Journal of Agricultural Sciences
(Tarim Bilimleri Dergisi)

J Agr Sci-Tarim Bili
e-ISSN: 2148-9297
jas.ankara.edu.tr

JAS

Soil Temperature Prediction via Self-Training: Izmir Case

Goksu TUYSUZOGLU[a*], Derya BIRANT[a], Volkan KIRANOGLU[a]

[a]*Dokuz Eylul University, Department of Computer Engineering, Buca, Izmir, TURKEY*

ARTICLE INFO
Research Article
Corresponding Author: Goksu TUYSUZOGLU, E-mail: goksu.tuysuzoglu@deu.edu.tr
Received: 30 July 2020 / Revised: 06 March 2021 / Accepted: 09 March 2021 / Online: 20 January 2022

ABSTRACT

This paper proposes a new model, called *Soil Temperature prediction via Self-Training* (STST), which successfully estimates the soil temperature at various soil depths by using machine learning methods. The previous studies on soil temperature prediction only use labeled data which is composed of a variable set *X* and the corresponding target value *Y*. Unlike the previous studies, our proposed STST method aims to raise the sample size with unlabeled data when the amount of pre-labeled data is scarce to form a model for prediction. In this study, the hourly soil-related data collected by IoT devices (Arduino Mega, Arduino Shield) and some sensors (DS18B20 soil temperature sensor and soil moisture sensor) and meteorological data collected for nearly nine months were taken into consideration for soil temperature estimation for future samples. According to the experimental results, the proposed STST model accurately predicted the values of soil temperature for test cases at the depths of 10, 20 30, 40, and 50 cm. The data was collected for a single soil type under different environmental conditions so that it contains different air temperature, humidity, dew point, pressure, wind speed, wind direction, and ultraviolet index values. Especially, the XGBoost method combined with self-training (ST-XGBoost) obtained the best results at all soil depths (R^2 0.905-0.986, MSE 0.385-2.888, and MAPE 3.109%-8.740%). With this study, by detecting how the soil temperature will change in the future, necessary precautions for plant development can be taken earlier and agricultural returns can be obtained beforehand.

Keywords: Soil temperature prediction, Self-training, Regression, Machine learning, Agriculture, Artificial Intelligence, STST

1. Introduction

As technology advances, new solutions are emerging to help simplify agriculture and improving productivity and consumer satisfaction with the increased interest in precision agriculture. The applications of modern information technologies, machine learning, and artificial intelligence offer favorable computational as well as analytical solutions by employing data from multiple sources for decision making in the management of crop production (Friedl 2018). Precision agriculture, soil temperature, and climatic parameters have complex inter-relationships and this complex problem can be efficiently solved using machine learning techniques. The estimation of greenhouse gas emission from agricultural soils (Hamrani et al. 2020), evaluation of farm efficiency (Nandy & Singh 2020), weed classification (Dadashzadeh et al. 2020), plant disease detection (Giraddi et al. 2020), and the determination of the concentration of chemical matters in a grain (Niedbała et al. 2020) are some of the agricultural issues where machine learning is frequently implemented, nowadays.

Soil is of great importance in the terrestrial ecosystem by affecting the physical, biological, and chemical processes. Many studies in agriculture are canalized into this field, especially in terms of the effects of soil moisture and soil temperature on crop yields and plant growth and due to their impact on organic and chemical substances found in soil (Hillel 2005; Yang et al. 2019). Soil temperature plays an important role in agriculture since it is closely related to the myriad events occurring in the soil. It is a very important ecological factor that affects plant life at all stages from seed germination to seedling growth and development. It has a great effect on germination speed and duration. Although other conditions are suitable, if the temperature is too low or too high, there will be little or no germination. If the soil temperature is suitable, biological and chemical activities in the soil continue. These activities stop when the temperature drops and the soil freezes. On the other side, soil resistance to the physical events such as erosion and subsidence can drop dramatically at high soil temperatures. For this reason, factors affecting soil warming and control of soil temperature are extremely important. If we predict further changes in soil temperature, we can develop new strategies in different application areas such as setting up automatic cooling / heating system or irrigation system and determining the planting dates of temperature-sensitive crops etc.

Because of the importance of the subject, various different regression and statistical analysis techniques were proposed considering machine learning such as support vector machines (SVM) (Xing et al. 2018), to estimate soil temperature, and decision tree regression (Pekel 2020) and the least-squares support vector machine (Ren et al. 2019) to predict soil moisture and collaborative data mining using the algorithms of local polynomial regression, neural networks, k-nearest neighbor, support vector machine (Anton et al. 2019) to estimate both soil temperature and soil moisture.

The soils of Turkey are divided into three major groups: zonal, intrazonal, and azonal. Zonal soils that are formed by the dissolution of rocks under the influence of climatic conditions and vegetation consist of brown forest soils, podsolic forest soils, Terra-Rossa's, chernozems, brown steppe soils, chestnut steppe soils. Vertisols, rendzinas, and volcanic sandy soils are in the category of intrazonal soils that reflect the characteristics of the bedrock. Alluvial/colluvial soils and loesses are examples of azonal soils that are not dependent on natural factors such as climate and vegetation but are formed due to accumulation by the effect of external factors such as streams and wind (Gönençgil et al. 2016; Akengin & Dölek 2019).

The soil temperature of Turkey at the depth of 0.5 cm increases from the Black Sea coast to the Aegean and Mediterranean coasts and decreases continuously from the coastal belt to the mountains and Eastern Anatolia. The lowest underground temperature varies between -3 °C and -6 °C in the higher parts of Eastern Anatolia. The temperature varying between 1-3 °C in Central Anatolia rises to 4-6 °C in the Black Sea, Marmara, and North Aegean coastal belt, and to 9-10 °C in the Southern Aegean and Mediterranean coastal belt. It is between 3-6 °C on the low plains of Southeastern Anatolia. Soil temperature varies between 2-14 °C at a depth of 100 cm in winter. The temperature, which is 2-3 °C in the higher parts of Eastern Anatolia, increases to 8-9 °C in the Black Sea coastal zone and to 11-13 °C in the Aegean and Mediterranean coasts. The temperature, which is between 4-6 °C in Central Anatolia, rises above 10 °C in Southeastern Anatolia. The soil temperature at 0.5 cm depth in July, which characterizes the summer period, varies between 20-25 °C in Eastern Anatolia and 25-30 °C in Central Anatolia. The temperature, which is around 25 °C in the Black Sea coastal zone, reaches 30-35 °C in the Aegean coasts and 35-38 °C in Southeastern Anatolia. The temperature changes between 13-27 °C in 100 cm deep soil in July. The lowest decrease in the soil temperature towards the bottom in July occurs in the Eastern Anatolia and the Black Sea coastal zone, the highest decrease in the Mediterranean and Southeastern Anatolia. As a matter of fact, the temperature decrease at the depth of 0.5 cm to 100 cm in July reaches 5 to 7 °C on the Black Sea coast, 7 to 10 °C in Central Anatolia and 10 °C in Southeastern Anatolia (Gönençgil et al. 2016).

Table 1 displays the recent machine learning studies taking soil temperature prediction in Turkey as the main subject. In addition to the past values of soil temperature, meteorological factors such as air temperature, relative humidity, and solar radiation, etc. were generally used as input for the applied models for estimation. The experimental studies were made from 5 cm to 100 cm depth in general. The performed method was mostly artificial neural networks (ANN) among them.

