% of the records are selected randomly to train the machine learning algorithms and develop the estimation model, and the 20% remaining are used for validation and evaluations.

2.2.2. Determining optimal bands for relative humidity and temperature

By using standard deviation and correlation statistics, parameters that have greater correlation and amplitude than the others are determined, and then a specific number of optimal parameters is used instead of all the parameters and executed according to formula (2), in this study. This formula is set for three parameters, but in this research, it is applied for five parameters (including four bands of satellite imagery and surface evaporation values):

$$x = \frac{(std(B1) + std(B2) + std(B3))}{|corr(B1, B2)| + |corr(B1, B3)| + |corr(B2, B3)|} \quad (2)$$

Table 1: The result of the calculations using equation (1). Extraction of four optimal bands using nine satellite image bands and with respect to temperature values.

B1	B2	B3	B4	B5	B6	B7	B9	B10	Temp(°C)
...
...
431.30	448.86	407.44	361.94	234.69	41.59	11.79	8.27	4.73	8.1
388.77	406.29	359.77	322.18	211.47	40.76	11.77	8.6	4.70	3.5
303.31	311.84	280.64	249.34	171.22	40.39	11.19	6.11	5.67	15.4
...

State	X Index
...	...
B1B2B7B10T	249.85
B1B2B9B10T	560.06
B1B3B4B5	55.44
...	...

B1	B2	B9	B10	Temp(°C)
...
...
431.30	448.86	8.27	4.73	8.1
388.77	406.29	8.6	4.70	3.5
303.31	311.84	6.11	5.67	15.4
...

Table 2: The result of the calculations using equation (1). Extraction of four optimal bands using nine satellite image bands and with respect to relative humidity values.

B1	B2	B3	B4	B5	B6	B7	B9	B10	RH
...
...
236.44	240.27	205.74	179.50	115.60	14.53	4.26	5.37	3.94	36.59
180.19	178.17	148.35	126.62	80.74	9.76	2.67	3.91	4.26	32.40
268.55	273.45	232.88	203.74	130.84	15.71	4.29	6.72	3.85	29.4
...

State	X Index
...	...
B4B6B9B10T	-714.96
B4B7B9B10RH	6906.29
B5B6B7B9	26.68
...	...

B1	B2	B9	B10	RH
...
...
236.44	240.27	5.37	3.94	36.59
180.19	178.17	3.91	4.26	32.40
268.55	273.45	6.72	3.85	29.4
...

2.2.3. Determining optimal bands for HT index

Major databases are increasingly expanding and publicizing, while making them more difficult to interpret. Principal Component Analysis (PCA) is a technique for reducing the size of such databases, increasing the capability of interpreting, and simultaneously minimizing data problems. The PCA technique does this by creating a new variable that maximizes the variance successively [32]. Table 3 illustrates the result of the PCA technique. All selected bands for relative humidity and temperature in the previous sections (1, 2, 4, 7, 9 and 10) are used to extract four optimal bands.

2.3. Machine Learning algorithms

In general, 7 machine learning methods are used to achieve the study objectives, each of which is briefly described in this section:

2.3.1. Artificial Neural Network (ANN)

Neural network is one of the techniques of machine learning, the application of which has been proved in numerous studies such as modeling and predicting many phenomena including climate phenomena. Sulaiman and Wahab [13, 33] describe the modeling and prediction of heavy rainfall.

Which, x is the benchmark for optimization or Optimum Index Factor (OIF), B_i is parameter (bands and relative humidity or temperature value), std and corr are standard deviation and correlation respectively. The x index in formula (2) is calculated for all possible states in selecting four bands among nine bands of satellite images. The state which has maximum value of the optimization index is considered as the best quad-combination of 9 bands of landsat-8 images. The selected state is the combination of bands 1, 2, 9, and 10 for temperature and 4, 7, 9, and 10 for relative humidity. Tables 1 and 2 show the results of the calculations for relative humidity and temperature respectively.

