



# Reference Evapotranspiration Prediction from Limited Climatic Variables Using Support Vector Machines and Gaussian Processes

Yasser Zouzou<sup>1</sup>, Hatice Citakoglu<sup>2\*</sup>

<sup>1</sup> Department of Civil Engineering, Erciyes University, Kayseri, Turkey (ORCID ID 0000-0002-9690-3882), [yasserzouzou@gmail.com](mailto:yasserzouzou@gmail.com)

<sup>2</sup> Department of Civil Engineering, Erciyes University, Kayseri, Turkey (ORCID ID 0000-0001-7319-6006), [hcitakoglu@erciyes.edu.tr](mailto:hcitakoglu@erciyes.edu.tr)

(1st International Conference on Applied Engineering and Natural Sciences ICAENS 2021, November 1-3, 2021)

(DOI: 10.31590/ejosat.999319)

**ATIF/REFERENCE:** Zouzou, Y., Citakoglu, H., (2021). Reference Evapotranspiration Prediction from Limited Climatic Variables Using Support Vector Machines and Gaussian Processes. *European Journal of Science and Technology*, (28), 346-351.

## Abstract

Climatic variables collected from weather stations evenly distributed in all regions of Turkey were used to study the potential of Gaussian Process Regression (GPR) and Support Vector Regression (SVR) in predicting reference evapotranspiration (ET<sub>0</sub>). The variables used as input features for the GPR and SVR models were solar radiation, mean temperature, wind speed, relative humidity, and month of the year. The corresponding ET<sub>0</sub> values were calculated using the Food and Agriculture Organization recommended equation FAO 56 PM using climatic measurements collected from the same stations. Results show that regression models with high accuracies are possible using GPR and SVR models. The most effective input variable for ET<sub>0</sub> prediction was found to be solar radiation. Relative humidity had the lowest impact on model accuracies.

**Keywords:** Reference Evapotranspiration, Gaussian Processes Regression, Support Vector Regression.

## Destek Vektör Makineleri ve Gauss Süreçleri Kullanılarak Sınırlı İklim Değişkenlerinden Referans Evapotranspirasyon Tahmini

### Öz

Türkiye'nin tüm bölgelerine eşit olarak dağılmış hava istasyonlarından toplanan iklim değişkenleri, Gaussian Proses Regresyon (GPR) ve Destek Vektör Regresyonunun (SVR) referans evapotranspirasyonu (ET<sub>0</sub>) tahmin etme potansiyelini incelemek için kullanılmıştır. GP ve SVR modelleri için girdi özellikleri olarak kullanılan değişkenler güneş radyasyonu, ortalama sıcaklık, rüzgar hızı, bağıl nem ve yılın ayıdır. Karşılık gelen ET<sub>0</sub> değerleri, aynı istasyonlardan toplanan iklim ölçümleri kullanılarak Gıda ve Tarım Örgütü tarafından önerilen FAO 56 PM denklemi kullanılarak hesaplanmıştır. Sonuçlar, GPR ve SVR modelleri kullanılarak yüksek doğruluğa sahip regresyon modellerinin mümkün olduğunu göstermektedir. ET<sub>0</sub> tahmini için en etkili girdi değişkeninin güneş radyasyonu olduğu bulunmuştur. Bağıl nem, model doğrulukları üzerinde en düşük etkiye sahiptir.

**Anahtar Kelimeler:** Referans Evapotranspirasyon, Gaussian Proses Regresyon, Destek Vektör Regresyon.

\* [hcitakoglu@erciyes.edu.tr](mailto:hcitakoglu@erciyes.edu.tr)

## 1. Introduction

Evapotranspiration is the loss of moisture from earth by evaporation of free water and transpiration from plants. Measuring evapotranspiration rates is crucial for planning efficient irrigation schemes, especially in arid and semi-arid regions where water resources are limited [1]. The Food and Agriculture Organization (FAO) recommends using the FAO56 Penman-Monteith equation, Eq. (1), for reference evapotranspiration calculation [2]. Reference evapotranspiration (ET<sub>0</sub>) is the evapotranspiration that occurs in a surface of specified crop type and properties (reference surface). Then, using pre-determined relationships between evapotranspiration and crop properties, the real evapotranspiration relative to local crop types can be found. Using this method, surface and crop related variables can be eliminated from the evapotranspiration equation, which renders it more generalized.

