TYPE Original Research
PUBLISHED 07 June 2023
DOI 10.3389/fpls.2023.1143462

Simulating soil salinity dynamics, cotton yield and evapotranspiration under drip irrigation by ensemble machine learning

Zewei Jiang[1], Shihong Yang[1,2,3]*, Shide Dong[4,5], Qingqing Pang[6], Pete Smith[7], Mohamed Abdalla[7], Jie Zhang[1], Guangmei Wang[4] and Yi Xu[1]

[1]College of Agricultural Science and Engineering, Hohai University, Nanjing, China, [2]State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, China, [3]Cooperative Innovation Center for Water Safety & Hydro Science, Hohai University, Nanjing, China, [4]CAS Key Laboratory of Coastal Environmental Processes and Ecological Remediation, Yantai Institute of Coastal Zone Research (YIC), Chinese Academy of Sciences (CAS), Shandong Key Laboratory of Coastal Environmental Processes, YICCAS, Yantai, Shandong, China, [5]Shandong Saline-Alkali Land Modern Agriculture Company, Dongying, China, [6]Nanjing Institute of Environmental Sciences, Ministry of Ecology and Environment, Nanjing, China, [7]Institute of Biological & Environmental Sciences, University of Aberdeen, Aberdeen, United Kingdom

Check for updates

OPEN ACCESS

EDITED BY
Shaohua Wan,
University of Electronic Science and Technology of China, China

REVIEWED BY
Geng(Frank) Bai,
University of Nebraska-Lincoln,
United States
Yong He,
Chinese Academy of Agricultural Sciences,
China

*CORRESPONDENCE
Shihong Yang
✉ ysh7731@hhu.edu.cn

RECEIVED 13 January 2023
ACCEPTED 11 May 2023
PUBLISHED 07 June 2023

CITATION
Jiang Z, Yang S, Dong S, Pang Q, Smith P, Abdalla M, Zhang J, Wang G and Xu Y (2023) Simulating soil salinity dynamics, cotton yield and evapotranspiration under drip irrigation by ensemble machine learning.
Front. Plant Sci. 14:1143462.
doi: 10.3389/fpls.2023.1143462

Cotton is widely used in textile, decoration, and industry, but it is also threatened by soil salinization. Drip irrigation plays an important role in improving water and fertilization utilization efficiency and ensuring crop production in arid areas. Accurate prediction of soil salinity and crop evapotranspiration under drip irrigation is essential to guide water management practices in arid and saline areas. However, traditional hydrological models such as Hydrus require more variety of input parameters and user expertise, which limits its application in practice, and machine learning (ML) provides a potential alternative. Based on a global dataset collected from 134 pieces of literature, we proposed a method to comprehensively simulate soil salinity, evapotranspiration (ET) and cotton yield. Results showed that it was recommended to predict soil salinity, crop evapotranspiration and cotton yield based on soil data (bulk density), meteorological factors, irrigation data and other data. Among them, meteorological factors include annual average temperature, total precipitation, year. Irrigation data include salinity in irrigation water, soil matric potential and irrigation water volume, while other data include soil depth, distance from dripper, days after sowing (for EC and soil salinity), fertilization rate (for yield and ET). The accuracy of the model has reached a satisfactory level, R^2 in 0.78-0.99. The performance of stacking ensemble ML was better than that of a single model, i.e., gradient boosting decision tree (GBDT); random forest (RF); extreme gradient boosting regression (XGBR), with R^2 increased by 0.02%-19.31%. In all input combinations, other data have a greater impact on the model accuracy, while the *RMSE* of the S1 scenario (input without meteorological factors) without meteorological data has little difference, which is -34.22%~19.20% higher than that of full input. Given the wide application of drip irrigation in cotton, we recommend the application of ensemble ML to predict soil salinity and crop evapotranspiration, thus serving as the basis for adjusting the irrigation schedule.

KEYWORDS

salinity, evapotranspiration, drip irrigation, cotton, ensemble machine learning

1 Introduction

Cotton (Gossypium barbadense L.) is a crucial economic crop in the world (Rodriguez-Sanchez et al., 2022). It is not only one of the main sources of natural fibers for textiles as well as edible oil (Ibrahim et al., 2022), but also plays an important role in national defense, medicine, the automobile industry, and other fields (Xu et al., 2021). Meanwhile, more than 1 billion hectares of soil in the world are threatened by soil salinization, thus land degradation, food reduction (even up to 50%), and environmental threats are increasing day by day (Wang et al., 2021). In arid or semi-arid areas where cotton is widely planted, more than half of the irrigation systems are related to salinization (Wang et al., 2018). Under the background of human growth and global warming, water resources are becoming increasingly scarce, and the traditional irrigation-drainage balance to wash salt method is difficult to maintain, especially in cotton-producing areas such as the arid areas in northwest China (Zong et al., 2022). Smart irrigation based on Internet to Things technology has also paved the way for developing precision irrigation technologies (Boursianis et al., 2021). Thus, drip irrigation, a more water-saving and efficient irrigation technology, has been widely promoted in cotton planting. The way salt moves and accumulates in the soil may be affected by poor drainage, irrigation practices, vegetation removal, and landscape remodelling through earthworks (Sun et al., 2012). Therefore, the dynamics of soil salt in cotton fields under drip irrigation is a problem worthy of study, because the accurate prediction of soil salinity is the most prominent and economic method to prevent soil salinization (Xiao et al., 2023).

Soil salinity is generally expressed by the percentage of salt content (Vermeulen and Van Niekerk, 2017), or electrical conductivity (EC) detected by electromagnetic induction or dielectric sensor. Conventional soil salinity measurement methods include direct assay in the laboratory through chemical methods (Peng et al., 2016), as well as indirect methods such as remote sensing inversion and soil reflectance conversion (Zhao et al., 2022). However, field measurement requires destructive sampling of soil, which is also time-consuming, laborious, and expensive. Electromagnetic induction (EMI), such as EM38 or EM31, can be used as mature methods and auxiliary data to quickly map soil properties related to salinity and measure apparent EC as well (Castrignanò et al., 2012; Narjary et al., 2019). Although the method of satellite remote sensing can be used to identify soil salinity in a large area, its accuracy still needs to be improved, and it is mainly concentrated in the surface soil. Previous studies (Yang et al., 2020) have shown that the increase of soil salinity doesn't affect crop production until it exceeds a certain threshold level. However, once the salinity threshold is exceeded, the cotton yield almost linearly decreases with the increase of soil salinity (Oster, 1994). At the same time, accurate prediction of cotton yield is of great significance for coping with climate change, cotton breeding (Ashapure et al., 2020), farmers and stakeholders to make wise decisions, such as water and fertilizer input, storage demand, cash flow calculation, crop insurance, etc (Xu et al., 2021). Similarly, the traditional yield measurement methods are either large-scale harvesting in the harvest season, destructive sampling, or remote sensing estimation. However, the former is too time-consuming, while the latter is difficult to improve the resolution and is vulnerable to weather problems such as clouds. In addition, evapotranspiration (ET) of cotton, the loss of water vapor flux transmitted from land and vegetation to air, represents the productivity of crops and is an important indicator for studying the relationship between crop yield and water content (Bhattacharya et al., 2011). ET also plays a key role in the water, carbon and energy cycle of terrestrial ecosystems (Zhang et al., 2022). ET not only affects cotton growth and development, but also influences atmospheric circulation and climate (Huang et al., 2019). The commonly used reference ET (ET_0) method requires a large amount of meteorological data input, which is difficult to achieve in remote areas (Mattar, 2018). However, the method of crop conversion coefficient (K_c) depends too much on the accuracy of K_c, with great uncertainty. To sum up, accurate and practical alternative prediction methods are urgently needed for soil salinity, EC, cotton yield, and ET of drip-irrigated cotton fields. Moreover, considering the advantages of traditional process-based models and data-driven models that can overcome the problem of time-consuming and costly fields trails (Jiang et al., 2023a; Jiang et al., 2023b), this provides a new choice for the prediction of those issues.

A common method is to use hydrological models, such as Hydrus-2D and DRAINMOD, to simulate the water, salt and heat transport process of cotton field system under drip irrigation (Li et al., 2018; Ning et al., 2021). They can be used as a powerful complement to experiments to assess soil hydraulic properties, boundary conditions, irrigation frequency and salinity, and crop types to optimize soil and water management practices (Devkota et al., 2022). However, they require more input data and higher user skills (Table S1), which may be difficult to obtain in many regions where data are scarce ad undeveloped. Although widely proven that hydrological models are more interpretable (Liu et al., 2021), they are not always easy to use models and are mainly concentrated on smaller spatiotemporal scales. Most Hydrus studies focused on a depth of 2 m and a range of one year. In addition, previous studies (Karandish and Šimůnek, 2016; Elnesr and Alazba, 2017) have compared hydrological models with ML and found that the former is very sensitive to boundary and initial conditions. If excessive relaxation occurs, the hydrological model may be unstable and can have difficulties in term of speed and convergence probability. Considering that accuracy is proportional to the hardware resources required, it require a higher level of human skills than data-driven models such as ML. Another potential alternative might be machine learning (ML), which is different from the limited regional empirical model and is not as complex and demanding as the hydrological model (Kisi, 2016). Being good at solving nonlinear and multivariable problems, ML has been widely used in hydrology, agriculture, environment, and other fields in recent years (Wan and Goudos, 2020; Jiang et al., 2022a). Among them, ML has been proved to be a powerful tool in crop ET and yield prediction (Dramsch, 2020; Filippi et al., 2022). However, although ML has been used for soil salinity prediction in remote sensing and other fields, it is more concentrated in reflectance and other aspects. Dynamic prediction of soil salinity or EC in drip-irrigated cotton fields based on ML has not been reported. In addition, cotton yield includes not only seed yield, but also lint yield, one of the most important criteria for selecting new lines in breeding (Rodriguez-Sanchez et al., 2022). Hence, we assume that soil salinity, EC, seed yield, lint yield and ET of cotton field under drip irrigation can be predicted by ML and simple input parameters.

