GMR

The Original

PREDICTION OF SOIL MOISTURE BY OPTIMIZABLE ENSEMBLE MACHINE LEARNING APPROACH USING SUB-HOURLY WEATHER PARAMETERS

S Anbuselvan[1], A Alagesan[2*], D Rajakumar[3], J Balamurugan[4], B Karthikeyan[5]

[1]Department of Agronomy, V.O. Chidambaranar Agricultural College and Research Institute, Tamil Nadu Agricultural University, Killikulam-628 252, Tamil Nadu, India, arjunananbuselvan@gmail.com, (https://orcid.org/0009-0004-2521-0693)
[2*]Department of crop management, Dr. M.S. Swaminathan Agricultural College and Research Institute, Tamil Nadu Agricultural University, Eachangkottai-614 902, Tamil Nadu, India, alagesan@tnau.ac.in, (https://orcid.org/0000-0002-2698-7187)
[3]Department of Agronomy, V.O. Chidambaranar Agricultural College and Research Institute, Tamil Nadu Agricultural University, Killikulam-628 252, Tamil Nadu, India, rajakumar.d@tnau.ac.in, (https://orcid.org/0000-0002-1085-1322)
[4]ICAR-Krishi Vigyan Kendra, Tamil Nadu Agricultural University, Pudukottai-622 303, Tamil Nadu, India, jbalamurugan73@yahoo.co.in, (https://orcid.org/0000-0002-3602-6670)
[5]Department of crop management, Agricultural College and Research Institute, Tamil Nadu Agricultural University, Chettinad, Sivagangai-630 102, Tamil Nadu, India, bkeyan473@gmail.com, (https://orcid.org/0000-0002-0853-1581)
***Corresponding author:** Dr. A Alagesan, Email: alagesan@tnau.ac.in, ORCID: (https://orcid.org/0000-0002-2698-7187)

ABSTRACT
Soil moisture influences seed germination, nutrient transport, microbial activity and crop productivity and its accurate prediction is becoming more important as rainfall patterns grow more irregular and temperatures rise. This study evaluated regression-based machine learning models for predicting soil moisture from meteorological data, including air temperature, relative humidity, wind speed, sunshine duration and soil temperature, recorded at an Automatic Weather Station at 15-minute intervals between October 2021 and November 2023. Twenty-eight sub-models spanning eight machine learning categories were compared using MAE, MSE, RMSE, RSE and R^2. An optimizable ensemble model tuned with grid search gave the best validation performance, with an MAE of 1.9756, MSE of 8.4283, RMSE of 2.9031 and R^2 of 0.83 and test performance of MAE 1.8240, MSE 7.5353, RMSE 2.7451 and R^2 of 0.85. Permutation importance and SHAP (Shapley Additive Explanations) analyses both ranked relative humidity, soil temperature and time of observation as the most influential predictors and SHAP dependence plots showed relationships between these variables and the predicted soil moisture that were consistent with known evaporative and diurnal processes. The model achieved this performance using only above-ground meteorological inputs, without soil moisture as a predictor. These findings indicate that an optimized ensemble model can predict soil moisture from routinely measured weather variables with reasonable accuracy and interpretability, providing a potential basis for future irrigation-management and drought-monitoring applications.

KEYWORDS: Grid search; Machine learning; Optimizable ensemble; Precision irrigation; SHAP analysis; Soil moisture prediction

INTRODUCTION
Soil moisture is a central variable in agroecosystems: it governs seed germination, nutrient transport, microbial activity, root water uptake and land–atmosphere interactions (Seneviratne *et al.,*2010; Dorigo *et al.,*2017). Climate change is intensifying evapotranspiration and increasing the variability of rainfall, disrupting the natural soil water balance and leading to either waterlogging or moisture stress that reduces crop performance (Pokhrel *et al.,*2021). Effective irrigation scheduling, rainwater harvesting, mulching and conservation agriculture all depend on timely and reliable soil moisture information (Adeyemi *et al.,*2018). Conventional measurement methods, such as gravimetric analysis and in-situ sensors, provide accurate estimates but are labour-intensive and difficult to scale for continuous, large-area monitoring (Shawon *et al.,*2025).
Machine learning and regression-based approaches in particular, offers an alternative by learning nonlinear relationships between soil moisture and readily available meteorological variables (Xu *et al.,*2020; Fang *et al.,*2017). Algorithms such as multiple linear regression, regression trees, random forests, gradient boosting and support vector regression have been applied to forecast soil moisture from inputs including air temperature, relative humidity, rainfall, solar radiation, wind speed and vapour pressure deficit (Breiman, 2001; Chen & Guestrin, 2016; Ali *et al.,*2015; Cai *et al.,*2019; Nguyen *et al.,*2022). Ensemble methods, which combine multiple weak learners, have generally been reported to reduce overfitting and improve prediction accuracy relative to single-model approaches (Ren *et al.,*2023; Houben *et al.,*2025). Model performance in this literature is typically assessed using MAE, MSE, RMSE, RSE and R^2 (Krause *et al.,*2005) and interpretability methods such as permutation importance and SHAP (Shapley Additive Explanations) are increasingly used to examine the contribution of individual predictors and support the physical interpretation of model outputs (Fisher *et al.,*2019; Lundberg & Lee, 2017).

Despite this body of work, comparative evaluation of optimizable ensemble regression learners for soil moisture prediction using only routinely measured, above-ground weather variables remain limited and few studies combine systematic hyperparameter optimization with both permutation-based and SHAP-based interpretation on sub-hourly data. This study addresses that gap by developing and evaluating an optimizable ensemble regression model for predicting soil moisture from sub-hourly meteorological variables and by interpreting its predictions using permutation importance and SHAP analysis.

MATERIALS AND METHODS

Weather data were obtained from an Automatic Weather Station (AWS) located at the ICAR-Krishi Vigyan Kendra, Tamil Nadu Agricultural University, Pudukottai, Tamil Nadu, India (10.459242° N, 78.7027° E) and comprised soil moisture, air temperature, relative humidity, wind speed, sunshine duration and soil temperature, each recorded at 15-minute intervals. After data cleaning, each variable had 29,229 observations (175,374 data points overall), covering the period from 08 October 2021 to 03 November 2023 and the descriptive statistics for all variables are given in Table 1. For the purpose of model development, 70% of this data, that is 20,460 data points per variable, was used for training the model, while the rest 30% or 8,768 data points per variable, were used for testing the model.

Air temperature, relative humidity, wind speed, sunshine duration, soil temperature and time of observation were used as predictor variables, with soil moisture as the response variable. The time predictor was derived from the 24-hour recording cycle: because observations were logged every 15 minutes (96 per day), time of day was rescaled to a continuous 0 to 1 range, so that each 15-minute step corresponded to an increment of 1/96 (~0.0104). The soil-moisture values used as the response variable are the sensor-derived soil-moisture percentages recorded by the AWS; Units of measurement were °C for temperature, % for relative humidity and soil moisture, m/s for wind speed and minutes for sunshine duration. All preprocessing, development and predictions involved in the analysis process were performed using MATLAB R2026a.

Table 1. Descriptive Statistics of Weather Parameters and Soil Moisture

(n = 29,229 observations; Pudukottai, Tamil Nadu, India; Oct 2021 – Nov 2023)

Parameter	Range		Central Tendency		Dispersion		
	Min	Max	Mean	Median	SD	SE	CV (%)
Air Temperature (°C)	17.20	40.40	27.729	27.100	4.101	0.024	14.79
Relative Humidity (%)	24.00	100.00	77.668	82.000	19.295	0.113	24.84
Wind Speed (m/s)	0.00	10.10	1.161	0.900	0.989	0.006	85.17
Sunshine Duration (min)	0.00	654.00	230.021	219.000	202.262	1.183	87.93
Soil Temperature (°C)	24.70	47.20	33.367	33.600	3.728	0.022	11.17
Soil Moisture (%)	8.00	47.00	18.191	17.000	7.033	0.041	38.66

Note: SD = Standard Deviation; SE = Standard Error; CV = Coefficient of Variation (SD/Mean × 100). Min/Max = observed minimum and maximum values.

Machine Learning Model Development

Regression-based machine learning models were developed to predict soil moisture using the meteorological variables as predictor variables. A total of 28 regression models belonging to eight machine learning categories were evaluated (Table 2).

Table 2: Machine Learning Models and Their Corresponding Sub-Model

Main Model	Sub-Models
Linear Regression	Interaction Linear, Stepwise Linear, Linear, Robust Linear
Tree	Fine Tree, Medium Tree, Coarse Tree

Support Vector Machine (SVM)	Fine Gaussian SVM, Medium Gaussian SVM, Cubic SVM, Quadratic SVM, Coarse Gaussian SVM, Linear SVM
Efficient Linear	Efficient Linear SVM, Efficient Linear Least Squares
Gaussian Process Regression (GPR)	Exponential GPR, Rational Quadratic GPR, Matern 5/2 GPR, Squared Exponential GPR
Kernel	Least Squares Kernel Regression Kernel, SVM Kernel
Ensemble	Bagged Trees, Boosted Trees
Neural Network	Wide Neural Network, Trilayered Neural Network, Medium Neural Network, Bilayered Neural Network, Narrow Neural Network

Model performance was assessed using five statistical evaluation metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Relative Squared Error (RSE) and the coefficient of determination (R^2) (Table 3). The mathematical expressions of these metrics are presented in Equations (1)-(5). In this study, validation performance represents the cross-validated results from the training data, while test performance represents the results from the independent 30% test dataset. This terminology is used consistently throughout the Results and Tables.

Mean Absolute Error (MAE)

$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid$ (Eqn. 1)

Where:

- n= number of observations
- y_i= actual (true) value
- $\hat{y}_i$= predicted value
- $\mid y_i - \hat{y}_i \mid$= absolute error for each prediction

Mean Absolute Error (MAE) measures the average absolute difference between predicted and actual values. The lesser the MAE, the greater is the prediction accuracy, while a greater MAE shows that the model is less efficient.