$$ET_0 = \frac{\Delta(R_n - G) + \frac{\rho_a c_p (e_s - e_a)}{r_a}}{\Delta + \gamma \left( a + \frac{r_s}{r_a} \right)} \quad (1)$$

R<sub>n</sub>: Net Radiation, G: Soil heat flux, (e<sub>s</sub>-e<sub>a</sub>): Vapour pressure deficit of the air, ρ<sub>a</sub>: Mean air density, c<sub>p</sub>: Specific heat of the air, Δ: Slope of saturation vapour pressure temperature relationship, γ: Psychrometric constant, r<sub>s</sub>, r<sub>a</sub>: Surface and aerodynamic resistances

The FAO56-PM ET<sub>0</sub> equation still requires numerous measured variables, some of which can only be acquired from major weather stations. The absence of these measurements in rural areas, especially in developing countries, limits the usability of the FAO56-PM equation. This complexity of evapotranspiration measurement has induced research in estimating its values from easily obtainable climatic measurements, such as temperature and wind speed. One of the earliest and most popular equations that estimate ET<sub>0</sub> using few variables, ambient air temperature and solar radiation, is the Hargreaves equation (HG) [3]. Various empirical equations were proposed for estimating evapotranspiration from limited data [4].

With the immergence of machine learning (ML), scientists studied the potential of ML in ET<sub>0</sub> estimation. Machine learning is a set of algorithms that can learn relationships between different variables and patterns in a dataset, eventually finding the equations that relate variables to each other without being explicitly being programmed for this purpose. ML models are based on statistical and probabilistic rules and can deal with large datasets and cases of nonlinearity. These properties of ML rendered its potential in ET<sub>0</sub> estimation using few variables very high. V. Nourani et al. (2014) thoroughly reviewed the use of artificial intelligence tools for reference evapotranspiration estimation, focusing on studies that apply a Wavelet transform on input variables [5]. Chia et al. (2020) provide a more recent review of ET<sub>0</sub> estimation using AI methods [6].

Citakoglu et al. (2014) estimated monthly mean reference transpiration in Turkey from several combinations of input variables using adaptive network based fuzzy inference system (ANFIS) and artificial neural network (ANN) models [7]. H. Citakoglu et al. (2014) concluded that solar radiation, air temperature, wind speed, and relative humidity are the most effective input features for estimating ET<sub>0</sub>.

Wen et al. (2015) set a measuring station in a specific site in Ejina basin in China to measure meteorological data required for daily ET<sub>0</sub> estimation using the FAO-56 PM equation [8]. Then, support vector regression (SVR) models were used to estimate daily ET<sub>0</sub> values. Results show that models that took in maximum and minimum daily temperatures and solar radiation as inputs gave the most accurate results.

Carter and Liang (2019) compared ten machine learning models to predict ET values using data collected from satellite sensors (GLASS, MODIS) and data from measuring towers obtained from Fluxnet and Ameriflux [9]. This study emphasized the potential of satellite data inclusion in ML models for ET<sub>0</sub> estimation.

Yu et al. (2020) investigated the uncertainty of artificial intelligence methods in estimating daily reference evapotranspiration [10]. Whereas other papers were concerned about the potential of different AI models for ET<sub>0</sub> estimation and their accuracies, this paper attacked the problem from a more scientific way. It was shown that wind speed, solar radiation, and max/min temperatures were the most contributing variables to the variation of ET<sub>0</sub> estimates. These results were scientifically verified through the explanation of energy transformation and evapotranspiration demands.

In this study, climatic data collected from weather stations around Turkey will be used to estimate ET<sub>0</sub> using Gaussian process regression (GPR). The use of this ML algorithm for ET<sub>0</sub> estimation is limited in academia. Moreover, to the knowledge of the author, this study is the first to apply GPR on climatic data collected from sparsely spread locations in Turkey.

## 2. Material and Method

### 2.1. Support Vector Machines (SVM)

Support vector machine (SVM) algorithm was first developed as a classifier. The SVM algorithm finds the best hyperplane that separates different variables. The best hyperplane is the one that has the greatest margin that separates it from all variables equally (Fig 1.). In most cases, variables cannot be separated by a hyperplane in the original dimensional space. Kernels are used to transform the data to a higher dimensional space where variables can be separated using a hyperplane. For this study, the radial basis function kernel was used as an SVM kernel. Using the same algorithm, the classification hyperplane can be used as a regression hyperplane by maximizing the number of variables that fall inside the margin.