Moreover, to compare the effects of different ML models, a new stacking ensemble ML algorithm was also introduced in this study. Since it integrates the basic ML model, it is usually found to have better prediction performance (Jiang et al., 2022b). Meanwhile, the input parameters of the model have a great impact on the results, so it is necessary to find the most suitable and convenient input combination for ML algorithm. The objectives of this research are: (1) to build a global data set of drip irrigated cotton fields, and verify the feasibility of using ML models to predict soil salinity, EC, cotton yield and ET based on basic soil data, meteorological data, irrigation data, and other data; (2) to compare the performance of three common ML models (gradient boosting decision tree, GBDT; random forest, RF; extreme gradient boosting regression, XGBR) and stacking ensemble ML algorithm; (3) to analyze the influence of different input combinations on the accuracy and stability of models.

2 Materials and methods

2.1 Machine learning models

In this study, three basic ML models, gradient boosting decision tree (*GBDT*), random forest (*RF*), extreme gradient boosting regression (*XGBR*), and a stacking ensemble ML algorithm were selected. This is because they represent three classical commonly used models, and previous studies (Liu et al., 2020; Jiang et al., 2022c) have found that they have good performances in solving regression problems. Initially we intended to include other basic ML models such as support vector machine (SVM) and multiple perceptions (MLP, a type of neural network-based ML, similar to ANN and DNN), but the preliminary results showed that their performance was significantly worse than those tree-based models, and therefore, decided not to include them in this study. More details can be found in the Supplementary Materials. All the codes were implemented in python 3.8 by applying the sklearn (https://scikit-learn.org/) and XGB (https://xgboost.readthedocs.io/en/latest/index.html) packages. The models were run on a laptop equipped with Intel core i5-8300H CPU and NAVID GTX 1050 GPU.

2.1.1 GBDT, RF, XGBR

The GBDT model proposed by (Friedman, 2002) is a widely used method using decision stumps or regression trees as basic learners to solve classification or regression problems. It constructs additive regression models by using the least squares method to sequentially plus a simple parameterized function to the current pseudo residuals, the gradient of the loss functional being minimized, in each iteration. The RF model, based on the theory of random partition selection and random subspace, is a simple bagging ensemble of tree predictors, and the results of each tree were weighted and averaged to achieve the final output (Breiman, 2001). The generalization error converges to the limit with the number of trees increasing and the exchange or change of covariables. Both GBDT and RF combine weak learners, but the difference is that the tree of the former is fitted on the residual of the previous tree, so the biases can be reduced, while the latter reduce the variance (Fan et al., 2018). The RF is obtained by training N decision trees on the training-testing set of N samples by putting back samples. The current approximation of GBDT and the margin function of RF are as Eqs. (1-2) respectively. As supposed by (Chen and Guestrin, 2016), the XGBR based on the lifting method, has integrated all the predictions of weak learners (classification and regression trees) through boosting and additive training strategies. It has been found that XGBR improves the objective optimization function by optimizing the loss function and complexity penalty, thereby preventing overfitting. Moreover, the functions in the XGBR model will automatically perform parallel computing in the training period to reduce calculation costs (Eq. (3)). The regularized objectives were minimized according to Eqs. (4-5).

$$F_m(x) = F_{m-1}(x) + \nu\ \gamma_{lm} I(x \in R_{lm}) \tag{1}$$

$$mg(X, Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j) \tag{2}$$

$$y_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{3}$$

$$L(\phi) = \sum_i l(y_i', y_i) + \sum_k \Omega(f_k) \tag{4}$$

$$where\ \Omega(f) = \gamma T + \frac{\lambda}{2}\|w\|^2 \tag{5}$$

where $F_m(x)$ is the function maps x to y, which are jointly fit to the training data in a forward "stage wise" manner to boost approximates. ν is the shrinkage parameter that controls the learning rate, and the smaller values represent better generalization error. γ_{lm} is a constant value predicted by the decision tree in the region R_{lm}. l is the node of the tree. mg (X, Y) is the margin function of X and Y, representing the extent to which the average number of votes of one tree exceeds the average number of votes of any other class. I () is the indicator function. A larger marginal value indicates the degree of confidence in classification or regression. $F = \{f(x) = w_{q(x)}\}(q: R^m \to T, w \in R^T)$ is the space of regression trees. q, w, and T are the independent tree structure (decision rules), leaf weight corresponding to each f_k, and the number of tree leaves. l, $y_i{}'$, yi and Ω represent differentiable convex loss function, prediction, target, and the regression model function, respectively.

2.1.2 Stacking ensemble model

The stacking or stacked generalization was proposed by (Wolpert, 1992). It constructs multi-level classifiers or regressors hierarchically, and uses the mutual complementarities between basic models to enhance the generalization ability. Based on the leave-one-out cross validation method, the stacking model uses the meta learner to optimally combine the prediction results of the basic models to obtain the final output, while the output of the basic learners will not be trained to avoid overfitting. More information can be found in previous studies (Chou et al., 2014; Gu et al., 2022). The weights of base learners were calculated as Eq. (6). In addition, the objective function was computed as follows [Eq. (7)] to estimate stacking weights by minimizing the mean square linear regression.

$$y_{p,i} = \sum_{m=1}^{M} \omega_m f_{m,i} \tag{6}$$

$$\Omega = \operatorname{argmin} \sum_{i=1}^{N} \left(y_{o,i} - \sum_{m=1}^{M} \omega_m f_{m,i}\right)^2,\ \left(\sum_{m=1}^{M} \omega_m = 1, \omega_m \geq 0\right) \tag{7}$$

where ω_m is the weight assigned to each base model, $f_{m,i}$ represents the prediction of the i_{th} observation by model m. $\Omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$ denotes the weight set assigned to the base model.

2.2 Model simulation and evaluation

2.2.1 Data collection

To obtain the global dataset of soil salinity, cotton yield and ET from cotton fields under drip irrigation, we conducted a comprehensive literature search on the Web of Science and the China Knowledge Resource Integrated (CNKI) database before October 2022. The following keywords were used: soil salt, soil salinity, drip irrigation, evapotranspiration, cotton, and yield. The complete search format was available in the Supplementary Materials. All the literature were downloaded and checked manually to ensure they are point measurements, follow a standard data collection protocol, and weather data are from the local or nearest meteorological station. Since it's hard to make sure all of the drip irrigation setups from global experiments are the same, we used the distance from drip to partly represent line spacing. More details of the uncertainty could be found in the discussion. The total number of papers is 1317, but only 230 papers

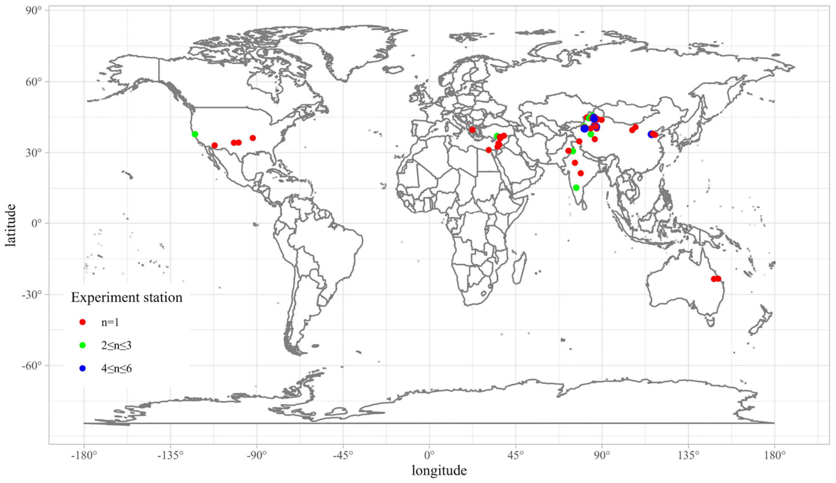

FIGURE 1
Location of experiment station used in this study. n represents the number of literatures, n > 1 represents multiple trials conducted at the same site.

(50 about ET, 112 about salt, and 68 about yield) were selected after excluding greenhouse, pot, laboratory, and modelling research. Then the abstracts were reviewed carefully to check whether they meet the inclusion criteria, only field trials with multiple replicate samples and details about the experiments. Finally, we collected 134 pieces of literature that are suitable for this study (Supplementary Materials). As shown in Figure 1, the experiment sites of cotton under drip irrigation cover several continents including Asia, Europe, America, Oceania, and Africa. Data source is from peer-reviewed literatures from database. Data type are xlsx or csv data which could be found in the revised Supplementary Materials now. Data structure is matrixes of 10 dimensions, 1514 columns (10, 1514) and (10, 1748) for EC and soil salinity, and matrixes of (8, 232), (8, 121), and (8, 312) for ET, lint yield and grain yield, respectively. The output is a single column vector, and the sample size is the same as the number of the input rows. Sensor type is float for most data except integer for year. Spatial and temporal resolution are globally and yearly for yield and ET (but the salinity includes some daily results). Data size has been mentioned before. Public availability is open access now.

2.2.2 Model inputs, outputs and K-fold cross-validation

On the basis of referring a previous study (Xiao et al., 2023), four types of data were selected for model inputs, soil data (bulk density), meteorological factors (average temperature, total precipitation, year), irrigation data (salinity in irrigation water (SIW), soil matric potential (SMP), irrigation) and other data (soil depth, distance from dripper, days after sowing (DAS) for EC and soil salinity, fertilization for yield and ET). The dataset was divided into several sub datasets since the model inputs available are not the same in different literature, and the outputs are soil EC, salinity, seed yield, lint yield and ET respectively (Table 1). Given the economic attributes of cotton, data of seed yield and lint yield were collected. In addition, due to the different emphasis of different studies on soil salinity, we have collected data on soil EC and salt content. This was because there is a positive correlation between them, but they cannot be accurately transformed now (Phonphan et al., 2014). It is obviously soil data and meteorological factors will influence the results; the irrigation data was selected to estimate the effects of drip irrigation. Soil salinity and EC were related to soil depth, dripper position and cotton growth stage. Cotton yield was significantly related to fertilization, which also affected crop growth and ET. Soil data, meteorological factors and irrigation data were mandatory, while the other data was different for salinity, yield and ET. The detailed model inputs of different outputs are also available in Table 2. The dataset was randomly split to the training and testing sets by a commonly used ratio of 70%:30% according to a previous study (Livera et al., 2019). Specifically, the sample sizes of training-testing subsets of EC, soil salinity, seed yield, lint yield, and ET are 1060 and 454, 1224 and 524, 261 and 112, 85 and 37, 162 and 70 respectively. Moreover, 10-fold cross-validation was conducted to avoid over-fitting and the hyperparameters were optimized by grid search and trial and error (Table S2).

TABLE 1 Collected global dataset.