Table 1- Recent studies for soil temperature prediction in Turkey*

Ref / Study Area	*Aim*	*Data*	*Methods*	*Performance Measure*
Alizamir et al. 2020b / The city of Mersin	Monthly ST prediction at depths of 5, 10, 50, and 100 cm	25-year (1986–2010) monthly values of AT, SR, RH, WS, and ST	ELM, ANN, CART, and GMDH	RMSE, NS, R^2
Kisi et al. 2017 / The cities of Adana and Mersin	Monthly ST prediction at the depths of 10, 50, and 100 cm	25-year (1986–2010) monthly values of AT, SR, RH, WS, and ST	ANN, ANFIS, and GP	RMSE, MARE, NS, R^2
Yener et al. 2017 / All of the 81 provinces in Turkey	Monthly ST prediction for shallow geothermal applications at depths of 5, 10, 20, 50, and 100 cm	Monthly values of AT and ST between 1960 and 2015	TIR, ANN, PDV, and soil heat calculator program	Maximum average percentage error
Citakoglu 2017 / 261 stations all over Turkey	Monthly ST prediction at depths of 5, 10, 20, 50, and 100 cm	AT, P, and ST values between 20 and 45 years of data (over the period from 1974 through 2010)	ANN, ANFIS, and MLR	MAE, RMSE, R^2
Kisi et al. 2015 / The city of Mersin	Monthly ST prediction at depths of 5, 10, 50, and 100 cm	25-year (1986–2010) monthly values of AT, SR, RH, WS, and ST	MLP, RBNN, and GRNN	RMSE, MAE, R^2
Bilgili M et al. 2013 / The 7 meteorological stations, namely, Afyonkarahisar, Aydın, Denizli, Kütahya, Uşak Manisa, and Muğla as the neighboring stations, and İzmir as the target station	Monthly ST prediction of a target station only using the data of neighboring stations at depths of 5, 10, 20, 50, and 100 cm	Monthly ST data between 2000 and 2006	SR analysis and ANN	MAPE, R
Bilgili M 2012 / Kütahya, Manisa, Usak, Afyonkarahisar, Izmir, Aydın, Denizli, and Mugla	Monthly ST prediction at depths of 5, 10, 20, 50 and 100 cm	Monthly AT and ST data between 2000 and 2006	ANN	MAE, R
Bilgili M 2011 / The city of Adana	Monthly ST prediction at depths of 5, 10, 20, 50, and 100 cm	Monthly values of AT, AP, RH, WS, R, and ST between 2000 and 2007	ANN	MAPE, R
Ozturk et al. 2011 / 66 Turkish state meteorological service locations	Monthly ST prediction at depths of 5, 10, 20, 50 and 100 cm	Altitude, latitude, longitude, monthly values of AT, SD, SR, AT between 2006-2008	ANN	RMSE, R
Bilgili M 2010 / The city of Adana	Monthly ST prediction at depths of 5, 10, 20, 50 and 100 cm	Monthly values of ST, AT, AP, WS, RH, R, SR, SD between 2000 and 2007	LR, NLR, and ANN	MAPE, R

*ELM, Extreme learning machine; ANN, Artificial neural networks; CART, Classification and regression trees; GMDH, Group method of data handling; ANFIS, Adaptive neuro-fuzzy inference system; GP, Genetic programming; TIR, Thermal infrared technique; PDV, Philip and de Vries model; SR, Stepwise regression; LR, Linear regression; NLR, Nonlinear regression; RMSE, Root mean square error; NS, Nash-Sutcliffe coefficient; R^2, Coefficient of determination; MARE, Mean absolute relative errors; MAPE, Mean absolute percentage error; R, Correlation coefficient; MAE, Mean absolute error; AT, Air temperature; SR, Solar radiation; RH, Relative humidity; WS, Wind speed; P, Precipitation; AP, Atmospheric pressure; R, Rainfall; SD, Sunshine duration.

Differently from the mentioned studies, this is the first study that is performed by one of the semi-supervised learning techniques known as “self-training” in the subject of soil temperature prediction. *Semi-supervised learning* is a machine learning approach in which there are a small number of samples whose output is known and a large number of samples with unknown labels to develop a classification/regression model during training (Belkin et al. 2006). One of the semi-supervised learning methods is *self-training* (Zhu & Goldberg 2009).

A novel model, called *Soil Temperature prediction via Self-Training* (STST) is proposed. The STST facilitates the capability of temperature estimation for new samples in case there are few numbers of labeled samples to discover the hidden patterns.

This is the first time that the analysis is being reported in detail to determine which regression method provides the most accurate predictions under the self-training framework and the variation in the performances. For this purpose, it compares the self-training versions of machine learning algorithms, including Random Forest (RF), Support Vector Regression (SVR), K-nearest Neighbors Regression (KNNReg), Extremely Randomized Trees (ETReg), Decision Tree Regression (DTReg), and Extreme Gradient Boosting (XGBoost).

This study is also original in that it investigates the performances of the semi-supervised machine learning algorithms on soil temperature prediction with different ratios of labeled data varying from 5% to 85% with an increment of 5. It should be highlighted that this paper is the first to propose a multi-depth self-training learning framework that considers estimating the soil temperatures at five different soil depths (10, 20, 30, 40, and 50 cm). It presents a new application of semi-supervised machine learning to provide a smart way of soil temperature prediction. The purpose is to estimate the soil temperature in Izmir, Turkey by investigating the dynamics of the past soil temperature & soil moisture data and the meteorological data.

2. Material and Methods

In this section, the materials used to collect data, the proposed "*Soil Temperature prediction via Self-Training*" (STST) model, and the machine learning methods used in the experiments are presented.

2.1. Data collection

The location of the experimental area is Izmir, Turkey during the dates of 01.09.2019 and 22.05.2020. Izmir is located in the Aegean region of Turkey between the latitude of 38° 24′ 46″ and the longitude of 27° 8′ 18″. It is located in the Mediterranean climate zone and it has hot and dry summers and warm and rainy winters. In the middle latitude zone, it is open to marine effects and has a climate affected by the tectonic characteristics of the coastal Aegean strip and the bay having inland sea character. Depending on the sunshine duration and sufficient amount of rainfall, the soil structure has an agriculturally suitable climate (Turkish State Meteorological Service, 2021). While the yearly mean value of rainfall is 700.2 mm, the yearly mean values of air temperature and soil temperature are 17.6 °C and 19.8 °C, respectively (Republic of Turkey Ministry of Agriculture and Forestry 2021). Its soil moisture and temperature regimes are xeric and thermal (Bolca et al. 2011; Kapur et al. 2018). Figure 1 shows the study area in the location map.

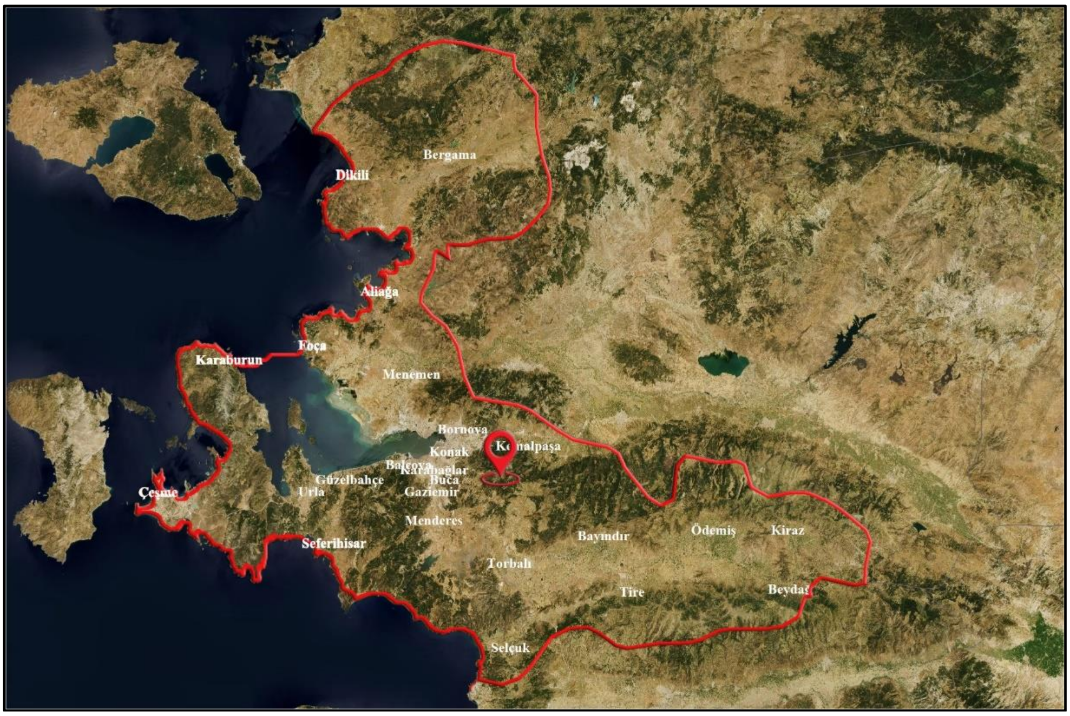

Figure 1- The location map of the study area

The distribution of the soil types are as follows: red-brown Mediterranean soils and limeless brown forest soils with the ratio of 16%, alluvial and colluvial soils with the ratio of 12%, brown forest soils with the ratio of 4%, red Mediterranean soils, and rendzinas with the ratio of 3%, chestnut soils with the ratio of 0.4% and regosols% with the ratio of 0.1. 22.5% of the soils in İzmir are deep, or very deep, 4% medium-deep, 38.5% shallow, and 35% very shallow (Dizdar 2003).