1) Mean Squared Error (MSE)

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

Where:

- n= number of observations
- y_i= actual value
- $\hat{y}_i$= predicted value
- $(y_i - \hat{y}_i)^2$ = squared error

Mean Squared Error (MSE) can be understood as the average of the squared deviations between observed and estimated values, where larger errors carry greater weight. Low MSE denotes higher accuracy of predictions, while high MSE signifies poor model performance.

2) Root Mean Squared Error (RMSE)

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

Where:

- n= number of observations
- y_i= actual value
- $\hat{y}_i$= predicted value

Root Mean Squared Error (RMSE) is the square root of MSE and represents the average prediction error in the same unit as the original data. A lower RMSE indicates better model accuracy, while a higher RMSE reflects larger prediction errors.

3) Relative Squared Error (RSE)

$\text{RSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ (Eqn. 4)

Where:

- y_i= actual value
- $\hat{y}_i$= predicted value
- $\bar{y}$= mean of actual values
- n= number of observations

Relative Squared Error (RSE) compares the model's prediction error with a baseline model, usually the mean of actual values. It shows how well the model performs relative to simply predicting the average. An RSE of 0 indicates perfect prediction, RSE = 1 means the model is equal to the mean prediction, RSE < 1 indicates better performance and RSE > 1 indicates poorer performance than the baseline.

4) Coefficient of Determination (R^2)

$R^2 = (1 - RSE) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ (Eqn. 5)

Where:

- y_i= actual value
- $\hat{y}_i$= predicted value
- $\bar{y}$= mean of actual values
- n= number of observations

R^2 (Coefficient of Determination) measures how well a model explains the variability in the data. An R^2 value of 1 indicates perfect prediction, $R^2 = 0$ means the model performs as well as predicting the mean and $R^2 < 0$ indicates poor performance worse than the mean. It represents the proportion of variation in the actual data explained by the model.

Hyperparameter Optimization

The regression models were initially compared using the evaluation metrics described above. The most suitable regression model was subsequently subjected to hyperparameter optimization using Bayesian optimization, grid search and random search techniques. These optimization methods were employed to identify the combination of hyperparameters that produced the best predictive performance. The comparative performance of the optimized models is presented in Tables 4A and 4B, while the optimization workflow is illustrated in Figure 1.

Model Validation and Interpretation

After selecting the optimal regression model, a series of validation and interpretability analyses were performed to evaluate its predictive reliability and to understand the influence of individual predictor variables. Partial dependence plots, proposed by Friedman (2001), were generated to illustrate the influence of individual predictor variables on the predicted soil moisture while keeping all other variables constant. Residual analysis was conducted by calculating the residuals ($e_i = y_i - \hat{y}_i$) to assess the distribution of prediction errors and identify any systematic deviations from randomness. Predicted versus observed plots were also generated for both the training and testing datasets to evaluate the agreement between measured and predicted soil moisture values. Feature importance was quantified using the Permutation Importance method proposed by Breiman (2001), in which the importance of each predictor variable was determined by measuring the increase in prediction error after randomly permuting its values. To further improve model interpretability, Shapley Additive Explanations (SHAP) proposed by Lundberg and Lee (2017) were employed. The SHAP value for feature i was calculated as

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}[f(S \cup \{i\}) - f(S)]$$

where $\phi_i$represents the contribution of feature i, S denotes a subset of predictor variables and $f(S)$represents the model prediction based on that subset. SHAP summary plots were used to rank predictor variables according to their overall importance and to visualize the magnitude and direction of their contributions to soil moisture prediction. In addition, SHAP dependence plots were generated to examine the relationships between individual predictor variables and their influence on the model output.

RESULT AND DISCUSSION:

Model Comparison and Baseline Performance

Eight regression-based machine learning categories, expanded into 28 sub-models, were compared using the validation RMSE, MSE, R^2 and MAE reported in Table 3. Among these, the bagged-tree ensemble gave the lowest validation error (RMSE = 3.1641, MSE = 10.0116, MAE = 2.1937) and the highest R^2 (0.7991). Therefore, it was selected for further hyperparameter optimization. The next-best models were the Gaussian process regression variants (RMSE 3.22-3.32), while the linear-regression and efficient-linear models performed weakest (RMSE > 5.6). This ordering is broadly consistent with reports that ensemble tree methods tend to outperform linear and support-vector regression approaches for soil moisture prediction (Ren *et al.,*2023; Houben *et al.,*2025

Table 3: Comparative Evaluation of Machine Learning Models Based on Validation and Test Performance Metrics

Model Type	RMSE (Validation)	MSE (Validation)	RSquared (Validation)	MAE (Validation)	MAE (Test)	MSE (Test)	RMSE (Test)	RSquared (Test)
Ensemble - Bagged tree	3.164105	10.01156	0.799104	2.19366	2.095726	9.285941	3.047284	0.808848
Gaussian Process Regression - Rational Quadratic GPR	3.324934	11.05518	0.778162	2.310792	2.081196	9.456655	3.075168	0.805334
Gaussian Process Regression - Exponential GPR	3.242519	10.51393	0.789023	2.255112	2.173207	9.878159	3.142954	0.796658
SVM - Fine Gaussian SVM	3.465965	12.01292	0.758944	2.295921	2.237702	11.31964	3.364468	0.766985
Gaussian Process Regression - Matern 5/2 GPR	3.576627	12.79226	0.743305	2.565989	2.549855	12.44366	3.527558	0.743847
Gaussian Process Regression - Squared Exponential GPR	3.778145	14.27438	0.713565	2.736186	2.567381	12.60342	3.55013	0.740558
Neural Network - Wide Neural Network	3.761101	14.14588	0.716143	2.811259	2.759	13.77102	3.710932	0.716523
Tree - Fine Tree	3.845584	14.78852	0.703248	2.261557	2.177212	14.09655	3.754537	0.709822
Neural Network - Trilayered Neural Network	4.005642	16.04517	0.678031	2.953168	2.781364	14.3987	3.794562	0.703602
Tree - Medium Tree	3.987919	15.9035	0.680874	2.6572	2.542233	14.63197	3.825176	0.6988
Neural Network - Bilayered Neural Network	4.182792	17.49575	0.648923	3.107059	3.030473	16.38803	4.048213	0.662652
Tree - Coarse Tree	4.273865	18.26592	0.633469	3.073827	2.994573	17.33079	4.163026	0.643245
Neural Network - Medium Neural Network	4.107079	16.8681	0.661518	3.09498	3.162661	17.54844	4.189086	0.638764
SVM - Medium Gaussian SVM	4.379312	19.17837	0.615159	3.097201	3.047183	18.30687	4.278653	0.623152
Kernel - Least Squares Kernel Regression Kernel	4.509839	20.33865	0.591877	3.38709	3.311975	19.30609	4.39387	0.602583
Neural Network - Narrow Neural Network	4.536947	20.58389	0.586956	3.461813	3.315382	19.40661	4.405294	0.600514
Kernel - SVM Kernel	4.699449	22.08482	0.556837	3.404045	3.3792	21.28615	4.613692	0.561823
SVM - Cubic SVM	4.721796	22.29536	0.552612	3.496472	3.47103	21.62486	4.650253	0.554851
Ensemble - Boosted Trees	4.848502	23.50797	0.52828	3.890601	3.857109	23.23377	4.820142	0.521732
SVM - Quadratic SVM	5.027572	25.27648	0.492792	3.745138	3.718461	24.67277	4.967169	0.49211
SVM - Coarse Gaussian SVM	5.104103	26.05187	0.477233	3.832195	3.783562	25.22204	5.022155	0.480803
Linear Regression -Stepwise Linear Regression	5.623052	31.61872	0.365526	4.515343	4.480725	31.01819	5.569398	0.361489
Linear Regression - Interaction Linear	5.622898	31.61698	0.365561	4.514326	4.481463	31.02594	5.570093	0.361329
Linear Regression - Linear	5.996201	35.95443	0.278524	4.840489	4.816296	35.57192	5.96422	0.26775
Linear Regression - Robust Linear	6.004966	36.05961	0.276414	4.810721	4.782346	35.60343	5.966861	0.267101
SVM - Linear SVM	6.126058	37.52858	0.246937	4.766266	4.723453	36.77308	6.064081	0.243024
Efficient Linear - Efficient Linear SVM	6.548913	42.88826	0.139387	5.434009	5.329285	41.68292	6.456231	0.141955
Efficient Linear - Efficient Linear Least Squares	6.979424	48.71236	0.022518	5.931027	5.863979	47.53765	6.894755	0.021435

Hyperparameter Optimization

The selected ensemble model was then tuned using grid search, random search and Bayesian optimization (Fig. 1; Table 4A). Grid search systematically evaluated the predefined hyperparameter combinations and identified the best-performing configuration among the combinations tested, giving the lowest validation RMSE (2.9031) among the three methods, followed closely by random search (2.9076) and Bayesian optimization (2.9220). The differences between the three optimization methods were small in absolute terms (validation R^2 was 0.83 for all three), so grid search should be regarded as the best-performing option among those tested here rather than as substantially superior. Using grid search, the optimized model achieved a validation MAE of 1.9756, MSE of 8.4283, RMSE of 2.9031 and R^2 of 0.83 and a test MAE of 1.8240, MSE of 7.5353, RMSE of 2.7451 and R^2 of 0.85 (Table 4A). For context, Teshome et al. (2024) reported test R^2 values of 0.80-0.87 for similar ensemble weather-based soil moisture models and Mao et al. (2024) reported RMSE values of 3.2-4.5 using various deep learning techniques; the performance obtained here falls within this reported range.

Table 4B additionally reports feature-selection scores (MRMR, F-test and RReliefF) for each predictor across the three optimization methods. Several predictors show an infinite (∞) F-test score. This is a known property of the univariate F-test, an infinite F-statistic can occur when the within-group variance for a predictor is zero, while a very small within-group variance relative to the between-group variance can produce an extremely large finite F-statistic. Thus, an infinite F-test score does not necessarily indicate that the affected predictors are more important than those with finite scores. The predictor ranking discussed in this study is therefore based on the permutation-importance and SHAP results described below, rather than on the F-test scores alone.