	Bulk density	SIW	SMP	Temp	Prec	Irrigation	Soil depth	Distance	Year	DAS	Fertilization	EC	Soil salinity	Seed yield	Lint yield	ET
Units	g cm^{-3}	g L^{-1}	-kPa	°C	mm	mm	cm	cm	y	d	kg N ha^{-1}	mS cm^{-1}	g kg^{-1}	kg ha^{-1}	kg ha^{-1}	mm
n	3987	3987	3987	3987	3987	3987	3262	3262	3987	3262	725	1514	1748	372	121	232
mean	1.49	1.75	5.74	10.69	160.10	466.60	60.67	6.33	3.26	118.42	228.08	2.00	7.17	5117.42	2470.18	571.94
max	1.75	9.00	50.00	29.40	1168.70	1032.00	650.00	99.00	22.00	360.00	480.00	15.68	46.03	7720.00	3899.00	1129.43
min	1.16	0.20	1.50	5.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.06	1266.83	1006.00	272.00
STD	0.10	2.15	3.89	3.53	150.80	249.22	52.37	15.60	3.98	96.60	124.00	2.23	7.17	1458.90	695.48	220.07

SIW, SMP, Temp, Prec, distance, DAS and n denote salinity of irrigation water, soil matric potential, average annual temperature, Annual precipitation, distance from dripper, days after sowing and number respectively.

TABLE 2 Input combinations for scenario simulation.

Scenarios	S0	S1	S2
Models	GBDT0, RF0, XGBR0, Stacking0	GBDT1, RF1, XGBR1, Stacking1	GBDT2, RF2, XGBR2, Stacking2
EC/Soil salinity	Bulk density, SIW, SMP, Temp, Prec, Irrigation, Soil depth, Distance, year, DAS	Bulk density, SIW, SMP, Irrigation, Soil depth, Distance, year, DAS	Bulk density, SIW, SMP, Temp, Prec, Irrigation
Seed/lint yield/ET	Bulk density, SIW, SMP, Temp, Prec, Irrigation, year, Fertilization	Bulk density, SIW, SMP, Irrigation, year, Fertilization	Bulk density, SIW, SMP, Temp, Prec, Irrigation

SIW, SMP, Temp, Prec, distance, and DAS denote salinity of irrigation water (g L^{-1}), soil matric potential (-kPa), average annual temperature (°), Total precipitation (mm), distance from dripper (cm), and days after sowing (d) respectively.

2.2.3 Scenarios simulation

To test the effects of the four kind of input data (soil data, meteorological data, irrigation data, and other data) on model accuracy, three scenarios (S0, S1, S2) of input combinations were set up for simulation. Among them, S0 is all data as input, S1 represents input without meteorological factors, S2 represents input without other data. The detailed information can be found in Table 2.

2.2.4 Evaluation criteria

In this study, we selected three commonly used criteria, i.e., mean absolute error (*MAE*), root mean square error (*RMSE*), and coefficient of determination (R^2) to evaluate the model performance (accuracy and stability) as follows.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_{O,i} - Y_{P,i}\right| \tag{8}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_{O,i} - Y_{P,i})^2}{n}} \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_{O,i} - Y_{P,i})^2}{\sum_{i=1}^{n}(Y_{O,i} - \overline{Y_O})^2} \tag{10}$$

where $Y_{O,i}$, $Y_{P,i}$, $\overline{Y_O}$ and n represent the observed, predicted, and mean of observed values (soil EC, salinity, seed yield, lint yield and ET). The closer R^2 is to 1, the closer *RMSE* and *MAE* are to 0, and the higher the accuracy of the model.

3 Results

3.1 Overview of the global datasets and the relationship among inputs and outputs

The overview of the global dataset obtained in this study can be found in Table 1. In general, the dataset includes five sub data sets, soil EC, soil salinity, seed yield, lint yield, and ET, and their sample sizes are 1514, 1748, 373, 121, and 232, respectively. To explore the relationship between model inputs and prediction objectives, and verify the feasibility of our selection of input parameters, we built the Spearman correlation heatmap between inputs and outputs (Figure 2). SMP was positively correlated with EC and salt. However, it is worth pointing out that the negative correlation between SIW and EC is not reliable, because the SIW research mainly focused on soil salinity, and most of the EC studies have no SIW description (mostly -5 kPa). It can be easily found that SIW, temperature, and distance from drippers were positively related to soil salinity (Figure 2B), while precipitation, irrigation, depth, year, and DAS were negatively correlated with salinity. This is understandable because the greater the salinity of the irrigation water, the farther away from the emitter, the less the soil was washed, and the greater the possibility of salt residue. Irrigation could wash away soil salt, and with the growth of cotton (DAS and year increasing), the soil salinity will reduce, which was consistent with previous studies (Phogat et al., 2012; Wang et al., 2018). When it comes to the relationship heatmap of cotton yield (Figure 2C, D), there was a positive correlation between fertilization, irrigation, soil bulk density and yield, no matter seed yield or lint yield, while a negative correlation was found between SMP, temperature and yield. The positive correlation between cotton yield and fertilization amount in a certain range has been confirmed by many studies (Dong et al., 2010; Ibrahim et al., 2022). The negative correlation between SMP and yield may be explained by the inhibition of salt and high temperature. As for the positive correlation between ET and irrigation, year, fertilization, and the negative correlation with rainfall (Figure 2E), it may be explained by the change of seed cotton yield (Figure 2C).

3.2 Performance of ML models modelling EC and soil salinity

Figure 3 show the results of EC and soil salinity simulated by different ML models (GBDT, RF, XGBR, and stacking). It could be found that most of the points are very close to the 1:1 line, which denotes that the ML algorithms could predict EC and soil salinity. As illustrated by Table 3, three basic models (GBDT, RF, XGBR) could capture the dynamics of both EC and soil salinity, with R^2 ranging from 0.89 to 0.98 and 0.78 to 0.91 during the training and testing periods respectively. While *MAE* of EC and soil salinity were 0.05-0.14 mS cm^{-1} and 0.63-1.49 g kg^{-1} (training), 0.27-0.34 mS cm^{-1} and 1.56-1.89 g kg^{-1} (testing). By contrast, the stacking ensemble ML model performed the best, with R^2, *MAE*, *RMSE* of EC and soil salinity in 0.92-0.98 and 0.87-0.98, 0.05-0.27 mS cm^{-1} and 0.61-1.55 g kg^{-1}, 0.34-0.71 mS cm^{-1} and 1.02-2.48 g kg^{-1}. During the training stage, the R^2 of stacking model increased by 1.10%-2.65% (EC) and

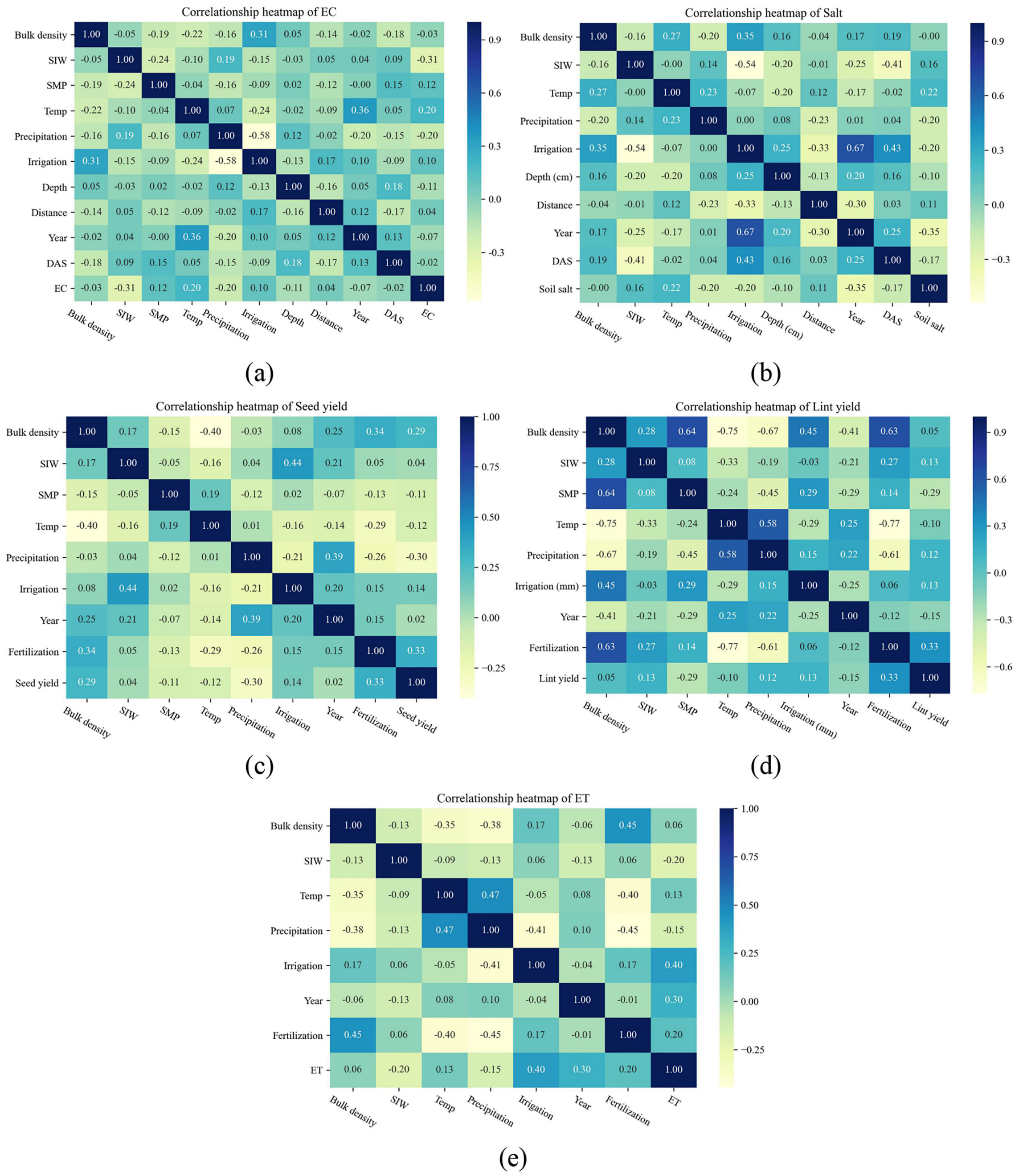

FIGURE 2
Spearman correlation heatmap of input parameters and conductivity (*EC*, **A**), soil salt **(B)**, seed yield **(C)**, lint yield **(D)**, and *ET* **(E)**. The *SIW*, *SMP*, *Temp*, *DAS* are salinity of irrigation water (g L^{-1}), soil matric potential (kPa), temperature (°C), and days after sowing respectively. The units of bulk density, precipitation, irrigation, and fertilization are g cm^{-3}, mm, mm, kg N ha^{-1}. The units of *EC*, soil salt, seed yield, lint yield, and *ET* are mS cm^{-1}, g kg^{-1}, kg ha^{-1}, kg ha^{-1}, and mm, respectively.