The properties of the collected soils for all depths are as follows: pH (KCL):7.2 pH, organic matter: 7.3%, total nitrogen: 975.0 kg/da, total phosphorus: 300.00 kg/da, total potassium 43.99 kg/da, clay: 9.1%, cation exchange capacity: 347.1 mmol+/kg. The soil texture is sandy loam.

A part of the dataset used in the experiments was collected using IoT devices (Arduino Mega, Arduino Shield) and various sensors (DS18B20 soil temperature (ST) sensor and soil moisture (SM) sensor for measuring hourly data at the soil depths of 10, 20, 30, 40, and 50 cm, and light-dependent resistor (LDR) sensor for hourly light intensity (LI)). The meteorological part of the dataset was obtained from the web page of weather.com that provides air temperature (AT), humidity (H), dew point (DP), air pressure (AP), wind speed (WS), wind direction (WD), and ultraviolet index (UV). Hourly values of the aforementioned features were taken into consideration. After data collection, the records with missing values were removed. Finally, 4500 instances were left for data analysis.

A sample fragment of the dataset at the depth of 50 cm between 01:00 and 04:00 p.m. on September 19, 2019, is given as hourly in Table 2. SM and LI sensors have raw analogue reading values of 0 to 1023 as shown in Table 2. If the value of SM is close to 1023, it means soil moisture is high, otherwise low. In the same manner, if the value of LI is high, the light intensity is a lot. For the prediction of ST of each depth, SM values from all depths are taken into consideration.

Table 2- A part of the dataset used in the experiments*

AT	*H*	*DP*	*AP*	*UV*	*WS*	*WD*	*LI*	*SM_{50}*	*SM_{40}*	*SM_{30}*	*SM_{20}*	*SM_{10}*	*ST_{50}*
(°C)	*(%)*	*(°C)*	*(mb)*		*(km/s)*								*(°C)*
28	40	13	1015.2	7/10	3	Northwest	900	322	335	268	270	282	22.56
29	32	11	1013.9	7/10	5	Northwest	874	316	337	267	267	279	23.00
29	34	11	1014.2	5/10	18	West	866	320	339	267	266	276	23.38
28	28	13	1014.2	4/10	23	West	843	317	339	275	264	272	23.82

*AT, Air temperature; H, Humidity; DP, Dew point; AP, Air pressure; WS, Wind speed; WD, Wind direction; UV, Ultraviolet index; LI, Light intensity; SM, Soil moisture; ST, Soil temperature.

2.2. Proposed method: Soil temperature prediction via self-training (STST)

The rationale behind self-training is to increase the training set size with unlabeled data when there is a very small number of pre-labeled data compared to unlabeled ones so that a more optimized classifier model can be constructed using the updated labeled training set. Considering the problem of soil temperature prediction, we may not always be able to access all past data whose results are known to estimate new values. Because of the high cost of manual labeling, it is hard to obtain sufficient, reliable, and up-to-date labeled data for effective soil temperature prediction. In such cases, by applying the proposed *Soil temperature prediction via self-training* (STST) model, we first estimate the values for the unlabeled historical samples and then use all past records to estimate the new values for the future records.

The proposed approach (STST) has a number of advantages that can be summarized as follows:

- The traditional soil temperature prediction is limited to using only labeled data to build a regression model. Differently from the previous studies, the proposed STST approach overcomes this limitation and deals with the design of prediction models in the presence of both labeled and unlabeled data. In addition to labeled data, the STST approach also exploits unlabeled data to help improve soil temperature prediction performance. Due to the STST approach, the unlabeled data samples provide additional knowledge that is relevant for prediction, and they can successfully be used to improve the generalization ability of the learning system.

- An important advantage of the STST approach is that it can be used with the combination of any supervised base learner such as SVR, KNNReg, and ETReg. The STST approach is entirely unaware of the regression method, in fact, it simply learns from the labeled and pseudo-labeled samples as if they were regular labeled samples.

- Another advantage is that the STST approach can be applied to any soil data without any prior information about the given dataset. It does not make any specific knowledge and specific assumptions for the given data.

- Soil temperature prediction at different depths is useful in agricultural management (Abyaneh et al. 2016; Huang et al. 2020). However, the measured soil temperature data at various depths are rarely available for many locations. In many real-world agricultural applications, a huge amount of unlabeled data is available. The proposed STST approach addresses this inherent bottleneck by automatically allowing the model to integrate the available unlabeled data at various soil depths. Since the proposed STST approach covers multiple soil depths, it enables enormous agricultural applications, and so it expands the application of machine learning algorithms in the field of agriculture.

Figure 2 expresses the main course of this study as a series of processes. The first part is the collection of soil data at different depths, meteorological data, and light intensity by using IoT devices and sensors or by pulling data from web pages. After data collection, if needed, missing data imputation can also be performed, otherwise, the machine learning process including self-training takes part. In the self-training phase, a machine learning model is firstly built with the initial labeled data for the purpose of classifying unlabeled instances, and then, it is re-trained by adding its own estimations to the labeled data. After that, future samples are assigned their predicted temperature values as an output. The final part is the reporting and presentation facility in order to adjust the findings in an interpretable format.

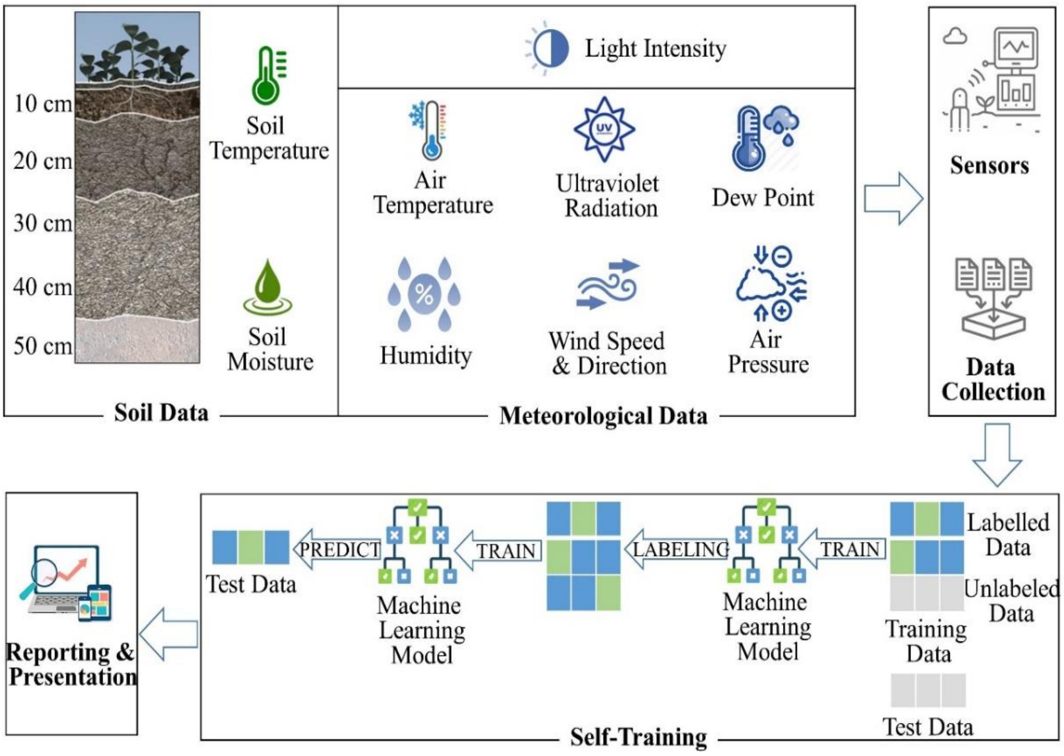

Figure 2- The general framework of the proposed "Soil temperature prediction via self-training (STST)" model

Algorithm 1 given below theoretically displays the pseudo-code of the training part of the proposed model. The aim is to obtain the soil temperature values (Y^*) of new instances in D_{Test} at all depths. In the first part, there are two sets of instances: the labeled dataset (D_L) and the unlabeled dataset (D_U). The hidden patterns affecting the soil temperature can be detected by analyzing the labeled instances. Therefore, using the instances in D_{L_d}, a classifier C_d is trained and a model that facilitates labeling other instances is obtained at the depth d. In this way, the pseudo-label (y_i) of each instance x_i in D_{U_d} is discovered. All the pseudo-labeled instances are then gathered together in D_d. Now, a new labeled training set D_{ALL} is ready by expanding the initially given D_{L_d} with the pseudo-labeled set D_d. The final step is to predict the temperature values (Y_d) of new instances at the depth d by using the new classifier model C_d trained with D_{ALL}. The resulting output is the predicted soil temperature values of all depths as the collection of each Y_d in the set Y*.

