PLANT SCIENCE TODAY
ISSN 2348-1900 (online)
Vol x(x): xx–xx
https://doi.org/10.14719/pst.7291

HORIZON
e-Publishing Group
HePG

RESEARCH ARTICLE

Soil temperature prediction based on ensemble tree bagger machine learning algorithm for agricultural decision making

A Alagesan[1], Thukkaiyannan P[2*], Satheeshkumar N[3], Thiruvarassan S[4], Ganesan K[5] & Ayyadurai P[6]

[1]ICAR-Krishi Vigyan Kendra, Tamil Nadu Agricultural University, Pudukottai 622 303, Tamil Nadu, India

[2]ICAR-Krishi Vigyan Kendra, Tamil Nadu Agricultural University, Tiruppur 641 667, Tamil Nadu, India

[3]Maize Research Station, Tamil Nadu Agricultural University, Vagarai 624 613, Tamil Nadu, India

[4]ICAR-Krishi Vigyan Kendra, Tamil Nadu Agricultural University, Villupuram 604 002, Tamil Nadu, India

[5]Directorate of Planning and Monitoring, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India

[6]Centre of Excellence in Millets, Tamil Nadu Agricultural University, Athiyandal 606 603, Tamil Nadu, India

*Email: thukkaiyannan@tnau.ac.in

OPEN ACCESS

ARTICLE HISTORY
Received: 09 January 2025
Accepted: 25 January 2025
Available online
Version 1.0 : 14 March 2025

Check for updates

Additional information
Peer review: Publisher thanks Sectional Editor and the other anonymous reviewers for their contribution to the peer review of this work.

Reprints & permissions information is available at https://horizonepublishing.com/journals/index.php/PST/open_access_policy

Publisher's Note: Horizon e-Publishing Group remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Indexing: Plant Science Today, published by Horizon e-Publishing Group, is covered by Scopus, Web of Science, BIOSIS Previews, Clarivate Analytics, NAAS, UGC Care, etc
See https://horizonepublishing.com/journals/index.php/PST/indexing_abstracting

CITE THIS ARTICLE
Alagesan A, Thukkaiyannan P, Satheeshkumar N, Thiruvarassan S, Ganesan K, Ayyadurai P. Soil temperature prediction based on ensemble tree bagger machine learning algorithm for agricultural decision making. Plant Science Today (Early Access). https:/doi.org/10.14719/pst.7291

Abstract

This study focuses on predicting surface soil temperature (ST) at a 5 cm depth, which significantly influences agricultural decisions such as sowing time, irrigation management and soil-plant-atmosphere dynamics. Machine learning (ML) algorithms were used to predict ST using above-ground weather variables viz., air temperature (T), relative humidity (RH), wind velocity (WV) and sunshine duration (SS) measured at 15-min intervals. Six regression-based ML models (Ensemble, Gaussian Process Regression, Support Vector Machine, Tree, Neural Network and Kernel) were trained and tested for predictive accuracy. The Ensemble Bagging Tree model showed the highest precision, with RMSE values of 2.04 and 1.9 for validation and testing, respectively. Various combinations of the weather variables were tested and the model performed best when using above mentioned variables. Among the predictors, T had the greatest impact on ST prediction, as indicated by mean absolute Shapley values. The Shapley values of the variables revealed that T had a critical role in the model output, with time, SS, RH and WV following in importance. Additionally, as a model explainable artificial intelligence (xAI) metrics, SHapley Additive exPlanations (SHAP) were analysed and found that SHAP dependency had a defined relationship between the predictors and ST at a 5 cm depth. This study highlights the effectiveness of machine learning in predicting soil temperature and emphasizes the role of weather variables in agricultural decision-making. decision-making.

Keywords

artificial intelligence; explainable ai; ensemble tree bagger; machine learning; regression learner; SHAP; Shapley; soil temperature

Introduction

Soil temperature (ST) plays an important role in the entire earth system process, which regulates the exchange of energy between atmosphere and the terrestrial land surface. ST also influences the climatic process, eco system functioning, land surface process, agricultural ecosystem and the energy balance of the planet earth (1). With respect to the agriculture ecosystem, the ST influences the seed germination process, establishment of seedling growth and development of crop plants. For instance, crops such as corn, soybeans and spinach exhibit marked temperature thresholds that

influence germination rates and subsequent seedling development. It also influences crop yield by way of its influence on plant respiration, metabolic activity, photo synthesis process and below ground plant parameters viz., root growth, root development and root expansion etc., which facilitate the uptake of required nutrients by the plant from the soil (2). The ST aids various agricultural decision-making process in agriculture land use management, identification of plant stress, selection of suitable crop variety, decisions regarding the timing of sowing and planting, scheduling of irrigation, precision farming and many other soil management strategies (3, 4). In addition, many earth system models viz., hydrology, atmospheric and numerical weather prediction models use the ST as one of the important predictor variables for increasing its model efficiency.

Keeping the above importance of ST for specific purposes, it is believed that, in the future, the accurate availability of ST will play a critical role in agriculture for making informed decisions related to productive agriculture. However, the availability of information on ST is very scarce across the agricultural production system and the countries worldwide, irrespective of their advancement (5). This is attributable to the instrumentation cost, maintenance and performance of sensor network in the real agricultural production system (5). Conventionally, the ST is being modelled through statistical techniques using historic weather data aided by the time series statistical model viz., ARIMA, SARIMA etc. Although these physical models predict ST, they depend on highly sensitive, real-time ST measurements obtained via costly sensor, making them impractical for widespread use in the real-world agricultural production systems. Due to recent advances in soft computing technologies and data driven machine learning techniques and artificial intelligence, decision makers can predict the ST with less computation cost, in comparison to physical models that require high-cost input data and time. During the past decades, scientists has developed different soft computing algorithm for prediction of ST by inputting various meteorological parameters viz., ST, relative humidity (RH), wind velocity (WV), solar radiation (SR), precipitation (P), atmospheric pressure (p) etc. throughout the world. Weather stations across the world normally record above ground weather parameters for prediction of different interrelated variables. However, majority of these weather stations lack the observation on the important micro climatic parameters of ST.

The use of regression learner-based machine learning approaches for soil temperature prediction offers several advantages, particularly when applied to sub-hourly timescales using meteorological data. First, these methods can effectively capture non-linear relationships and complex interactions among variables, enabling more accurate predictions compared to traditional linear regression models. Additionally, ensemble-based approaches, such as bagging trees, combine predictions from multiple models, reducing overfitting and improving generalization performance. These algorithms also allow for high temporal resolution, making them suitable for sub-hourly soil temperature prediction. By leveraging weather parameters like air temperature, humidity, wind velocity and sunshine duration, machine learning models can provide real-time or near-real-time estimates of soil conditions, which is critical for agricultural applications such as irrigation management and sowing time decisions. Furthermore, the ability to train models on different resamples of data enhances robustness and stability, even in the presence of noise or variability in input variables. This capability ensures reliable predictions under uncertain conditions, making machine learning approaches a valuable tool for soil-plant-atmosphere system dynamics analysis. Overall, these advantages underscore the potential of regression learner-based methods in advancing our ability to model soil temperature with high precision and relevance to environmental and agricultural research.