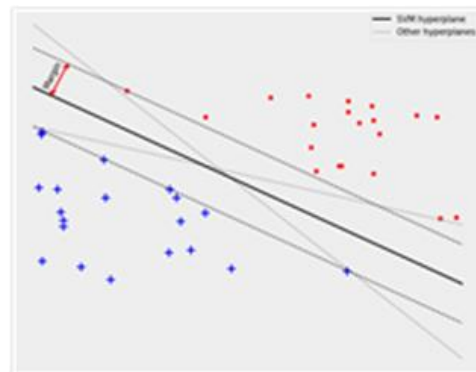


Fig. 1 Support vector machine (SVM) classification

In order to create a regularized model and avoid overfitting, an allowance for the regression hyperplane to miss some values is given by introducing a threshold parameter  $\epsilon$ . By defining a maximum threshold value, all training predictions will be constrained to not be farther from true values than the threshold allows. The thresholds of all variables are multiplied by a parameter C that penalizes high threshold parameters in the loss function. Therefore, large C values result in low training error and possible overfit, and small C values result in higher tolerance for errors and less overfit. For this study, Python's scikit-learn library, which is an open-source machine learning library, was used. The RBF kernel equation used in this library (Eq. 2) contains one optimizable variable,  $\gamma$ . This variable determines the proximity correlated variables to each other. For instance, a small  $\gamma$  corresponds to variables away from each other being correlated, thus a smoother function.

$$K_{RBF}(x_i, x_j) = \exp(-\gamma \cdot (x_i - x_j)^2) \quad (2)$$

The three previously mentioned SVR parameters are chosen using cross validation. In cross validation, the available data is divided to a number of folds, in this study five folds, whereby in every iteration one fold is used as a test set and the other folds are used to train the model. Using a set of preliminary values for each variable, a grid search is applied by creating a model for each variable combination. From the mean score of the 5 tests, the best parameter combination is determined.

## 2.2. Gaussian Processes Regression

Gaussian process is a joint distribution of variables, where every subset of these variables is normally distributed. Gaussian processes are used to solve regression problems. To define a Gaussian process, a prior multivariate Gaussian distribution is required [11]. This prior is defined by a mean vector and a covariance matrix. The prior mean vector value has minor effect on the regression model and is usually assumed zero. The covariance matrix is created using kernel functions that take in every two data point combination and return the corresponding value for the covariance matrix. Kernels ensure that the resulting covariance matrix is positive semi-definite. The choice of the kernel and its parameters is the determining factor of the accuracy and quality of the regression model. Using the defined kernel function, a modified covariance matrix is created using training and test data, which represents the posterior [12]. Predictions made using Gaussian processes are accompanied with their uncertainty values. Predictions near training data have low uncertainty, and predictions further away have higher uncertainty values. It is worth mentioning that the regression curve tends to return to the mean value of the prior distribution when data is unavailable (Fig. 2). Therefore, it is not recommended to extrapolate from the resulting regression function.

Radial basis function (RBF) kernel, Eq. (3), was seen as most suitable for ET0 prediction. The RBF kernel assumes that points next to each other have higher correlations than points away from each other, which is the case for reference evapotranspiration. There are two parameters that should be optimized in an RBF kernel, variance and length scale. Variance determines the prior variance in case no data is available. Length scale determines the spacing between correlated variables, i.e., a low value of length scale corresponds to a noisier function and a higher value corresponds to a smoother function.

$$K_{RBF}(x_i, x_j) = \sigma^2 \exp\left(-\frac{(x_i - x_j)^2}{2l^2}\right) \quad (3)$$

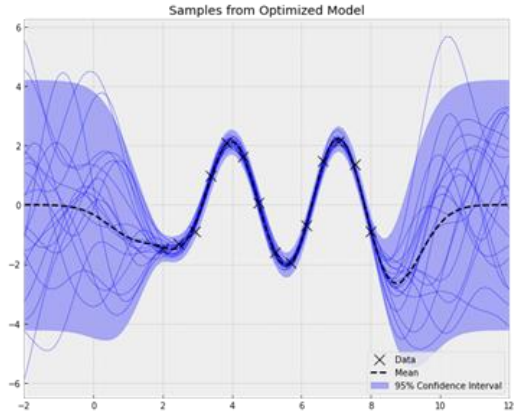


Fig. 2 Sine function regression using Gaussian Processes

The posterior covariance matrix will have zero variance for training data points, which is unrealistic and causes overfitting. To solve this issue, Gaussian noise, Eq. (4), is added to the derived covariance matrix to allow for variance round training data points. The variance of the added Gaussian noise should also be optimized.

$$\epsilon = N(0, \sigma^2) \quad (4)$$

GPy Python package was used in this study. GPy is an open-source Gaussian process framework written in Python by The Sheffield Machine Learning group. Parameter optimization is done using the Limited Broyden – Fletcher – Goldfarb – Shanno (LBFGS) algorithm.