0.34%-10.07% (soil salinity), with *MAE* and *RMSE* decreased by 40.86%-62.19% and 3.62%-59.44% (EC), 18.68%-33.06% and 4.95-57.79% (soil salinity), respectively, compared with basic models. While in the testing period, the corresponding R^2 increased by 1.68%-4.64% (EC) and 3.20%-10.52% (salinity), *MAE* and *RMSE* decreased by 0.40%-19.22% and 0.09%-1.37% (EC), 0.47%-18.02% and 1.20%-17.82% (salinity), respectively. In addition, the calculating time of XGBR and stacking model was almost the smallest among these three models, especially during the testing stage.

3.3 Simulation accuracy of seed yield and lint yield

Figure 4 offers the scatter plots of seed yield (a) and lint yield (b) predicted by four ML models. It shows that although the performance was worse than soil salinity, the points are still near the 1:1 line, especially the blue points (stacking ensemble model). Table 4 illustrated the modelling performance of cotton yield by different ML models. Specifically, the R^2, *MAE* and *RMSE* of seed

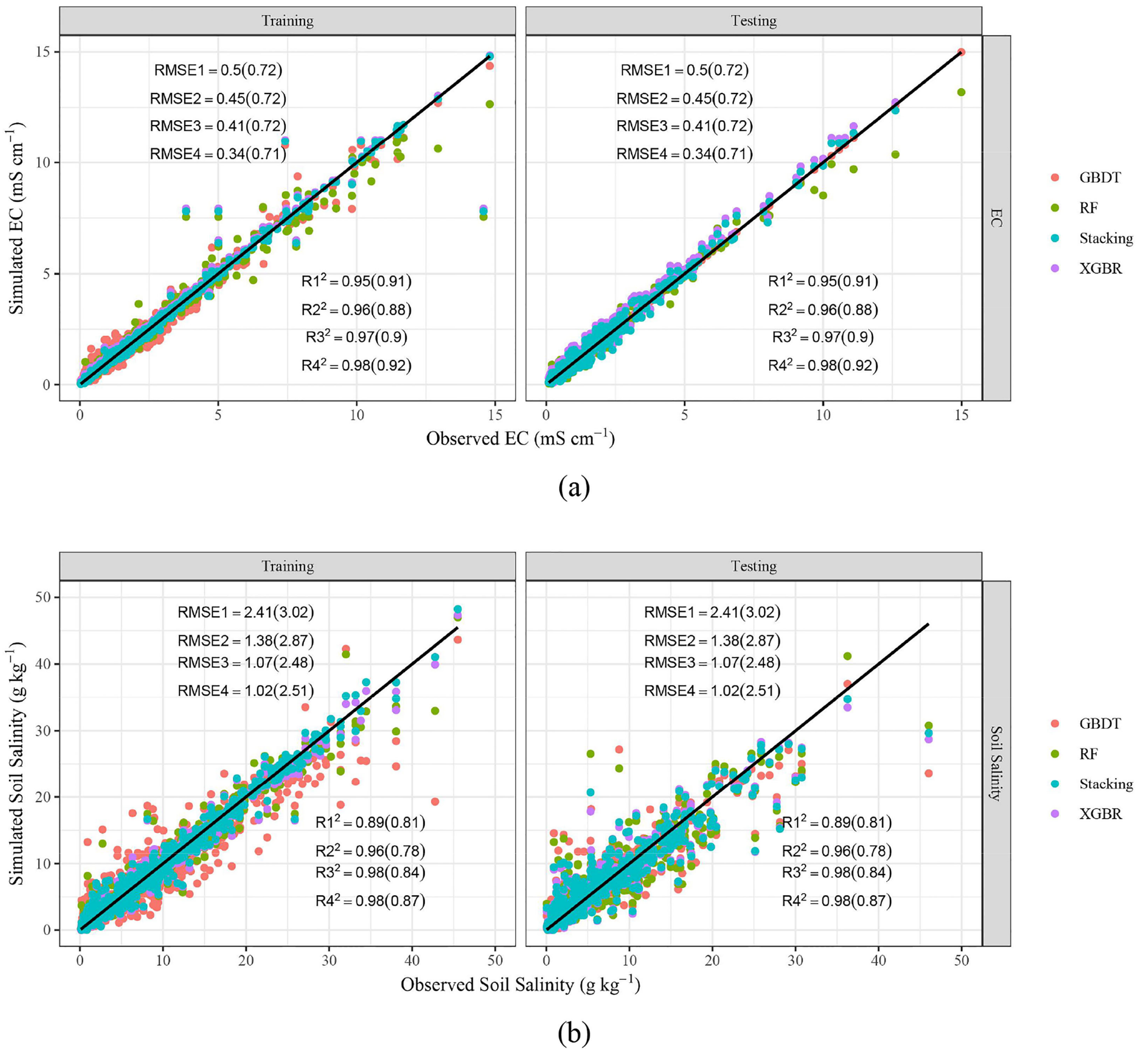

FIGURE 3
Scatter plots of predicted and observed EC **(A)** and Soil salinity **(B)**. R_1^2, R_2^2, R_3^2, R_4^2, $RMSE_1$, $RMSE_2$, $RMSE_3$, $RMSE_4$ represent R^2 and RMSE (unit: mS cm^{-1}, a, g kg^{-1}, b) of GBDT, RF, XGBR, Stacking respectively. The values outside and within brackets represent the results of the training and testing periods.

yield and lint yield predicted by three basic models (GBDT, RF, and XGBR) were in 0.94-0.99, 15.63-216.47 kg ha^{-1}, 81.30-318.01 kg ha^{-1} and 0.96-0.98, 0.32-92.08 kg ha^{-1}, 32.34-122.99 kg ha^{-1} (training stage), 0.72-0.85, 418.08-442.00 kg ha^{-1}, 571.34-627.83 kg ha^{-1} and 0.74-0.86, 190.27-231.51 kg ha^{-1}, 258.99-301.83 kg ha^{-1} (testing stage), respectively. Similarly, the stacking ensemble ML model also obtained the best performance, with R^2, *MAE*, and *RMSE* in, 0.86-0.99, 28.66-396.61 kg ha^{-1}, 83.62-557.80 kg ha^{-1} (seed yield), and 0.88-0.99, 0.13-169.24 kg ha^{-1}, 31.44-230.44 kg ha^{-1} (lint yield) respectively. The R^2 of seed yield and lint yield were 0.02%-19.31% and 1.02%-18.11% higher than basic models, while those *MAE* and *RMSE* were and 5.14%-92.78% and 2.37%-74.43%, 11.05%-99.86% and 2.77%-74.43% lower, respectively.

3.4 Comparison of different ML models predicting ET

The ET data used in this study were the seasonal cumulative ET, which were calculated by the water balance method. Specifically, it is the result of irrigation amount, precipitation, the change of soil water storage minus surface runoff and the downward flux below the crop root zone (Supplementary Materials). As displayed in Figure 5, the stacking ensemble ML algorithm was the closest to the observed ET points, followed by GBDT, XGBR, and RF. The points were almost all distributed around the 1:1 line, which denotes the acceptable accuracy of models. The comparison of model performance was shown in Table 4. It could be found that

TABLE 3 Performance of four ML models for predicting EC and soil salinity during training and testing stages.

Periods	Training				Testing			
Models	GBRT	RF	XGBR	Stacking	GBRT	RF	XGBR	Stacking
EC								
Time (s)	0.20	0.36	0.10	0.10	0.24	0.27	0.07	0.01
R^2	0.95	0.96	0.97	0.98	0.91	0.88	0.90	0.92
MAE	0.09	0.14	0.10	0.05	0.34	0.27	0.30	0.27
RMSE	0.50	0.45	0.41	0.34	0.72	0.72	0.72	0.71
Soil salinity								
Time (s)	0.36	0.29	0.54	0.18	0.24	0.26	0.07	0.01
R^2	0.89	0.96	0.98	0.98	0.81	0.78	0.84	0.87
MAE	1.49	0.75	0.63	0.61	1.89	1.71	1.56	1.55
RMSE	2.41	1.38	1.07	1.02	3.02	2.87	2.48	2.51

predicting ET was much easier than other outputs based on the soil data, meteorological data and fertilization, with R^2, *MAE*, *RMSE* in 0.99 and 0.87-0.97, 1.72-16.17 mm and 30.21-46.07 mm, 6.61-23.02 mm and 41.47-69.94 mm during the training and testing periods, respectively. Compared with those three basic models (GBDT, RF, and XGBR), the stacking ensemble model obtained the best model performance also, with R^2 increased by 1.15%-11.41%, MAE and RMSE decreased by 7.93%-89.39% and 0.27%-71.29% respectively.

3.5 Results of scenario simulation

To test the impacts of input parameters on model performance, the results of EC, soil salinity, although the *RMSE* of S1 in the training phase was close to or even less than the full inputs scenario, in most cases, especially during the testing period, the *RMSE* of S0 was still the smallest. The RMSE of S1 and S2 in the testing phase increased by 7.80%-19.20% and 23.22%-27.55% than S0, respectively. As seen in Figure 6C, D, the model performance with different input combinations varies greatly. In particular, the model accuracy under S2 scenario was significantly reduced. During the testing period, the *RMSE* of yield are 42.32%-63.63% and 26.70%-51.02% (seed), 14.06%-55.05% and 15.99%-46.04% (lint) higher than S0 and S1 respectively. Overall, the *RMSE* of stacking model was the smallest, indicating its best accuracy regardless of input combinations. It could be found that the prediction of ET depends on the meteorological factors and fertilization data also, especially during the testing periods. The *RMSE* of ET predicted by ML models under S0 decreased by 0.29%-52.09% and 6.92%-481.39% than S1 and S2, respectively.