Although it was difficult to model and predict climatic phenomena, machine learning methods, especially artificial neural networks, are reliable and can be used for climate phenomena such as precipitation and surface evaporation. The data structures and functionality of neural nets are designed to simulate associative memory. Neural nets learn by processing examples, each of which contains a known "input" and "result," forming probability-weighted associations between the two, which are stored within the data structure of the net itself. (The "input" here is more accurately called an input set, since it generally consists of multiple independent variables, rather than a single value.) [9].

2.3.2. Neuro-Fuzzy (ANFIS)

ANFIS is an artificial neural network based on the Takagi-Sugeno fuzzy system [34]. Since this system combines neural networks and fuzzy logic concepts, we can use both of them at the same frame. Neuro-fuzzy hybridization results in a hybrid intelligent system that synergizes these two techniques by combining the human-like reasoning style of fuzzy systems with the learning and connectionist structure of neural networks. Neuro-fuzzy hybridization is widely

termed as Fuzzy Neural Network (FNN) or Neuro-Fuzzy System (NFS) in the literature. Neuro-fuzzy system (the more popular term is used henceforth) incorporates the human-like reasoning style of fuzzy systems through the use

of fuzzy sets and a linguistic model consisting of a set of IF-THEN fuzzy rules. The main strength of neuro-fuzzy systems is that, they are universal approximations with the ability to solicit interpretable IF-THEN rules [35].

Table 3: the result of the PCA technique. Extraction of four optimal parameters using six optimal satellite image bands and with respect to HT index values.

B1	B2	B4	B7	B9	B10	HT	P1	P2	P3	P4	HT
...
76.34	71.53	50.92	3.00	0.44	9.11	0.55	-1.26	-0.31	0.04	0.12	0.55
204.18	206.54	154.01	0.93	1.99	5.23	1.11	1.32	-1.59	-1.10	0.50	1.11
117.83	122.82	123.97	7.68	0.21	11.64	0.23	-0.13	2.17	0.01	-0.01	0.23
...

2.3.3. Classification-KNN-Coarse

The KNN algorithm is one of the simplest data mining and classification algorithms. This algorithm performs simple classification operations and returns reliable results as predictions. In a literal sense, this method chooses the tracks in which the selected neighborhood has the highest number of records attributed to them. Therefore, traces that are more closely related to each other in the K nearest neighbor are considered as the new record category [36].

2.3.4. Robust Regression-Linear

In robust statistics, robust regression is a form of regression analysis designed to overcome some limitations of traditional parametric and non-parametric methods. Regression analysis seeks to find the relationship between one or more independent variables and a dependent variable. Certain widely used methods of regression, such as ordinary least squares, have favorable properties if their underlying assumptions are reliable, but can give misleading results if those assumptions are not supportive. Therefore, ordinary least squares are said not to be reasonable for violations of its assumptions. Robust regression methods are designed not to be overly affected by violations of assumptions by the underlying data-generating process [37].

2.3.5. Gaussian SVM

One of the most common methods in the data classification domain is the support vector machine or SVM algorithm. In simple terms, support vectors are a collection of points in the n-dimensional data that defines the boundaries of the categories. The classification of the data is based on these points, and the output of the classification may be changed by moving one of them. SVM is basically a binary separator. A multi-class pattern recognition can be achieved by combining two-class vector machines [38].

2.3.6. Matern 5/2 GPR

Gaussian process consists of a set of random variables as one of the new methods of data mining with its normal characteristics, and has a high ability to solve nonlinear problems by using kernel functions. The Gaussian

regression models are based on the assumption that the regulatory observation should carry information about one another. Gaussian processes are a way to specify the priority directly on the function space [39].

2.3.7. Ensemble-Bagged Trees

Bagging is an ensemble technique which is used when the goal is to reduce the variance of a decision tree [40]. Here, the idea is to create several subsets of data from training sample chosen randomly with replacement. At this point, each collection of subset data is used to train their decision trees. As a result, we end up with an ensemble of different models. Average of all the predictions from different trees are more robust than a single decision tree.

3. Results and discussion

In this study, after preparing (or so-called GIS-Ready) the temperature and relative humidity data for 40 stations, along with corrected reflectance values of Landsat-8 satellite imagery, four optimal bands of the nine bands (1, 2, 3, 4, 5, 6, 7, 9, and 10) are selected in Landsat-8 satellite imagery to continue the work. Two ways are applied: first one is Optima Index Factor (OIF) to estimate temperature and relative humidity values of satellite imagery, and the second one is PCA method for the HT index. In each of which, the data is divided into two categories: train (80%) and test (20%).