Algorithm 1. Soil temperature prediction via self-training (STST)

```
Inputs:
  $D_L$: the labeled dataset $D_L = \{(x_i, y_i)\}_{i=1}^{l}$ with $l$ intances
  $D_U$: the unlabeled dataset $D_U = \{x_j\}_{j=l+1}^{l+u}$ with $u$ intances
  $D_{Test}$: test instances
Outputs:
  $Y^*$: predicted values

foreach depth $d$ do
    $C_d$ = Train ($D_{L_d}$)
    foreach $x_i$ in $D_{U_d}$
        $y = C_d(x_i)$
        $D_d$.Add($x_i$,$y$)
    end foreach
    $D_{ALL} = D_{L_d} \cup D_d$
    $C_d$ = Train ($D_{ALL}$)
    foreach $x_i$ in $D_{Test}$
        $y = C_d(x_i)$
        $Y_d = Y_d \cup y$
    end foreach
    $Y^* = Y^* \cup Y_d$
end foreach

End Algorithm
```

2.3. Machine learning methods

The machine learning methods performed in the experiments are SVR, RF, KNNReg, ETReg, DTReg, and XGBoost. The parameters of all the applied methods were left as their default values in the sklearn library of Python.

2.3.1. Support vector regression

The main aim is to minimize the error by maximizing the margin around the separating hyperplane. The general formula can be written as an optimization problem as in Equation 1 where w is the normal vector to the bounding planes, C is the penalty associated with the instances which are either misclassified or violate the maximal margin, $\phi(X_i)$ is a function which maps data point X_i into a higher dimensional space, b shows the positions of bounding planes relative to the origin and ξ is a slack variable for soft margins defined for linearly non-separable cases.

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{subject to } Y_i(w^T\phi(X_i) + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \tag{1}$$

In non-linear problems, the kernel functions, $K(X_i, X_j) = \phi(X_i)^T\phi(X_j)$, are used to transform the data into a higher dimensional feature space to make it possible to perform the linear separation. Two kernel functions are generally used for these cases as polynomial kernel (*poly kernel*) or Gaussian radial basis function (*RBF kernel*). For two samples X_i and X_j, the poly kernel function and RBF kernel function can be written as in Equation 2 and Equation 3, respectively.

$$K(X_i, X_j) = (aX_i^T X_j + b)^d \tag{2}$$

$$K(X_i, X_j) = \exp\left(-\frac{\|X_i - X_j\|^2}{2\sigma^2}\right) \tag{3}$$

2.3.2. K-nearest neighbors regression

It is one of the instance-based lazy learners where the method memorizes the training dataset instead of learning a discriminative function for predicting future samples. It compares a given test instance with training instances that are similar to it. The parameter k refers to the number of samples to be considered in the determination of the numeric outcome for a new sample. The nearest neighbors are calculated using one of the distance metrics such as *Manhattan*, *Euclidean*, *Chebychev*, *Cosine*, etc. In the experimental studies, Euclidean distance given in Equation 4, where two points are described as $X_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ for m features is used. The outcome of the k nearest samples to a specific instance are averaged.

$$\text{Dist}(X_1, X_2) = \sqrt{\sum_{i=1}^{m} (x_{1i} - x_{2i})^2} \tag{4}$$

2.3.3. Decision tree regression

It is one of the supervised learning methods where the regression scenario is represented as a tree-based system in which each branch points out a possible outcome. The depth-first strategy in a top-down recursive and divide-and-conquer manner is applied to predict unknown target values for test cases. Each node of the tree refers to a specific attribute as the branches show their values. The leaves carry the results (the value of the class label).

The construction of a decision tree is started with a root node. The determination of the initial node is based on the mutual information which gives the highest benefit for learning. For this purpose, *information gain* given in Equation 5, which is calculated for each attribute, is used that S is the instances of the parent node, A is an attribute to perform the split, S_{Left} and S_{Right} are the samples found in the left and right child nodes, respectively, and I is the impurity measure. Information gain evaluates the gain of each feature in the context of a target variable. It is performed by taking the mutual information (i.e. the determination of the statistical dependence) between two random variables. I is the mean squared error (MSE) of the children nodes, which is given in Equation 8.

$$\text{InfoGain}(S, A) = I(S) \text{-} \left(\frac{|S_{Left}|}{|S|} I(S_{Left}) + \frac{|S_{Right}|}{|S|} I(S_{Right}) \right) \tag{5}$$

2.3.4. Random forest

It is a meta estimator based on decision trees applied to many bootstrapped subsamples of a dataset. First of all, a specified number of decision tree regressors are built. A subset of features is randomly selected to be used as candidates at each split so that the constructed decision trees do not rely on the same set of features and high correlation among trees can be prevented. The bootstrapped instances also prevent the individual trees from overfitting. The numeric predictions of each estimator are averaged and assigned to the test sample as the final output. Equation 6 is used to make a prediction for a new sample x, where B is the number of bootstrapping, T_i is the bootstrapped tree constructed by a set of samples of size n, which is the total number of instances in the training data, and selecting m variables from all features at random for iteration i. The best split point among m variables is determined using the mean squared error as in Equation 8.

$$\hat{f}_B(x) = \frac{1}{B} \sum_{i=1}^{B} T_i(x) \tag{6}$$

2.3.5. Extremely randomized trees

It is an ensemble of decision trees where cut-points are randomly determined while splitting nodes, on the other hand, the whole samples are used as given at the beginning instead of performing bootstrapping. The final output is assigned by averaging the results of ensemble iterations. Two important parameters are the number of randomly selected features and the minimum sample size for splitting a node.

2.3.6. Extreme gradient boosting

XGBoost is also one of the ensemble models of decision trees. It is based on gradient boosting in which errors are minimized by the gradient descent algorithm. By adding models on top of each other iteratively, the errors of the previous model are corrected by the next predictor, until the training data is accurately predicted or reproduced by the model. Instead of assigning different weights to the classifiers after every iteration, gradient boosting fits the new model to new residuals of the previous prediction and then minimizes the loss when adding the latest prediction. XGBoost uses this algorithm with an additional custom regularization term in the objective function. XGBoost uses a loss function to build trees by minimizing the value in Equation 7, where the first part represents the loss function calculating the pseudo residuals of the predicted $(\hat{y}_i)$ and the real value (y_i) of the i^{th} instance in each leaf, the second part includes T as the number of leaves, γ as the penalty parameter used for pruning, λ as a regularization term, and w as the leaf weights.

$$\mathcal{L}(\phi) = \sum_i \ell(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad \text{where } \Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{7}$$

3. Results and Discussion

From here onwards, the abbreviation of the self-training (ST) method followed by the abbreviation of the base classifier technique is used to refer to the related approach. For example, ST-SVR refers to the self-training method with the SVR base classifier.

Three performance metrics (mean squared error (MSE) given in Equation 8, coefficient of determination (R^2) given in Equation 9, and mean absolute percentage error (MAPE) given in Equation 10, where *ST* is the measured value of soil temperature, $\widetilde{ST}$ is the predicted value of soil temperature, $\overline{ST}$ is the mean of the observed data, and *n* is the number of samples) were calculated to evaluate the usability of the proposed methodology and to select the best one in terms of the given criteria. The results show the outputs of ten-fold cross-validation. By considering the ratio of initially given training data as *p*%, the experimental results were obtained for all varying values of *p* as 5% to 85% with an increment of 5.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(ST_i - \widetilde{ST_\iota}\right)^2 \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(ST_i - \widetilde{ST_\iota})^2}{\sum_{i=1}^{n}(ST_i - \overline{ST_\iota})^2} \tag{9}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{ST_i - \widetilde{ST_\iota}}{ST_i}\right| * 100 \tag{10}$$

Table 3 displays the MSE values of the methods under the self-training framework at the depth of 50 cm. Even though there are very few known cases of soil temperature, in the beginning, the constructed models achieved a remarkable performance by predicting too close to the real values for test cases. The self-training model led to increasing the amount of labeled data by pseudo-labeling the past training data by applying one of the regression methods described in Section 2.3. As a result, the constructed regression model had the advantage of discovering more patterns hidden in data due to more sets of instances with known outputs.