Materials and Methods

The meteorological weather parameters were obtained from historical records of Tamil Nadu Agricultural University, specifically for Pudukottai district of Tamil Nadu, India. The dataset consists of T, RH, WV, sunshine duration and ST with a 15-min time scale resolution recorded through an automatic weather station. The data set consist of all the above parameters from 08.10.2021 to 03.11.2023. Each variable has 29229 data points and in total about 175374 data were used for the entire process. For model training purpose 75 % of the data were used, accounting about 21922 data per variable and 25 % of the data points were used for testing the model that accounted for 7307 data points per variable. The used data points were recorded at 15 min time interval, ensuring 96 data points per day. The 24 hr of a day with 15 min time interval was converted into 0 to 1 time scale automatically by the model for its algorithm compatibility with an incremental value of 0.01 for each 15 min for training and testing the machine learning model. The unit of weather variables were 15 min interval for time, temperature in ^{0}C, wind velocity in m/s and sunshine duration in accumulated min. The data were subjected to different machine learning algorithm for the prediction of soil temperature using the rest of variables as an input component. This study trained six machine learning models for prediction of soil temperature and the performance of all the trained models were compared using the statistical parameters, to select the best using the soft computing models with software MATLAB 2024b.

The regression learner-based machine learning algorithm used in this study for the prediction of ST were Ensemble Bagging Tree (ETB), Support Vector Machine (SVR), Gaussion Process Regression (GPR), Neutral Network (NN), Decision Tree (DT) and Regression Tree (RT). All the above models were trained, tested and its performance were interpreted based on the statistical metrics as defined in Equation 1 to 5 as detailed below.

i) Mean Absolute Error (MAE) (Eqn. 1)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

The MAE is always positive and like the RMSE, but less sensitive to outliers. Look for smaller values of the MAE.

Where.

Yi = represents the actual or observed value for *the* i^{th} data point.

$\hat{y}_i$ = represents the predicted value for the i^{th} *data point.*

ii) Mean Squared Error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad \text{(Eqn. 2)}$$

The MSE is the square of the RMSE. Look for smaller values of the MSE.

iii) Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad \text{(Eqn. 3)}$$

The RMSE is always positive and its units match the units of the response. Look for smaller values of the RMSE.

iv) Relative Squared Error Score (R^2) or Coefficient of determination

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \qquad \text{(Eqn. 4)}$$

Where.

SSR - sum of squared residuals between predicted and actual values

SST - total sum of squares, which measures the total variance in the dependent variable

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad \text{(Eqn. 5)}$$

Coefficient of determination. The app calculates ordinary (unadjusted) R^2 values. R^2 is always smaller than 1 and usually larger than 0. It compares the trained model with the model where the response is constant and equals the mean of the training response. Look for an R^2 value close to 1.

The RMSE metrics is used as a tool to assesses the performance of different models during its training and testing phase. The data sets were trained with fivefold cross validation. The above models were trained and best performing model was selected for further data processing, model training, validation and model testing. Different input combination of weather variables was trained, tested and their performance evaluated. The best input combination of weather variables was selected for further evaluation of the model performance to fine-tune the model to achieve a high degree of prediction accuracy.

Ensemble Tree Bagging is a type of ensemble learning technique used to improve model performance by creating multiple replicas (or bootstrap samples) of a training dataset. For a given weak learner, such as a decision tree, the process involves generating numerous bootstrap replicas, each created by randomly selecting N out of N observations with replacement, where N is the total number of observations in the dataset. This method helps to reduce variance in predictions and often improves accuracy. Furthermore, random forests, which involve an additional step where each tree randomly selects predictors for decision splits, are known to enhance the predictive power of bagged trees (6). By default, the number of predictors selected at each split is typically set to the square root of the total number of predictors for classification tasks and one-third of that number for regression tasks. This configuration often achieves optimal predictive performance. After training a model, predicting new data involves taking an average over predictions from all individual trees in the ensemble using the predict function. The default minimum number of observations per leaf for bagged trees is 1 for classification and 5 for regression. Trees with these settings are usually deep and tend to be close to optimal in terms of predictive power (7). Increasing the leaf size can reduce training and prediction time, as well as memory usage, without significantly compromising performance.

The 'MinLeafSize' parameter in templateTree or TreeBagger allows users to specify the minimum number of observations per leaf, providing flexibility in balancing computational efficiency with model performance. Additionally, out-of-bag (OOB) observations, which are omitted on average 37% of the time for each decision tree, play a crucial role in enabling properties like variable importance analysis and highlighting outliers in the data. A notable feature of TreeBagger is the proximity matrix, accessible via the Proximity property. This matrix measures how close two observations are based on their placement in the same leaf across different trees in the ensemble. By normalizing these proximities over all trees, a symmetric matrix with diagonal elements of 1 and off-diagonal values ranging from 0 to 1 is obtained. This matrix can be used for tasks such as identifying outliers and uncovering clusters through techniques like multidimensional scaling (8).

SHapley Additive exPlanations (SHAP) is an explainable artificial intelligence (XAI) technique that uses Shapley value from the concept of cooperative game theory (9, 10). SHAP is a groundbreaking methodology aimed at enhancing model interpretability. Consider a group of players working together to clear a game. How would the final reward be split if each player contributed differently? Shapley value can be used in this situation to guarantee that the allocation of rewards to each player is fair by calculating the marginal contribution of each player (11). Shapley values satisfy the four axioms for calculating each player's marginal contribution viz., 1. Efficiency: the final reward must be shared among the players in cooperation, 2. Symmetry: players who made the same contribution as each other will receive the same amount of reward, 3. Dummy: players who did not contribute to the game clearance are known as dummy players and will receive no reward and 4. Additivity: if the game has multiple parts, the player's reward allocation must consider the individual contribution to each part rather than the collective contribution to the game as a whole (12). Its primary objective of using SHAP is to provide a clear

framework for making complex models, such as those based on machine learning or deep learning, more accessible for researchers to understand. In the field of explainable AI (XAI), one of the most challenging tasks has been selecting the most appropriate algorithm for a specific model type. To address this challenge, Lundberg and colleagues developed SHAP, an innovative framework that assigns importance values to individual features in the context of a particular prediction. By providing these importance scores, SHAP helps researchers identify which factors have the greatest influence on a given outcome, thereby supporting the identification of key drivers of predictions. This advancement contributes significantly to our ability to understand how complex models operate and make decisions. As a result, SHAP has the potential to be used in predicting ST by analyzing the contribution different weather parameters. a) Shapley importance: Shapley importance plot, by utilizing Shapley values across a series of query points, we can assess which predictors have the greatest or least impact on regression model predictions. For each query point, the Shapley value of a predictor quantifies how that variable contributes to deviations in the model's predicted output from its average baseline prediction. The sign of this value indicates whether the effect is positive or negative, while the absolute value reflects the magnitude of the influence. Calculating the mean of these absolute Shapley values across all query points provides a comprehensive measure of each predictor's significance in shaping model predictions, enabling a deeper understanding of their individual contributions to the regression model's outcomes. b) Shapley summary: For regression models, Shapley values can be used on a set of query points to assess the influence of individual predictors on model predictions. At each query point, the Shapley value for a predictor quantifies the deviation in the predicted response from the average prediction. The sign of this value indicates whether the effect is positive or negative, while its absolute value reflects the magnitude of the impact. Consequently, Shapley values near zero suggest that the specific predictor has minimal influence on the model's predictions for that query point. c) Shapley dependency: For regression models, Shapley values can be employed on query points to analyze the influence of individual predictors on predictions. The value explains how much a prediction deviates from the average due to each predictor. The sign of the Shapley value indicates whether the effect is positive or negative, while its absolute value represents the magnitude of impact. Therefore, when Shapley values are close to zero, it suggests that a particular predictor has minimal influence on the model's predictions for that specific query point.