In the next step, train data is introduced into each of the modeling methods to model the relationship between satellite images and temperature, relative humidity and HT index values; and then it generates a simulator or decision function for each of them. Since 7 machine learning methods are applied, 7 functions are trained for each of the three quantities (temperature, relative humidity and HT index). Models' accuracies are evaluated according to (i) the Root Mean Square Error (RMSE) statistic and (ii) comparison with the generated map of the interpolation method. The RMSE can be expressed as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_{i,observed} - X_{i,predicted})^2} \quad (3)$$

Where N is the number of data points used in the study, and X represents considered quantity.

3.1. Modeling daily temperature and relative humidity

In this part, the 7 machine learning methods are applied to estimate temperature and relative humidity of 40 stations in the studied area using four optimal bands obtained by OIF (Equation.2). Tables 4 and 5 show test results of the applied models in estimating temperature and relative humidity data. The minimum and maximum RMSE values for temperature are 9.1826 and 11.3870 for Classification-KNN-Coarse and Neuro-Fuzzy (ANFIS) methods respectively. Therefore, according to RMSE values, Classification-KNN-Coarse method is known as the best method among the others for modeling temperature. Figure (2) illustrates the error variation of validation results especially for the artificial neural network model. The minimum and maximum RMSE values for relative humidity are 14.8639 and 90.5108 for Ensemble-Bagged Trees and Neuro-Fuzzy (ANFIS) methods respectively. Therefore, it is clearly evident that the Ensemble-Bagged Trees model has the best performance among the other models for modeling relative humidity. Figure (3) illustrates the error variation of validation results especially for the Matern 5/2 GPR models. It is also clear that all models generally offer the same precision.

Table 4: Test results of the applied models in estimating temperature

Machine Learning Algorithms	RMSE (mm)	100 - Normalized RMSE (%)
Artificial neural network	10.84755	76.21151
Neuro-Fuzzy(ANFIS)	11.38702	75.02846
Classification-KNN-Coarse	9.182639	79.86263
Robust Regression-Linear	11.13681	75.57717
Gaussian SVM	11.27292	75.27868
Matern 5/2 GPR	11.29329	75.23402
Ensemble-Bagged Trees	11.17591	75.49143

Table 5: Test results of the applied models in estimating relative humidity

Machine Learning Algorithms	RMSE (mm)	100 - Normalized RMSE (%)
Artificial neural network	15.1214	84.66701
Neuro-Fuzzy(ANFIS)	90.5108	8.222673
Classification-KNN-Coarse	17.76584	81.98556
Robust Regression-Linear	15.54172	84.2408
Gaussian SVM	15.09846	84.69027
Matern 5/2 GPR	14.89611	84.89545
Ensemble-Bagged Trees	14.8639	84.92811

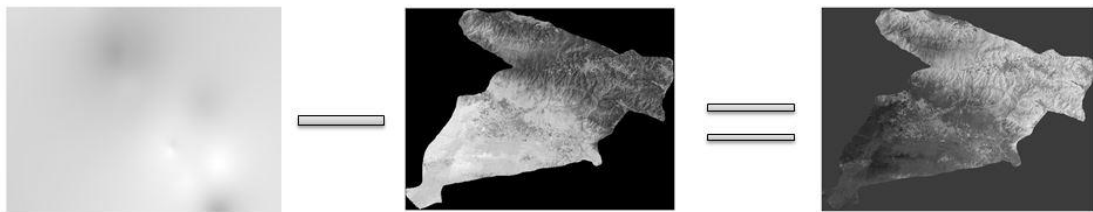


Fig. 2: The error variation of the test temperature results especially for the artificial neural network model. From the left, the map obtained from the IDW interpolation method and the ANN machine learning algorithm and spatial distribution of errors