## 2.3. Materials

This study is conducted on a dataset consisting of 3300 measurements of five climatic variables and the corresponding reference evapotranspiration obtained from the General Directorate of Turkish State Meteorological Service (TSMS). The measured variables are solar radiation (R\_s), mean temperature (T\_avg), relative humidity (RH), wind speed (WS), and the number of the month (Mo) in which the measurements were taken. These climatic variables were collected from 275 weather stations around Turkey during a period ranging between 20 to 45 years on some stations. The reference evapotranspiration corresponding to each set of climatic variables was calculated using the FAO-56 PM equation, whereby the required variables for the equation were collected from the same weather stations.

The dataset can be accepted as a general case in Turkey, as the weather stations are evenly distributed around the country. The dataset was split into a training set and a testing set by a ratio of 1:4. The distribution of variables of both sets are shown in Fig. 3. It can be seen that both, training and test datasets, have similar distributions.

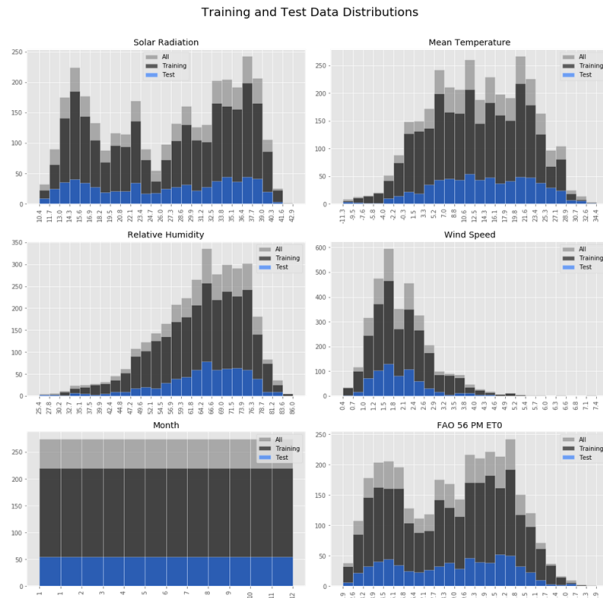


Fig 3. Distributions of variables

### 2.4. Model Evaluation

Model evaluation is done using three parameters: mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination ( $R^2$ ), given in Eq. (5-7). MAE and RMSE are two measures of error, therefore perfect models would have MAE and RMSE values equal to zero. The coefficient of determination measures the ratio of variation represented by the regression line to the total variation of data for a linear regression. A regression line that is the mean value of data would have  $R^2=0$ , whereas a perfect model would have  $R^2=1$ .

$$MAE = \frac{1}{N} \sum_{i=1}^n |Y_{i,measured} - Y_{i,predicted}| \quad (5)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^n (Y_{i,measured} - Y_{i,predicted})^2} \quad (6)$$

$$R^2 = \frac{\sum_{i=1}^n (Y_{i,predicted} - Y_{mean})^2}{\sum_{i=1}^n (Y_{i,measured} - Y_{mean})^2} \quad (7)$$

## 3. Results and Discussion

Several combinations of input variables were used to create Gaussian Process and Support Vector Regression models. Resulting scores for both training and test datasets are given in Table 1–2 sorted from the worst to the best model, with the relative estimation-measurement plots given in Fig 4–6 for GPR and SVR models, respectively. All models, GPR and SVR, were created using the radial basis function kernel; however, the kernel

parameters were optimized for each input case. It can be seen that solar radiation on its own can produce good results, which is scientifically expected as solar radiation is the factor that determines energy balance. Function samples from the Gaussian process model with solar radiation as an input are represented in Fig 4–6 with a 95% confidence range.