Table 3- The comparison of the applied methods under self-training framework in terms of MSE values

				Methods		
%	*ST-RF*	*ST-SVR*	*ST-KNNReg*	*ST-ETReg*	*ST-XGBoost*	*ST-DTReg*
5	2.293	18.594	6.374	5.128	**2.189**	4.573
10	1.466	16.584	4.020	3.444	**1.279**	3.041
15	1.256	15.459	3.134	3.070	**1.007**	2.448
20	1.197	14.537	2.779	2.607	**0.934**	2.234
25	1.035	14.065	2.411	2.457	**0.772**	2.052
30	0.891	13.553	2.331	2.189	**0.736**	1.702
35	0.816	13.195	2.068	2.140	**0.638**	1.425
40	0.747	12.920	1.991	2.047	**0.582**	1.402
45	0.728	12.554	1.921	1.868	**0.552**	1.530
50	0.670	12.259	1.728	1.687	**0.522**	1.219
55	0.611	12.004	1.732	1.626	**0.506**	1.203
60	0.567	11.730	1.672	1.593	**0.482**	1.197
65	0.539	11.557	1.617	1.452	**0.481**	1.118
70	0.490	11.310	1.559	1.255	**0.419**	0.960
75	0.508	11.113	1.546	1.226	**0.432**	1.008
80	0.467	10.889	1.489	1.160	**0.386**	1.019
85	0.443	10.721	1.457	1.268	**0.385**	1.005

Furthermore, the most noticeable thing is the reduction in error values as the percentage (*p*) value increases. It is because as the number of training instances real labels of which are known rises, the pseudo-labeled training data is less required to predict the outcome of new cases. More accurate estimations can be obtained as a result.

The best results for all the ratio values were achieved when ST-XGBoost was performed. XGBoost is one of the ensemble learning methods which boosts high performance compared to single learners so it was expected that its results were good. ST-RF is another ensemble learning method that followed ST-XGBoost in terms of low error values. The highest errors were obtained when ST-SVR was the learner. It can be inferred that ST-SVR required more labeled training data in order to find the optimal hyperplane to make more accurate predictions. It could not manage to perform well when the value of *p* is small. The performance of ST-SVR may be improved by changing the kernel type or updating the parameters instead of using default values.

Table 4 shows the MAPE values of the applied methods at the depth of 50 cm. As in Table 3, there is a tendency to decrease in the error values when the percentage of the initially given labeled training data is increased. ST-XGBoost performs the best (7.436% - 3.456%) especially for the p values between 5% and 60%. In addition to ST-XGBoost, ST-RF also achieves the predictions with the least errors (3.434% - 3.059%) for the p values of 65% to 85%. ST-DTReg is the leading one among the single learners with error rates of 9.706% - 4.001%.

Table 4- The comparison of the applied methods under self-training framework in terms of MAPE values (%)

%	*Methods*					
	ST-RF	*ST-SVR*	*ST-KNNReg*	*ST-ETReg*	*ST-XGBoost*	*ST-DTReg*
5	7.544	26.706	13.921	10.899	**7.436**	9.706
10	6.157	24.932	11.410	8.933	**5.939**	8.115
15	5.522	24.107	9.821	8.110	**5.334**	7.120
20	5.125	23.025	9.247	7.549	**4.819**	6.712
25	4.746	22.592	8.440	7.226	**4.395**	6.134
30	4.493	22.141	8.328	6.827	**4.223**	5.948
35	4.267	21.688	7.621	6.555	**4.082**	5.445
40	4.191	21.405	7.456	6.309	**3.914**	5.437
45	3.990	20.906	7.156	5.973	**3.722**	5.177
50	3.828	20.641	6.843	5.827	**3.572**	4.812
55	3.717	20.229	6.820	5.741	**3.581**	4.843
60	3.547	20.053	6.720	5.432	**3.456**	4.622
65	**3.434**	19.842	6.468	5.188	3.486	4.352
70	3.288	19.554	4.338	4.974	**3.275**	4.275
75	**3.258**	19.295	6.266	4.759	3.271	4.243
80	**3.126**	19.095	6.194	4.657	3.156	4.129
85	**3.059**	18.890	6.077	4.665	3.109	4.001

The comparisons of the methods according to R^2 values are indicated in Figure 3 at the depth of 50 cm. R^2 expresses the proportion of the variance for a dependent variable that is explained by the model's inputs. In this study, it is the relationship between soil temperature and other independent variables such as soil moisture, air temperature, ultraviolet radiation, etc. as in the work of Shamshirband et al. (2020) and Tabari et al. (2011). According to the results, the general impression is that there is a steady increase as the value of *p* increases. The same condition as in MSE is valid for R^2 that ST-SVR has the lowest coefficient of determination (0.331 to 0.614) while the best results are found in ST-XGBoost (0.921 to 0.987). Similarly, at the more sensitive depth of 10 cm, the best prediction accuracy was achieved by the ST-XGBoost algorithm. The best learner, ST-XGBoost, has the R^2 values in the interval of 0.927 to 0.986, while the values of the ST-SVR method is in the interval of 0.321 to 0.606.

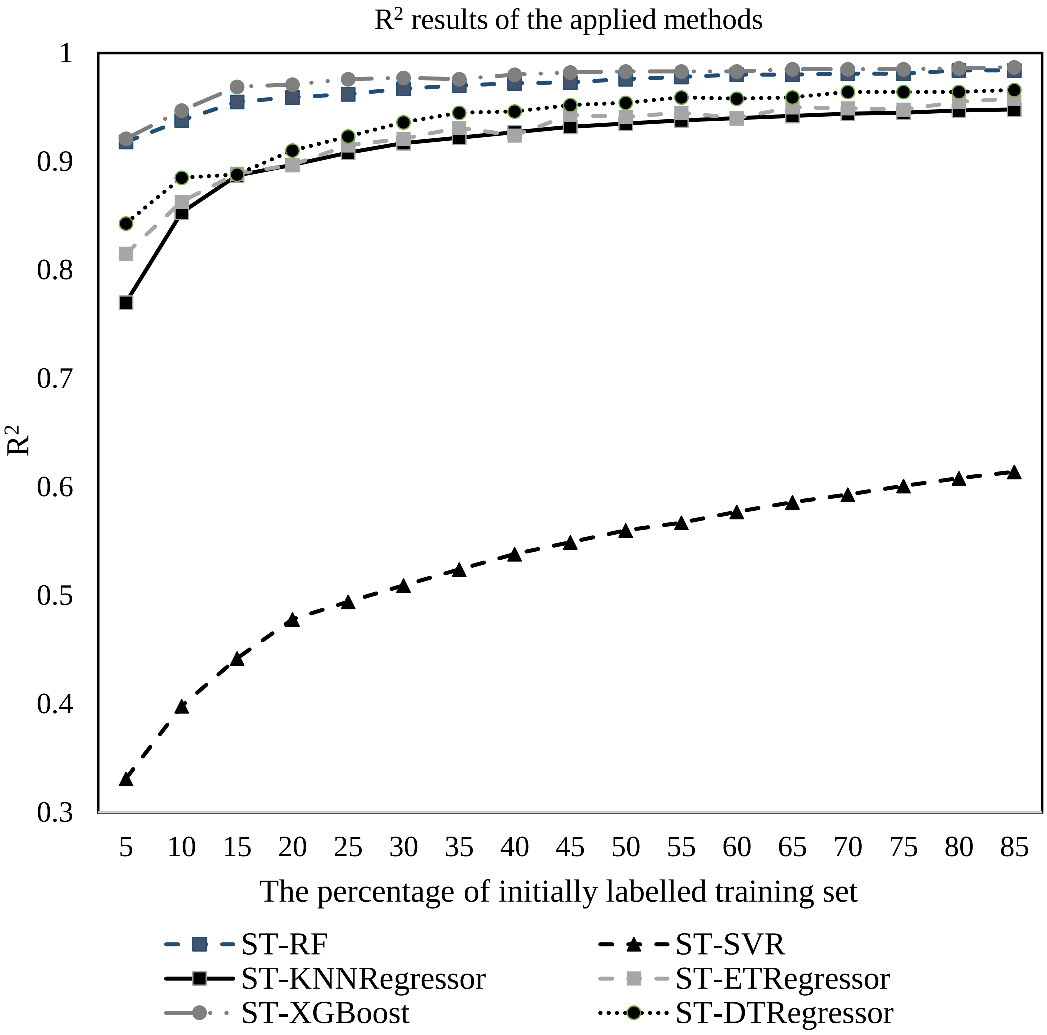

Figure 3- The comparison of the applied methods under self-training framework in terms of R^2 values (ST-RF: the RF method combined with self-training, ST-SVR: the SVR method combined with self-training, ST-KNNRegressor: the KNNRegressor method combined with self-training, ST-ETRegressor: the ETRegressor method combined with self-training, ST-XGBoost: the XGBoost method combined with self-training, ST-DTRegressor: the DTRegressor method combined with self-training)