Table 4: Comparative Analysis of Model Performance and Feature Importance Across Optimization Methods

4A: Model Performance and Efficiency

Category	**Parameter**	**Grid Search**	**Random Search**	**Bayesian Optimization**
Validation Performance	RMSE	2.9031	2.9076	2.9220
	R^2	0.83	0.83	0.83
	MSE	8.4283	8.4542	8.5380
	MAE	1.9756	1.9777	1.9918
	MAPE	13.0%	13.0%	13.1%
Test Performance	RMSE	2.7451	2.7460	2.7662
	R^2	0.85	0.85	0.85
	MSE	7.5353	7.5404	7.6516
	MAE	1.8240	1.8245	1.8401
	MAPE	12.0%	12.0%	12.1%
Efficiency	Prediction Speed (obs/sec)	~11000	~19000	~5100
	Training Time (sec)	352.83	244.34	261.45
	Model Size (Compact)	155 MB	95 MB	353 MB
	Model Size (Coder)	61 MB	38 MB	139 MB
Hyperparameters	Ensemble Method	Bag	Bag	Bag
	Minimum Leaf Size	1	1	2
	Number of Learners	210	129	500
	Predictors to Sample	6	6	6
Search Space	Ensemble Methods	Bag, LSBoost	Bag, LSBoost	Bag, LSBoost
	Learners Range	10-500	10-500	10-500
	Learning Rate	0.001-1	0.001-1	0.001-1
	Leaf Size Range	1-10230	1-10230	1-10230
	Predictors Range	1-6	1-6	1-6

4B: Feature Importance Across Selection Methods

Feature	MRMR (Grid search)	F-Test (Grid search)	RReliefF (Grid search)	MRMR (Random search)	F-Test (Random search)	RReliefF (Random search)	MRMR (Bayesian optimization)	F-Test (Bayesian optimization)	RReliefF (Bayesian optimization)
ST	0.2712	∞	0.0060	0.2712	∞	0.0060	0.2712	∞	0.0060
T	0.0920	∞	0.0034	0.0920	∞	0.0034	0.0920	∞	0.0034
SS	0.4259	436.6615	0.0032	0.4259	436.6615	0.0032	0.4259	436.6615	0.0032
RH	0.0983	∞	0.0014	0.0983	∞	0.0014	0.0983	∞	0.0014
WS	0.1057	602.7559	-0.0035	0.1057	602.7559	-0.0035	0.1057	602.7559	-0.0035
TIME	0.0036	0.1314	-0.0014	0.0036	0.1314	-0.0014	0.0036	0.1314	-0.0014

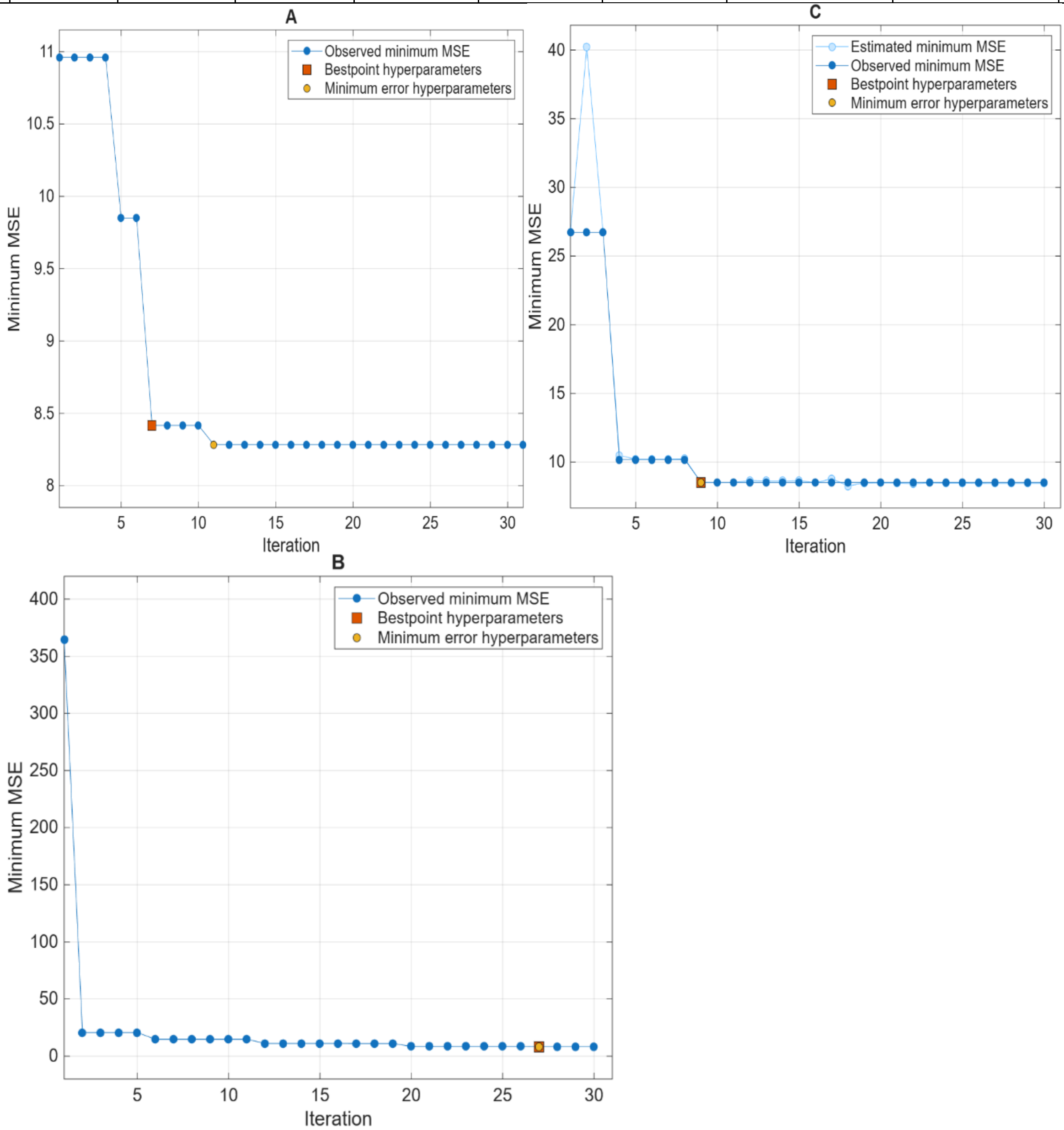


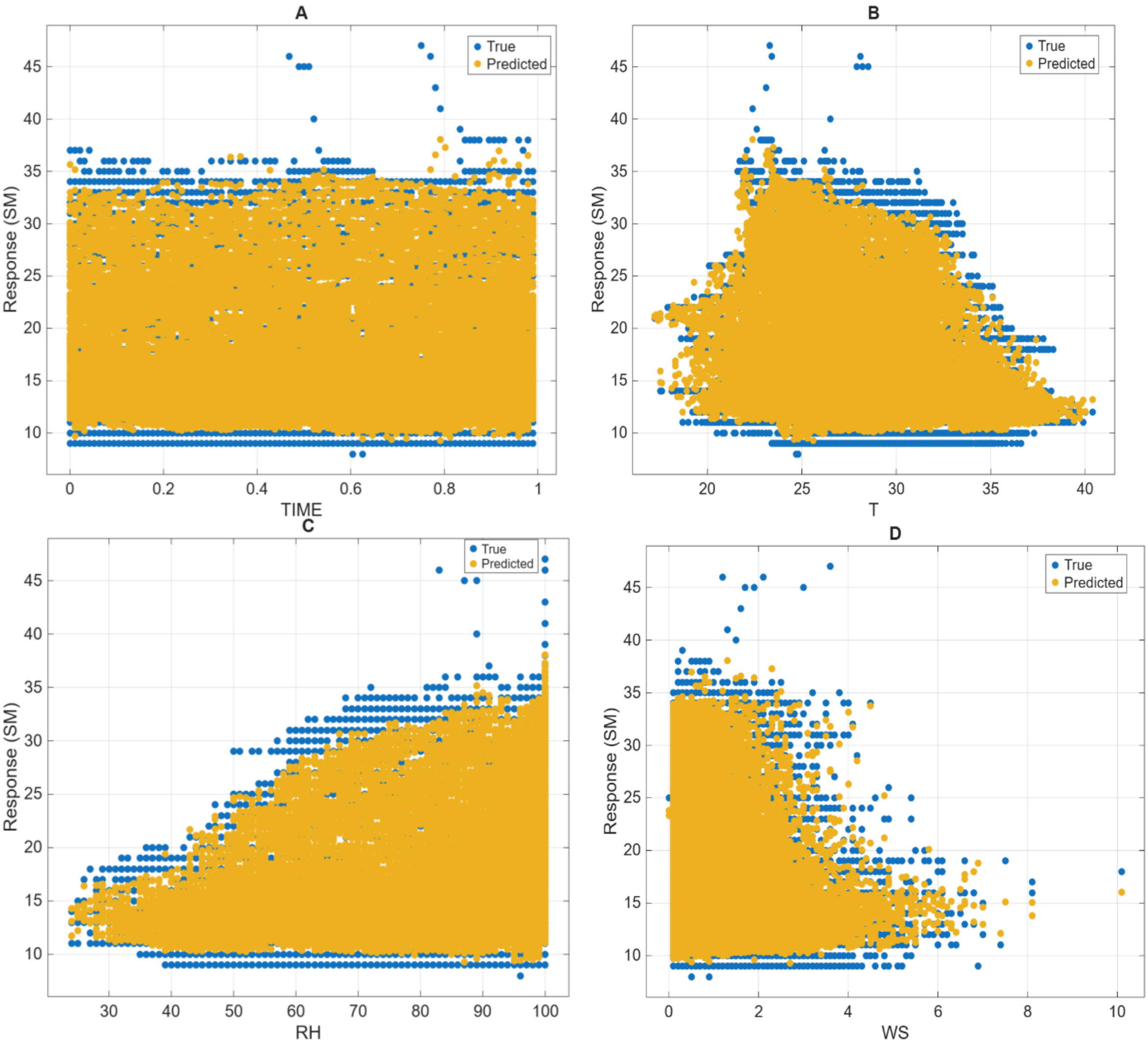
Fig. 1. Comparison of Mean Squared Error (MSE) values for the optimizable ensemble model using different hyperparameter tuning methods. (A) Grid Search, (B) Random Search and (C) Bayesian Optimization

Partial Dependence Analysis

Partial dependence plots (Fig. 2), following the approach of Friedman (2001), show how the fitted model's predictions change with each predictor individually, holding the others constant. The association between higher relative humidity and higher predicted soil moisture is consistent with the expected relationship between atmospheric humidity, vapour pressure deficit and evaporative demand: a narrower vapour pressure deficit is generally linked to lower evaporative water loss. The reverse pattern for soil temperature, where higher values were associated with lower predicted soil moisture, is consistent with faster evaporation at higher soil temperatures. These are model-derived associations rather than direct evidence of causal mechanisms. The model was evaluated using only above-ground weather inputs, without soil moisture supplied as a predictor, which is relevant for settings where direct soil measurements are difficult to obtain (Basir *et al.,*2024).

