Agricultural Systems 218 (2024) 103955

Contents lists available at ScienceDirect

Agricultural Systems

journal homepage: www.elsevier.com/locate/agsy

Optimizing machine learning for agricultural productivity: A novel approach with RScv and remote sensing data over Europe

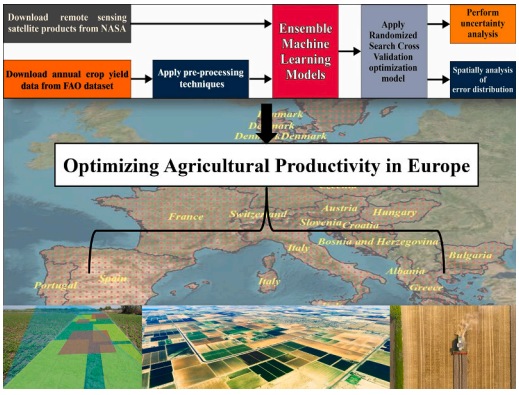
Seyed Babak Haji Seyed Asadollah [a,b,*], Antonio Jodar-Abellan [c], Miguel Ángel Pardo [a]

[a] Department of Civil Engineering, University of Alicante, 03690 Alicante, Spain
[b] Department of Environmental Resources Engineering, State University of New York, College of Environmental Science and Forestry, 1 Forestry Drive, Syracuse, NY 13210, United States
[c] Soil and Water Conservation Research Group, Centre for Applied Soil Science and Biology of the Segura, Spanish National Research Council (CEBAS-CSIC), Campus de Espinardo 30100, P.O. Box 164, Murcia, Spain.

HIGHLIGHTS

- Application of a novel meta-heuristic optimiser in machine learning algorithms was assessed.
- Applied methodology was used to predict crop yield of four major crop types using remote sensing data.
- Crop yield records were obtained from 20 European countries over the past 20 years.
- Assessing different ensemble algorithms, the Ada-boost shows the highest accuracy.
- Predictive models show better accuracy in Wheat compare to Barley, Oats and Rye.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Zhao Zhang

Keywords:
Crop yield
Remote sensing
Machine learning
Randomized search
Agricultural prediction

ABSTRACT

CONTEXT: Accurate estimating of crop yield is crucial for developing effective global food security strategies which can lead to reduce of hunger and more sustainable development. However, predicting crop yields is a complex task as it requires frequent monitoring of many weather and socio-economic factors over an extended period. Satellite remote sensing products have become a reliable source for climate-based variables. They are easier to obtain and provide detailed spatial and temporal coverage.

OBJECTIVE: The aim of this study is to assess the effectiveness of implement a novel optimization algorithm, called Randomized Search cross validation (RScv), on various machine learning algorithms and measure the prediction accuracy enhancement.

METHODS: Annual yields of four crops (Barley, Oats, Rye, and Wheat) were predicted across 20 European countries for 20 years (2000–2019). Two NASA missions, namely GPCP and GLDAS satellites, provided us with climate- and soil-based input variables. Those variables were employed as the input of four ensemble Machine Learning (ML) algorithms (Ada-Boost (AB), Gradient Boost (GB), Random Forest (RF) and Extra Tree (ET)) which are faster and more adoptable compare to classic AI algorithms.

* Corresponding author.
E-mail addresses: sbhs1@alu.ua.es (S.B.H.S. Asadollah), ajodar@cebas.csic.es (A. Jodar-Abellan), mpardo@ua.es (M.Á. Pardo).

https://doi.org/10.1016/j.agsy.2024.103955
Received 13 January 2024; Received in revised form 2 April 2024; Accepted 14 April 2024
Available online 29 April 2024

RESULTS AND CONCLUSIONS: Main results show that applying RScv improves the prediction ability of all ML models over the four crops. In particular, the RScv-AB reaches the overall highest accuracy for predicting yields ($R^2_{max} = 0.9$). Spatial evaluation of predicting errors depicts that the proposed models were more shifted toward underestimation. An uncertainty analysis was also carried out which shows that applying ML algorithms creates higher and lowers uncertainty in Barley and Wheat respectively.
SIGNIFICANCE: Considering the robustness of the optimised ML models and the global coverage of remote sensing data, our current methodology demonstrates great transferability and can be applied in other regions across the globe with higher temporal extents. In addition, this tool could be beneficial to decision makers in various sectors to improve the water allocations, deal with climate change effects and keep sustainable agricultural development.

1. Introduction

Food security denotes the access of individuals to safe, nutritious and sufficient food, which can satisfy their dietary requirement to preserve a healthy and active life style (Van Wart et al., 2013; Boix-Fayos and de Vente, 2023; Luo et al., 2023). It is a vital criterion which has been directly noted in two disciplines out of seventeen UN's Sustainable Development Goals (SDGs). As a subsection of zero hunger goal of SGDs, a strategic plan for food security and agricultural productivity is need to be devised so that by 2030 the overall crop yields become double which will make the world one step closer to end the hunger (Han and Niles, 2023; Konefal et al., 2023). Considering the importance of this factor, the rapid growth of the world population, increase the need for larger amount of food sources is expected to be nearly 100% in the not-distant future (Roser et al., 2013; Hu et al., 2019). This significant surge will challenge the current statues of food security, and may lead to socio-economic crises, especially in not developed countries (Fukase and Martin, 2020; Cavan et al., 2023). In addition, the climate change phenomena imposes higher side effects to agricultural processes as the weather extremities, such as floods or droughts, become much more frequent which directly affect the crops yields (Challinor et al., 2014; Kang et al., 2009; Derdour et al., 2023). Considering all these detrimental issues, the development of tools, which can give accurate estimations of crops yields, plays an important role in developing better food strategies and aiding decision-makers across the world (Spanaki et al., 2022; Alexandridis et al., 2023).

In the early 20th century, several numerical and mathematical regression models were developed to predict the crop yields in different parts of the world (Dourado-Neto et al., 1998; Hochmuth et al., 1998; Xevi et al., 1996). While these approaches showed a good range of accuracy, they also came with series disadvantages which limited their overall application, especially to agricultural-based issues (Hunink et al., 2017; Mohamadou et al., 2020). Not only these models could not deal with non-linearity or complex data structures (Zhang et al., 2020), they also lack the ability to be adapted to large-scale datasets (Khairunniza-Bejo et al., 2014; Lencastre et al., 2023).

As an alternative, in recent years, Artificial Intelligence (AI) models have shown their potential as a valuable research tool in various fields, including sensitivity evaluation, classification, and regression analysis (AghaKouchak et al., 2022; Kumar and Kumar, 2022). These models can learn the non-linearity among engaged parameters and reduce the associated noises (Asadollah et al., 2023). On the contrary to old models, the AI algorithms perform significantly better in term of complexity and large size datasets, as they learn more pattern and features (Davenport, 2019; Lu, 2019). AI models also depict the ability to learn from data and improve their performance over time, saving time and resources, whereas conventional models are considered static approaches and less efficient (Zhang et al., 2020; Sarker, 2022).

The agricultural sector has also embraced AI and leveraged its capabilities for a multitude of purposes, being yield estimation the most extensively used application (Van Wart et al., 2013; Abbaspour-Gilandeh et al., 2022; Poornappriya and Gopinath, 2022). Crop yield calculation involves many criteria and parameters, making the application of AI an effective and reliable analytical technique (Khairunniza-Bejo et al., 2014). Researchers have repeatedly used classic AI algorithms, such as Artificial neural network (ANN) and Adaptive neuro-fuzzy inference systems (ANFIS), as predictive tools for this specific area of research (Naderloo et al., 2012; Bi and Hu, 2021; Shastry et al., 2015). While AI models outperform the conventional prediction approaches, they also represent some challenges which may have a negative impact on their final outputs. As the most important issue, the prediction accuracy of these algorithms depends highly on the quantity and quality of input data (Weber et al., 2023). Furthermore, the overfitting issue is also another concern of AI as the model captures all patterns in a dataset which may include the outliers and biased samples (Ying, 2019). All these issues, if not addressed properly, may lead to a decrease in model robustness and its applicability. This intensifies the role of the pre-processing phase in every AI-related simulation so that the prediction model can obtain the best and most reliable output as possible (Mo et al., 2022; Shobha and Nickolas, 2018).