Results and Discussion

ST plays an important role in agricultural ecosystem as a vital variable in controlling the soil-plant-atmosphere continuum. In general, majority of the meteorological observatory located in a particular agricultural eco system region normally records the weather variable viz., air temperature, RH, WV, sunshine duration, rainfall and other above ground weather parameters. Using these above ground weather parameters the meteorologists used to predict the future value of the above ground variables using historic weather data. When measure ST datasets is available along with above ground weather parameters, it is possible to model the data sets and ST can be predicted for future using machine learning algorithm by proper training and testing the data sets. Such trained models can be used for predicting the ST in different locations, were only above ground historic weather data available.

Application of machine learning for the prediction of ST was assessed using data training, validation and testing process. In this study six regression learner-based machine learning algorithm were trained, tested and evaluated for their performance and its metrics were presented in Table 1. Upon scrutinizing and evaluating the different models trained and tested, it was found that the regression learner-based machine learning algorithm, called Ensemble-Bagged Tree, had the highest prediction accuracy compared to the five models. The least performance accuracy was observed for the trained and tested model of Kernel based regression learner. Hence, the Model Ensemble Bagged Tree was selected for further testing process and model performance evaluation.

The ST prediction process was divided into five components based on the number of input variables used for data training and testing. The particulars of input variable combinations used in training and testing phase for the prediction of ST is presented in Table 2, along with the corresponding data validation and testing metrics. From this model metrics, it is observed that the model performance accuracy improves with the addition of more predictor variables. The highest model prediction accuracy was observed for the predictor input variables used viz., time of observation, air temperature, RH, WV and sunshine duration.

This study refers the work of (13) which modelled the ST data using multi-layer perception, Radial Basis Neural Network and Generalized Regression Neural Network (GRNN) and concluded that the GRNN model performed well for the prediction of ST at 5 cm depth with a RMSE value of 2.06. It is evident from the Table 1, 2 that the proposed

Table 1. Evaluation of different regression learner machine learning algorithms using different model metrics

Model Number	Model Type	RMSE (Validation)	MSE (Validation)	R Squared (Validation)	MAE (Validation)	MAE (Test)	MSE (Test)	RMSE (Test)	R Squared (Test)
1	Ensemble Bagged Tree	2.04	4.15	0.70	1.46	1.4	3.96	1.99	0.72
2	Gaussian Process Regression	2.23	4.96	0.64	1.68	1.63	4.8	2.19	0.66
3	Support vector machine	2.31	5.35	0.61	1.65	1.6	5.19	2.28	0.63
4	Tree	2.42	5.83	0.58	1.61	1.52	5.41	2.33	0.61
5	Neural Network	2.4	5.75	0.59	1.83	1.8	5.58	2.36	0.6
6	Kernel	2.49	6.21	0.55	1.94	1.93	6.16	2.48	0.56

Ensemble Tree Bagging machine learning algorithm-based model has successfully modelled the ST at 5 cm depth with a relatively smaller RMSE as presented in Table 2. ST at 5 cm depth was predicted (14), using daily weather parameter viz., maximum temperature, minimum temperature, average temperature, atmospheric pressure and solar radiation and found that the ELM model performed well with a performance metric RMSE value of 1.22 for training data set and 1.23 for testing data set at Bandar Abbas Station. For SAE-ELM, the corresponding value was 1.20 and 1.18, respectively. Whereas the same author conducted a study for Kerman Station and the RMSE value for training and testing data sets were 2.1 and 2.2 respectively for the ELM model and 2.1 and 2.1 for the SAE-ELM model. ST was modelled at different soil depth using four different models (15) viz., Random Forest, Extreme learning machine, generalized regression neural network and Back propagation neural network and reported the RMSE values as 2.31, 2.36, 2.48 and 2.39, respectively, for each model. Another study modelled the daily ST at ISFAHAN and Urmia Stations at five different soil depths using the Bi-linear (BL model) for training and testing and reported RMSE values of 1.72 and 1.75 for Isfahan and 1.43 and 1.33 for Urmia Station, respectively (16). The same author reported a RMSE value of 1.45 and 1.49 for Anfis model for ST at 5 cm depth in a different study reported a range of RMSE value for various input parameters and reported a RMSE value of 2.19 for training and 3.47 for testing data set using the input variable temperature, dew, evapo-transpiration, radiation, WV and humidity at 10 cm depth (17). ST was modelled based on the look back period (18). The author used Bagging Regressor, Random Forest Regressor, Ada Boost Regressor using air temperature, WV and RH and reported that the variable air temperature has the highest variable importance weights followed by WV and RH. The ST was modelled using Artificial neural network model over the training and test modelling phases and reported a RMSE value of 4.40 and 4.88 respectively at 5 cm depth (19). The ST was modelled using GANs-LSTM model and predicted the ST at 5 cm depth at Change Bai Mountain and HaiBai Stations and reported an estimated RMSE value of 2.71 and 2.40 for respective stations (20). Another study predicted the ST using Convolutional Neural Network and reported a range of RMSE value for using different sub models which ranges from 0.46 to 0.74. (22) estimated the ST using Novel Genetic based negative correlation learning algorithm and reported the ST prediction at 5 cm depth with a RMSE value of 1.98 and 2.07 for training and testing data respectively (21).

ST was modelled by (23) using climatic parameters and employed various models viz., Gaussian Process regression, M5P model, Random Forest and Multi-layer perception model, with the aim to identify the best performing model. Previous study predicted the ST using Tree based hybrid data mining models (Gradient Boosted Trees (GBT), Decision Trees (DT), Hybrid (DT-GBT) found that GBT model performed well with the RMSE value of 3.12 (16). The ST was modelled using Gaussian Process Regression (GPR), Multilayer Perception Artificial Neural Network (MLPANN), Random Forest Regression (RFR), Support Vector Regression (SVR) and evaluated the performance based on RMSE and found that GPR model produced more accurate value at 5 cm soil depth with RMSE value of 1.81 (24). Previous study predicted the ST dynamics at hill slope using four machine learning algorithm and found that Extreme Gradient Boosting System (XGBOOST) performed better for hill slopes followed by Random Forest Regression, Multi-layer Perception and Support Vector Regression (25).

After training the model for the prediction of ST, response plots were generated for each input variables viz., data set number, time of observation, air temperature, RH, WV and sunshine duration. The response plots were generated along with residual bars and presented in Fig. 1. The above prediction was done after model validation without providing the corresponding soil temperature observation. The prediction was done using the trained model only. The residuals are the difference between actual and predicted values and standard error over the residuals were computed and plotted against target versus predicted values.

Fig. 2A shows the predicted response between observed and predicted ST after validating the model and Fig. 2B showed the corresponding information after testing the model. Both response plots obtained during validation and testing phases showed that a significant number of data points were centred on the 1: 1 line. It is also observed that there is a reasonable presence of outliers, which scattered away from the 1: 1 line. A similar trend was observed in these scattered plots obtained by testing the model. When comparing the two scattered plots, the response obtained during the test phase has slightly improved compared to the same observed during the data validation process.