4 Discussion

4.1 Comparison of different ML models

The soil salinity was affected by soil texture, soil moisture and irrigation, thus those factors should be considered when predicting soil salinity or EC (Wang et al., 2022). Proper evaluation of soil salinity requires traditional laboratory analysis, which is often cumbersome, expensive, time-consuming and laborious (Xiao et al., 2023). The prediction by process-based model requires

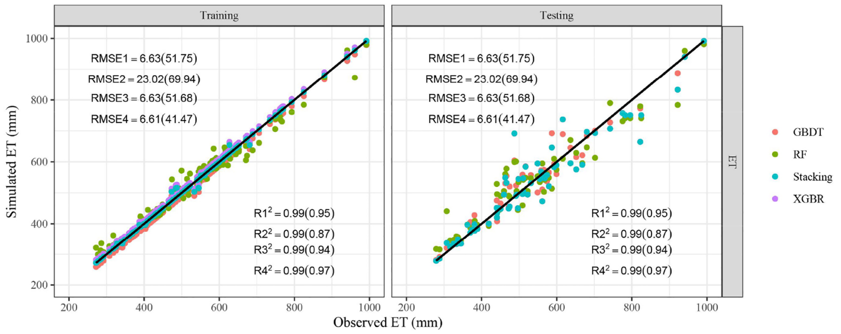

FIGURE 5
Comparison of predicted and observed ET. R_1^2, R_2^2, R_3^2, R_4^2, $RMSE_1$, $RMSE_2$, $RMSE_3$, $RMSE_4$ represent R^2 and *RMSE* (unit: mm) of GBDT, RF, XGBR, Stacking respectively. The values outside and within brackets represent the results of the training and testing periods.

TABLE 4 Performance of four ML models for predicting seed yield, lint yield and ET of cotton during training and testing stages.

Periods	Training				Testing			
Models	GBRT	RF	XGBR	Stacking	GBRT	RF	XGBR	Stacking
Seed yield								
Time	0.81	0.14	0.04	0.01	0.83	0.14	0.04	0.01
R^2	0.99	0.94	0.99	0.99	0.85	0.72	0.82	0.86
MAE	15.63	216.47	30.20	28.86	419.74	442.00	418.08	396.61
RMSE	81.30	318.01	84.08	83.62	573.40	627.83	571.34	557.80
Lint yield								
Time	0.46	0.02	0.03	0.01	0.51	0.02	0.04	0.01
R^2	0.98	0.96	0.98	0.99	0.85	0.74	0.86	0.88
MAE	0.32	92.08	0.32	0.13	190.36	231.51	190.27	169.24
RMSE	32.34	122.99	32.52	31.44	259.10	301.83	258.99	230.44
ET								
Time	0.63	0.94	0.06	0.01	0.64	0.72	0.04	0.01
R^2	0.99	0.99	0.99	0.99	0.95	0.87	0.94	0.97
MAE	2.01	16.17	2.01	1.72	32.85	46.07	32.81	30.21
RMSE	6.63	23.02	6.63	6.61	51.75	69.94	51.68	41.47

complex input data and high knowledge, or the spectral reflectance of soil salinization recorded by optical remote sensing satellite (Gorji et al., 2020) or unmanned aerial vehicle (UAV) multispectral data (Zhao et al., 2022). Recently, these methods have also made significant progress, becoming more user-friendly and accurate. More studies (Ivushkin et al., 2017; Ning et al., 2021) have used process-based or remote sensing models to simulate the dynamics of soil salinity in cotton fields, providing guidance for irrigation schedules and water salt balance. What's more, the leverage of massive data brought by remote sensing has also made ML more widely used. For example, a study (Qi et al., 2022) has proposed a soil salinity monitoring method based on satellite-ground spectral fusion and satellite-UAV collaboration in cotton planting areas, thereby drawing a soil salinity distribution map for cotton fields. In this study, we establish a method to predict soil salinity, EC, seed yield, lint yield and ET based on four ML methods (GBDT, RF, XGBR, stacking ensemble model) and soil data, meteorological data, irrigation data and other data. Overall, the model obtained an acceptable performance especially for the stacking ensemble model. It could be found that both XGBR and GBDT showed a better prediction performance than RF, which was basically consistent with a previous study (El Bilali et al., 2021). In other words, the ability of more complex integration algorithms such as boosting (XGB) and stacking ensemble model to deal with complex problems is significantly higher than that of bagging (RF), which has been proved by previous study (Pham and Won, 2022). The possible reason is that the GBDT is based on the boosting method which establishes a decision tree in each iteration, and the subsequent tree corrects the errors of the previous tree and thereby, continuously approaching the true value (Obsie et al., 2020). This is different from the RF model based on the bagging method which selects training data by putting back the sample and can simultaneously establish multiple independent decision trees (Du et al., 2022). This study mainly focused on tree-based models rather than, among others, SVM, MLP and artificial neural networks. This is because when we test the model performance, tree-based models obtained a much better result than SVM and MLP (Figure S1). However, future studies will be critical to investigate this further.

When it comes to the prediction of cotton yield (seed yield and lint yield), the model still performed well, especially the XGB and stacking. Previous studies (Xu et al., 2021; Kaur Dhaliwal et al., 2022) on the importance score of variables for predicting cotton yield by ML found that the yield had the highest response to management variables (nitrogen fertilizer application amount, covering crops, no tillage years), followed by soil and climate variables. This is consistent with our study, which proved the importance of fertilizer, meteorological data and soil bulk density by a global dataset. In addition, the nonlinear relationship among cotton yield and fertilization rate, year has been found (Nouri et al., 2020; Xie et al., 2021). This also partly proves the feasibility of using ML, which is good to solving nonlinear problems, to predict cotton yield. Last but not least, the prediction effect of ET is pretty good, even better than soil salinity and cotton yield. It is understandable since numerous studies (Xu et al., 2018; Zhang et al., 2022) have found that ML is a good tool to predict ET based on meteorological data. Previous studies (Xu et al., 2017) have also found that tree ensemble and boosted regression tree performed well in predicting ET than other ML models, which is in line with this study. The ensemble ML method shows an improvement in the predictability of crop yields, compared to the linear relationship of traditional

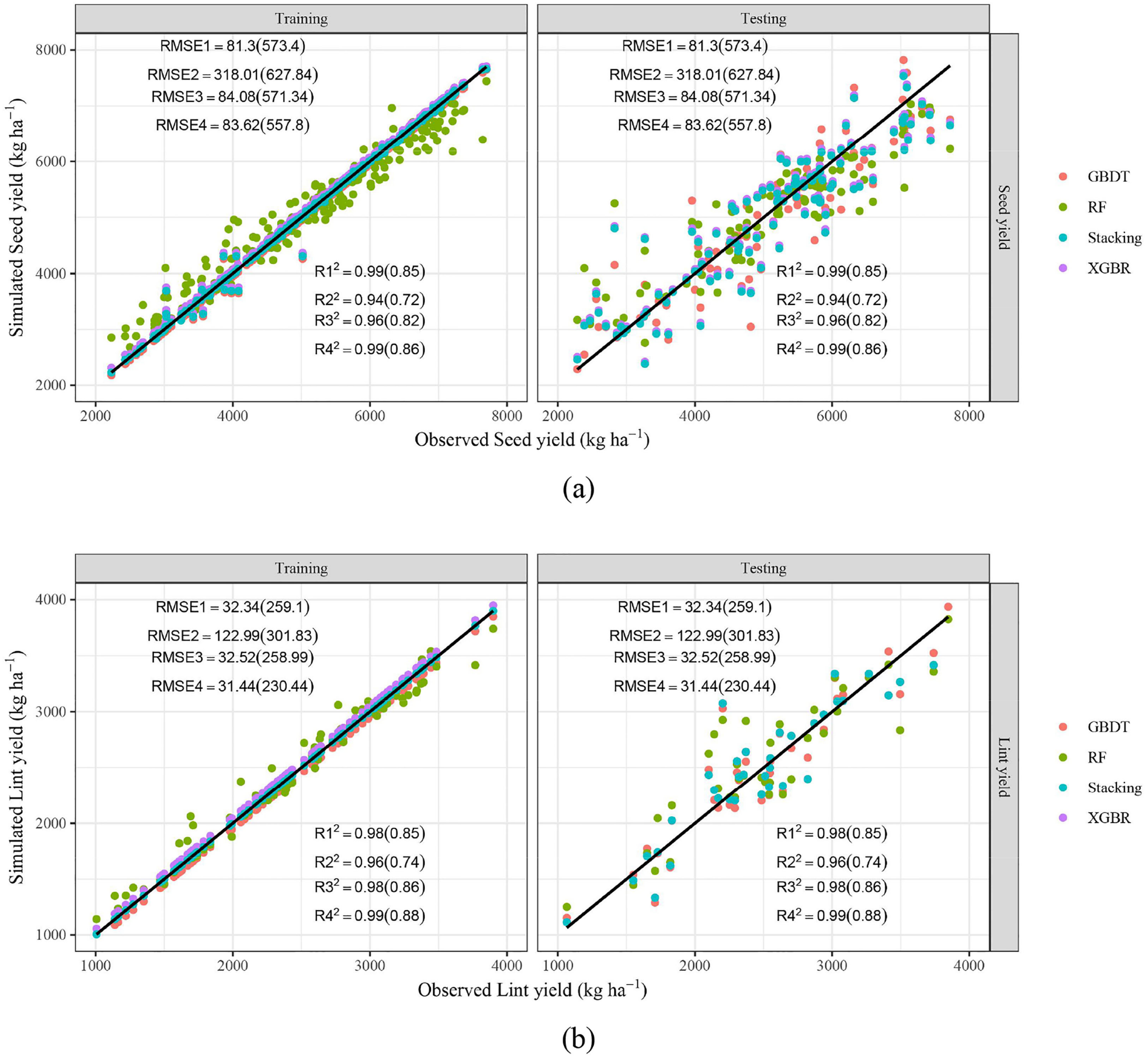

FIGURE 4
Scatter plots of predicted and observed seed yield **(A)** and lint yield **(B)** of cotton. R_1^2, R_2^2, R_3^2, R_4^2, $RMSE_1$, $RMSE_2$, $RMSE_3$, $RMSE_4$ represent R^2 and *RMSE* (unit: kg ha^{-1}) of GBDT, RF, XGBR, Stacking respectively. The values outside and within brackets represent the results of the training and testing periods.

models, which is consistent with previous studies (e.g., Kaur Dhaliwal et al., 2022). The R^2 of process-based models typically ranged from 0.83 to 0.96 (Masasi et al., 2020). This might be because cotton production is affected by many factors, including crop management, soil, and climate parameters. The relationship between yield response and environmental factors are not linear and the process-based models could explain the variance of the linear portion of the response variables better (Xie et al., 2021). Accurate simulation of cotton yield using crop models often requires substantial expertise, intensive data, and extensive calibration compared with ML (Dumont et al., 2015). However, it cannot be ignored that the performance of ML models is limited by the lack of spatial and temporal data covering a wide range of output and prediction variables (Shahhosseini et al., 2019). We suggest that future research combine ML with process-based models to overcome data limitations.