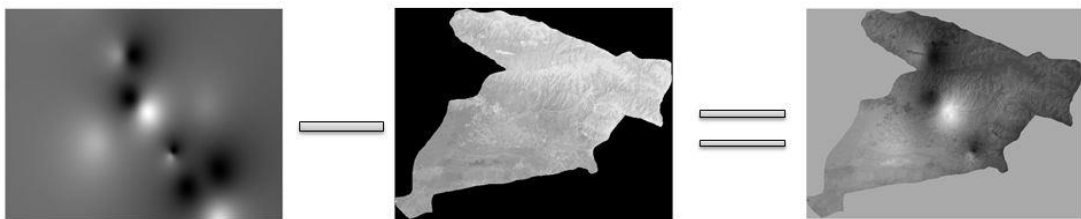


Fig. 3: The error variation of the test relative humidity results especially for the Matern 5/2 GPR model. From the left, the map obtained from the IDW interpolation method and the Matern 5/2 GPR machine learning algorithm and spatial distribution of errors

3.2. Modeling daily HT index

At this stage, machine learning methods were applied to estimate hybrid HT index of 40 stations on the study area using four optimal bands gained from the optimal bands used in the previous section by the PCA method. Table 6 gives the test results of the applied models in estimating HT

data. RMSE values range from 0.1993 to 0.5782 mm for the 7 models. The minimum RMSE values were found for Matern 5/2 GPR (test set) while the Neuro-Fuzzy (ANFIS) model provided the worst accuracy. Therefore, according to RMSE values, the Matern 5/2 GPR method is known as the best method among the other applied methods for HT index modeling.

Table 6: Test results of the applied models in order to estimate HT index values

Machine Learning Algorithms	RMSE (mm)	100 - Normalized RMSE (%)
Artificial neural network	0.215028	92.32042
Neuro-Fuzzy (ANFIS)	0.578229	79.34897
Classification-KNN-Coarse	0.24697	91.17966
Robust Regression-Linear	0.19977	92.86536
Gaussian SVM	0.21819	92.2075
Matern 5/2 GPR	0.19938	92.87929
Ensemble-Bagged Trees	0.20216	92.78

Figure (4) illustrates the error variation of test results especially for the ANN model. From the figure, it is clear that all models generally provided similar accuracy and compared to the results obtained for relative humidity and temperature quantities in the previous section, satellite images can be considered as highly capable tools of estimating phenomena such as the HT index.

Generally, calculating Root Mean Squared Error (RMSE) indicated that the combination of temperature and relative humidity parameters can be better derived from satellite imagery. In other words, there are many environmental parameters that are recorded as a unique quantity combined

with other parameters in satellite imagery, and these quantities are not capable of derivation in most cases. Thus, significant changes are seen in the mean square error of the HT index compared to the temperature and relative humidity. Therefore, it is easy to see the ability of different machine learning methods and satellite images to estimate combined indices in Table 6. For each quantity, most models have very close RMSE values, which proves that machine learning decision making models are valid in modeling climatic phenomena such as HT index using remote sensing data, and applying these decision models for modeling and data mining is inevitable in the future.

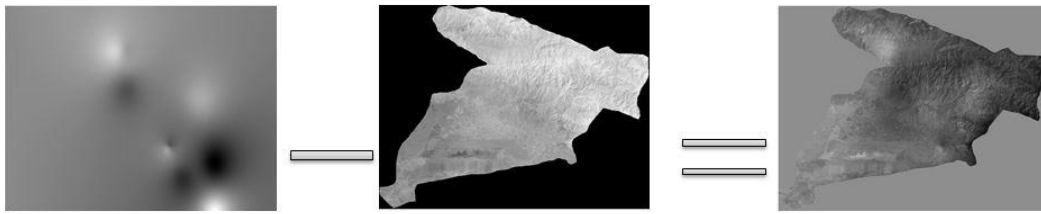


Fig. 4: The error variation of the test HT index results spatially for the ANN model. From the left, the map obtained from the IDW interpolation method and the ANN machine learning algorithm and spatial distribution of errors

In Figures 2, 3 and 4, the maps are obtained from both the interpolation method and the machine learning methods. These figures indicate the validity of the result for the combined HT index compared with the relative temperature and relative humidity. In these figures, the spatial distribution of the difference between the interpolation and machine learning methods is also seen. It is obvious that by changing the methods of database preparation in order to train the models, the modeling results can be changed effectively. Therefore, in this section, satellite imagery and machine learning algorithms are re-established in the estimation of combined environmental quantities.