Researchers have demonstrated that newly developed Machine Learning (ML) algorithms are better approaches compared to classic AI algorithms. Specifically, ensemble ML algorithms such as Ada-Boost (AB), Gradient Boosting (GB) and Random Forest (RF) proved to have better accuracy and output reliability as they combine several simple algorithms which not only captures more patterns but substantially reduce the chances of overfitting (Pede and Mountrakis, 2022). Jeong et al. (2016) showed that RF provided more correct predictions of crop yields for wheat, maize and potato on both global and regional scales compared to several classic AI algorithms. Many studies compared the application of ensemble ML models in predicting the yield of different crop types. For instance, Keerthana et al. (2021) assessed the performance of DT, RF, GB, AB, K-Nearest Neighbour (KNN) and Bagging regressors for predicting crop yields using average ground-based rainfall and temperature observations, as well as the value of the pesticide used. The above-mentioned studies used ground-based climate variables which were usually obtained from hydrological stations. While this data sampling approach has very good accuracy, it is time-consuming, costly and more importantly covers a limited regional and temporal ranges. To overcome these issues, application of Remote Sensing (RS) sources has gained significant popularity because of their high temporal extend and global coverage (Leo et al., 2023; Zhang et al., 2020). Among various remote sensing approaches, researchers consider satellite-based technology as the most robust and comprehensive (Sakamoto, 2020; Naimaee et al., 2024).

In a related study, Ju et al. (2021) compared the performance of seven ML algorithms in prediction of corn, paddy rice and soybean in three different countries between 2003 and 2016. Using remote sensing-based vegetation indices extracted from Moderate resolution imaging spectroradiometer (MODIS), county-level land cover distribution and several time-series climate data, the SVM shows higher accuracy compare to ANN, DT, RF, Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). Likewise, Huber et al. (2022) used MODIS products time series of parameters such as reflectance bands, temperature and vapour pressure to forecast soybean yield in the United States (US). The satellite-based image products were processed and

subsequently used as inputs for AI algorithms, including CNN, LSTM and XGBoost. Dang et al. (2021) employed different indices obtained from remote sensing data as inputs for three data-driven models (SV, RF and Deep Neural Network). This approach predicted the autumn crop yield in China and compared the accuracy of these models. Global Land Data Assimilation System (GLDAS), Tropical Rainfall Measuring Mission (TRMM) and MODIS were the main remote sensing sources, which, based on prediction accuracy evaluators, proven to be excellent inputs as it was also stated by Syed et al. (2008) and Asadollah et al. (2023). Although the use of MODIS products as an input for predictive methods and algorithms has proven to be a beneficial choice (Hunink et al., 2017; Sakamoto, 2020), it was not included in the current study as its temporal range (2012 ~ present) was not fit with the one chosen here (2000–2019).

Based on the reported studies, several satellite missions (such as TRMM and CHIRPS) offer rainfall estimations over the globe. However, the products of NASA's Global Precipitation Climatology Project (GPCP) have improved rainfall observation, so this study used the GPCP precipitation product (Sadeghi et al., 2019; Adler et al., 2020; Kotsias et al., 2020). For other engaged variables such as soil properties and temperature, the different missions of GLDAS has been employed, an excellent example of corresponding ground-based observations (Padhee and Dutta, 2020; Wu et al., 2021; Asadollah et al., 2023).

As one of the major novelties of this study, Randomized Search Cross Validation (RScv) has been used to tune the associated hyperparameters of employed ML algorithms. RScv is considered as a novel optimization algorithm which has shown some successful application in different disciplines (Sharma et al., 2023). Priscilla and Prabha (2020) applied the RScv on XGBoost ensemble algorithm and shows that this experiment improves the ability of fraud detection in credit cards. In a similar study, Vishnu et al. (2023) evaluated the application of RScv on GB, RF and DT to predict the gastric cancer and shows that this optimization technique increases the accuracy especially over RF.

The aim of this research is to assess the efficacy of employing novel RScv optimization algorithm on four machine learning algorithms in predicting the annual crop yield of Barley, Oats, Rye and Wheat over the past 20 years. To reduce the uncertainties associated with yield prediction, we expanded the study region beyond a single country and included 20 countries across Europe. This enables the evaluation of various climate and topographic conditions. Also, considering this number of countries reduces the chance of spatial autocorrelation, which may involves a negative impact on our prediction performance. Since a wide range of parameters were potential ML algorithms inputs, first a collinearity assessment was performed so that our dataset does not affect by such deficiencies. Next, combining these climate variables were used as inputs of four ML algorithms and a prediction task was performed before and after application of RScv optimization. Because of this study only used remote sensing products, a marginal decrease in predictive models' accuracy is expected because of the bias and the measurement-error associated with satellite observations. However, using this integration of machine learning and remote sensing approaches may be significantly helpful for forecasting the annual yield of Barley, Oats, Rye and Wheat as four strategic crop types across the globe (Van Wart et al., 2013). Therefore, our results could assist national, local authorities and farmers in decision-making processes, especially in non-cultivated areas where crop yield needs to be estimated in advance.

2. Materials and methods

2.1. Dataset description

In this study, crop yield data for Barley, Oats, Rye and Wheat was obtained from the Food and Agriculture Organization Corporate Statistical Database (FAOSTAT, 2023). This FAOSTAT dataset includes crop yields observations from 1965 to 2019. The choice of the 20-years ranges from 2000 to 2019 was based on the primary focus of using satellite remote sensing products. Only countries with a complete record in the chosen annual range were selected for analysis, which resulted in 20 countries. The crop yields were evaluated in terms of hectogram per hectare (hg/ha), and Table 1 lists the selected countries and their statistical characteristics.

2.2. Description of satellite sources

2.2.1. GPCP

The Global Precipitation Climatology Project (GPCP) is a significant contributor to collecting satellite-based observations that aid in understanding Earth's energy and water cycles (Sadeghi et al., 2019; Behrangi and Song, 2020). This project operates under the authority of the Global Water and Energy Experiment (GEWEX) Data and Assessment Panel (GDAP). The latest version of GPCP, version 3.2, uses modern merging approaches to gather rainfall records and integrates multiple input datasets (Adler et al., 2020; Asadollah et al., 2023). One dataset utilized is the Goddard Profiling Algorithm (GPROF), which is combined with Special Sensor Microwave Imager/Sounder (SSMI/SSMIS) data for calibration. The spatial resolution of GPCP is $0.5^{\circ} \times 0.5^{\circ}$ over the longitude and latitude axis, and it covers both a daily and monthly basis (NASA, 2023).

2.2.2. GLDAS

In this study, the products of Global Land Data Assimilation System (GLDAS) version 2 have been used, comprising GLDAS version −2.0, −2.1 and − 2.2 (Beaudoing and Rodell, 2016). GLDAS v-2.0 is aligned with Princeton meteorological data with the temporal range of 1948 to 2014, while version 2.1 covers from 2000 to the present day (Wu et al., 2021). These products were extracted from the NASA's database and depict a spatial resolution of $0.25^{\circ} \times 0.25^{\circ}$ (NASA, 2023). The GLDAS product used covers a monthly temporal range from 2000 to 2022, aligning well with the annual period of the current study. This remote sensing source offers a total of 30 climatic and soil variables covering a wide range of components of the hydrological cycle which may have an effect on crop productivity. To ensure consistency with previous research (Syed et al., 2008; Asadollah et al., 2023) and optimise the inputs, several elimination criteria were applied to the 30 parameters, taking into account both theoretical and practical considerations mentioned in the introduction section. As part of imposed restrictions, snow- and radiation-based parameters as well as the heat flux variables were excluded from consideration because of their overall lack of relevancy with agricultural productivity. Furthermore, we selected all

Table 1
Crop yields, in hg/ha, of four major crops calculated in 20 European countries.