Predicted versus Actual Soil Moisture Values

Predicted-versus-actual soil moisture values (Fig. 3A, B) clustered closely around the 1:1 line in both the validation and test sets, consistent with the evaluation approach of Hastie et al. (2009). A small number of points deviated more noticeably, including an over-prediction at (31, 37.31) and an under-prediction at (19, 10.01) in the validation set and mild under- and over-predictions at (26, 22.94) and (8, 9.11) in the test set. This close agreement between predicted and observed values is consistent with the ensemble model capturing much of the nonlinear relationship between the meteorological predictors and soil moisture (Fang *et al.,*2017; Mao *et al.,*2024).



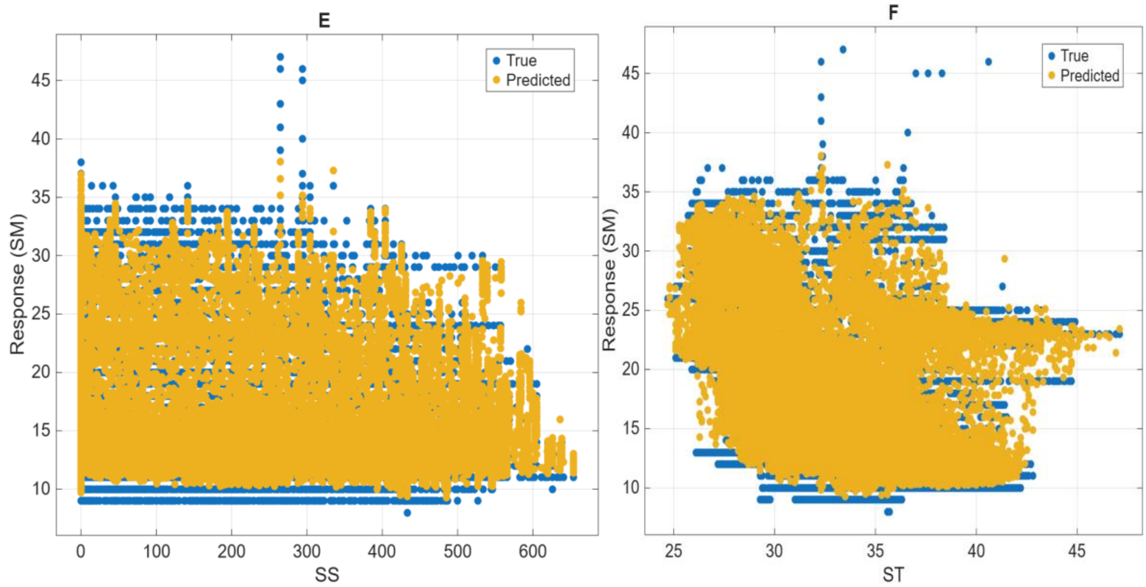


Fig. 2. Response plots of the optimizable ensemble model showing the influence of different predictor variables. (A) Time, (B) Air Temperature, (C) Relative Humidity (RH), (D) Wind Speed (WS), (E) Sunshine Duration and (F) Soil Temperature.

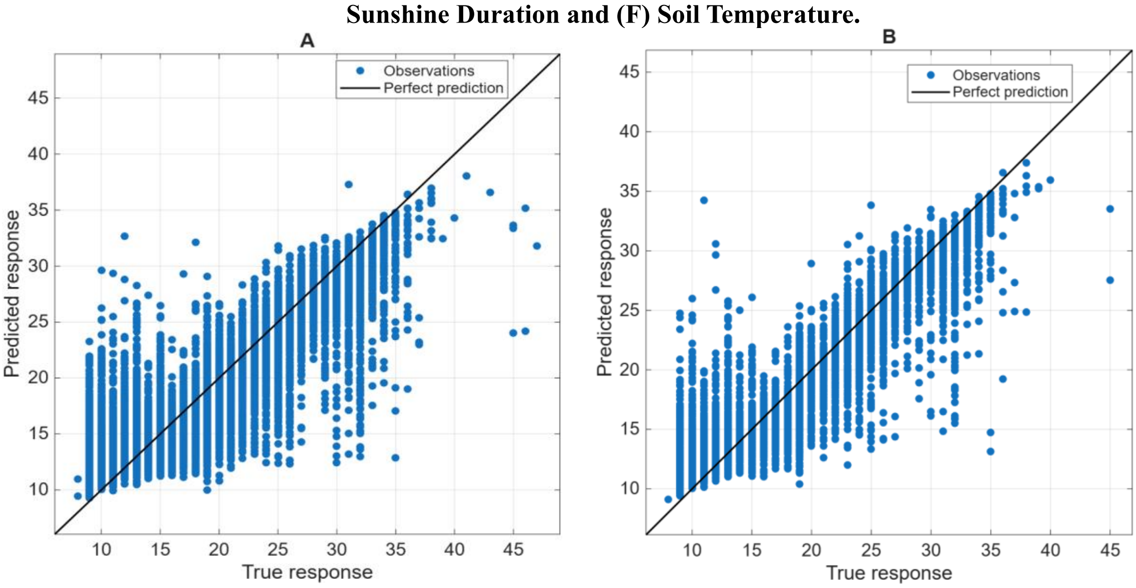


Fig. 3. Predicted vs. actual values of the optimizable ensemble model. (A) Validation dataset and (B) Testing dataset.

Residual Analysis

Residuals (predicted minus actual values) were examined from several angles (Fig. 4). Against the actual response, residuals stayed close to zero in both the validation and test sets, with only a few larger deviations, such as (12, −20.65) and (35, 22.14) in the validation set (Fig. 4A) and (11, −23.25) and (35, 21.87) in the test set (Fig. 4B). The residuals did not show a strong systematic pattern, indicating that no major unexplained structure was evident in the model errors, broadly consistent with the residual behaviour reported by Ren et al. (2023) and Houben et al. (2025).

Against the predicted response (Fig. 4C and 4D), a mild funnelling pattern appeared, meaning errors grew slightly larger within certain prediction ranges, a form of mild heteroscedasticity likely caused by nonlinearity among the environmental variables and an uneven spread of soil moisture samples (Mao *et al.,*2024) Such behaviour is common in environmental datasets because soil moisture exhibits greater variability under extremely dry or wet conditions than under moderate moisture conditions. Krause et al. (2005) stated that this type of residual behaviour is acceptable in hydrological modelling, provided that the residuals are randomly distributed and do not show a clear systematic trend.

Breaking residuals down by individual predictor (Fig. 5) gave further detail. Time of observation (Fig. 5A-5B) showed some of the largest residuals, e.g. (0.0625, -20.65) and (0.98, 22.14), which may reflect diurnal fluctuations in solar radiation, evapotranspiration and irrigation events that a single time variable cannot fully represent (Basir *et al.,*2024). Air temperature's residuals (Fig. 5C-5D) were consistent with an indirect effect operating through evapotranspiration rather than a direct one (Liu *et al.,*2025). Relative humidity (Fig. 5E-5F) produced its largest residuals near saturation (100% RH), where the relationship with soil moisture is harder for the model to capture (Lakra *et al.,*2025). Wind speed (Fig. 5G-5H) showed only modest scatter at low speeds (0.4-0.8 m/s), consistent with its comparatively low predictive contribution relative to humidity and temperature (Basir *et al.,*2024). Sunshine duration (Fig. 5I-5J) produced its largest residuals on sunless days, e.g. (0, 22.14) and (0, −20.65), when rainfall and humidity rather than sunshine were more likely associated with soil-moisture variation (Lakra *et al.,*2025). Soil temperature (Fig. 5K-5L) showed its largest residuals at the highest temperatures tested, around 32 °C, which may reflect more complex moisture-loss dynamics under hot conditions (Nagappan *et al.,*2020). Overall, residuals remained close to zero across predictors, which is consistent with reasonably stable model behaviour (Mao *et al.,*2024).