4. Conclusion

With regard to the importance of changing the future climate of the planet and the wide effects on the various aspects of meteorological and hydrological issues, extensive efforts have been made in order to extract climatic data, more accurately and at the same time less costly and without human and physical errors, in the future. On the other hand, the changes in climate variables in regional scales are not explicitly identified, as they depend on a large number of

local factors. Hence, addressing the improvement and expansion of machine learning methods and modeling algorithms using remote sensing data is inevitable. One of the outputs of this research is simulation models for data mining through satellite imagery, which is shown in Figure 5. This figure shows an example of these products.

In this research, in order to study the appropriate methods for modeling and data mining for the surface evaporation, we employed important methods of machine learning and the time series of remote sensing, meteorological data and their integration, as well as the impact of the use of methods such as PCA and OIF which were used to prepare data as a result of training the models. This research proved that the combined HT index is estimated with higher precision instead of relative humidity and temperature values through satellite imagery. According to the results, the artificial neural network model had acceptable performance in both methods, and it was quite evident that the impact of the methods of database preparation could be impressively significant. Since the discussion of data preparation for training modeling algorithms has not yet been sufficiently considered, and given the significant effect on the results of the obtained models, it is suggested to pay more attention to

this issue in future studies. For example, combined indices such as HT can be extracted with high precision of satellite

images, and then the temperature and relative humidity quantities can be derived from that.

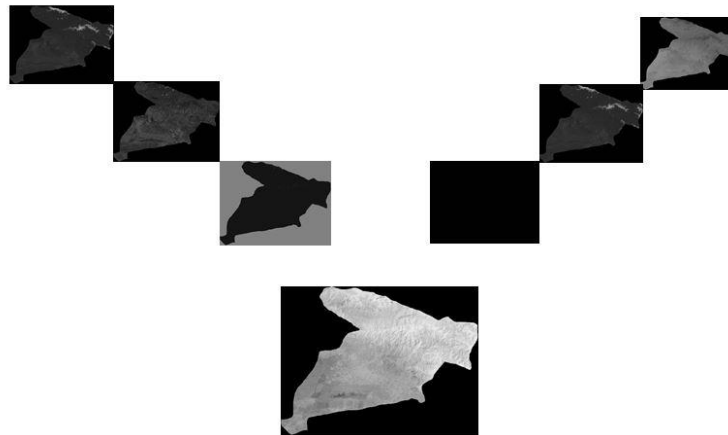
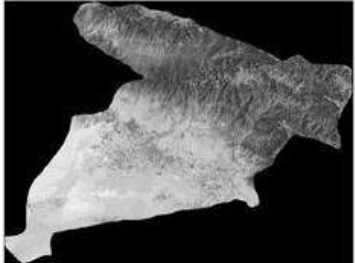
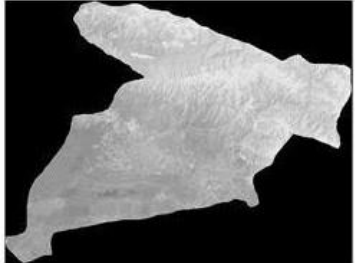
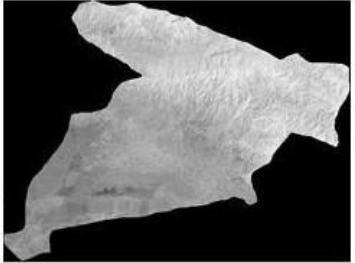


Fig. 5: An example of this study’s products. Top, six optimized bands selected by using the index X (formula.2), Down, HT values map prepared from these optimal bands through the ANN algorithm.

Table 7: Test final results of the research

Parameter	Temperature	Relative Humidity	HT
The result map			
The best ML algorithm	Classification-KNN-Coarse	Ensemble-Bagged Trees	Matern 5/2 GPR
100 - Normalized RMSE (%)	79.86	84.92	92.87

References

[1] Q. Weng, *Advances in environmental remote sensing: sensors, algorithms, and applications*, CRC Press, 2011.