Country	Barley (hg/ha)	Oats (hg/ha)	Rye (hg/ha)	Wheat (hg/ha)
Albania	27,476.25	18,795.65	21,334.95	36,469.65
Austria	49,521.70	38,880.50	40,838.10	51,909.70
Belgium	77,296.30	54,950.80	44,052.90	85,862.55
Bosnia	29,822.70	24,843.75	27,864.00	34,204.10
Bulgaria	35,096.10	18,570.30	17,618.45	37,846.70
Croatia	38,683.65	28,003.90	29,588.60	48,724.15
Czech	45,160.70	31,796.70	44,612.15	52,877.75
Denmark	53,884.30	47,949.35	53,882.00	72,991.00
France	63,439.10	44,924.20	46,591.65	69,932.10
Germany	62,309.35	45,942.40	51,609.75	74,945.95
Greece	26,044.20	19,494.10	21,827.90	26,058.55
Hungary	39,398.55	24,899.30	25,337.15	43,811.40
Italy	37,172.35	23,183.65	29,417.70	36,439.10
Netherlands	64,451.45	53,383.70	42,631.70	85,720.30
Poland	33,194.90	25,369.50	25,585.15	41,060.20
Portugal	18,499.80	11,702.30	9248.05	17,212.80
Slovenia	40,926.75	28,141.90	33,044.90	46,610.50
Spain	28,404.30	19,265.50	19,181.20	30,003.55
Sweden	44,087.65	39,399.65	56,461.05	60,788.50
Switzerland	63,277.65	50,992.00	58,931.60	57,606.25

measurements within the depth range of 10 to 40 cm for further analysis for soil parameters that involved measurements at different depths (Beaudoing and Rodell, 2020). Table 2 shows 15 candidate predictor variables and their description obtained from both GPCP and GLDAS sources using NASA's GES DISC platform (NASA, 2023).

2.3. Description of machine learning algorithms

2.3.1. Gradient boosting

Introduced by Friedman (2001), Gradient Boosting Regression (GB) is a technique which combines several Weak Learners (WLs), such as classic DT, and structures one strong prediction model (Lin et al., 2012). In GB, the new WLs are built sequentially aiming to minimise the earlier learner's residual. The GB's prediction approach can be represented by Eq. 1.

$$\widehat{y}_i = \sum_{k=1}^{K} h_k(x_i) \tag{1}$$

Where $\widehat{y}_i$ is the predicted value for i^{th} sample; K is the total number of WLs and $h_k(x_i)$ corresponds to prediction value of k^{th} WL for the i^{th} sample. This iterative generation process is referred to as functional gradient descent (described in Eq. 2), which involves optimizing the simulation at each step by incorporating a new learner if the loss function (L) is lower than in the previous step (Yang et al., 2020). The update of the residuals at each iteration can be described as:

$$r_{im} = -\left[\frac{\partial \mathrm{L}(y_i, F(x_i))}{\partial (F(x_i))}\right] \tag{2}$$

Where r_{im} is the residual which corresponds to i^{th} sample and m^{th} iteration. The L measures the difference between observed (y_i) and prediction ($F(x_i)$) values.

2.3.2. Ada-boost

Ada-Boost Regression (AB) is another boosting ensemble machine learning approach that, on some basis, operates like GB (Shanmugasundar et al., 2021). While GB considers a forward-staged sequential learner-generating approach, AB employs serialization and applied weights to each weak learner (Taufiqurrahman et al., 2020; Wang et al., 2020). Using the Eq. 3, AB method starts by fitting an initial WL to the dataset and calculating the residual error by subtracting the y_i from $\widehat{y}_i$.

$$\widehat{y}_i = \sum_{t=1}^{T} w_t h_t(x_i) \tag{3}$$

Where w_t corresponds to weights assigned to t^{th} weak learner. Subsequently, copies of the initial WLs are created and fitted to the dataset once more. However, the algorithm assigns weights to these newly generated WLs based on the measured residual error as described in Eq. 4 and 5 (Xiao et al., 2019).

$$\vartheta_t = \frac{1}{2} ln\left(\frac{1-\gamma_t}{\gamma_t}\right) \tag{4}$$

$$w_{t+1}(i) = \frac{w_t(i) exp(\vartheta_t y_i h_t(x_i))}{Z_t} \tag{5}$$

In these equations, ϑ_t and γ_t are respectively denoting the assigned weight and weighted error corresponds to t^{th} weak learner. $w_t(i)$ shows the weight of i^{th} sample at the t^{th} iteration and Z_t is the normalization parameter which regulates the $w_t(i)$ distribution. This weight-assigning procedure continues until the error reaches the desired limit and then these corresponding weights are fixated to each developed WL and the regression task is performed (Shanmugasundar et al., 2021).

2.3.3. Random Forest

Random Forest (RF) is one of the most popular ensemble machine learning algorithms employed in both classification and regression tasks (Ahmad et al., 2018). Initially proposed by Breiman (2001a), this algorithm operates based on a statistical concept known as bagging, achieved through a bootstrap aggregating procedure. In this algorithm, the original dataset is divided into several randomly extracted subsamples and a forest of decision trees are trained based on these new shorter datasets (Sakamoto, 2020). Following the principles of bootstrapping, each individual WL undergoes training with multiple trials of the sample dataset. The main idea of this approach comprises combining the trained WLs to build a predictive model that yields improved regression results compared to relying on a single WL. The primary objective of RF is to minimise variance without significantly altering the bias, while maximising prediction accuracy to the greatest extent possible (Shanmugasundar et al., 2021). The overall prediction procedure for RF can be described by the following equation (Eq.6).

$$\widehat{y}_i = \frac{1}{B}\sum_{j=1}^{B} f_j(x_i) \tag{6}$$

In Eq. 6, B corresponds to the total number of trees in the structured RF model and $f_i(x_i)$ shows the prediction result of j^{th} tree over i^{th} observation.

2.3.4. Extra tree

The Extra Tree (ET), which was proposed by Geurts et al. (2006), is considered an extension of RF which operates based on the overall principle described by Eq. 6. While RF and ET operate based on the same mathematical equation, two major differences may lead to better applicability of ET compare to RF (John et al., 2016). First, ET does not use the bagging approach to generate sub-samples from the dataset and uses the whole original dataset to train the trees in the presumed structured forest. Second, in the node split stage, ET chooses the best feature and its corresponding value randomly. Taking these two steps made ET less prone to over fitting and shows better prediction accuracy (Ahmad et al., 2018).

Table 2
Description of the 15 candidate input variables.

Variable name	Description	Unit
Tair	Air Temperature	K
ESoil	Direct Evaporation from Bare Soil	W m^{-2}
Evap	Evapotranspiration	kg m^{-2} s^{-1}
GwRunOff	Baseflow-groundwater runoff	kg m^{-2}
PotEvap	Potential evaporation rate	W m^{-2}
Precip	Precipitation (from GPCP)	mm day^{-1}
Qair	Specific humidity	kg kg^{-1}
RunOff	Storm surface runoff	kg m^{-2}
SoilMoi	Soil moisture at 10 cm depth	kg m^{-2}
SoilTmp	Soil temperature	K
SurPre	Surface air pressure	Pa
SurTmp	Average surface skin temperature	K
Tprecip	Total precipitation rate (from GLDAS)	kg m^{-2} s^{-1}
Tveg	Transpiration	W m^{-2}
Wind	Wind speed	m s^{-1}

2.4. Randomized search cross validation (RScv)

The Randomized Search Cross Validation (RScv) is a novel metaheuristic optimization approach which has become popular in term of improving the machine learning prediction ability (Vishnu et al., 2023). First step of utilizing the RScv is to determine the wanted hyperparameters and their corresponding search ranges. Then the model will structure several sets of combination based on the number of engaged hyperparameters. After this, RScv will initially pick a random sample of hyperparameters from the universe, and then for each hyperparameter employs the same sampling method for extracting a random range from the defined search range. Using these initial setups, the RScv will perform the model training and measures the prediction performance

using cross-validation technique. Considering a user-defined number of iterations, the RScv will repeat the random extraction of hyperparameters and their respective value ranges from the initial universe and calculate the performance metrics. Once the iteration is terminated, based on the highest accuracy the best value for each hyperparameter is extracted and the model is RScv using all samples in the training dataset ignoring the cross-validation technique.