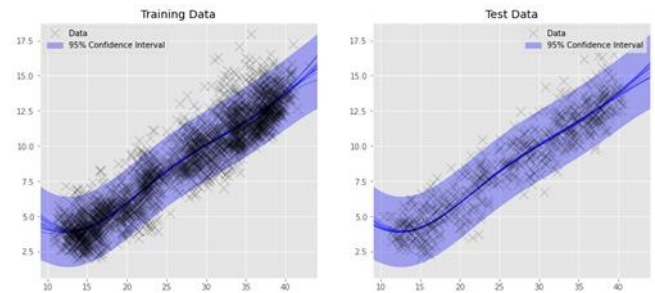


Fig 4. Function samples from model 3.

Table 1. Models and Variables

Model	Variables	Algorithm
Model 1	$T_{avg}, WS$	GPR
		SVR
Model 2	$T_{avg}, RH, WS$	GPR
		SVR
Model 3	$R_s$	GPR
		SVR
Model 4	$T_{avg}, Mo$	GPR
		SVR
Model 5	$T_{avg}, WS, Mo$	GPR
		SVR
Model 6	$R_s, T_{avg}, Mo$	GPR
		SVR
Model 7	$R_s, T_{avg}, WS, Mo$	GPR
		SVR

Relative humidity was seen to have minor effect on the model results and was therefore eliminated from other models. Adding the month of the year to the model with mean temperature and wind speed caused a significant improvement in the model accuracy. This can be explained by the correlation between month of the year and sunshine hours and solar radiation. A model including solar radiation, wind speed, mean temperature, and month variables showed very high accuracy and low MAE and RMSE errors. Results show that GPR and SVR models have similar accuracies and errors, although GPR models were marginally better on test data, hence the better regularization of GPR models.

Table 2. Model Errors for Training and Test Scores

Algorithm	Training Scores			Test Scores		
	MAE	RMSE	R <sup>2</sup>	MAE	RMSE	R <sup>2</sup>
GPR	1.441	1.771	0.749	1.561	1.91	0.691
SVR	1.422	1.787	0.744	1.577	1.974	0.670
GPR	1.275	1.594	0.796	1.428	1.779	0.732
SVR	1.260	1.629	0.787	1.418	1.814	0.721
GPR	0.976	1.259	0.873	0.907	1.167	0.885
SVR	0.974	1.265	0.872	0.915	1.179	0.882
GPR	0.716	0.977	0.923	0.694	0.946	0.924
SVR	0.698	0.960	0.926	0.678	0.930	0.927
GPR	0.578	0.798	0.949	0.601	0.855	0.938
SVR	0.536	0.767	0.953	0.602	0.866	0.936
GPR	0.564	0.753	0.955	0.508	0.663	0.963
SVR	0.544	0.741	0.956	0.489	0.641	0.965
GPR	0.240	0.321	0.992	0.255	0.350	0.990
SVR	0.242	0.323	0.992	0.271	0.368	0.988