[2] S. Behzadi, Z. Mousavi, E. Norouzi, *Mapping Historical Water-Supply Qanat Based On Fuzzy Method. An Application to the Isfahan Qanat (Isfahan, Iran)*, *International Journal of Numerical Methods in Civil Engineering*, 3(4) (2019) 24-32.

[3] Ö.J.I.S. Kişi, *Modeling monthly evaporation using two different neural computing techniques*, 27(5) (2009) 417-430.

[4] Ö. Kişi, *Evolutionary neural networks for monthly pan evaporation modeling*, *Journal of Hydrology*, 498 (2013) 36-45.

[5] O. Kisi, *Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree*, *Journal of Hydrology*, 528 (2015) 312-320.

[6] S. Kim, J. Shiri, O.J.W.R.M. Kisi, *Pan Evaporation Modeling Using Neural Computing Approach for Different Climatic Zones*, 26(11) (2012) 3231-3249.

[7] A. Guven, O. Kisi, *Monthly pan evaporation modeling using linear genetic programming*, *Journal of Hydrology*, 503 (2013) 178-185.

[8] Ö. Kişi, M.J.J.o.h. Tombul, *Modeling monthly pan evaporations using fuzzy genetic approach*, 477 (2013) 203-212.

[9] M.K. Goyal, B. Bharti, J. Quilty, J. Adamowski, A. Pandey, *Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS*, *Expert Systems with Applications*, 41(11) (2014) 5267-5276.

[10] L. Wang, B. Hu, O. Kisi, M. Zounemat-Kermani, W.J.Q.J.o.t.R.M.S. Gong, *Prediction of diffuse photosynthetically active radiation using different soft computing techniques*, 143(706) (2017) 2235-2244.

[11] A. Malik, A. Kumar, O.J.C. Kisi, E.i. Agriculture, *Monthly pan-evaporation estimation in Indian central Himalayas using different heuristic approaches and climate based models*, 143 (2017) 302-313.

[12] D.P. Solomatine, K.N.J.H.S.J. Dulal, *Model trees as an alternative to neural networks in rainfall—runoff modelling*, 48(3) (2003) 399-411.

[13] J. Sulaiman, S.H. Wahab, *Heavy Rainfall Forecasting Model Using Artificial Neural Network for Flood Prone*

Area, in: IT Convergence and Security 2017, Springer, 2018, pp. 68-76.

[14] X. Lu, Y. Ju, L. Wu, J. Fan, F. Zhang, Z. Li, Daily pan evaporation modeling from local and cross-station data using three tree-based machine learning models, *Journal of Hydrology*, 566 (2018) 668-684.

[15] S.A. Naghibi, H.R. Pourghasemi, B.J.E.m. Dixon, assessment, GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran, 188(1) (2016) 44.

[16] M. Kühnlein, T. Appelhans, B. Thies, T.J.R.S.o.E. Nauss, Improving the accuracy of rainfall rates from optical satellite sensors with machine learning—A random forests-based approach applied to MSG SEVIRI, 141 (2014) 129-143.

[17] T. Lillesand, R.W. Kiefer, J. Chipman, *Remote sensing and image interpretation*, John Wiley & Sons, 2014.

[18] S. Ghimire, R.C. Deo, N.J. Downs, N.J.R.S.o.E. Raj, Self-adaptive differential evolutionary extreme learning machines for long-term solar radiation prediction with remotely-sensed MODIS satellite and Reanalysis atmospheric products in solar-rich cities, 212 (2018) 176-198.

[19] T. Xu, Z. Guo, S. Liu, X. He, Y. Meng, Z. Xu, Y. Xia, J. Xiao, Y. Zhang, Y.J.J.o.G.R.A. Ma, Evaluating different machine learning methods for upscaling evapotranspiration from flux towers to the regional scale, 123(16) (2018) 8674-8690.