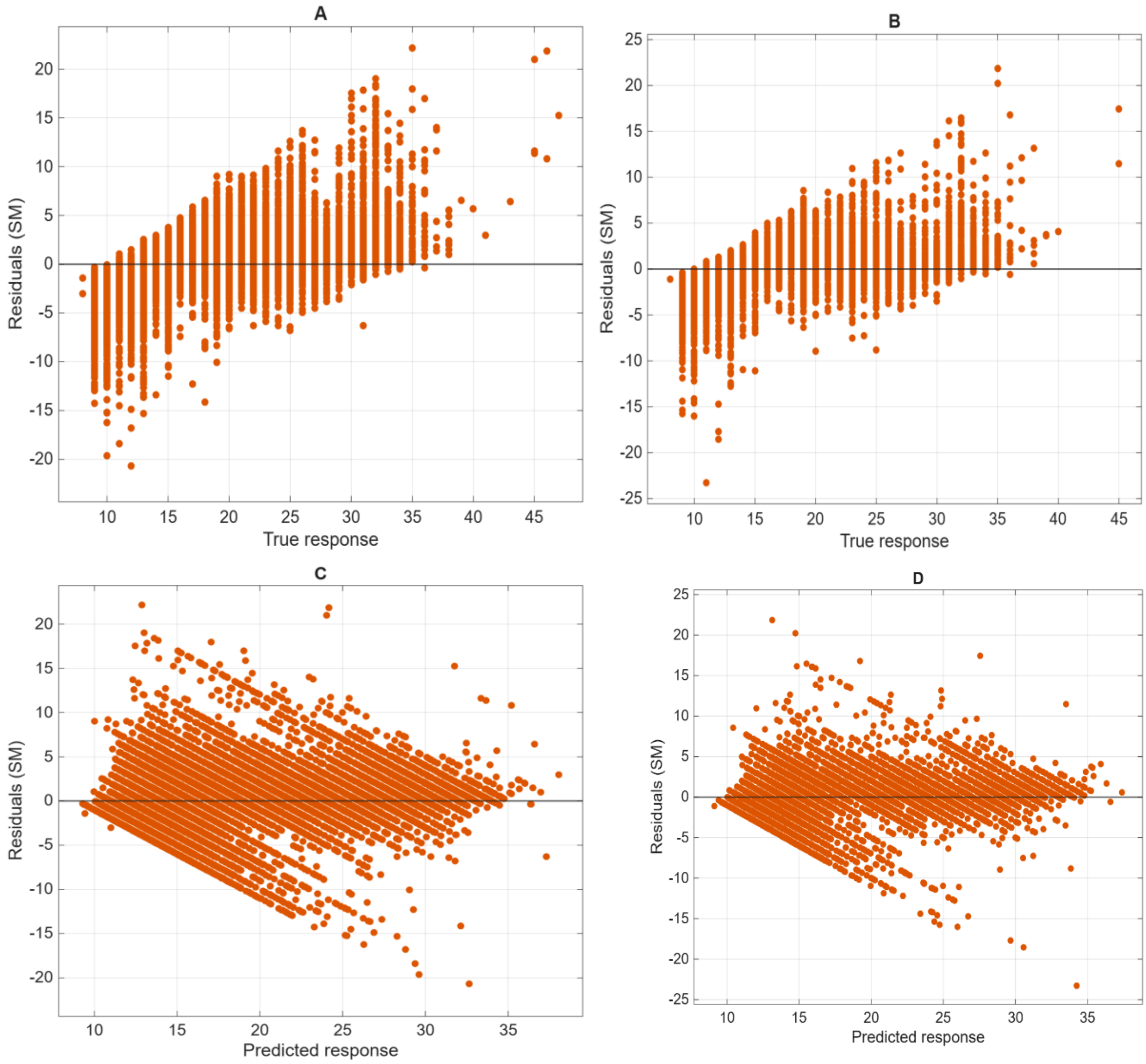
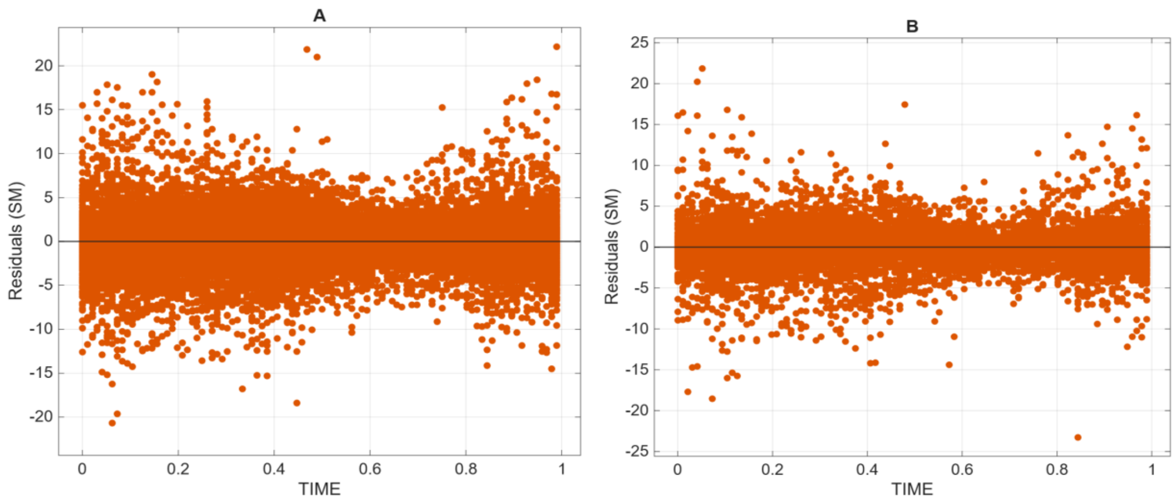
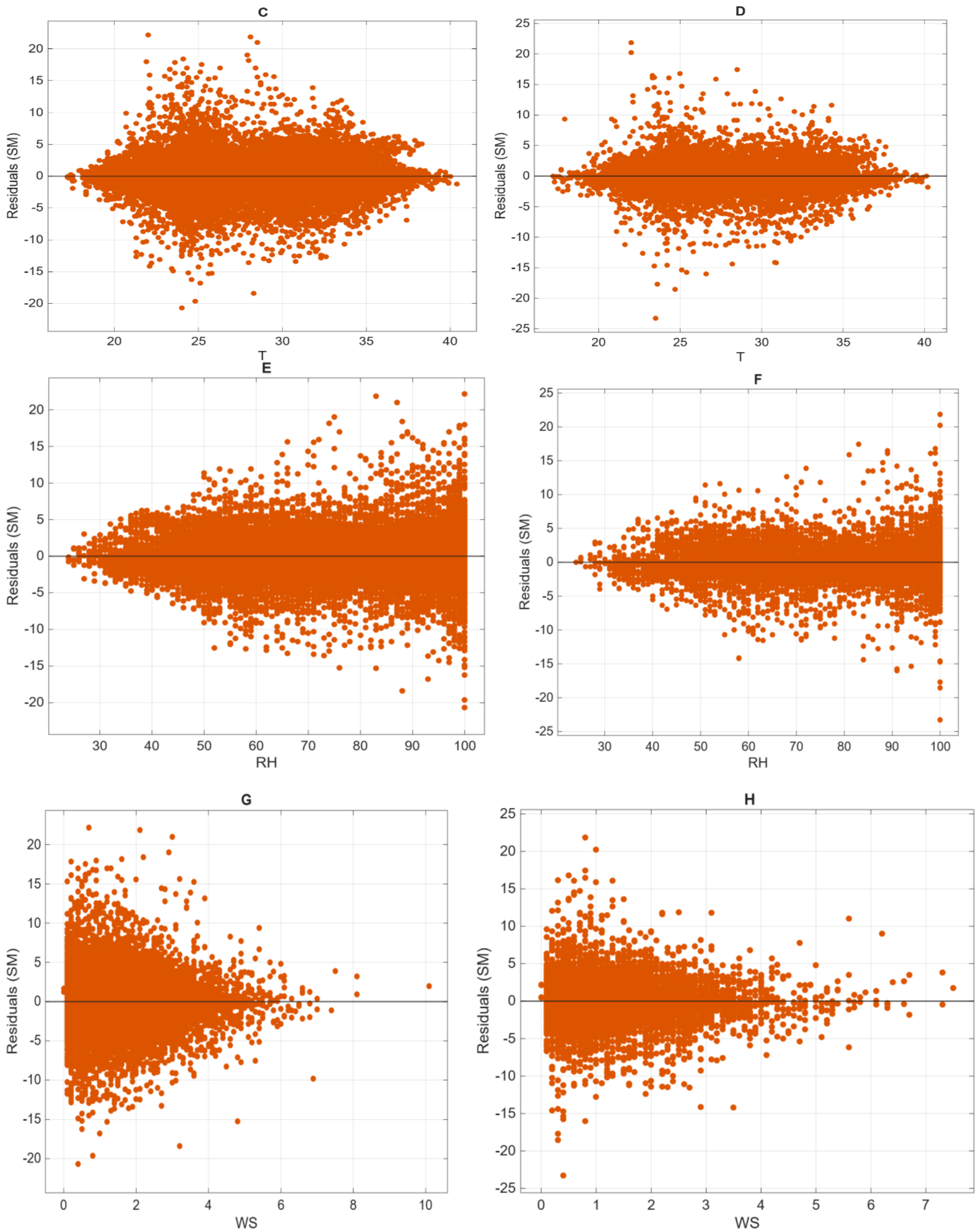


Fig. 4. Residual plots of the optimizable ensemble model for validation and testing datasets. (A) Residual plot for true response (validation), (B) Residual plot for true response (testing), (C) Residual plot for predicted response (validation) and (D) Residual plot for predicted response (testing).