4.2 Effects of input combinations on model performance

The performance of ML models is greatly affected by regional climate conditions and the combination of input meteorological factors (He et al., 2020). This study also found this, and not only in predicting ET, but also in soil salinity, EC, seed yield, and lint yield. It is obvious that when all the data, namely soil data, meteorological data, irrigation data and other data, are used as inputs, the accuracy of all ML models is better than that of S1 and S2 scenarios with the

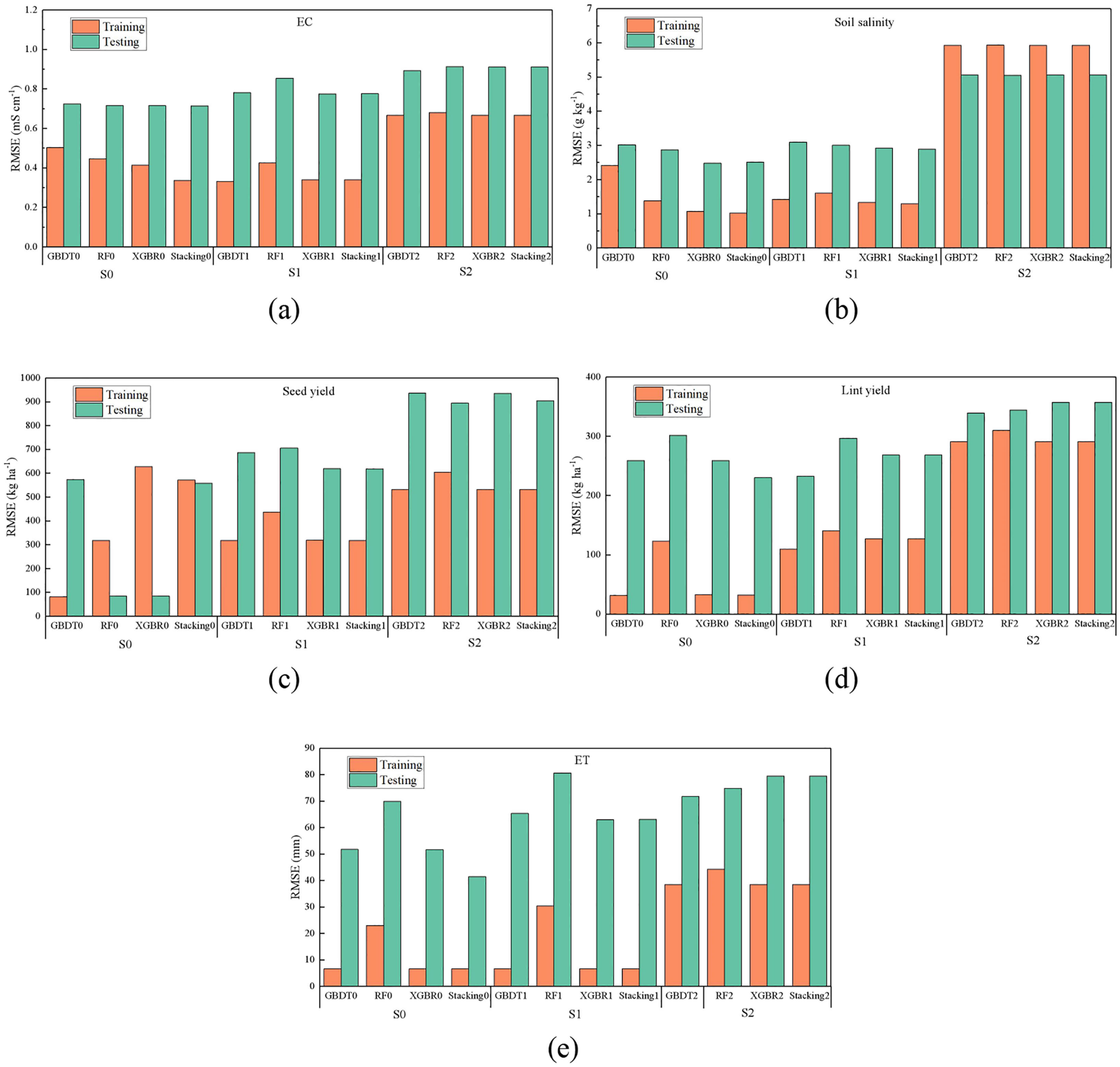

FIGURE 6
Comparison of RMSE of EC **(A)**, soil salinity **(B)**, seed yield **(C)**, lint yield **(D)**, and ET **(E)** under three scenarios (S1, S2, S3).

combination of some input parameters (Figure 6). Actually, it is easily understandable since the movement of soil salt will be affected by management measures such as irrigation, salt washing and meteorological factors (Ren et al., 2019), as well as the location of emitters and soil layers. The cotton yield and ET are obviously be affected by fertilization and meteorological factors (Huang et al., 2019). In addition, many previous studies (Dong et al., 2020; Xiao et al., 2023) have also proved this point. The results of Spearman correlation matrix (Figure 2) also showed that there was no significant correlation between most of the input parameters, which also partly supported the necessity of using more comprehensive input parameters to predict soil salinity, cotton yield and ET. As for other data such as soil depth, it was also found (Su et al., 2022) that the migration difficulty of different irons determines their accumulation depth under irrigation and evaporation. When it comes to the difference between S1 and S2, the RMSE of ML models of S1 are lower than S2 in most cases. This partly proved that the importance of other data even rather than meteorological data. It is reasonable when considering the large spatial and temporal variability of soil salinity or EC (soil depth, distance from dripper, and DAS), while cotton yield and ET are obviously affected by crop growth and fertilizer application. Moreover, the prediction accuracy of ET by ML models is higher than soil salinity, which might be because the correlation of ET and input parameters are larger (Figure 2). However, it should also be pointed out that the future studies should be critical to find the relationship between input combinations with model performance, since there might be some potential overfitting.

4.3 Implications, prospects and limitations

This study proved the ML models, especially the XGBR and stacking ensemble ML algorithm, are useful tools to predict soil salinity, EC, cotton yield and ET. It is encouraging because all the input parameters are not difficult to obtain, the use of the model is relatively simple, and the accuracy and stability are satisfactory. Based on this study, we are confident that we can realize the real-time prediction of soil salinity, ET and cotton yield under drip irrigation in the future. This can be used to guide practice and agricultural production, and optimize the management measures. The model obtained a satisfactory performance overall, even better than Hydrus, for example, previous studies (Hu et al., 2017; Liu et al., 2021) have found R^2 in simulating soil salinity ranged from 0.53-0.98, which was smaller than this study. However, there are some factors limit and degrade the current research, which can be further studied. Firstly, the currently predicted soil salinity and EC dynamics are still on a relatively coarse temporal and spatial scale, that is, based on DAS, soil depth, distance from emitters, etc., and whether the model has a more refined (hourly scale) and wider range prediction capability has not been confirmed. Although the distance from drippers reflects some spatial variability, different drip irrigation settings also bring some uncertainty. Secondly, as the types of data input are still limited, whether the model accuracy can be further improved remains to be discussed. This is mainly because the basic parameters measured in different studies are difficult to be completely consistent. The present selection of input parameters takes more into account the availability of data and their relevance to the results. This makes it easier for the model to obtain the input data in actual use, but it may also bring some uncertainties. For example, the temperature and precipitation used in this study are the annual average temperature and cumulative precipitation respectively. In the arid areas where the cotton is planted, the rainfall is very small. However, to make more accurate predictions, it may be better to select the precipitation and temperature during the growth period. But these data are usually lacking in many pieces of literature. In addition, it would be more encouraging to get the dynamic prediction of soil salinity, EC and ET with time series. Although the current research can also fill the gap for the observation data, it is more about the prediction of a certain state. How to achieve more accurate prediction combined with increasingly powerful and robust deep learning (DL) model may be another direction. Initially, we tested the results of the neural network during model selection but found the simulation results not satisfactory (data not shown). We suggest further future research such as integrating basic DL models to investigate this. Last but not least, ML algorithms are often criticized by black boxes because they can obtain good prediction results, but it is difficult to reveal how they are implemented. We have made some analysis of the spearman correlation between the input and output of the model, but it may be possible to further study it through the interpretive ML algorithm in the future (Jones et al., 2022). Moreover, in recent years, graph convolutional network (GCN), knowledge distillation (KD), edge artificial intelligence algorithms and other technologies have shown good potential in unsupervised learning (Gou et al., 2022; Wu et al., 2023; Zhang et al., 2023), but how to apply these methods to the agricultural field remains to be further studied.

5 Conclusions

This study displayed the application of different ML models (GBDT, RF, XGBR, and stacking ensemble) to accurately predict soil salinity, EC, seed yield, lint yield, and ET in drip irrigated cotton fields. Based on the global data set collected from 134 literatures, we verified the feasibility of predicting those outputs based on soil data, meteorological data, irrigation data and other data. SIW, temperature, and distance from drippers were positively related to soil salinity, while precipitation, irrigation, depth, year, and DAS were negatively correlated with salinity. The ML models have achieved satisfactory performance in both training and testing stages, and the accuracy of the models are higher in predicting cotton yield and ET than that of soil salinity. The stacking ensemble ML could improve model performance. Taking the prediction of soil salinity and EC for example, the R^2 increased by 1.68%-10.52%, *MAE* and *RMSE* decreased by 0.40%-19.22% and 0.09%-17.82%, respectively. When the input parameters of the model are reduced (especially the other data and meteorological data), the accuracy of the model is significantly reduced. Therefore, under the condition of complete input parameters, it is recommended to apply ML algorithm, especially the stacking ensemble ML, to predict soil salt dynamic, ET and cotton yield in drip irrigated cotton fields in arid regions of the world.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

ZJ provided the idea, collected the data, run the model, writing the original draft. SY provided the funding and writing- review and editing. SD and QP conducted part of the field data and validated the draft. PS, MA, and YX helped revised the paper and provided suggestions. JZ provided the methodology and edited the paper. GW provided the funding. All authors contributed to the article and approved the submitted version.