RScv represents some note benefits compare to other similar methods for hyperparameters tuning in ML models. It provides efficiency by requiring fewer iterations compared to Grid Search (GS) models, making it more scalable for exploring high-dimensional hyperparameters spaces or computationally expensive models (Sharma et al., 2023). RScv encourages exploration of the entire hyperparameters space by randomly sampling hyperparameters, potentially leading to better generalization and discovery of promising regions. However, its randomness may cause suboptimal configurations, and it sacrifices exhaustiveness for efficiency. Despite these limitations, RScv is motivated by its suitability for large hyperparameters spaces, limited computational resources, the desire for exploration, and the focus on efficiency in hyperparameters optimization tasks (Vishnu et al., 2023).

2.5. Uncertainty analysis method

Uncertainty is usually originated from sampling errors, preprocessing phase and prediction model structure from which this study only focused on the later source (Ramirez-Villegas et al., 2017). In order to assess the uncertainties associated with the predictive algorithms, for every single crop, first the average and standard deviation of each sample in the testing phase were calculated. Considering these two obtained statistics, by applying Monte-Carlo simulation using the normal distribution, 1000 new yield values (number of rows) were produced for each set of samples (number of columns). After this new dataset was created, each column was sorted separately in ascending mode. Then, the rows were ranked from 1 to 1000 and their probability ($P(i)$) was calculated using Eq. 7.

$$P(i) = (rank\ (i)/1001\ ) \times 100 \tag{7}$$

To perform this uncertainty analysis method, the 95 Percent Prediction Uncertainty (PPU) needs to be determined which is also known as 95% prediction confidence interval. Using eq. (1), the nearest probability to 2.5 and 97.5% were respectively selected as lower and upper boundaries of 95 PPU band. The level of uncertainty was determined using the R-factor criteria which was calculated using Eq. 8.

$$R-factor = \gamma/\sigma \tag{8}$$

In eq. (8), σ is the standard deviation of original values in the testing phase and γ was calculated using Eq. 9.

$$\gamma = \sum_{i=1}^{n} \left(95PPU^{i}_{upper\ boundary} - 95PPU^{i}_{lower\ boundary}\right)/n \tag{9}$$

In eq. (9), n is the total number of samples in the testing phase so γ can be considered as the average between difference of upper and lower boundaries.

2.6. Methodology hierarchy

The annual crop yields of Barley, Oat, Rye and Wheat for the selected 20 countries (Table 1) were extracted from the FAO dataset as dependent variables. To predict the behaviour (the values) of these dependent variables, several environmental and climate components (described in Table 2) were obtained from GPCP and GLDAS satellite sources (NASA, 2023) and considered as predictor or independent variables. Since GPCP has a coarser spatial resolution, its longitude and latitude grid points were considered as the data extraction points in each of the 20 European countries. This grid-based point extraction procedure was successfully applied in Sharafati et al. (2020) for predicting groundwater levels. The location of these points is showed in Fig. 1.

In particular, each point of this grid (Fig. 1) depicts different values of the considered climate parameters (Table 2) which were obtained through two procedures: i) directly extracted from satellite products; or ii) by using interpolation techniques. Likewise, the number of points used in each country is noted in Table 3.

The general flow-chart of the method performed in the present study is showed in Fig. 2 and explained within the following three steps.

Step 1: Pre-processing the data: The Dependent variables (four crop yields) and seven independents (satellite products) were processed into annual basis. In particular, these independent climate variables, with a monthly time resolution, were extracted from GPCP and GLDAS sources in each point of the constructed grid across Europe (Fig. 1). Next, these predictor variables were averaged for each country obtaining an annual basis because of the FAO's crop yield dataset depicts an annual time resolution (Fig. 2). Thus, each of the 20 countries obtained 20 observations (one per year from 2000 to 2019), resulting in an output data frame with 400 rows (samples). Meanwhile, the data frame columns were composed by seven predictor variables (which multicollinearity was assessed before) and the four mentioned dependent variables. Therefore, a dataset with 28 columns was obtained.

Step 2: The constructed dataset was divided into training and testing phases, based on an 80–20 proportion, and then used as the input of four different machine learning algorithms once without and once with RScv application.

Step 3: Output accuracies were evaluated over each crop type and among four algorithms before and after application of RScv. Particularly, each algorithm underwent optimisation and evaluation using four statistical metrics: Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Variance Score (VarS).

Step 4: Since the real-time simulations were performed over a large region and wide temporal ranges, the evaluation of associated prediction uncertainties for each crop was also performed to make the results more applicable.

3. Results

3.1. Feature selection procedure

Selecting the best input features is an essential task to have a robust AI-based prediction model. Numerous studies were focused on different methodologies to extract the best set of features (Mazumder, 2020). For example, Belloni et al. (2012) assessed applying Least Absolute Shrinkage and Selection Operator (Lasso) and post-Lasso for feature engineering purposes. Belloni and Chernozhukov (2013) investigated the application of ordinary least squares (OLS) methods for input selection, whereas Hanke et al. (2023) compared the application of forward stepwise selection (FSS) and Best subset selection (BSS) with Lasso models. In a more comprehensive research, Hastie et al. (2020) employed various sampling alternatives and evaluated the application of BSS, FSS and Lasso showing comparisons based on signal-to-noise. As these methods usually focus on a linear relationship between input and target parameters, this study uses a built-in feature of the proposed ensemble ML algorithms, which deals with the parameter selection. Because of tree-based ML models, an importance rank could be generated, which denotes the ability of each specific parameter to perform the branch split and sub-branch generation in a decision tree. Fig. 3 shows these importance values for each crop type.

As Fig. 3 shows, in all studied crops Tair, ESoil, Precip, Qair, SoilMoi, Tveg and Wind with fluctuating data shows the higher values importance. Therefore, these seven have been selected as the optimised variables for future investigations. To avoid further issues in the prediction process, it is important to evaluate these optimised inputs for the case of multicollinearity (Hunink et al., 2017; Mahato and Gupta, 2022). Thus,

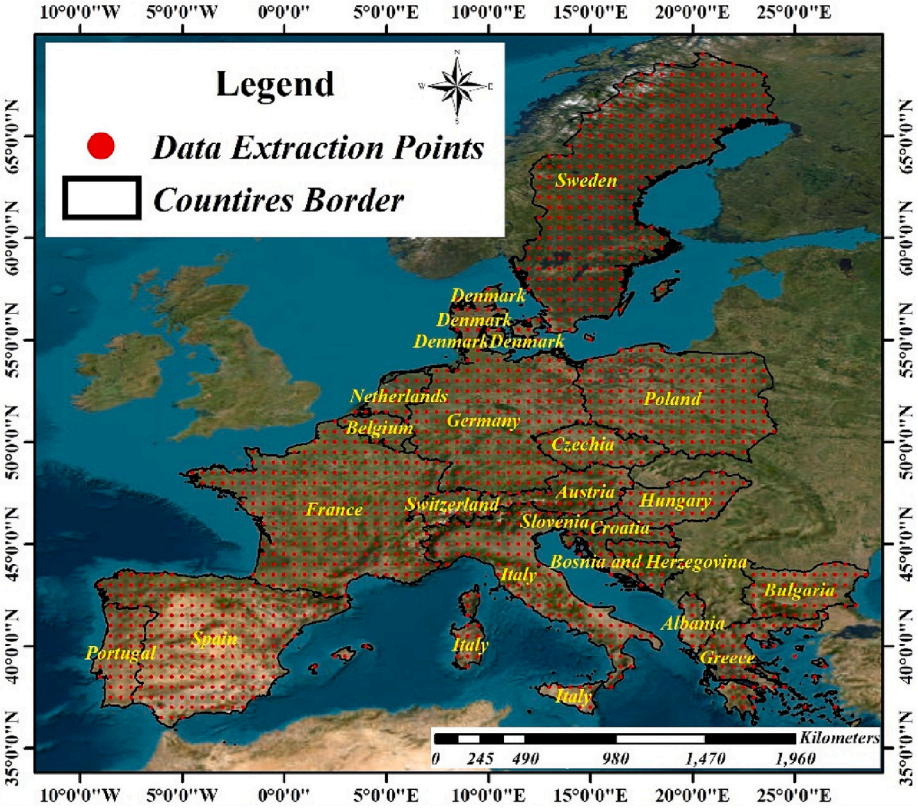


Fig. 1. Spatial resolution of the grid points across the 20 studied countries.