The model performance was evaluated using residual plots over the true response, predicted response and the record number of input data set and depicted in Fig. 3. The residual plots depict the difference between the predicted response and true response.

Table 2. Evaluation of regression learner Ensemble Tree Bagging Algorithm using various input variable combinations for the prediction of soil temperature

Model	Predictor	RMSE (Validation)	MSE (Validation)	R Squared (Validation)	MAE (Validation)	MAE (Test)	MSE (Test)	RMSE (Test)	R Squared (Test)
M1	Time	3.40	11.54	0.17	2.85	2.87	11.71	3.42	0.16
M2	Time+T	2.74	7.50	0.46	2.12	2.11	7.44	2.73	0.47
M3	Time+T+RH	2.42	5.87	0.58	1.82	1.82	5.86	2.42	0.58
M4	Time+T+RH+WV	2.41	5.79	0.59	1.81	1.77	5.56	2.36	0.59
M5	Time+T+RH+WV+SS	2.03	4.13	0.70	1.46	1.41	4.02	2.01	0.72

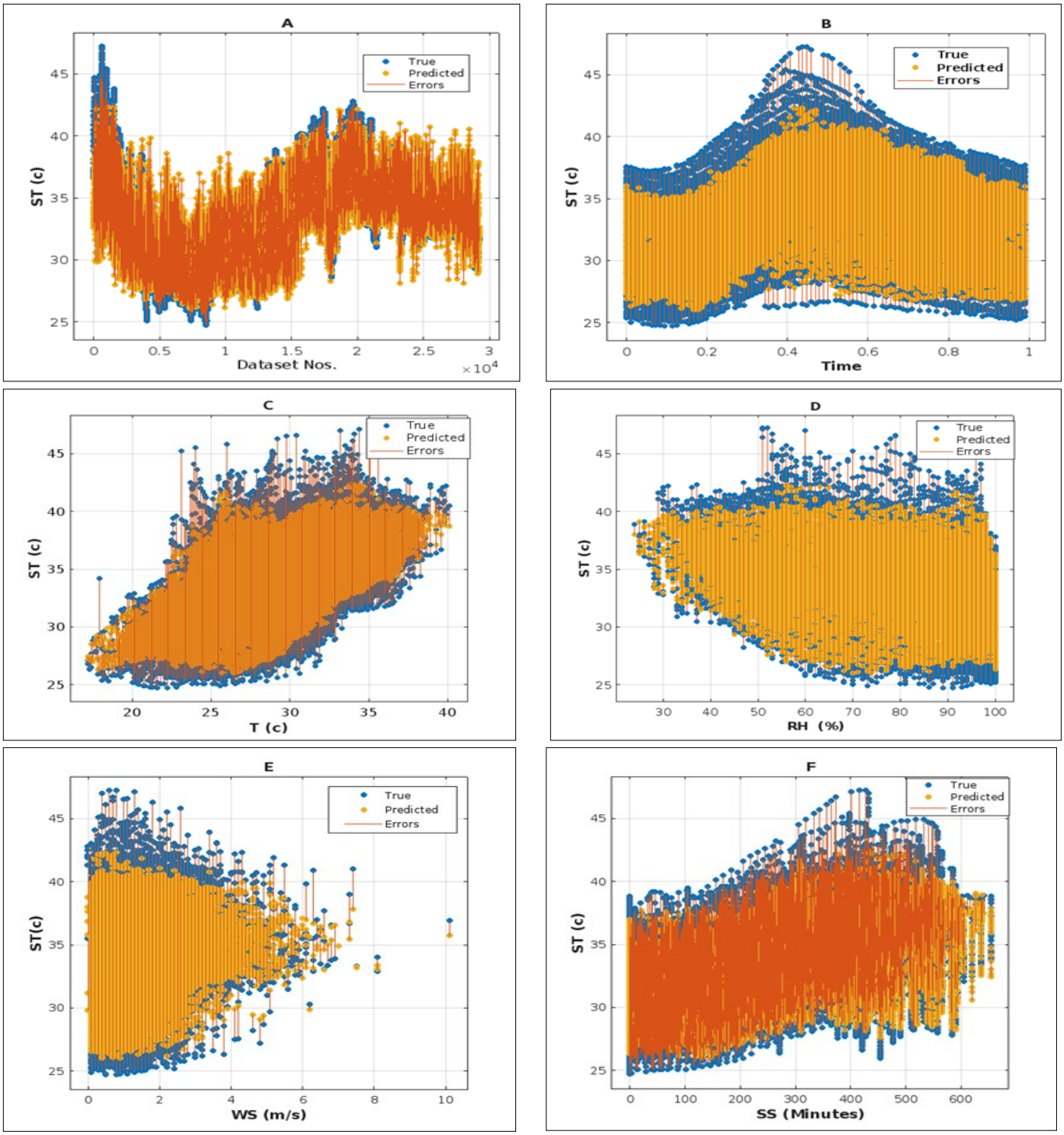

Fig. 1. Model predicted response for soil temperature, ST (°C) for different model input variables; A) Data set Nos, B) Time, C) Air temperature, T(°C), D) Relative humidity, RH (%), E) Wind Speed, WS (m/s) and F) Sun Shine duration, SS (Minutes).

The application of residual plots enhances model evaluation and guides subsequent modelling decisions, ultimately improving the reliability of predictions in studies like soil temperature assessments (26, 27). These plots provide insight into the differences between predicted and actual temperature values over a range of conditions, helping to identify patterns that might suggest improvements in model specifications. It is observed that most of the residuals are scattered around the value zero symmetrically. A clear pattern is observed in the residual plots for the true response, predicted response and record number against the residuals of ST.

Similarly, the residual plots for the predictor variables were generated and presented in Fig. 4. With respect to time variables most of the residuals over ST were centred around the zero value throughout the period (Fig. 4A). The highest residuals, which are away from the zero value ranges between 11 to -09. The residuals of ST vs T are depicted in Fig. 4B. Most of the residuals are centred around the desirable residual value of zero and the extreme residual values were ranges between 11 to -09. The highest data points which centred around zero value were observed at data ranges between 20 to 38 °C. With respect to RH, majority of the residuals over the ST were centred near the desired value of zero. However, there exists a trend pattern as whenever the RH is low, the residual against ST is also low. When RH increases, the residual values over ST deviate away from the desired value of zero in both positive and negative