Figure 4 displays the best-performed model, ST-XGBoost, in terms of the values of MSE at all depths. It is apparent that the model predicts soil temperature well at the depth of 50 cm while the values at the depth of 30 cm are generally estimated worse compared to others. Besides, as in Table 3, MSE values at depths of 10, 20, 30, and 40 cm are decreased when the size of the labeled training set is increased.

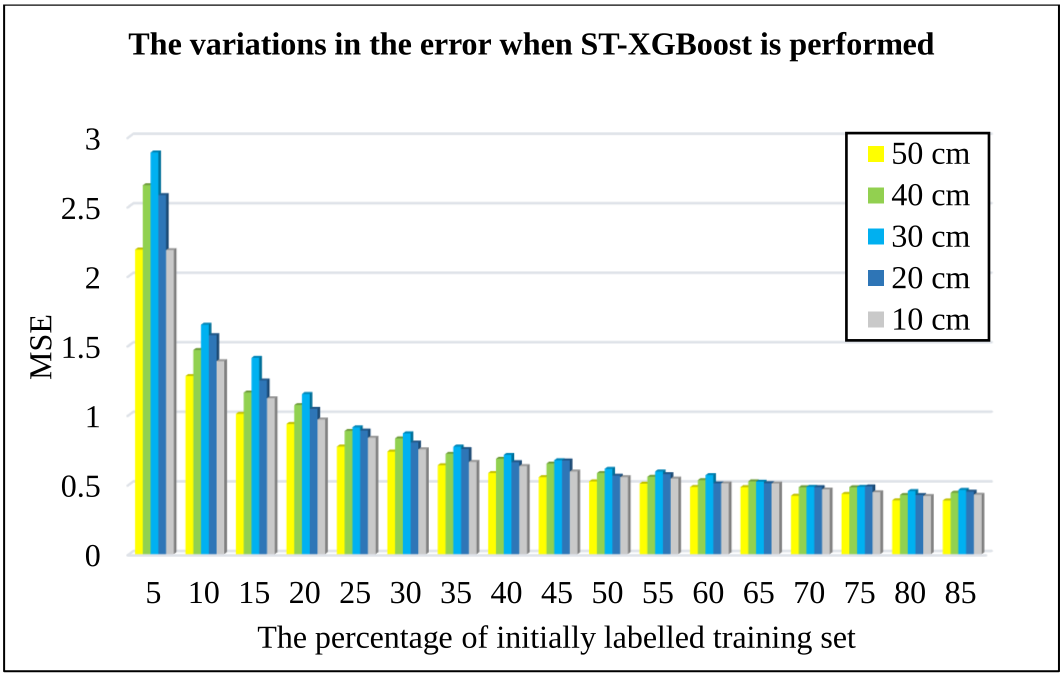

Figure 4- The results of MSE when the best predictor (ST-XGBoost) was performed at all soil depths

Figure 5 demonstrates the predicted and the measured values of test samples of 10-fold cross-validation steps separately when ST-XGBoost is performed at the depth of 50 cm and p is 85% of the whole data. In each iteration, whole data including 4500 samples are divided into training and test (90% of data as the training set and 10% of data as the test set) and this procedure is repeated ten times. In that way, the selection of the samples with different characteristics are increased instead of experimenting only with predefined test cases. The advantage of this is that, perhaps, the samples that are easy to predict in one iteration are collected in the test set, while in another iteration, the opposite (difficult samples for prediction) can be observed. Since the common result of all of them is obtained, a better inference is made than depending on a result of a single sample set. In this direction, Figure 5 shows clearly that the ST-XGBoost model obtains soil temperature values that are very close to the real measurements in each fold.

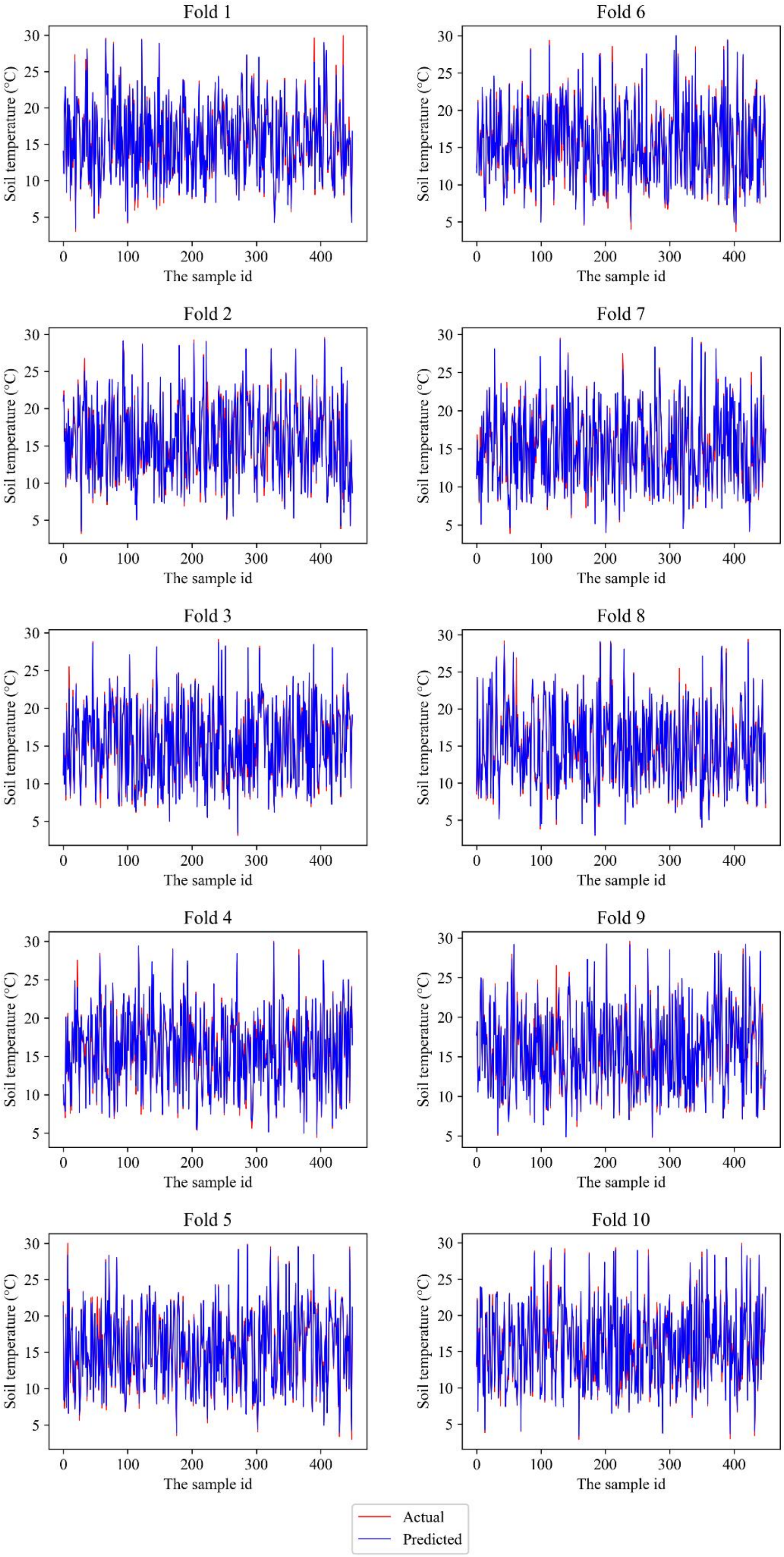

Figure 5- The predicted and the real values of the soil temperature when the ST-XGBoost model was applied at the depth of 50 cm and *p* is 85%