Table 3
Number of extracted points in each country.

Number	Country	Number of points	Number	Country	Number of points
1	Albania	12	11	Germany	176
2	Austria	42	12	France	260
3	Bosnia	23	13	Croatia	26
4	Belgium	15	14	Hungary	44
5	Bulgaria	51	15	Italy	135
6	Switzerland	20	16	Netherlands	19
7	Denmark	27	17	Poland	163
8	Spain	209	18	Portugal	37
9	Greece	59	19	Sweden	316
10	Czechia	40	20	Slovenia	10

Fig. 4 shows the scatter-plot matrix between these variables.

As Fig. 4 shows, both negative and positive linear association can be seen between pairs of variables. While the majority of pairs have low level of multicollinearity, some high correlation can be seen in the scatter-plot matrix. Qair and Tair ($Correlation = 0.81$) and Tair and SoilMoi ($Correlation = -0.72$) respectively present the highest positive and negative cases of multicollinearity. The Variance Inflation Factor (VIF) method, also noted in Fig. 4, was used to evaluate the effect of these pairs on estimation variance because of collinearity. Based on previous studies, VIF over 10 is considered as severe case of multicollinearity, however, more conservative threshold of 5 was also reported in some literatures (O'brien, 2007; Thompson et al., 2017). As Fig. 4 shows, the highest calculated VIF is 2.90 which is far from the most restrictive thresholds, so all the considered input parameters were used in the prediction phase.

3.2. Real-time prediction

In line with the previous section, the input data were evaluated based on multicollinearity which led to seven best variables and exclusion of eight relatively unimportant parameters. This optimised input dataset was then used in four selected machine learning algorithms and the yields of four crops were predicted. Next, the RScv was applied to the ML algorithms separately and the corresponding hyperparameters were put under tunning procedure. For RScv, the number of iterations was set to 100 epochs and 10-fold cross-validation was utilized to generate different subsets of the training dataset. The applied 10-fold cross-validation technique, iteratively divides the input dataset into training and testing batches with 90 and 10% proportions. Thus, the model can perform the learning and validation processes across the entire samples. This increases the data diversity and significantly reduces the possibility of over-fitting (Breiman, 2001b). In this study, the ML algorithms were employed from the Ensemble subcategory of Python's Scikit-learn (Sklearn) library. Table 4 shows the optimised values of each hyperparameter for the studied ML algorithms.

For GB, the "Squared error" loss function was considered being the most suitable choice using the trial-and-error procedure. Since weak learners in Sklearn's GB are considered decision trees, the node-splitting approach is an important factor, so the Friedman's Mean Squared Error (MSE) was selected to handle this procedure. For AB, the Exponential function proved to be the most suitable loss function. Considering the parameters proposed by Sklearn to optimise the performance of RF and ET, the quality of split in tree nodes was evaluated using the Squared Error function (Pedregosa et al., 2011).

As Table 4 shows, the number of hyperparameters differs from one algorithm with another. While based on SKlearn library each of the utilized ML algorithm has more defined hyperparameters, only the most effective ones was reported in Table 4. Other parameters were excluded as they do not improve the accuracy and depict no impact on the dataset of this case study. It is noteworthy to mention that the "n_estimators" are also considered as a regularization criterion which prevents the occurrence of over-fitting in ML algorithms. Since this criterion controls the model's overall complexity, obtaining the highest accuracy (using lower

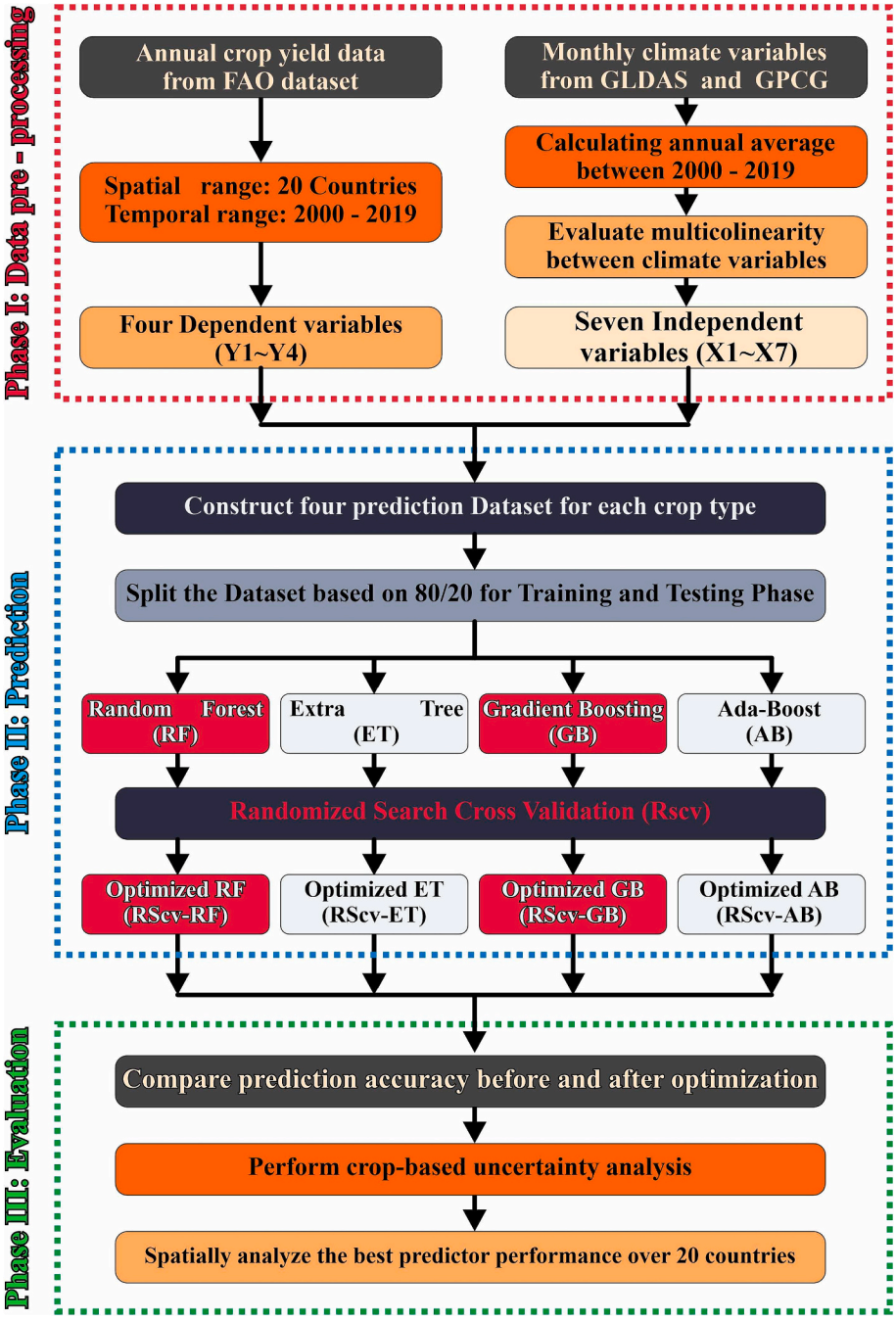


Fig. 2. Overall flowchart of the developed methodology.

rates for "n_estimators") is the main goal of a proper hyper-parameter tuning. Based on Table 4, the range of "n_estimators" for all algorithms are in relatively low ranges, which confirms the applicability of RScv.