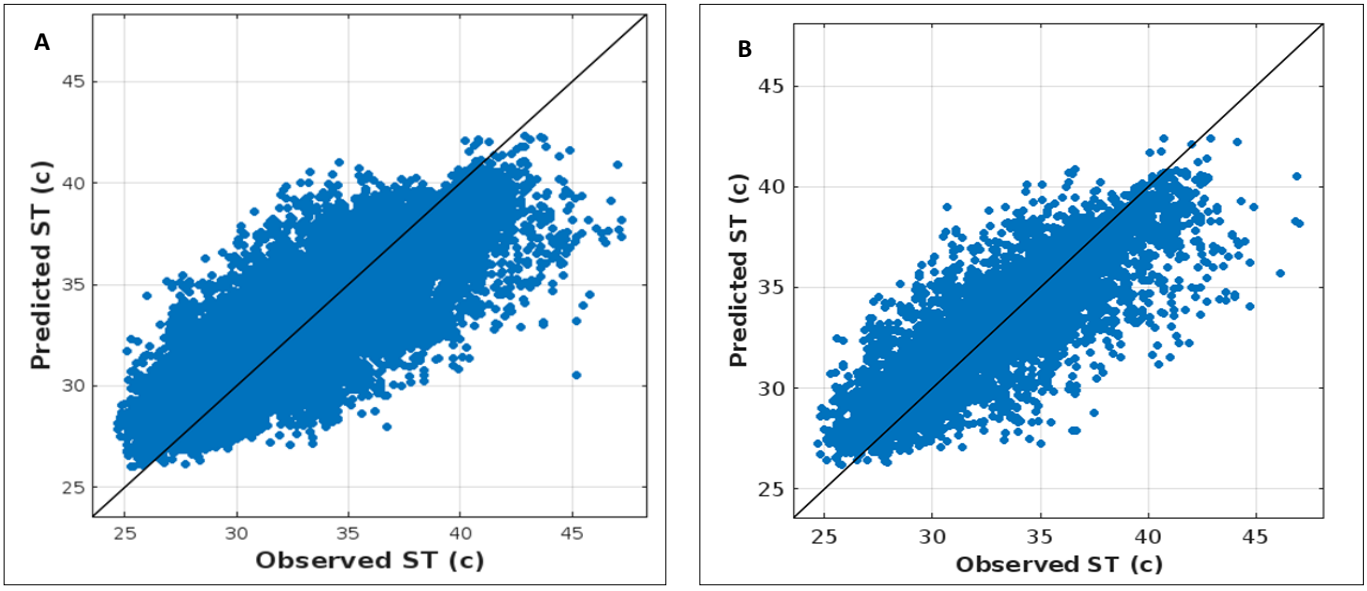

Fig. 2. Observed and predicted ST after data training and validation (A) and model testing (B).

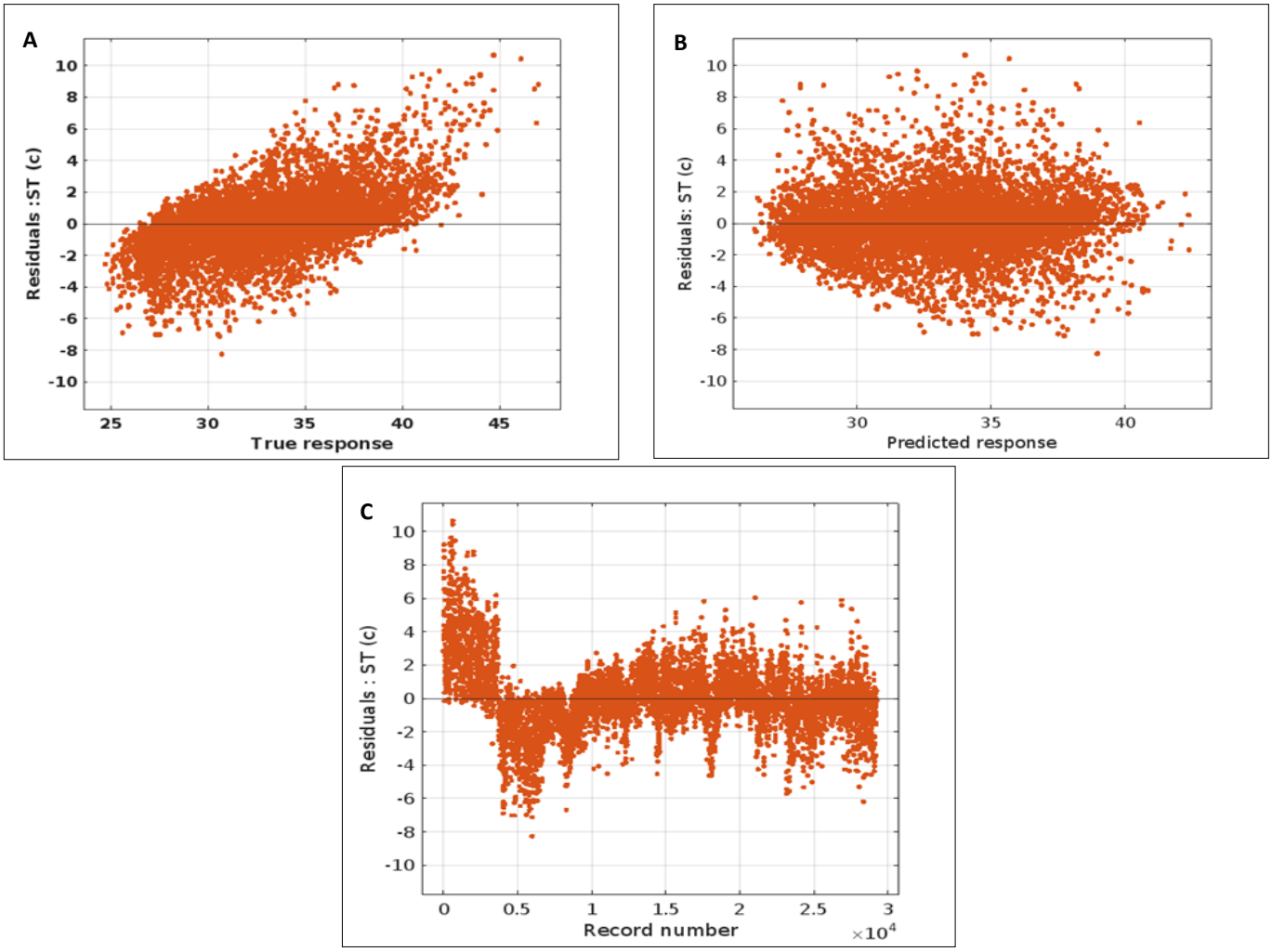

Fig. 3. Residual plot for predicted ST Vs True response (A), Predicted response (B) and Record number (C).

residual values (Fig. 4C). The residuals of WV over ST are depicted in Fig. 4D. It is observed that most of the residuals were centred around the desirable value of zero. Whenever the WV is low, the residuals of ST are very high, which ranges from 11 to -09.

The residual values of sunshine duration over ST are depicted in Fig. 4E. It is observed that much of the data set were cantered around the desired residual value of zero. The observed residual value was ranged from 11 to -09.

This approach aligns with findings from related agricultural research, such as the positive effects of organic amendments on soil properties and productivity, like those observed with sugar beet residual and vinasse treatments (28). Furthermore, similar methodologies used in herbicide bio-efficacy studies underscore the importance of residual analysis in understanding soil dynamics and microbial interactions, further enhancing the robustness of soil temperature predictions (29).