In order to show the trend in the results as the soil depth changes, Table 5 demonstrates the MSE values of each applied method by taking their average for all *p* values from 5 to 85. Especially for ensemble learning methods, there is a general pattern that the error first increases for the soil depth of 10 to 30 and then a decreasing trend follows as the depth increases from 40 to 50 cm. Their common characteristic is that they can better estimate the deepest soil temperature. On the contrary, there is a decreasing trend in most of the cases for single learners from shallow to deeper parts. ST-SVR and ST-DTReg manage to predict the best at the depth of 50 cm while ST-KNNReg performs well at the depth of 40 cm. The main inference for the proposed model is that the more accuracy is generally obtained the deeper the soil. It is clear that the effect of the meteorological parameters on the soil temperature is greater in the regions close to the soil surface. While the soil in the shallow regions is affected more by external factors such as rainfall and wind, as the soil depth increases, a more stable environment is found in terms of the soil temperature modelling, and regression models, therefore, create better predictive models with less error.

Table 5- The comparison of the mean values of the applied methods in terms of MSE values under self-training framework at different soil depths

Methods	*Depth (cm)*				
	50	*40*	*30*	*20*	*10*
ST-RF	0.866	0.927	0.958	0.941	0.931
ST-ETReg	2.130	2.339	2.492	2.502	2.321
ST-SVR	13.120	13.986	14.412	14.311	14.384
ST-XGBoost	0.724	0.831	0.894	0.836	0.764
ST-KNNReg	2.343	2.286	2.288	2.328	2.637
ST-DTReg	1.714	1.848	1.957	1.941	1.920

Table 6 shows the R^2 results at different soil depths to compare our study with the recent studies in the literature. The values are the best R^2 values obtained with the best parameter combinations in the mentioned studies and our results obtained on our dataset are the optimal R^2 values of the best method, ST-XGBoost, at the specified depths. It is apparent that our proposed method generally outperforms the other models and it manages to estimate soil temperature accurately.

Table 6- The accuracy of the results (R^2) obtained in this study and the similar results obtained in the literature

			R^2	
Ref	*Algorithm*	*Depth (cm)*	*Existing Method*	*Proposed Method (STST)*
Alizamir et al. 2020a	• Deep Echo State Network (Deep ESN)	20	0.970	**0.985**
	• Multilayer Perceptron Neural Network (MLPNN)	10	0.890	**0.986**
	• M5Prime Tree	10	0.870	**0.986**
	• Random Forest	10	0.900	**0.986**
Li et al. 2020	• Integrated Bidirectional Long Short-Term Memory Network (BiLSTM)	50	0.920	**0.986**
	• Long Short-Term Memory (LSTM)		0.880	
	• Bidirectional Long Short-Term Memory Network (BiLSTM)		0.860	
	• Deep Neural Network (DNN)		0.870	
	• Random Forest (RF)		0.860	
	• Support Vector Regression (SVR)		0.790	
	• Linear Regression (LR)		0.420	
Penghui et al. 2020	• Adaptive Neuro-Fuzzy Inference System with Grasshopper Optimization Algorithm (ANFIS-mSG)	10	0.977	**0.986**
Guan et al. 2020	• The Hybrid of Multilayer Perceptron by Invasive Weed Optimization (MLP-IWO)	20	0.962	**0.985**
Alizamir et al. 2020b	• Extreme Learning Machine (ELM)	10	**0.986**	**0.986**
	• Artificial Neural Networks (ANN)	50	0.984	**0.986**
	• Classification and Regression Trees (CART)	10	0.984	**0.986**
	• Group Method of Data Handling (GMDH)	10	**0.988**	0.986
	• Multi-Linear Regression (MLR)	50	**0.988**	0.986
Huang et al. 2020	• Multivariate Linear Regression	10	0.915	**0.986**
		20	0.889	**0.985**
		40	0.799	**0.986**
Behmanesh & Mehdizadeh 2017	• Gene Expression Programming (GEP)	10	0.974	**0.986**
	• Artificial neural networks (ANN)		0.980	
	• Multiple linear regression (MLR)		0.971	
Abyaneh et al. 2016	• Artificial Neural Networks (ANN)	10	0.968	**0.986**
		20	0.926	**0.985**
		30	0.893	**0.985**
		50	0.872	**0.986**
Ozturk et al. 2011	• Artificial Neural Networks (ANN)	10	0.960	**0.986**
		20	0.981	**0.985**
		50	0.966	**0.986**

The main findings of the study can be concluded as follows: *1)* It was observed that "self-training" smartly provides many advantages for predicting soil temperature, including reducing cost and providing additional information present in unlabeled data. *2)* The proposed STST approach has the potential to expand the application of machine learning in the agriculture sector, thanks to its advantages. *3)* The ST-XGBoost method outperformed the other methods (ST-RF, ST-ETReg, ST-SVR, ST-KNNReg, and ST-DTReg) in terms of prediction accuracy. *4)* The prediction error changes according to the soil depth. *5)* The accuracy of soil temperature prediction increased as the number of labeled data samples increased.

4. Conclusions

In this study, soil temperature at various soil depths was predicted using the proposed model, *Soil Temperature prediction via Self-Training* (STST). The past soil temperature & soil moisture data and meteorological data of Izmir, Turkey were considered in the interval of 01.09.2019 and 22.05.2020. The experimental results showed that self-training empowered the regression methods by presenting a more labeled pool of data for training a model for prediction. In this way, test samples were estimated more accurately using the information hidden in the expanded labeled instances instead of using few samples with known past values. Especially ensemble learning methods (ST-XGBoost and ST-RF) managed to capture the dynamics better behind the soil temperature prediction compared to other ones under the self-training framework. The best model, ST-XGBoost respectively obtained the results in the range of 0.385-2.888, 3.109%-8.740%, and 0.905-0.986 at depths of 5, 10, 20, 30, 40, and 50 cm for the performance metrics MSE, MAPE, and R^2. In addition, the best predictions were generally made at the depth of 50 cm with the mean MSE values of 0.866, 2.130, 13.120, 1.724, and 1.714 for ST-RF, ST-ETReg, ST-SVR, ST-XGBoost, and ST-DTReg, respectively.

This study contributes to the agricultural field in a way that plant growth can be handled more efficiently by taking the predicted soil temperature values into account. An automated irrigation system or cooling/heating system can be set up according to the variation in the temperature of the predicted time intervals. In the same manner, as future work, the proposed model may be customized and updated in order to estimate the soil moisture, which is another important parameter in plant production.

References

Abyaneh H Z, Varkeshi M B, Golmohammadi G & Mohammadi K (2016). Soil temperature estimation using an artificial neural network and co-active neuro-fuzzy inference system in two different climates. *Arabian Journal of Geosciences* 9(5): 377. https://10.1007/s12517-016-2388-8

Akengin H & Dölek İ (2019). Physical Geography of Turkey. Pegem Akademi, Ankara (In Turkish)

Alizamir M, Kim S, Zounemat-Kermani M, Heddam S, Shahrabadi A H & Gharabaghi B (2020a). Modelling daily soil temperature by hydro-meteorological data at different depths using a novel data-intelligence model: deep echo state network model. *Artificial Intelligence Review* 54: 2863-2890. https://doi.org/10.1007/s10462-020-09915-5

Alizamir M, Kisi O, Ahmed A N, Mert C, Fai C M, Kim S, Kim N W & El-Shafie A (2020b). Advanced machine learning model for better prediction accuracy of soil temperature at different depths. *Plos One* 15(4): e0231055. https://doi.org/10.1371/journal.pone.0231055

Anton C A, Matei O & Avram A (2019). Collaborative data mining in agriculture for prediction of soil moisture and temperature. In: *Computer Science On-line Conference (CSOC 2019)*, 24-27 April, Zlin, Czech Republic pp. 141-151. https://doi.org/10.1007/978-3-030-19807-7_15

Behmanesh J & Mehdizadeh S (2017). Estimation of soil temperature using gene expression programming and artificial neural networks in a semiarid region. *Environmental Earth Sciences* 76(2): 76. https://doi.org/10.1007/s12665-017-6395-1