C
D
E
F
G
H
Residuals (SM)
T
RH
WS

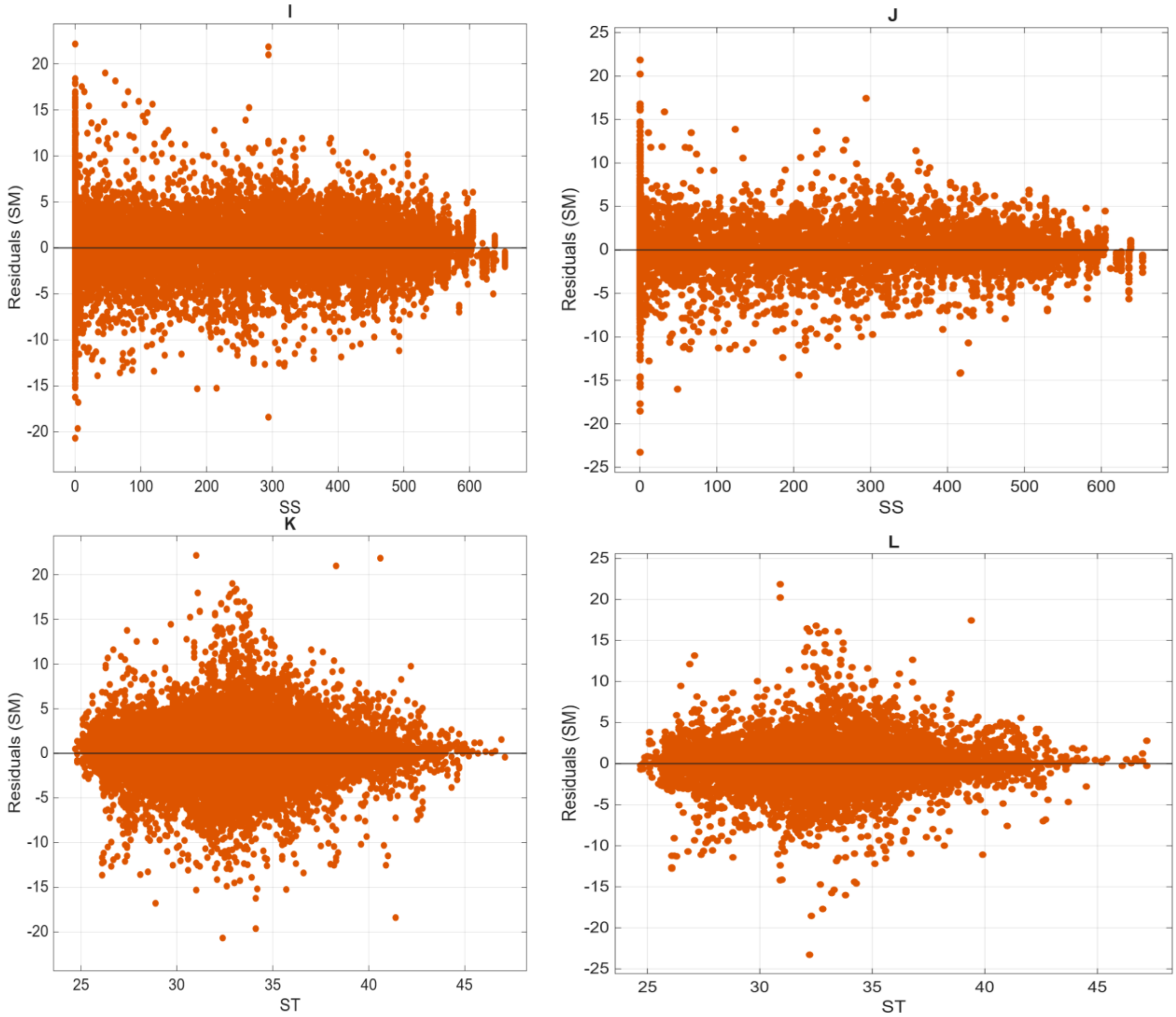


Fig. 5. Residual plots for individual predictors of the optimizable ensemble model for both validation and testing datasets. (A) Time (validation), (B) Time (testing), (C) Air Temperature (validation), (D) Air Temperature (testing), (E) Relative Humidity (RH) (validation), (F) Relative Humidity (RH) (testing), (G) Wind Speed (WS) (validation), (H) Wind Speed (WS) (testing), (I) Sunshine Duration (validation), (J) Sunshine Duration (testing), (K) Soil Temperature (validation) and (L) Soil Temperature (testing).

Permutation Feature Importance

Permutation importance (Fig. 6A-6B), which measures how much performance drops when a predictor's values are shuffled (Fisher *et al.,*2019), placed relative humidity and soil temperature clearly ahead of the rest, scoring 46.15 and 45.32 in training and 42.27 and 41.48 in testing, respectively. This consistency between training and test sets is itself a good sign that the model generalized reasonably well rather than overfitting the training data (Basir *et al.,*2024). Physically, this ranking is reasonable: higher humidity narrows the vapor pressure gap between soil and air, which is associated with slower evaporation and higher preserved moisture (Lakra *et al.,*2025), while warmer soil is associated with faster moisture evaporation (Nagappan *et al.,*2020); both patterns have also been reported by Lakra et al. (2025) and Huang et al. (2023).

Time of observation came next (35.42 training, 31.55 test), reflecting daily variation that may be tied to irrigation timing, plant growth stage and season, consistent with Basir et al. (2024). Air temperature scored lower than humidity (20.50 training, 17.70 test), consistent with an indirect influence on soil moisture working through the surface energy balance, whereas humidity is more directly linked to atmospheric vapor pressure deficit (Liu *et al.,*2025). Sunshine duration scored modestly (15.53 training, 13.65 test), consistent with its role in solar radiation and evaporation (Lakra *et al.,*2025), while wind speed had the lowest contribution among the predictors considered, by a wide margin (3.11 training, 1.83 test), matching Basir et al. (2024) and Senanayake et al. (2021), who likewise found wind speed to have the lowest predictive contribution in weather-based soil moisture models. This low predictive contribution reflects wind speed's limited role in this model and dataset and does not imply that wind speed is physically irrelevant to soil-moisture dynamics in general.

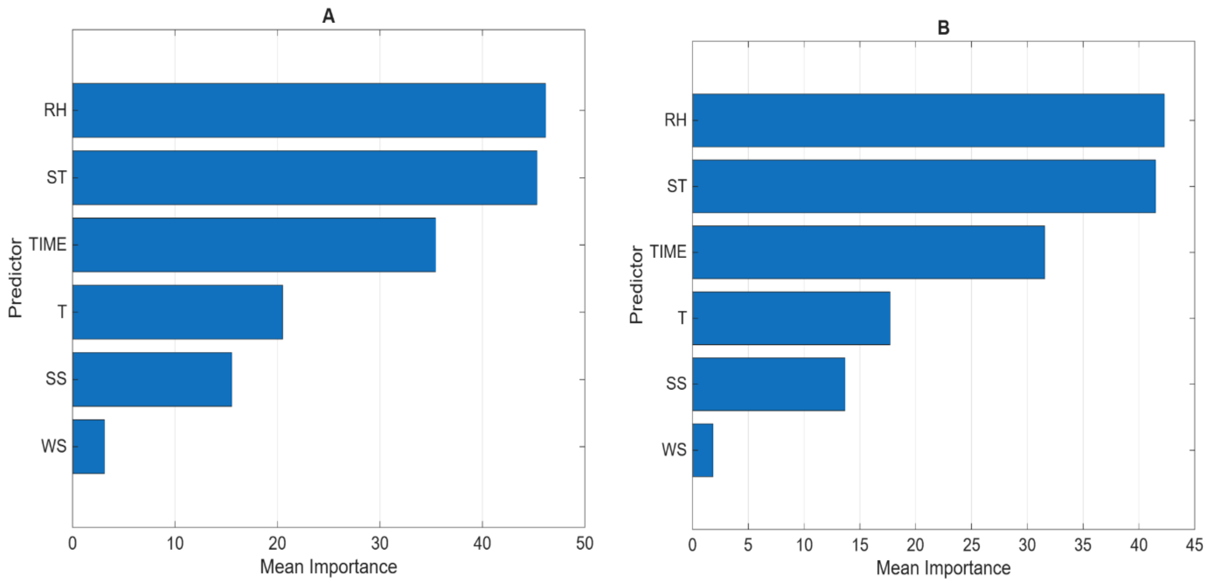


Fig. 6. Permutation importance of predictor variables for the optimizable ensemble model. (A) Training dataset and (B) Testing dataset.

SHAP-Based Interpretability Analysis

Permutation importance describes a predictor's overall contribution to model performance but not its effect on any individual prediction. SHAP (Shapley Additive Explanations) addresses this using the concept of cooperative games (Lundberg & Lee, 2017; Huang *et al.,*2023) and the SHAP-based ranking (Fig. 7A) was consistent with the permutation-importance ranking: relative humidity was again the most influential predictor with a mean SHAP value of 2.7692, followed by soil temperature (2.0025), time (1.6938), air temperature (0.8407), sunshine duration (0.6664) and wind speed (0.1545). The agreement between the two methods adds confidence to the ranking. Relative humidity's high contribution is consistent with its role in limiting evaporation (Huang *et al.,*2023; Lakra *et al.,*2025); the similarly high contribution of soil temperature is consistent with faster evaporation and biological activity in warmer soil (Nagappan *et al.,*2020); and the more moderate, indirect contribution of air temperature is consistent with its correlation with humidity (Liu *et al.,*2025). Wind speed's low SHAP contribution, like its low permutation-importance score, is consistent with Yin et al. (2026), who similarly reported humidity and soil temperature as the most influential predictors and wind speed contributing very little (Yin *et al.,*2026).

The SHAP summary plot (Fig. 7B) further showed the direction of each predictor's contribution to individual predictions. High relative humidity values were consistently associated with higher predicted soil moisture and low values with lower predicted soil moisture, spanning the widest SHAP-value range of any predictor, roughly −6.60 to +10.32 (Lakra *et al.,*2025). Soil temperature showed the reverse pattern, with low soil-temperature values associated with higher predicted soil moisture and high values with lower predicted soil moisture. Time of day, air temperature and sunshine duration showed more mixed, condition-dependent contributions (Basir *et al.,*2024; Liu *et al.,*2025), while wind-speed contributions clustered tightly near zero, consistent with its minor role (Basir *et al.,*2024). Overall, the SHAP summary plot is consistent with the permutation-importance ranking and with reasonable physical relationships between the input variables and soil moisture (Lundberg *et al.,*2020; Senanayake *et al.,*2021).