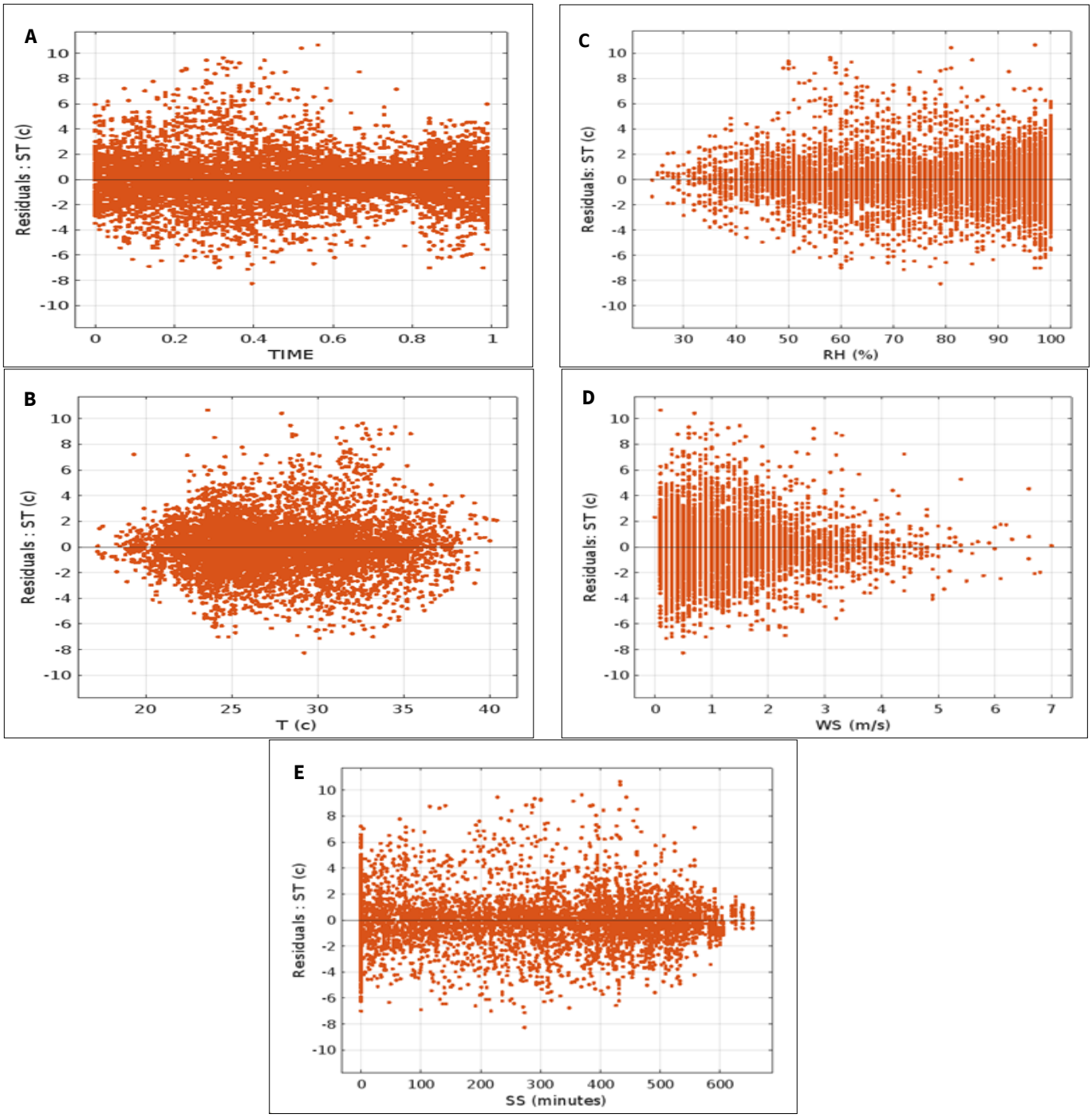

Fig. 4. Predictor residual plots for individual input variables. (A) Time, (B) Air temperature T(°C), (C) Relative humidity RH (%), (D) Wind Speed WS (m/s) and (E) Sun Shine duration, SS (Minutes).

After training the model, using the training and validation data set SHAP (Shapley Additive Explanation) importance was computed based on the mean of absolute Shapley value against the predictor and presented in Fig. 5A. For regression models, Shapley values can be used on a set of query points to assess the influence of individual predictors on model predictions. At each query point, the Shapley value for a predictor quantifies the deviation in the predicted response from the average prediction. The sign of this value indicates whether the effect is positive or negative, while its absolute value reflects the magnitude of the impact. Consequently, Shapley values near zero suggest that the specific predictor has minimal influence on the model's predictions for that query point. It is observed that the input variable air temperature has the highest absolute Shapley value, followed by time, SS, RH and WV.

The SHAP values, which explain the output of the trained machine learning model, were calculated for each input variable and its impact on the model output was plotted and presented in Fig. 5B. The SHAP value explains how each input variable affects the final prediction value. It also explains the interaction between different input variables.

Shapley values are based on game theory and assign a value for each input variable. It is observed that the Shapley value for each input variable ranges from positive to negative values. The positive value impacts the prediction of the model positively and vice versa. The air temperature has a wide range of positive and negative values, followed by a narrow range for the input variables time, sunshine duration, RH and WV.

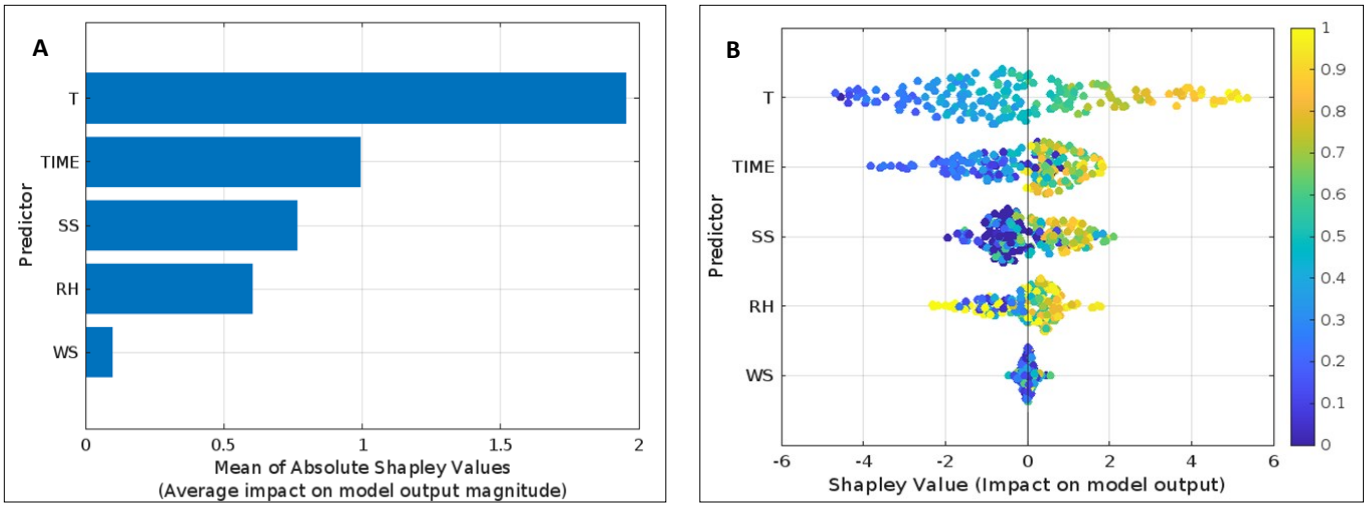

Fig. 5. Shapley importance value for the various input variables (A), Shapley summery swarm plot for different input variables for the prediction of soil temperature (B).

Shapley values are used to interpret the impact of each individual predictor on model prediction. The sign of the Shapley value is used to identify the direction of the Shapley value deviation from zero and the absolute Shapley value indicates its magnitude. Fig. 6A depicts the Shapley dependency of the variable over its Shapley value. The value deviates from zero on both the positive and negative sides, with ranges from 2.2 to -3.8. This value exhibits diurnal variation over 24 hr time. Fig. 6B depicts the Shapley dependency of air temperature, which has both positive and negative Shapley values. Only a very small number of data points had a value near zero. When the air temperature is less than 28 °C, the Shapley value tends to decrease on the negative side. If the value is more than 28 °C, the Shapley value tends to increase its magnitude positively. Fig. 6C shows the Shapley dependency of RH. The Shapley value scatters over both positive and negative sides. When the RH is less than 65 %, Shapley value tends to decrease towards negative value and vice versa. When the RH reaches very high levels beyond 95 %, the Shapley value returns to the negative side.