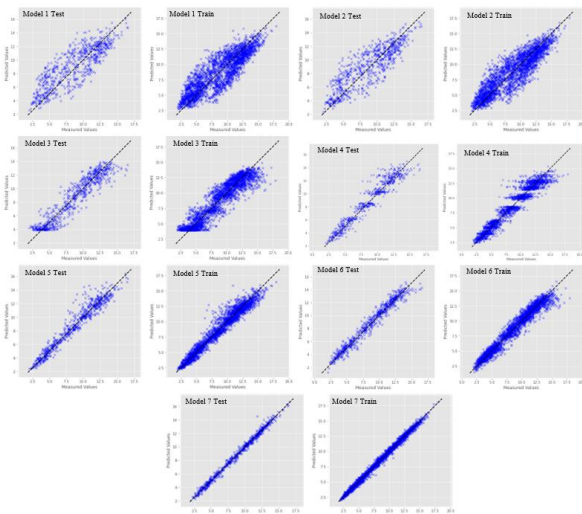


Fig. 5 GPR Prediction-Measurement plots

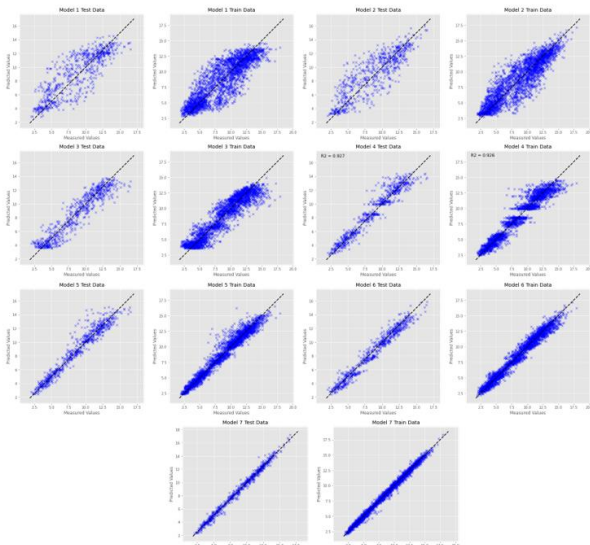


Fig. 6 SVR Prediction-Measurement plots

#### 4. Conclusions and Recommendations

Artificial intelligence has taken a large portion of research in climate science. The nonlinearity of meteorological variables and their dependency on many other properties and variables render machine learning models beneficial and efficient in this field. Calculating reference evapotranspiration using minimum climatic variables has always been a requirement, especially in rural areas where obtainable measurements are limited. Using machine learning algorithms, accurate ET<sub>0</sub> estimations are obtainable.

In this paper, the potential of Gaussian Processes in predicting ET<sub>0</sub> values from easily measurable climatic variables was investigated. Prediction accuracies exceeding 99% were obtainable from only four input features, in comparison with the numerous input variables for the FAO 56 PM equation. Besides being accurate, Gaussian Process models provide an uncertainty value with every prediction, making it more suitable for real-world usage.

The dataset used to conduct this study consists of climatic variables measured from all around Turkey. Moreover, both test and train datasets contained data from all Turkey’s regions. This, along with the obtained accuracy results, proves that a single model is suitable for use in different areas of relatively different climates. Therefore, predictions to be done in areas in Turkey other than the studied areas will be as accurate as the ones in this study.

#### References

- [1] Droogers, P., Allen, R. G., Estimating reference evapotranspiration under inaccurate data conditions. *Irrigation and Drainage Systems*, 16: 33–45, 2002.
- [2] Allen, R. G., Pereira, L. S., Raes, D., Smith, M., Crop evapotranspiration guidelines for computing crop water requirements. *FAO Irrigation & drainage Paper 56*. FAO, Food and Agriculture Organization of the United Nations, Roma, 50 pp, 1998.

- [3] Hargreaves, G. H., Samani, Z. A., Reference Crop Evapotranspiration From Temperature. *American Society of Agricultural Engineers*, 96–99, 1985.
- [4] Xu, C. Y., Singh, V. P., Cross comparison of empirical equations for calculating potential evapotranspiration with data from Switzerland. *Water Resources Management*, 16 (3): 197–219, 2002.
- [5] Nourani, V., Hosseini Baghanam, A., Adamowski, J., Kisi, O., Applications of hybrid wavelet-Artificial Intelligence models in hydrology: A review. *Journal of Hydrology*, 514: 358–377, 2014.
- [6] Chia, M. Y., Huang, Y. F., Koo, C. H., & Fung, K. F., Recent advances in evapotranspiration estimation using artificial intelligence approaches with a focus on hybridization techniques—a review. *Agronomy*, 10(1), 101, (2020).
- [7] Citakoglu, H., Cobaner, M., Haktanir, T., Kisi, O., Estimation of Monthly Mean Reference Evapotranspiration in Turkey. *Water Resources Management*, 28 (1): 99–113, 2014.
- [8] Wen, X., Si, J., He, Z., Wu, J., Shao, H., & Yu, H., Support-Vector-Machine-Based Models for Modeling Daily Reference Evapotranspiration With Limited Climatic Data in Extreme Arid Regions. *Water Resources Management*, 29 (9): 3195–3209, 2015.
- [9] Carter, C., Liang, S., Evaluation of ten machine learning methods for estimating terrestrial evapotranspiration from remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 78 (October 2018): 86–92, 2019.
- [10] Yu, H., Wen, X., Li, B., Yang, Z., Wu, M., & Ma, Y., Uncertainty analysis of artificial intelligence modeling daily reference evapotranspiration in the northwest end of China. *Computers and Electronics in Agriculture*, 176, 105653, 2020.
- [11] Özkaya, U., Öztürk, Ş., Gaussian Regression Models for Day-Level Forecasting of COVID-19 in European Countries. In *Understanding. COVID-19: The Role of Computational Intelligence*, 2022, pp. 339-356, Springer, Cham.
- [12] Özkaya, U., Yiğit, E., Seyfi, L., Öztürk, Ş., Singh, D., Comparative Regression Analysis for Estimating Resonant Frequency of C-Like Patch Antennas. *Mathematical Problems in Engineering*, 2021.