Once the hyperparameters were fully tunned, the prediction was again performed and the results of RScv application were compared with classic algorithms. Table 5 shows the crop-based performance evaluation of ML models before and after the application of RScv using R^2 and RMSE metrics.

As Table 5 shows, the RScv could enhance the prediction accuracy over all ML models and crops types at different rates. Based on RMSE and R^2, the highest level of improvement can be seen in RScv-GB with nearly 16 and 12% of enhancement over Oats and Barley. Based on the table outputs, RScv-AB with the average accuracy of $\overline{R^2} = 0.865$ and $\overline{RMSE} = 6283.63$ has the best prediction performance. RScv-ET is the second best algorithm which shows the averaged metrics of $\overline{R^2} = 0.852$ and $\overline{RMSE} = 6685.57$. Furthermore, the RScv-GB with $\overline{R^2} = 0.837$ and $\overline{RMSE} = 7125.87$ shows the worst prediction performance. In term of crops, Barley obtaining $\overline{R^2} = 0.816$ and $\overline{RMSE} = 7795.54$ shows the lowest overall accuracy, while the highest performance was obtained in Wheat with $\overline{R^2} = 0.882$ and $\overline{RMSE} = 7549.35$.

Fig. 5 depicts the prediction performance of optimised ML algorithms using scatter plots of testing phase. As mentioned before, the regression and statistical metrics of R-squared, RMSE, MAE and VarS have also been included in the plots. It is worthy to mention that, to have better comparable visualization of prediction performances, the values have been normalized between 0 and 1. Based on these figures, for

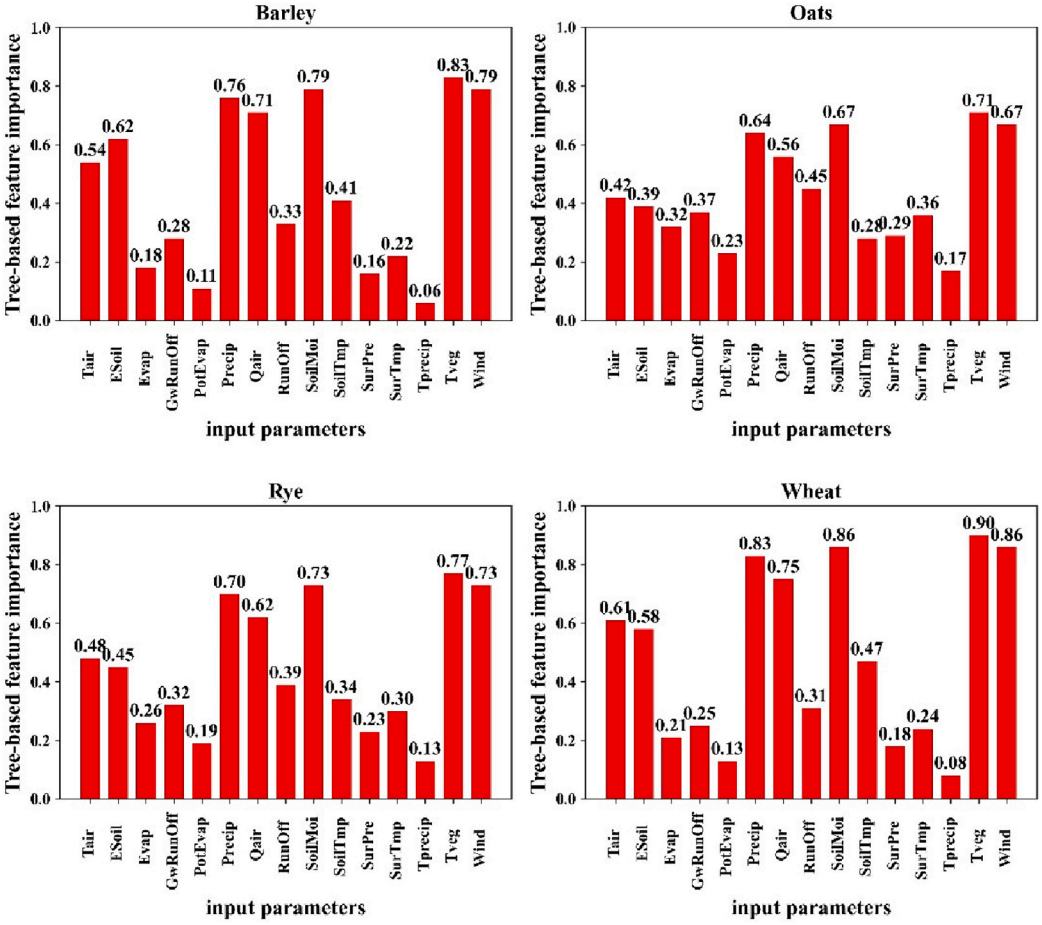


Fig. 3. Comparison between relevancy of different input parameters using tree-based feature importance values for each crop type.

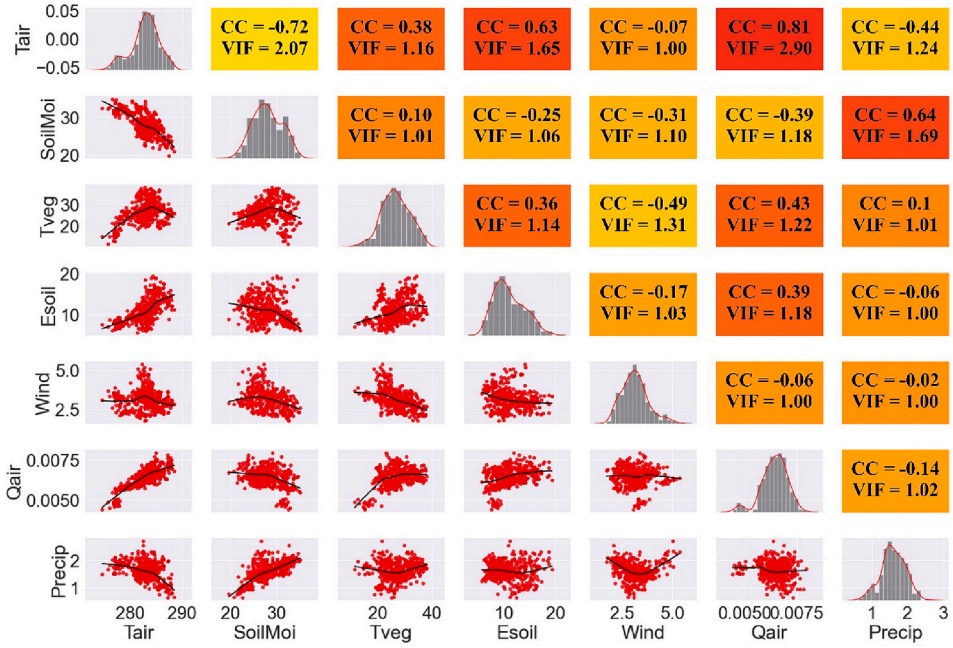


Fig. 4. Scatter-plot matrix between the input variables.

Barley the RScv-ET shows the highest accuracy regarding the $R^2 = 0.84$ *and VarS* $= 0.835$, but considering the RMSE ($= 0.107$) and MAE ($= 0.083$) the RScv-AB takes the upper hand. For Oats, RScv-AB has the best values for $R^2 = 0.875$ *and VarS* $= 0.779$, however, RScv-RF shows better RMSE ($= 0.119$) and MAE ($= 0.093$) metrics. The prediction

Table 4
RSCV-based Optimised hyperparameters for boosting-based models (ABR and GBR) as well as tree-based algorithms (ETR and RFR).