Fig. 6D depicts Shapley dependency of WV. The values range between positive and negative sides and it always has smaller Shapley values across different WV. Fig. 6E depicts the Shapley value of sunshine duration. It was observed that sunshine values scattered both positive and negative values. When accumulated sunshine duration increases, the value of Shapley become positive and its magnitude increases. A similar observation was reported by (30) while testing the predicted ST using Explainable AI (Ex AI) and LSTM and integrated that the model performance Shapley Additive Explanation (SHAP), Permutation Importance (PI) and Partial Dependence Plots (PPP) and stated that the temperature of air at 3 m above surface has a greater influence on ST and SHAPLEY value varies significantly over different seasons.

Conclusion

Surface ST, particularly at a 5 cm depth, plays an important role in agricultural decision-making, influencing actions such as time of sowing, water management and precision agriculture practices. Predicting the ST at 15 min interval provides essential support for timely decision making. In this study, we applied machine learning approach to predict surface ST using above ground meteorological parameters viz., air temperature, RH, WV, sunshine duration and the respective time scale of 15 min interval. This study employed six machine learning algorithm for the prediction of surface ST viz., Ensemble, Gaussian Processor Regression, Support Vector Model, Tree, Neural Network and Kernel based algorithm for the prediction of ST at 5 cm depth at the time scale of 15 min interval and found that the superiority of Ensemble (BT) in the prediction of ST with a RMSE value of 2.04 and 1.99 respectively for the validation and test data sets. Further, it was found that temperature plays a critical role in the prediction of ST as indicated by the absolute Shapley value, followed by time, sunshine duration, RH and WV. The impact of different input variables on the model output followed the same trend as the mean absolute Shapley value of these variables. The Shapley dependency analysis revealed a distinct pattern, which is consistent with findings reported in recent studies.

The analysis revealed that the Regression Learner-Based Ensemble Tree Bagging Machine Learning Algorithm demonstrates a significant capacity for accurately predicting soil temperature by utilizing weather data. The findings indicate that this algorithm not only outperformed traditional predictive models but also exhibited greater resilience to variability in the input data, suggesting its potential utility in diverse agricultural and environmental contexts. Furthermore, this research highlights the importance of integrating advanced machine learning techniques into environmental studies, which could enhance predictive accuracy and inform better decision-making in agricultural practices. Future research should explore the algorithms adaptability across different climates and soil types to ascertain its generalizability. Additionally, the implications for practice include the necessity for ongoing collaboration between data scientists and agronomists to refine the model further and develop user-friendly applications that farmers and environmental managers can readily incorporate into their planning processes.

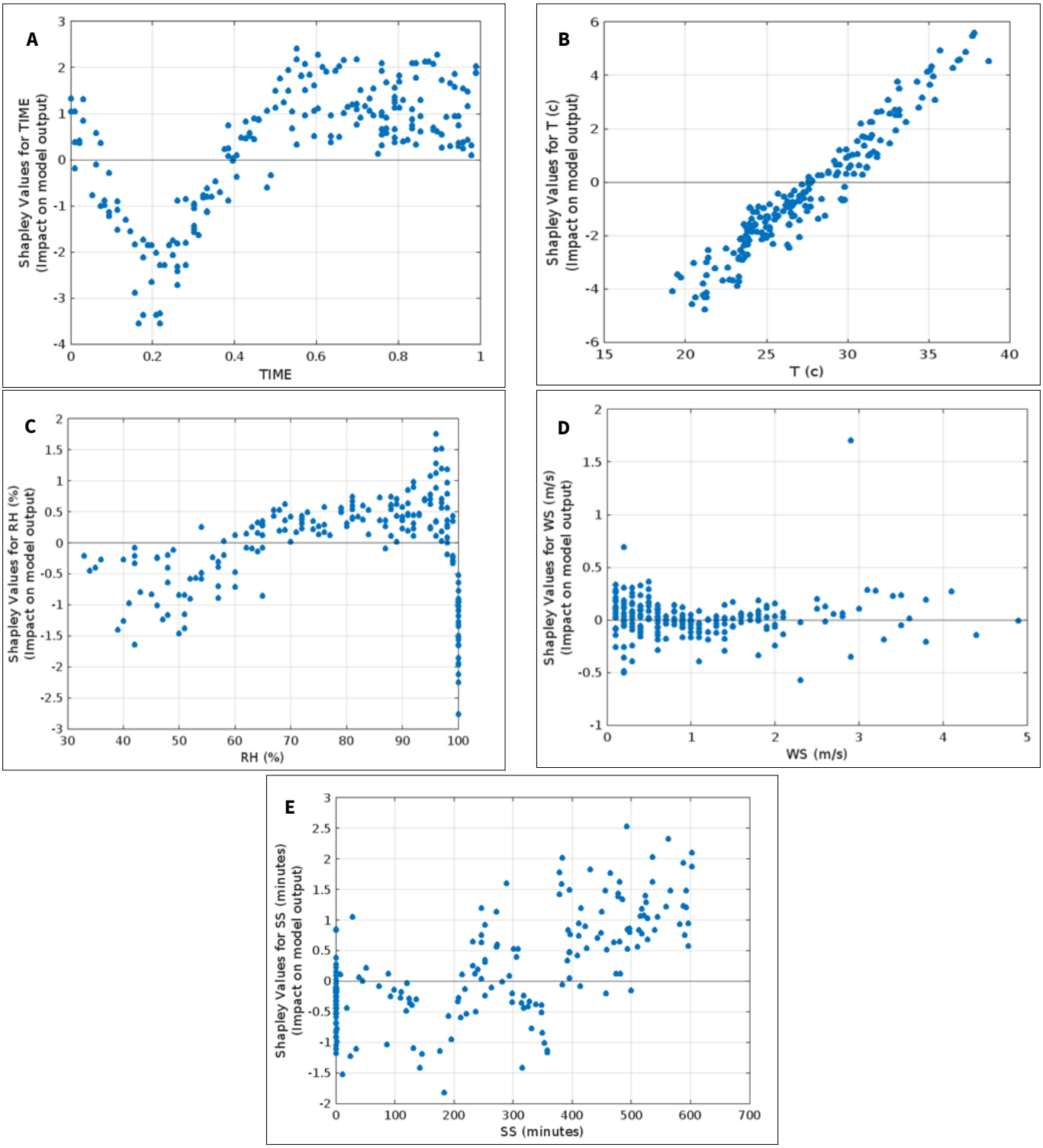

Fig. 6. Shapley dependency of different input variable over Shapley value (A-E).

Acknowledgements

The authors acknowledge the sample data source from Tamil Nadu Agricultural University ICAR-KVK, Vamban, Pudukkottai 622 303 and software facility used from Department of Physical Sciences & Information Technology, Agricultural Engineering College & Research Institute, Tamil Nadu Agricultural University, Coimbatore 641 003.