SHAP dependence plots (Fig. 8) show how each predictor's contribution to the model output changed across its own range. Relative humidity (Fig. 8C) showed a consistently positive relationship, with SHAP values increasing from about -6.66 to +10.32 as humidity rose toward saturation, consistent with Lakra et al. (2025). Soil temperature (Fig. 8F) showed the reverse pattern, with positive contributions at cooler values (26-28 °C, average +9.01) and negative contributions at hotter values (38-40 °C, -3.78), consistent with Nagappan et al. (2020). Time of day (Fig. 8A) showed a non-linear pattern, with positive contributions (+5.09) during mid-morning hours and negative contributions (-4.27) in the late evening; this pattern may reflect diurnal or irrigation-related temporal structure in the data rather than a direct physical effect of clock time (Basir *et al.,*2024). Air temperature (Fig. 8B) contributed negatively (-5.54) around 20-22 °C and slightly positively (+3.06) around 30-32 °C, consistent with its contribution depending on concurrent humidity and radiation conditions (Liu *et al.,*2025). Sunshine duration (Fig. 8E) contributed positively (+3.28) around 300-350 minutes and negatively (-2.52) beyond 600 minutes, which may be related to increased evapotranspiration on longer sunshine days. Wind speed (Fig. 8D) showed no discernible pattern, with contributions ranging narrowly from -2.07 to +1.57, consistent with its minimal predictive contribution (Basir *et al.,*2024). Taken together, relative humidity and soil temperature were the most influential predictors in the fitted model, time of day was a meaningful secondary predictor, air temperature and sunshine duration made moderate contributions and wind speed had the lowest contribution among the predictors considered; this pattern is consistent with Lundberg et al. (2020) and Senanayake et al. (2021).

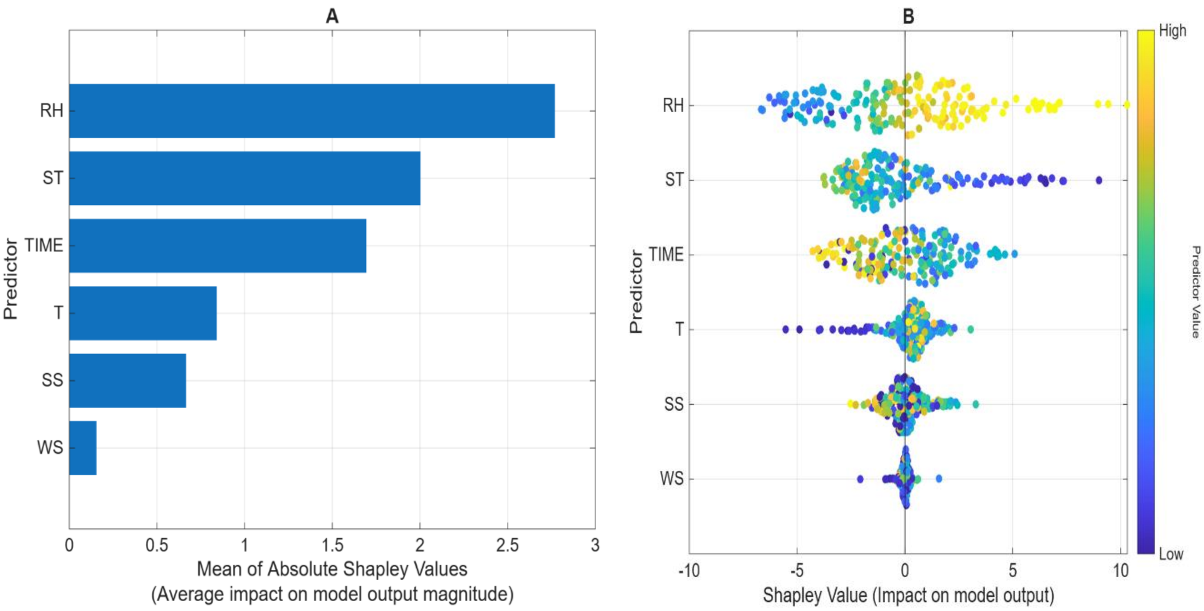
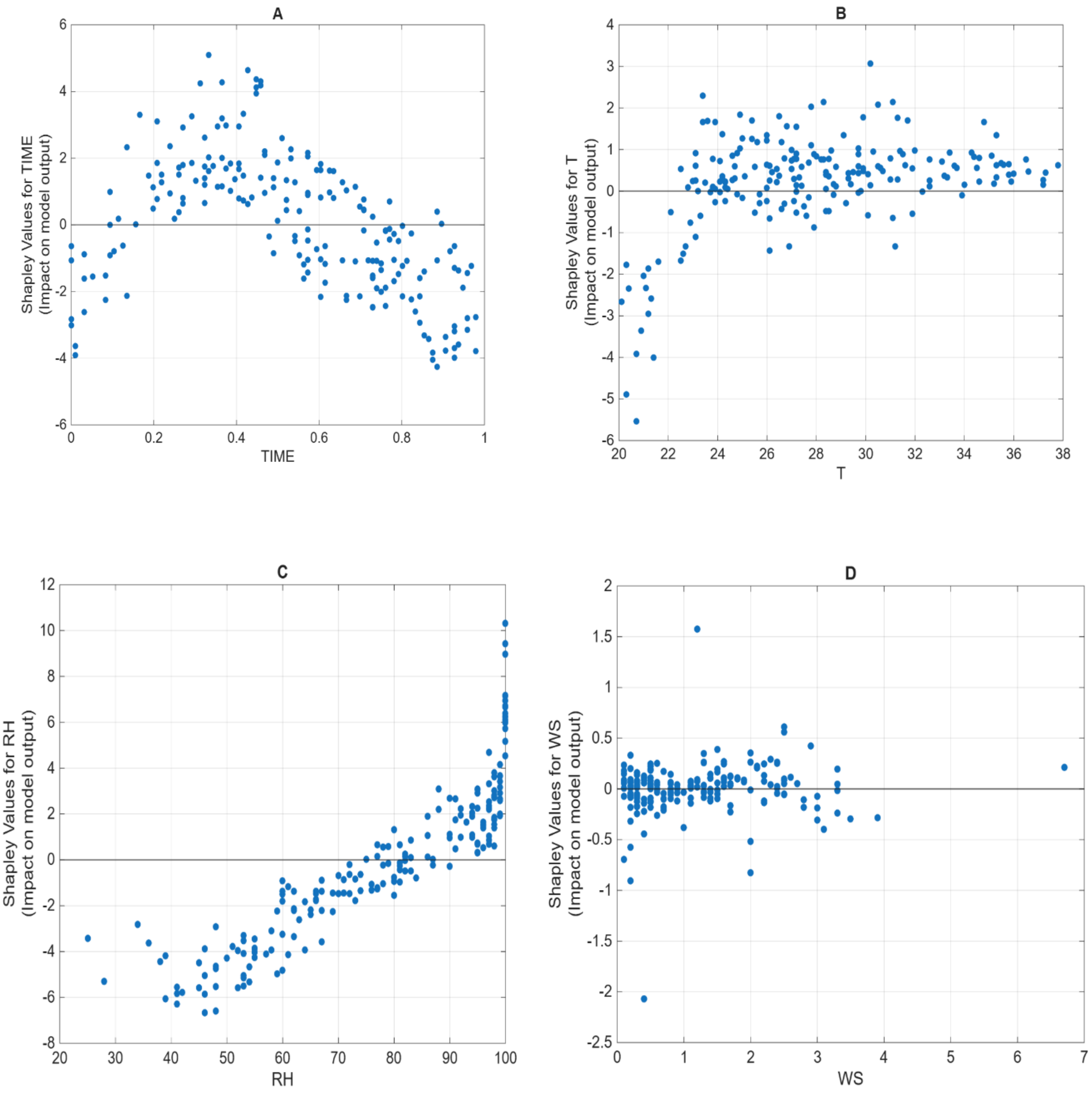


Fig. 7. SHAP (Shapley Additive Explanations) analysis of the optimizable ensemble model. (A) SHAP feature importance and (B) SHAP summary plot.



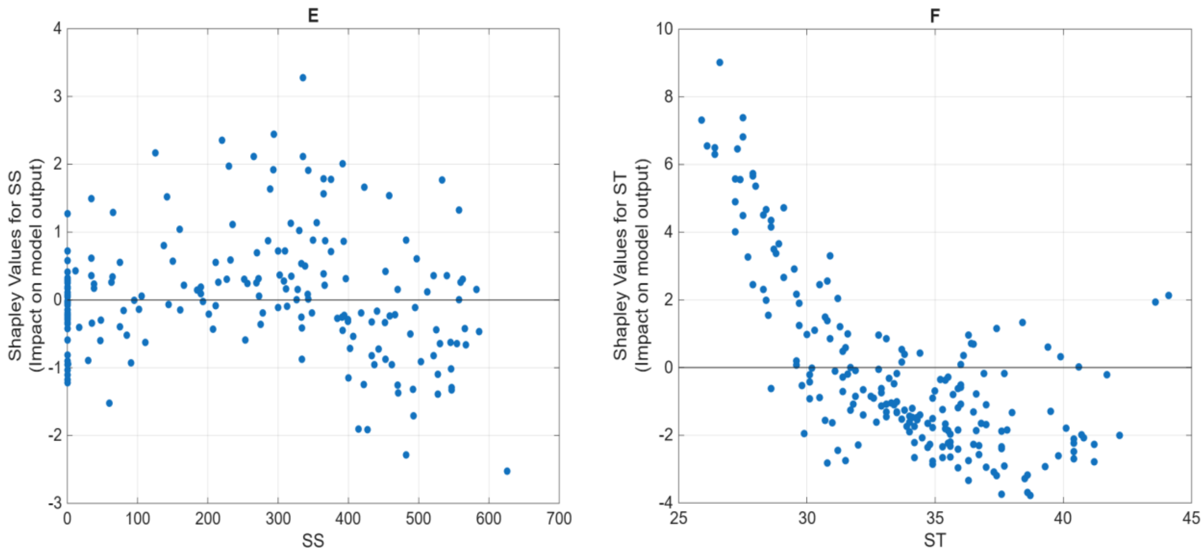


Fig. 8. SHAP (Shapley Additive Explanations) dependency plots for each predictor variable of the optimizable ensemble model. (A) Time, (B) Air Temperature, (C) Relative Humidity (RH), (D) Wind Speed (WS), (E) Sunshine Duration and (F) Soil Temperature.