Algorithms	*Hyperparameters*	*Barley*	*Oats*	*Rye*	*Wheat*
AB	***max_depth***	10	10	15	12
	n_estimators	23	92	22	11
	random_state	87	87	45	52
GB	***max_depth***	11	5	9	5
	max_leaf_nodes	12	10	10	12
	min_samples_leaf	21	22	9	22
	n_estimators	118	71	74	54
	random_state	19	6	45	97
ET	***max_depth***	17	16	15	15
	n_estimators	27	11	11	75
	random_state	116	86	68	187
RF	***max_depth***	13	9	14	10
	max_leaf_nodes	32	54	61	96
	n_estimators	44	11	11	84
	random_state	2	75	2	55

output of Rye reveals that the RScv-AB with $R^2 = 0.846, RMSE = 0.111, MAE = 0.082$ *and* $VarS = 0.84$ outperform all its alternative in term of yield prediction. Same results could be seen in Wheat as RScv-AB with $R^2 = 0.901, RMSE = 0.084, MAE = 0.066$ *and* $VarS = 0.88$ represent the highest prediction accuracy.

While scatterplots visualize the linear relationship between the observed and predicted variables, Fig. 6 used Boxplots to show the statistical distribution of the mentioned variables. Boxplots are useful tools for evaluating the variability among the predictive algorithms and their outliers and statistical quartiles. Likewise, the IQR value is an important criterion which was calculated from subtraction of 75% and 25% quartiles and used in this study to compare the similarity of boxplots. Based on this figure, for Barley (Figure6-a) RScv-AB and RScv-ET have nearly same and simultaneously the highest similarity with observed samples($IQR = 27017.75$). For Oats (Figure6-b), RScv-AB ($IQR = 23189.56$) shows the best approximation of observed ($IQR = 23189.56$) while RScv-RF ($IQR = 18932.31$) has the worst estimation. In a similar fashion, for Rye (Figure6-c) again the RScv-AB ($IQR = 23701.25$) mimics the observed samples ($IQR = 24901.25$) being better than its alternative. Similar to two previous crops, RScv-AB predictions ($IQR = 29724.25$) also represent the best estimation of observed data ($IQR = 37613.75$). Based on these graphical and statistical accuracy evaluators, the optimised version of AdaBoost is the better predictive algorithm over all studied crops, so its outputs have been used for further investigations.

3.3. Spatial distribution of prediction error

For spatial analysis of RScv-AB predictions, each crop-based dataset was sorted in ascending mode using the data column. Since for each year from 2000 to 2019 we have a total of 20 countries (Samples), a cross-validation approach was carried out which use 20-fold data split without considering the shuffling approach. Based on this, the sample from each fold (year) was predicted using the pre-trained RScv-AB model and the residuals were calculated by subtracting the predicted from the actual yields and then the measured residuals were averaged over the studied 20 years. Fig. 7 demonstrates the spatial distribution of prediction errors (residuals) across the 20 studied European countries.

Overall, the numerical range of generated errors are much higher in Barley compare to other three crops. However, in Barley, majority of errors are fluctuating around the [−500, +500] except for countries like Netherlands or Albania and Bosnia. In Oats, the distribution of residuals shows more tendency toward negative errors, which denotes higher chance of underestimation, especially in countries like Spain, Austria and Bosnia. Rye shows completely different error distribution compare to Oats, as the trend in errors shift toward positive or overestimation. This change of pattern is more apparently manifested in central European countries and the Norway to the north. Wheat also shows more negative error values but the rate of underestimation is much higher than the obtained in Oats. Overall, based on all four crops, it could be understood that while most RScv-AB residuals are near to 0, more tendency toward underestimation is recognised.

3.4. Uncertainty analysis

While the proposed RScv-AB shows excellent prediction capability, evaluation of uncertainty analysis seems a necessity. First, as described in section 2.5, the average and standard deviation of four utilized algorithms for each testing sample (a total of 80 samples) per crop were calculated. Next, using the mean and variance of 0 and 1 a total of 1000 random numbers were generated and then merged with calculated average and standard deviation so that 1000 new yields could be produced. Using the described method, the upper and lower boundary as well as the $R-factor$ for each crop have been quantified and depicted in Fig. 8.

From this figure, it can be understood that the values of $R-factor$ are nearly in the same numerical range and the observed line is inside the 95PPU in most of the cases. However, the level of uncertainty in Barley ($R-factor = 0.549$) is slightly higher than its alternatives, while Wheat ($R-factor = 0.505$) has the lowest associated uncertainty rate. Larger $R-factor$ denotes wider interval range in which a reference value is

Table 5
Summary of machine learning prediction performance before and after application for RSCV based on R^2 and RMSE accuracy metrics and their level of improvement to each crop yield (expressed in tons/ha) and for all the 20 European countries.

		RMSE			R^2		
Models	***Crops***	***Before***	***After***	***Improvement***	***Before***	***After***	***Improvement***
AB	***Barley***	8067.4	7285.5	9.69%	0.795	0.837	5.21%
	Oats	5780.1	5195.6	10.11%	0.843	0.875	3.77%
	Rye	6955.6	5837.6	16.07%	0.782	0.846	8.17%
	Wheat	8125.2	6815.9	16.11%	0.863	0.901	4.51%
ET	***Barley***	7973.5	7370.2	7.57%	0.806	0.840	4.25%
	Oats	5596.7	5323.5	4.88%	0.851	0.871	2.39%
	Rye	7421.2	6519.0	12.16%	0.748	0.808	8.05%
	Wheat	8100.5	7529.5	7.05%	0.863	0.888	2.88%
GB	***Barley***	9472.1	8100.3	14.48%	0.712	0.795	11.58%
	Oats	6900.1	5824.3	15.59%	0.771	0.836	8.52%
	Rye	6750.1	6527.0	3.30%	0.794	0.807	1.58%
	Wheat	9267.3	8051.9	13.12%	0.810	0.862	6.47%
RF	***Barley***	9107.0	8426.1	7.48%	0.740	0.791	7.01%
	Oats	6005.5	5390.6	10.24%	0.827	0.863	4.33%
	Rye	7033.4	6375.9	9.35%	0.775	0.816	5.23%
	Wheat	8211.3	7800.2	5.01%	0.863	0.876	1.54%