Authors' contributions

The concept of study was conceived by the first author AA and the author accessed the data, checked quality, processed and performed the machine learning algorithm. The corresponding author PT interpreted the results in a cohesive manner and scripted the manuscript in a systematic way. The co-authors NS, ST, KG and PA have contributed to result interpretation and manuscript writing. All authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest: Authors do not have any conflict of interests to declare.

Ethical issues: None

References

1. Katterer T, Andren O. Predicting daily soil temperature profiles in arable soils in cold temperate regions from air temperature and leaf area index. Acta Agric Scand Sec. B Soil Plant Sci. 2009;59:77–86
2. Pregitzer KS, King JS. Effects of soil temperature on nutrient uptake. In: BassiriRadH.(Ed.). Nutrient Acquisition by Plants: An Ecological Perspective. 2005; Springer, pp.277–310. https://doi.org/10.1007/3-540-27675-0_10
3. Singh RB. Environmental consequences of agricultural development: a case study from the Green Revolution state of Haryana, India. Agriculture, Ecosystems & Environment. 2000; 82 (1–3):97–103. https://doi.org/10.1016/S0167-8809(00)00219-X
4. Sun B, Zhang L, Yang L, Zhang F, Norse D, Zhu Z. Agricultural non-point source pollution in China: Causes and mitigation measures. Ambio. 2012; 41(4):370–79.
5. Hu G, Zhao L, Wu X, et al. An analytical model for estimating soil temperature profiles on the Qinghai-Tibet Plateau of China. J Arid Land. 2016;8:232–40. https://doi.org/10.1007/s40333-015-0058-4
6. Breiman L. Bagging Predictors. Machine Learning 26, 1996, pp. 123–140.
7. Breiman L. Random Forests. Machine Learning 45, 2001, pp. 5–32.
8. Breiman L. https://www.stat.berkeley.edu/~breiman/RandomForests/
9. Lundberg S, Lee SI. A unified approach to interpreting model predictions, arXiv - cs - AI 2017. https://doi.org/10.48550/arXiv.1705.07874
10. Lundberg SM, SI Lee. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.
11. Shapley L. A Value for n-Person Games. In: Kuhn, H. and Tucker, A., Eds., Contributions to the Theory of Games II, Princeton University Press, Princeton, 1953:307-17. https://doi.org/10.1515/9781400881970-018
12. Winter E, Chapter 53 the shapley value, in: Handbook of Game Theory with Economic Applications, 3, Elsevier, 2002, pp. 2025–54.
13. Kisi O, Sanikhani H, Cobaner M. Soil temperature modeling at different depths using neuro-fuzzy, neural network, and genetic programming techniques. Theor Appl Climatol. 2016;1–16. http://doi.org/10.1007/s00704-016-1810-1
14. Nahvi B, Habibi J, Mohammadi K, Shamshirband S, Al Razgan OS. Using self-adaptive evolutionary algorithm to improve the performance of an extreme learning machine for estimating soil temperature. Computers and Electronics in Agriculture. 2016;124:150–60. https://doi.org/10.1016/j.compag.2016.03.025
15. Feng Y, Cui N, Hao W, Gao L, Gong D. Estimation of soil temperature from meteorological data using different machine learning models. Geoderma. 2019;338:67–77. https://doi.org/10.1016/j.geoderma.2018.11.044
16. Mehdizadeh S, Ahmadi F, Kozekalani Sales A. Modelling daily soil temperature at different depths via the classical and hybrid models. Meteorological Applications. 2020;27:(4). https://doi.org/10.1002/met.1941
17. Alizamir M, Kisi O, Ahmed AN, Mert C, Fai CM, Kim S, et al. Advanced machine learning model for better prediction accuracy of soil temperature at different depths. PLoS ONE. 2020;15(4). https://doi.org/10.1371/journal.pone.0231055
18. Kovacevic T, Mrcela L, Mercep A, Kostanjcar Z. Impact of look-back period on soil temperature estimation using machine learning models. I2MTC 2020 - International Instrumentation and Measurement Technology Conference, Proceedings. 2020; https://doi.org/10.1109/I2MTC43012.2020.9128504
19. Bayatvarkeshi M, Bhagat SK, Mohammadi K, Kisi O, Farahani M, Hasani A, et al. Modeling soil temperature using air temperature features in diverse climatic conditions with complementary machine learning models. Computers and Electronics in Agriculture. 2021;185. https://doi.org/10.1016/j.compag.2021.106158
20. Li Q, Hao H, Zhao Y, Geng Q, Liu G, Zhang Y, et al. GANs-LSTM Model for soil temperature estimation from meteorological: A new approach. IEEE Access. 2020; 8:59427–43. https://doi.org/10.1109/ACCESS.2020.2982996
21. Hao H, Yu F, Li Q. Soil temperature prediction using convolutional neural network based on ensemble empirical mode decomposition. IEEE Access. 2021;9:4084–96. https://doi.org/10.1109/ACCESS.2020.3048028
22. Kazemi SMR, Bidgoli BM, Shamshirband S, Karimi SM, Ghorbani MA, Chau KW, et al. Novel genetic-based negative correlation learning for estimating soil temperature. Engineering Applications of Computational Fluid Mechanics. 2018;12(1):506–16. https://doi.org/10.1080/19942060.2018.1463871
23. Sihag P, Esmaeilbeiki F, Singh B, Pandhiani SM. Model-based soil temperature estimation using climatic parameters: the case of Azerbaijan Province, Iran. Geology, Ecology, and Landscapes. 2020;4(3):203–15. https://doi.org/10.1080/24749508.2019.1610841
24. Han M, Zhang H, DeJonge KC, Comas LH, Trout TJ. Estimating maize water stress by standard deviation of canopy temperature in thermal imagery. Agricultural Water Management. 2016;177:400–09. https://doi.org/10.1016/j.agwat.2016.08.031
25. Aliva Nanda, Sumit Sen, Awshesh AS, Sudheer KP. Soil temperature dynamics at Hillslope Scale—field observation and machine learning-based approach. Water. 2020;12(3):713. https://doi.org/10.3390/w12030713
26. Veronika E, Sandrine B, Gerald AM, CA Senior, Björn S, Ronald JS, et al. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geoscientific model development. 2016;9:1937-58. https://doi.org/10.5194/gmd-9-1937-2016
27. Bobak S, Kevin S, Ziyu W, Ryan PA, Nando de F. Taking the human out of the loop: a review of Bayesian optimization. Proceedings of the IEEE. 2015;104:148-75. https://doi.org/10.1109/jproc.2015.2494218
28. Sary DH, Abd EL-Rahman ZM, Awad FA. Effects of sugar beet residual and vinasse on soil properties and wheat productivity under calcareous soil conditions. Journal of Soil Sciences and Agricultural Engineering. 2024;15(10):271-8. https://doi.org/10.21608/jssae.2024.314327.1245
29. Sravani P, Chovatia PK, Muchhadiya RM, Sakarvadia HL, Kachhadiya SP, Solanki RM, et al. Evaluating herbicides bio-efficacy on yield, quality parameters of sweet sorghum and their residual effects on soil microbial diversity. Journal of Experimental Agriculture International. 2024;46(10):9-24. https://doi.org/10.9734/jeai/2024/v46i102920
30. Geng Q, Wang L, Li Q. Soil temperature prediction based on explainable artificial intelligence and LSTM. Frontiers in Environmental Science. 2024;12:1426942. https://doi.org/10.3389/fenvs.2024.1426942