Funding

We thank the China Scholarship Council (CSC) for providing a scholarship (202206710073) to Zewei Jiang. This work was supported by the Fundamental Research Funds for the Central

Universities (B220203009), the Postgraduate Research & Practice Program of Jiangsu Province (KYCX22_0669), the Water Conservancy Science and Technology Project of Jiangxi Province (201921ZDKT06, 202124ZDKT09), the National Natural Science Foundation of China (51879076), the Fundamental Research Funds for the Central Universities (B210204016), Science & Technology Specific Projects in Agricultural High-tech Industrial Demonstration Area of the Yellow River Delta, Grant No: 2022SZX01.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1143462/full#supplementary-material

TABLE S1
The original dataset collected in this study.

References

Ashapure, A., Jung, J., Chang, A., Oh, S., Yeom, J., Maeda, M., et al. (2020). Developing a machine learning based cotton yield estimation framework using multi-temporal uas data. *Isprs J. Photogrammetry Remote Sens.* 169, 180–194. doi: 10.1016/j.isprsjprs.2020.09.015

Bhattacharya, B. K., Mallick, K., Nigam, R., Dakore, K., and Shekh, A. M. (2011). Efficiency based wheat yield prediction in a semi-arid climate using surface energy budgeting with satellite observations. *Agric. For. Meteorol.* 151, 1394–1408. doi: 10.1016/j.agrformet.2011.06.002

Boursianis, A. D., Papadopoulou, M. S., Gotsis, A., Wan, S., Sarigiannidis, P., Nikolaidis, S., et al. (2021). Smart irrigation system for precision agriculture–the AREThOU5A IoT platform. *IEEE Sens. J.* 21, 17539–17547. doi: 10.1109/JSEN.2020.3033526

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Castrignanò, A., Wong, M. T. F., Stelluti, M., De Benedetto, D., and Sollitto, D. (2012). Use of EMI, gamma-ray emission and GPS height as multi-sensor data for soil characterisation. *Geoderma* 175-176, 78–89. doi: 10.1016/j.geoderma.2012.01.013

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (San Francisco, California, USA: Association for Computing Machinery), 785–794. doi: 10.1145/2939672.2939785

Chou, J., Tsai, C., Pham, A., and Lu, Y. (2014). Machine learning in concrete strength simulations: multi-nation data analytics. *Construct. Building Mater.* 73, 771–780. doi: 10.1016/j.conbuildmat.2014.09.054

Devkota, K.P., Devkota, M., Rezaei, M., and Oosterbaan, R. (2022). Managing salinity for sustainable agricultural production in salt-affected soils of irrigated drylands. *Agric. Systems.* 198, 103390. doi: 10.1016/j.agsy.2022.103390

Dong, H., Kong, X., Li, W., Tang, W., and Zhang, D. (2010). Effects of plant density and nitrogen and potassium fertilization on cotton yield and uptake of major nutrients in two fields with varying fertility. *Field Crops Res.* 119, 106–113. doi: 10.1016/j.fcr.2010.06.019

Dong, J., Wu, L., Liu, X., Li, Z., Gao, Y., Zhang, Y., et al. (2020). Estimation of daily dew point temperature by using bat algorithm optimization based extreme learning machine. *Appl. Thermal Eng.* 165, 114569. doi: 10.1016/j.applthermaleng.2019.114569

Dramsch, J. S. (2020). 70 years of machine learning in geoscience in review. *Adv. Geophys.* 61, 1–55. doi: 10.1016/bs.agph.2020.08.002

Du, Z., Yang, L., Zhang, D., Cui, T., He, X., Xiao, T., et al. (2022). Corn variable-rate seeding decision based on gradient boosting decision tree model. *Comput. Electron. Agric.* 198, 107025. doi: 10.1016/j.compag.2022.107025

Dumont, B., Basso, B., Leemans, V., Bodson, B., Destain, J. P., and Destain, M. F. (2015). A comparison of within-season yield prediction algorithms based on crop model behaviour analysis. *Agric. For. Meteorol.* 204, 10–21. doi: 10.1016/j.agrformet.2015.01.014

El Bilali, A., Taleb, A., and Brouziyne, Y. (2021). Groundwater quality forecasting using machine learning algorithms for irrigation purposes. *Agric. Water Manage.* 245, 106625. doi: 10.1016/j.agwat.2020.106625

Elnesr, M. N., and Alazba, A. A. (2017). Simulation of water distribution under surface dripper using artificial neural networks. *Comput. Electron. Agric.* 143, 90–99. doi: 10.1016/j.compag.2017.10.003

Fan, J., Yue, W., Wu, L., Zhang, F., Cai, H., Wang, X., et al. (2018). Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of china. *Agric. For. Meteorol.* 263, 225–241. doi: 10.1016/j.agrformet.2018.08.019

Filippi, P., Whelan, B. M., Vervoort, R. W., and Bishop, T. F. A. (2022). Identifying crop yield gaps with site-and season-specific data-driven models of yield potential. *Precision Agricul.* 23, 578–601. doi: 10.1007/s11119-021-09867-y

Friedman, J. H. (2002). Stochastic gradient boosting. *Comput. Stat Data Anal.* 38, 367–378. doi: 10.1016/S0167-9473(01)00065-2

Gorji, T., Yildirim, A., Hamzehpour, N., Tanik, A., and Sertel, E. (2020). Soil salinity analysis of urmia lake basin using landsat-8 OLI and sentinel-2A based spectral indices and electrical conductivity measurements. *Ecol. Indic.* 112, 106173. doi: 10.1016/j.ecolind.2020.106173

Gou, J., Sun, L., Yu, B., Wan, S., Ou, W., and Yi, Z. (2022). Multi-level attention-based sample correlations for knowledge distillation. *IEEE Trans. Ind. Inf.* 19, 1–11. doi: 10.1016/10.1109/TII.2022.3209672

Gu, J., Liu, S., Zhou, Z., Chalov, S. R., and Zhuang, Q. (2022). A stacking ensemble learning model for monthly rainfall prediction in the taihu basin, china. *Water* 14, 492. doi: 10.3390/w14030492

He, C., Liu, J., Xu, F., Zhang, T., Chen, S., Sun, Z., et al. (2020). Improving solar radiation estimation in china based on regional optimal combination of meteorological factors with machine learning methods. *Energy Conversion Manage.* 220, 113111. doi: 10.1016/j.enconman.2020.113111

Hu, Q., Yang, Y., Han, S., Yang, Y., Ai, Z., Wang, J., et al. (2017). Identifying changes in irrigation return flow with gradually intensified water-saving technology using HYDRUS for regional water resources management. *Agric. Water Manage.* 194, 33–47. doi: 10.1016/j.agwat.2017.08.023

Huang, G., Wu, L., Ma, X., Zhang, W., Fan, J., Yu, X., et al. (2019). Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *J. Hydrol.* 574, 1029–1041. doi: 10.1016/j.jhydrol.2019.04.085

Ibrahim, I. A. E., Yehia, W. M. B., Saleh, F. H., Lamlom, S. F., Ghareeb, R. Y., El-Banna, A. A. A., et al. (2022). Impact of plant spacing and nitrogen rates on growth characteristics and yield attributes of Egyptian cotton (Gossypium barbadense l.). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.916734

Ivushkin, K., Bartholomeus, H., Bregt, A. K., and Pulatov, A. (2017). Satellite thermography for soil salinity assessment of cropped areas in Uzbekistan. *Land Degradation Dev.* 28, 870–877. doi: 10.1002/ldr.2670

Jiang, Z., Yang, S., Chen, X., Pang, Q., Xu, Y., Qi, S., et al. (2022c). Controlled release urea improves rice production and reduces environmental pollution: a research based on meta-analysis and machine learning. *Environ. Sci. pollut. Res.* 29, 3587–3599. doi: 10.1007/s11356-021-15956-2

Jiang, Z., Yang, S., Liu, Z., Xu, Y., Shen, T., Qi, S., et al. (2022b). Can ensemble machine learning be used to predict the groundwater level dynamics of farmland under future climate: a 10-year study on huaibei plain. *Environ. Sci. pollut. Res.* 29, 44653–44667. doi: 10.1007/s11356-022-18809-8

Jiang, Z., Yang, S., Liu, Z., Xu, Y., Xiong, Y., Qi, S., et al. (2022a). Coupling machine learning and weather forecast to predict farmland flood disaster: a case study in Yangtze river basin. *Environ. Model. Softw.* 155, 105436. doi: 10.1016/j.envsoft.2022.105436

Jiang, Z., Yang, S., Smith, P., and Pang, Q. (2023b). Development of DNDC-BC model to estimate greenhouse gas emissions from rice paddy fields under combination of biochar and controlled irrigation management. *Geoderma* 433, 116450. doi: 10.1016/j.geoderma.2023.116450

Jiang, Z., Yang, S., Smith, P., et al. (2023a). Ensemble machine learning for modeling greenhouse gas emissions at different time scales from irrigated paddy fields. *Field Crops Res.* 292, 108821. doi: 10.1016/j.fcr.2023.108821

Jones, E. J., Bishop, T. F. A., Malone, B. P., Hulme, P. J., Whelan, B. M., and Filippi, P. (2022). Identifying causes of crop yield variability with interpretive machine learning. *Comput. Electron. Agric.* 192, 106632. doi: 10.1016/j.compag.2021.106632

Karandish, F., and Šimůnek, J. (2016). A comparison of numerical and machine-learning modeling of soil water content with limited input data. *J. Hydrol.* 543, 892–909. doi: 10.1016/j.jhydrol.2016.11.007

Kaur Dhaliwal, J., Panday, D., Saha, D., Lee, J., Jagadamma, S., Schaeffer, S., et al. (2022). Predicting and interpreting cotton yield and its determinants under long-term conservation management practices using machine learning. *Comput. Electron. Agric.* 199, 107107. doi: 10.1016/j.compag.2022.107107

Kisi, O. (2016). Modeling reference evapotranspiration using three different heuristic regression approaches. *Agric. Water Manage.* 169, 162–172. doi: 10.1016/j.agwat.2016.02.026