[20] C. Doña, N.-B. Chang, V. Caselles, J.M. Sánchez, L. Pérez-Planells, M.d.M. Bisquert, V. García-Santos, S. Imen, A.J.R.S. Camacho, Monitoring hydrological patterns of temporary lakes using remote sensing and machine learning models: Case study of la Mancha Húmeda Biosphere Reserve in central Spain, 8(8) (2016) 618.

[21] Q. Zhou, A. Flores, N.F. Glenn, R. Walters, B.J.P.o. Han, A machine learning approach to estimation of downward solar radiation from satellite-derived data products: An application over a semi-arid ecosystem in the US, 12(8) (2017) e0180239.

[22] K. Kuwata, R. Shibasaki, Estimating crop yields with deep learning and remotely sensed data, in: *Geoscience and Remote Sensing Symposium (IGARSS)*, 2015 IEEE International, IEEE, 2015, pp. 858-861.

[23] J. Rogan, J. Franklin, D. Stow, J. Miller, C. Woodcock, D. Roberts, Mapping land-cover modifications over large areas: A comparison of machine learning algorithms, *Remote Sensing of Environment*, 112(5) (2008) 2272-2283.

[24] G.B. Anderson, M.L. Bell, R.D. Peng, Methods to calculate the heat index as an exposure metric in environmental health research, *Environmental health perspectives*, 121(10) (2013) 1111-1119.

[25] M.S. Jin, Developing an index to measure urban heat island effect using satellite land skin temperature and land cover observations, *Journal of Climate*, 25(18) (2012) 6193-6201.

[26] L.P. Rothfus, N.S.R. Headquarters, *The heat index equation (or, more than you ever wanted to know about heat index)*, Fort Worth, Texas: National Oceanic and Atmospheric Administration, National Weather Service, Office of Meteorology, 9023 (1990).

[27] L. Bastistella, P. Rousset, A. Aviz, A. Caldeira-Pires, G. Humbert, M. Nogueira, Statistical Modelling of Temperature and Moisture Uptake of Biochars Exposed to Selected Relative Humidity of Air, *Bioengineering*, 5(1) (2018).

[28] T.U.S.G.S.E.E.A.o. <http://earthexplorer.usgs.gov>.

[29] F. Ling, G.M. Foody, H. Du, X. Ban, X. Li, Y. Zhang, Y.J.R.S. Du, Monitoring thermal pollution in rivers downstream of dams with Landsat ETM+ thermal infrared images, 9(11) (2017) 1175.

[30] M. Goodson, *Preparing Your Dataset for Machine Learning: 8 Basic Techniques That Make Your Data Better*, (2017).

[31] R.S. Michalski, J.G. Carbonell, T.M. Mitchell, *Machine learning: An artificial intelligence approach*, Springer Science & Business Media, 2013.

[32] I.T. Jolliffe, J.J.P.T.R.S.A. Cadima, Principal component analysis: a review and recent developments, 374(2065) (2016) 20150202.

[33] A. Jalilzadeh, S. Behzadi, Machine Learning Method for predicting the depth of shallow lakes Using Multi-Band Remote Sensing Images, *Soft Computing in Civil Engineering*, 3(2) (2019) 59-68.

[34] J.-S.J.I.t.o.s. Jang, man., cybernetics, ANFIS: adaptive-network-based fuzzy inference system, 23(3) (1993) 665-685.

[35] A. Abraham, Adaptation of fuzzy inference system using neural learning, in: *Fuzzy systems engineering*, Springer, 2005, pp. 53-83.

[36] K.J.I.J.o.A.R.i.C. Khamar, C. Engineering, Short text classification using kNN based on distance function, 2(4) (2013) 1916-1919.

[37] R.A. Maronna, R.D. Martin, V.J. Yohai, M. Salibián-Barrera, *Robust statistics: theory and methods (with R)*, Wiley, 2018.

[38] C.J.J.D.m. Burges, k. discovery, A tutorial on support vector machines for pattern recognition, 2(2) (1998) 121-167.

[39] M. Pal, S.J.C. Deswal, *Geotechnics*, Modelling pile capacity using Gaussian process regression, 37(7-8) (2010) 942-947.

[40] P. Van der Linden, J. Mitchell, editors, *ENSEMBLES: Climate change and its impacts-Summary of research and results from the ENSEMBLES project*, (2009).