CONCLUSION

This study evaluated regression-based machine learning models, including an optimizable ensemble approach, for predicting soil moisture from sub-hourly meteorological variables. Among 28 sub-models compared, the optimizable ensemble model tuned with grid search gave the best performance, with a validation MAE of 1.9756, MSE of 8.4283, RMSE of 2.9031 and R^2 of 0.83 and comparable test performance (MAE 1.8240, MSE 7.5353, RMSE 2.7451, R^2 0.85), indicating consistent predictive performance between the validation and test data. Permutation importance and SHAP analysis consistently identified relative humidity, soil temperature and time of observation as the most influential predictors, with relative humidity associated with higher predicted soil moisture and soil temperature associated with lower predicted soil moisture, consistent with expected evaporative and diurnal processes; air temperature and sunshine duration made moderate contributions and wind speed contributed comparatively little. These results suggest that an optimized ensemble model can estimate soil moisture from routinely measured, above-ground weather variables with reasonable accuracy, providing a potential basis for future irrigation-management and drought-monitoring applications. However, because the present analysis was conducted at a single study location, further validation across different locations, seasons and environmental conditions is required to assess the generalizability and broader applicability of the proposed approach.

ACKNOWLEDGEMENTS

SA acknowledge the sample data source from Tamil Nadu Agricultural University ICAR-KVK, Vamban, Pudukkottai-622 303. The MATLAB software Version R2026a by Mathworks Inc was provided by the Department of Physical Sciences and Information Technology, Agricultural Engineering College and Research Institute, Tamil Nadu Agricultural University, Coimbatore – 641003 is highly acknowledged.

Authors' contributions

SA wrote the manuscript draft and AA revised it. DR, JB and BK verified the contents. All authors read and approved the final manuscript.

Compliance with ethical standards

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

Ethical issues: None

Declaration: The authors declare that no generative artificial intelligence tools were used in the preparation of this manuscript. Only Grammarly software was used for language editing and improvement of grammar, without affecting the scientific content or interpretation of the study.

REFERENCES

1. Adeyemi, O., Grove, I., Peets, S., Domun, Y., & Norton, T. (2018). Dynamic neural network modelling of soil moisture content for predictive irrigation scheduling. Sensors, 18(10), 3408. https://doi.org/10.3390/s18103408
2. Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., & Notarnicola, C. (2015). Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. Remote Sens., 7(12), 16398-16421. https://doi.org/10.3390/rs71215841

3. Alibabaei, K., Gaspar, P. D., & Lima, T. M. (2021). Modeling soil water content and reference evapotranspiration from climate data using deep learning method. Appl. Sci., 11(11), 5029. https://doi.org/10.3390/app11115029
4. Basir, S., Noel, S., Buckmaster, D., & Ashik-E-Rabbani, M. (2024). Enhancing subsurface soil moisture forecasting: A long short-term memory network model using weather data. Agriculture, 14(3), 333. https://doi.org/10.3390/agriculture14030333
5. Breiman, L. (2001). Random forests. Mach. Learn., 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
6. Cai, Y., Zheng, W., Zhang, X., Zhangzhong, L., & Xue, X. (2019). Research on soil moisture prediction model based on deep learning. PLoS One, 14(4), e0214508. https://doi.org/10.1371/journal.pone.0214508
7. Chatterjee, S., Dey, N., & Sen, S. (2020). Soil moisture quantity prediction using optimized neural supported model for sustainable agricultural applications. Sustain. Comput. Inform. Syst., 28, 100279. https://doi.org/10.1016/j.suscom.2018.09.002
8. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM. https://doi.org/10.1145/2939672.2939785
9. Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., et al. (2017). ESA CCI Soil Moisture for improved Earth system understanding: State-of-the-art and future directions. Remote Sens. Environ., 203, 185-215. https://doi.org/10.1016/j.rse.2017.07.001
10. ElSaadani, M., Habib, E., Abdelhameed, A. M., & Bayoumi, M. (2021). Assessment of a spatiotemporal deep learning approach for soil moisture prediction and filling the gaps in between soil moisture observations. Front. Artif. Intell., 4, 636234. https://doi.org/10.3389/frai.2021.636234
11. Fang, K., Shen, C., Kifer, D., & Yang, X. (2017). Prolongation of SMAP to spatiotemporally seamless coverage of continental U.S. using a deep learning neural network. Geophys. Res. Lett., 44(21), 11030-11039. https://doi.org/10.1002/2017GL075619
12. Filipovic, N., Brdar, S., Mimic, G., Marko, O., & Crnojevic, V. (2022). Regional soil moisture prediction system based on long short-term memory network. Biosyst. Eng., 213, 30-38. https://doi.org/10.1016/j.biosystemseng.2021.11.019
13. Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res., 20(177), 1-81. https://doi.org/10.48550/arXiv.1801.01489
14. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Ann. Stat., 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451
15. Gao, L., & Ding, Y. (2020). Disease prediction via Bayesian hyperparameter optimization and ensemble learning. BMC Res. Notes, 13(1), 203. https://doi.org/10.1186/s13104-020-05050-0
16. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7
17. Houben, D., Jacobs, A., Boeckx, P., Mayer, J., Rumpel, C., & Schirrmann, M. (2025). Machine-learning based spatiotemporal prediction of soil moisture in a grassland hillslope. Vadose Zone J., 24(1), e70011. https://doi.org/10.1002/vzj2.70011
18. Huang, F., Zhang, Y., Zhang, Y., Nourani, V., Li, Q., & Li, L. (2023). Towards interpreting machine learning models for predicting soil moisture droughts. Environ. Res. Lett., 18(7), 074002. https://doi.org/10.1088/1748-9326/acdbe0
19. Krause, P., Boyle, D. P., & Base, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci., 5, 89-97. https://doi.org/10.5194/adgeo-5-89-2005
20. Lakra, D., Srivastava, P. K., & Pandey, A. C. (2025). Soil moisture retrieval over agricultural regions using SAR data and machine learning techniques. Front. Remote Sens., 6, 1513620. https://doi.org/10.3389/frsen.2024.1513620
21. Liu, X., Peterson, J., & Zhang, Z. (2025). High-resolution soil moisture estimation in the spatiotemporal variability of surface soil moisture using machine learning methods integrating satellite and ground-based soil moisture and environmental data. Water Resour. Manag., 39(1), 201-220. https://doi.org/10.1007/s11269-024-04069-3
22. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell., 2(1), 56-67. https://doi.org/10.1038/s42256-019-0138-9
23. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst., 30, 4765-4774. https://doi.org/10.48550/arXiv.1705.07874
24. Mao, F., Ji, Z., Xia, L., & Yue, H. (2024). A comprehensive study of deep learning for soil moisture prediction. Hydrol. Earth Syst. Sci., 28(4), 917-933. https://doi.org/10.5194/hess-28-917-2024
25. Nagappan, M., Vinodhini, G., & Murugappan, A. (2020). Prediction of reference evapotranspiration for irrigation scheduling using machine learning. Hydrol. Sci. J., 65(16), 2669-2677. https://doi.org/10.1080/02626667.2020.1830996
26. Nguyen, T. T., Ngo, H. H., Guo, W., Chang, S. W., Nguyen, D. D., Nguyen, C. T., et al. (2022). A low-cost approach for soil moisture prediction using multi-sensor data and machine learning algorithm. Sci. Total Environ., 833, 155066. https://doi.org/10.1016/j.scitotenv.2022.155066

27. Paul, S., & Singh, S. (2020). Soil moisture prediction using machine learning techniques. In Proceedings of the 2020 3rd International Conference on Computational Intelligence and Intelligent Systems (pp. 1-7). ACM. https://doi.org/10.1145/3440840.3440854
28. Pokhrel, Y., Hanasaki, N., Yeh, P. J., Biemans, H., & Wada, Y. (2021). Global terrestrial water storage and drought severity under climate change. Nat. Clim. Change, 11, 226-233. https://doi.org/10.1038/s41558-020-00972-w
29. Ren, Y., Ling, F., & Wang, Y. (2023). Research on provincial-level soil moisture prediction based on extreme gradient boosting model. Agriculture, 13(5), 927. https://doi.org/10.3390/agriculture13050927
30. Senanayake, I. P., Yeo, I. Y., Walker, J. P., & Willgoose, G. R. (2021). Estimating catchment scale soil moisture at a high spatial resolution: Integrating remote sensing and machine learning. Sci. Total Environ., 776, 145924. https://doi.org/10.1016/j.scitotenv.2021.145924
31. Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., et al. (2010). Investigating soil moisture-climate interactions in a changing climate: A review. Earth Sci. Rev., 99(3-4), 125-161. https://doi.org/10.1016/j.earscirev.2010.02.004
32. Shawon, S. M., Neha, N. I., Jui, A. N., Dey, N., & Zubair, H. T. (2025). Advances in soil moisture measurement techniques and prediction using artificial intelligence: An extensive and systematic review. Smart Agric. Technol., 12, 101613. https://doi.org/10.1016/j.atech.2025.101613
33. Teshome, F. T., Bayabil, H. K., Schaffer, B., Ampatzidis, Y., & Hoogenboom, G. (2024). Improving soil moisture prediction with deep learning and machine learning models. Comput. Electron. Agric., 226, 109414. https://doi.org/10.1016/j.compag.2024.109414
34. Xu, L., Abbaszadeh, P., Moradkhani, H., Chen, N., & Zhang, X. (2020). Continental drought monitoring using satellite soil moisture, data assimilation and an integrated drought index. Remote Sens. Environ., 250, 112028. https://doi.org/10.1016/j.rse.2020.112028
35. Yin, X., Li, Z., Du, T., & Kang, S. (2026). Unraveling soil moisture dynamics with dual-scale interpretable machine learning: Cover cropping and irrigation insights in semi-arid agriculture. Vadose Zone J., 25(1), e70077.