Li, S., Luo, W., Jia, Z., Tang, S., and Chen, C. (2018). The effect of natural rainfall on salt leaching under watertable management. *Land Degradation & Dev.* 29, 1953–1961. doi: 10.1002/ldr.2956

Liu, Y., Ao, C., Zeng, W., Srivastava, A. K., Gaiser, T., Wu, J., et al. (2021). Simulating water and salt transport in subsurface pipe drainage systems with HYDRUS-2D. *J. Hydrol.* 592, 125823. doi: 10.1016/j.jhydrol.2020.125823

Liu, Y., Xia, X., Yao, L., Jing, W., Zhou, C., Huang, W., et al. (2020). Downscaling satellite retrieved soil moisture using regression tree-based machine learning algorithms over southwest france. *Earth Space Sci.* 7, e2020EA001267. doi: 10.1029/2020EA001267

Livera, A., Theristis, M., Makrides, G., Ransome, S., Sutterlueti, J., Georghiou, G. E., et al. (2019). "Optimal development of location and technology independent machine learning photovoltaic performance predictive models," in *2019 IEEE 46th Photovoltaic Specialists Conference (PVSC)* Chicago, IL, USA. 1270–1275 (IEEE). doi: 10.1109/PVSC40753.2019.8980474

Masasi, B., Taghvaeian, S., Gowda, P. H., Marek, G., and Boman, R. (2020). Validation and application of AquaCrop for irrigated cotton in the southern great plains of US. *Irrigation Sci.* 38, 593–607. doi: 10.1007/s00271-020-00665-4

Mattar, M. A. (2018). Using gene expression programming in monthly reference evapotranspiration modeling: A case study in Egypt. *Agric. Water Manage.* 198, 28–38. doi: 10.1016/j.agwat.2017.12.017

Narjary, B., Meena, M. D., Kumar, S., Kamra, S. K., Sharma, D. K., and Triantafilis, J. (2019). Digital mapping of soil salinity at various depths using an EM38. *Soil Use Manage.* 35, 232–244. doi: 10.1111/sum.12468

Ning, S., Zhou, B., Shi, J., and Wang, Q. (2021). Soil water/salt balance and water productivity of typical irrigation schedules for cotton under film mulched drip irrigation in northern xinjiang. *Agric. Water Manage.* 245, 106651. doi: 10.1016/j.agwat.2020.106651

Nouri, A., Lee, J., Yoder, D. C., Jagadamma, S., Walker, F. R., Yin, X., et al. (2020). Management duration controls the synergistic effect of tillage, cover crop, and nitrogen rate on cotton yield and yield stability. *Agriculture Ecosyst. Environ.* 301, 107007. doi: 10.1016/j.agee.2020.107007

Obsie, E. Y., Qu, H., and Drummond, F. (2020). Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. *Comput. Electron. Agric.* 178, 105778. doi: 10.1016/j.compag.2020.105778

Oster, J. D. (1994). Irrigation with poor quality water. *Agric. Water Manage.* 25, 271–297. doi: 10.1016/0378-3774(94)90064-7

Peng, J., Ji, W., Ma, Z., Li, S., Chen, S., Zhou, L., et al. (2016). Predicting total dissolved salts and soluble ion concentrations in agricultural soils using portable visible near-infrared and mid-infrared spectrometers. *Biosyst. Eng.* 152, 94–103. doi: 10.1016/j.biosystemseng.2016.04.015

Pham, K., and Won, J. (2022). Enhancing the tree-boosting-based pedotransfer function for saturated hydraulic conductivity using data preprocessing and predictor importance using game theory. *Geoderma* 420, 115864. doi: 10.1016/j.geoderma.2022.115864

Phogat, V., Mahadevan, M., Skewes, M., and Cox, J. W. (2012). Modelling soil water and salt dynamics under pulsed and continuous surface drip irrigation of almond and implications of system design. *Irrigation Sci.* 30, 315–333. doi: 10.1007/s00271-011-0284-2

Phonphan, W., Tripathi, N. K., Tipdecho, T., and Eiumnoh, A. (2014). Modelling electrical conductivity of soil from backscattering coefficient of microwave remotely sensed data using artificial neural network. *Geocarto. Int.* 29, 842–859. doi: 10.1080/10106049.2013.868040

Qi, G., Chang, C., Yang, W., and Zhao, G. (2022). Soil salinity inversion in coastal cotton growing areas: an integration method using satellite-ground spectral fusion and satellite-UAV collaboration. *Land Degradation Dev.* 33, 2289–2302. doi: 10.1002/ldr.4287

Ren, D., Wei, B., Xu, X., Engel, B., Li, G., Huang, Q., et al. (2019). Analyzing spatiotemporal characteristics of soil salinity in arid irrigated agro-ecosystems using integrated approaches. *Geoderma* 356, 113935. doi: 10.1016/j.geoderma.2019.113935

Rodriguez-Sanchez, J., Li, C., and Paterson, A. H. (2022). Cotton yield estimation from aerial imagery using machine learning approaches. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.870181

Shahhosseini, M., Martinez-Feria, R. A., Hu, G., and Archontoulis, S. V. (2019). Maize yield and nitrate loss prediction with machine learning algorithms. *Environ. Res. Lett.* 14, 124026. doi: 10.1088/1748-9326/ab5268

Su, F., Wu, J., Wang, D., Zhao, H., Wang, Y., He, X., et al. (2022). Moisture movement, soil salt migration, and nitrogen transformation under different irrigation conditions: field experimental research. *Chemosphere* 300, 134569. doi: 10.1016/j.chemosphere.2022.134569

Sun, J., Kang, Y., Wan, S., Hu, W., Jiang, S., and Zhang, T. (2012). Soil salinity management with drip irrigation and its effects on soil hydraulic properties in north China coastal saline soils. *Agric. Water Manage.* 115, 10–19. doi: 10.1016/j.agwat.2012.08.006

Vermeulen, D., and Van Niekerk, A. (2017). Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates. *Geoderma* 299, 1–12. doi: 10.1016/j.geoderma.2017.03.013

Wan, S., and Goudos, S. (2020). Faster r-CNN for multi-class fruit detection using a robotic vision system. *Comput. Networks* 168, 107036. doi: 10.1016/j.comnet.2019.107036

Wang, X., Liu, S., Zhang, S., Li, H., Maimaitiaili, B., Feng, G., et al. (2018). Localized ammonium and phosphorus fertilization can improve cotton lint yield by decreasing rhizosphere soil pH and salinity. *Field Crops Res.* 217, 75–81. doi: 10.1016/j.fcr.2017.12.011

Wang, N., Peng, J., Xue, J., Zhang, X., Huang, J., Biswas, A., et al. (2022). A framework for determining the total salt content of soil profiles using time-series sentinel-2 images and a random forest-temporal convolution network. *Geoderma* 409, 115656. doi: 10.1016/j.geoderma.2021.115656

Wang, F., Yang, S., Wei, Y., Shi, Q., and Ding, J. (2021). Characterizing soil salinity at multiple depth using electromagnetic induction and remote sensing data with random forests: a case study in tarim river basin of southern xinjiang, China. *Sci. Total Environ.* 754, 142030. doi: 10.1016/j.scitotenv.2020.142030

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks* 5, 241–259. doi: 10.1016/S0893-6080(05)80023-1

Wu, Y., Zhang, L., Gu, Z., Lu, H., and Wan, S. (2023). Edge-AI-Driven framework with efficient mobile network design for facial expression recognition. *ACM Trans. Embedded Comput. Syst.* 22, 1–17. doi: 10.1145/3587038

Xiao, C., Ji, Q., Chen, J., Zhang, F., Li, Y., Fan, J., et al. (2023). Prediction of soil salinity parameters using machine learning models in an arid region of northwest china. *Comput. Electron. Agric.* 204, 107512. doi: 10.1016/j.compag.2022.107512

Xie, X., Wu, T., Zhu, M., Jiang, G., Xu, Y., et al. (2021). Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land. *Ecol. Indic.* 120, 106925. doi: 10.1016/j.ecolind.2020.106925

Xu, W. C., Chen, P. C., Zhan, Y. L., Chen, S. D., Zhang, L., Lan, Y., et al. (2021). Cotton yield estimation model based on machine learning using time series UAV remote sensing data. *Int. J. Appl. Earth Observ. Geoinf.* 104, 102511. doi: 10.1016/j.jag.2021.102511

Xu, T., Guo, Z., Liu, S., He, X., Meng, Y., Xu, Z., et al. (2018). Evaluating different machine learning methods for upscaling evapotranspiration from flux towers to the regional scale. *J. Geophys. Res. Atmos.* 123, 8674–8690. doi: 10.1029/2018JD028447

Xu, K., Metzger, S., and Desai, A. R. (2017). Upscaling tower-observed turbulent exchange at fine spatio-temporal resolution using environmental response functions. *Agric. For. Meteorol.* 232, 10–22. doi: 10.1016/j.agrformet.2016.07.019

Yang, G., Li, F., Tian, L., He, X., Gao, Y., Wang, Z., et al. (2020). Soil physicochemical properties and cotton (Gossypium hirsutum l.) yield under brackish water mulched drip irrigation. *Soil Tillage Res.* 199, 104592. doi: 10.1016/j.still.2020.104592

Zhang, C., Brodylo, D., Rahman, M., Rahman, M. A., Douglas, T. A., and Comas, X. (2022). Using an object-based machine learning ensemble approach to upscale evapotranspiration measured from eddy covariance towers in a subtropical wetland. *Sci. Total Environ.* 831, 154969. doi: 10.1016/j.scitotenv.2022.154969

Zhang, Y., Zhang, F., Jin, Y., Cen, Y., Voronin, V., and Wan, S. (2023). Local correlation ensemble with GCN based on attention features for cross-domain person re-ID. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 1–22. doi: 10.1145/3542820

Zhao, W., Zhou, C., Zhou, C., Ma, H., and Wang, Z. (2022). Soil salinity inversion model of oasis in arid area based on UAV multispectral remote sensing. *Remote Sens.* 14, 1804. doi: 10.3390/rs14081804

Zong, R., Han, Y., Tan, M., Zou, R., and Wang, Z. (2022). Migration characteristics of soil salinity in saline-sodic cotton field with different reclamation time in non-irrigation season. *Agric. Water Manage.* 263, 107440. doi: 10.1016/j.agwat.2021.107440