Belkin M, Niyogi P & Sindhwani V (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research* 7(Nov): 2399-2434

Bilgili M (2010). Prediction of soil temperature using regression and artificial neural network models. *Meteorology and Atmospheric Physics* 110(1-2): 59-70. https://doi.org/10.1007/s00703-010-0104-x

Bilgili M (2011). The use of artificial neural networks for forecasting the monthly mean soil temperatures in Adana, Turkey. *Turkish Journal of Agriculture and Forestry* 35(1): 83-93. https://10.3906/tar-1001-593

Bilgili M (2012). Regional assessment of monthly soil temperatures in the Aegean Region of Turkey. *Arabian Journal for Science and Engineering* 37(3): 765-775. https://doi.org/10.1007/s13369-012-0199-0

Bilgili M, Sahin B & Sangun L (2013). Estimating soil temperature using neighboring station data via multi-nonlinear regression and artificial neural network models. *Environmental Monitoring and Assessment* 185(1): 347-358. https://doi.org/10.1007/s10661-012-2557-5

Bolca M, Kurucu Y, Dengiz O & Nahry A D H (2011). Terrain characterization for soils survey of Kucuk Menderes plain, South of Izmir, Turkey, using remote sensing and GIS techniques. *Zemdirbyste-Agriculture* 98: 93-104

Citakoglu H (2017). Comparison of artificial intelligence techniques for prediction of soil temperatures in Turkey. *Theoretical and Applied Climatology* 130(1-2): 545-556. https://doi.org/10.1007/s00704-016-1914-7

Dadashzadeh M, Abbaspour-Gilandeh Y, Mesri-Gundoshmian T, Sabzi S, Hernández-Hernández J L, Hernández-Hernández M & Arribas J I (2020). Weed classification for site-specific weed management using an automated stereo computer-vision machine-learning system in rice fields. *Plants* 9(5): 559. https://doi.org/10.3390/plants9050559

Dizdar M Y (2003). Turkey's land resources. Retrieved in January, 10, 2021 from https://www.zmo.org.tr/ (In Turkish)

Friedl M A (2018). Remote sensing of croplands. *Comprehensive Remote Sensing* 6: 78-95. https://doi.org/10.1016/B978-0-12-409548-9.10379-3

Giraddi S, Desai S & Deshpande A (2020). Deep Learning for Agricultural Plant Disease Detection. In: *Proceedings of the 1st International Conference on Data Science, Machine Learning and Applications (ICDSMLA 2019)*, 21-22 Nov, Pune, India, pp. 864-871. https://doi.org/10.1007/978-981-15-1420-3_93

Gönençgil B, Biricik A S, Atalay İ, Aydınözü D, Çoban A & Ertek A (2016). Turkey physical geography. Retrieved in October, 9, 2021 from http://auzefkitap.istanbul.edu.tr/ (In Turkish)

Guan Y, Shen Y, Mohammadi B & Sadat M A (2020). Estimation of soil temperature based on meteorological parameters by the HYBRID INVASIVE Weed Optimization Algorithm Model. *IOP Conference Series: Earth and Environmental Science* 428: 012059. https://10.1088/1755-1315/428/1/012059

Hamrani A, Akbarzadeh A & Madramootoo C A (2020). Machine learning for predicting greenhouse gas emissions from agricultural soils. *Science of the Total Environment* 741: 140338. https://doi.org/10.1016/j.scitotenv.2020.140338

Hillel D (2005). Thermal properties and processes. In: D Hillel (Ed.), *Encyclopedia of Soils in the Environment*, Academic Press, pp. 156-163

Huang R, Huang J X, Zhang C, Ma H Y, Zhuo W, Chen Y Y, Zhu D H, Wu Q & Mansaray L R (2020). Soil temperature estimation at different depths, using remotely-sensed data. *Journal of Integrative Agriculture* 19(1): 277-290. https://doi.org/10.1016/S2095-3119(19)62657-2

Kapur S, Aydın M, Akça E & Reich P (2018). Climate change and soils. In: S Kapur, E Akça & H Günal (Eds.), *The Soils of Turkey*, Springer, pp. 45-55. https://doi.org/10.1007/978-3-319-64392-2_4

Kisi O, Tombul M & Kermani M Z (2015). Modeling soil temperatures at different depths by using three different neural computing techniques. *Theoretical and Applied Climatology* 121(1-2): 377-387. https://doi.org/10.1007/s00704-014-1232-x

Kisi O, Sanikhani H & Cobaner M (2017). Soil temperature modeling at different depths using neuro-fuzzy, neural network, and genetic programming techniques. *Theoretical and Applied Climatology* 129(3-4): 833-848. https://doi.org/10.1007/s00704-016-1810-1

Nandy A & Singh P K (2020). Farm efficiency estimation using a hybrid approach of machine-learning and data envelopment analysis: evidence from rural eastern India. *Journal of Cleaner Production* 267: 122106. https://doi.org/10.1016/j.jclepro.2020.122106

Niedbała G, Kurasiak-Popowska D, Stuper-Szablewska K & Nawracała J (2020). Application of artificial neural networks to analyze the concentration of ferulic acid, deoxynivalenol, and nivalenol in winter wheat grain. *Agriculture* 10(4): 127. https://doi.org/10.3390/agriculture10040127

Ozturk M, Salman O & Koc M (2011). Artificial neural network model for estimating the soil temperature. *Canadian Journal of Soil Science* 91(4): 551-562. https://doi.org/10.4141/cjss10073

Pekel E (2020). Estimation of soil moisture using decision tree regression. *Theoretical and Applied Climatology* 139(3): 1111-1119. https://doi.org/10.1007/s00704-019-03048-8

Penghui L, Ewees A A, Beyaztas B H, Qi C, Salih S Q, Al-Ansari N, Bhagat S K, Yaseen Z M & Singh V P (2020). Metaheuristic optimization algorithms hybridized with artificial intelligence model for soil temperature prediction: Novel model. *IEEE Access* 8: 51884-51904. https://10.1109/ACCESS.2020.2979822

Ren C, Liang Y J, Lu X J, & Yan H B (2019). Research on the soil moisture sliding estimation method using the LS-SVM based on multi-satellite fusion. *International Journal of Remote Sensing* 40(5-6): 2104-2119. https://doi.org/10.1080/01431161.2018.1475778

Republic of Turkey Ministry of Agriculture and Forestry (2021). Technical instruction of soil and land classification standards. Retrieved in January, 12, 2021 from https://www.tarimorman.gov.tr/ (In Turkish)

Shamshirband S, Esmaeilbeiki F, Zarehaghi D, Neyshabouri M, Samadianfard S, Ghorbani M A, Mosavi A, Nabipour N & Chau K W (2020). Comparative analysis of hybrid models of firefly optimization algorithm with support vector machines and multilayer perceptron for predicting soil temperature at different depths. *Engineering Applications of Computational Fluid Mechanics* 14(1): 939-953. https://doi.org/10.1080/19942060.2020.1788644

Tabari H, Sabziparvar A A & Ahmadi M (2011). Comparison of artificial neural network and multivariate linear regression methods for estimation of daily soil temperature in an arid region. *Meteorology and Atmospheric Physics* 110(3): 135-142. https://doi.org/10.1007/s00703-010-0110-z

Turkish State Meteorological Service (2021). Climate status of Izmir province. Retrieved in January, 10, 2021 from https://izmir.mgm.gov.tr/ (In Turkish)

Xing L, Li L, Gong J, Ren C, Liu J & Chen H (2018). Daily soil temperatures predictions for various climates in United States using data-driven model. *Energy* 160: 430-440. https://doi.org/10.1016/j.energy.2018.07.004

Yang H, Zhou J, Feng J, Zhai S, Chen W, Liu J & Bian X (2019). Ditch-buried straw return: A novel tillage practice combined with tillage rotation and deep ploughing in rice-wheat rotation systems. *Advances in Agronomy* 154: 257-290. https://doi.org/10.1016/bs.agron.2018.11.004

Yener D, Ozgener O & Ozgener L (2017). Prediction of soil temperatures for shallow geothermal applications in Turkey. *Renewable and Sustainable Energy Reviews* 70: 71-77. https://doi.org/10.1016/j.rser.2016.11.065

Zhu X & Goldberg A B (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 3(1): 1-130. https://doi.org/10.2200/S00196ED1V01Y200906AIM006