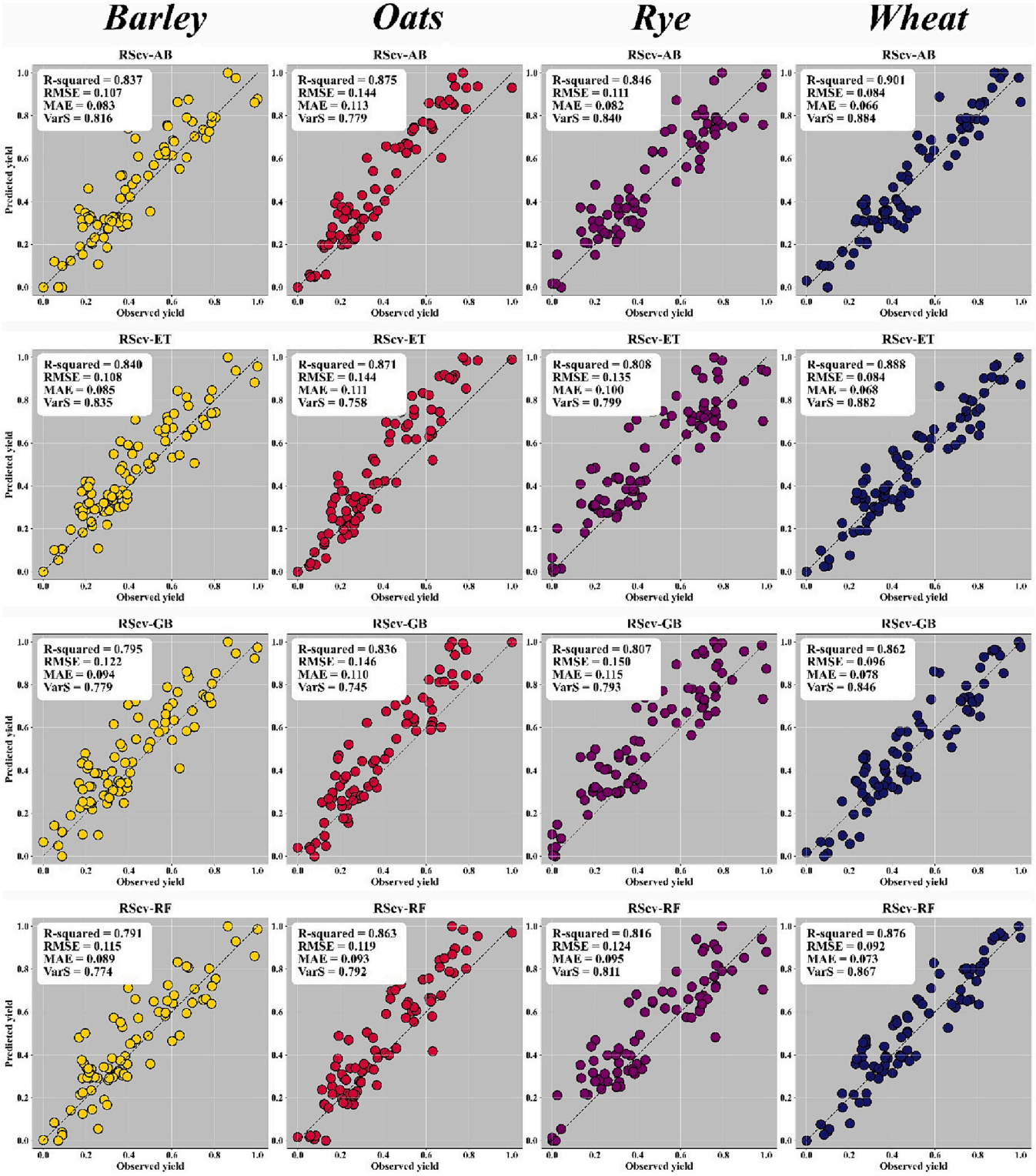


Fig. 5. Scatter plot of observed vs. predicted crop yield (hg/ha) using optimised RScv-ABR, −ETR, -GBR and -RFR.

probably going to lie, while smaller values reflect narrower intervals that encompass more accurate estimation.

4. Discussion

Since the primary aim of this research was focused on the utilisation of ML to estimate annual yields of four major crops (Barley, Oats, Rye, and Wheat) over Europe, it was necessary to compare our results with other related approaches. Thus, Cao et al. (2022) predicted the winter wheat yield by adopting MLR, Support Vector Regression (SVR), RF and XGBoost algorithms in the north of China. Using satellite and observational climate/atmospheric data, the latter model (with R^2 = 0.85) proved to be the best predictor. Li et al. (2021) used the RF to predict different crop yields in China using soil characteristics, climate variables and vegetation indices. The employed RF showed an R^2 of 0.56 as the highest obtained performance metric. In another study, Gopal and

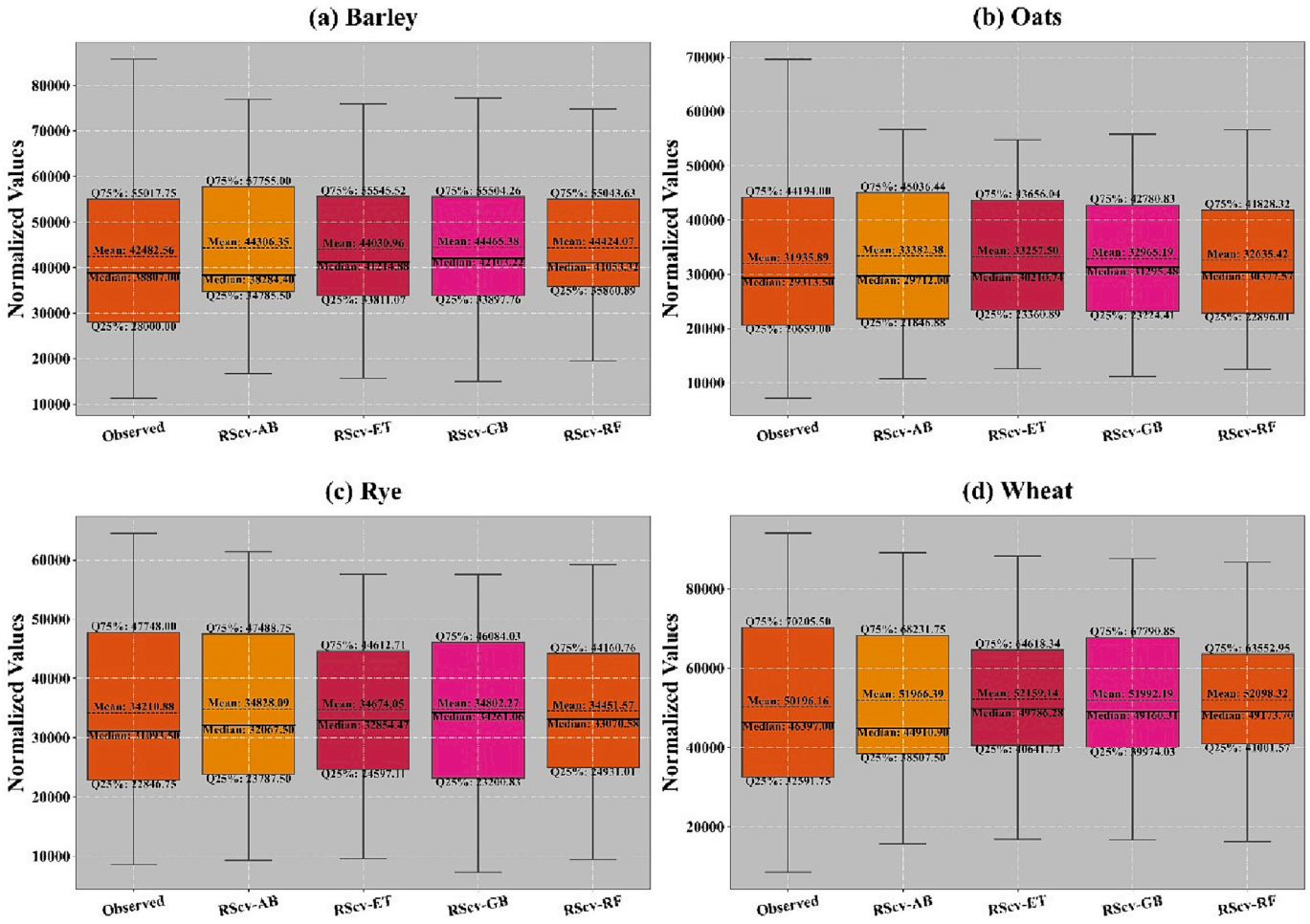


Fig. 6. Descriptive statistical analysis of observed vs. predicted crop yield (hg/ha) using Boxplot.

Bhargavi (2019) evaluated in India the performance of ANN, SVR, KNN and RF and demonstrated that the latter (with $R^2 = 0.88$) was the most accurate simulator of crop yield. In China, similar R^2 and RMSE metrics were reached by Dang et al. (2021) estimating autumn crop yields through the SVR, RF, and DNN models along with several vegetation indices. In an extend area of United States, formed by 13 adjacent states, Huber et al. (2022) predicted annual soybean yields considering Extreme Gradient Boosting (XGBoost), CNN and CNN-LSTM among other methods. During their studied period (from 2017 to 2021) best crop estimations were accomplished by the XGBoost ($R^2 = 0.79$ and RMSE = 4.25) followed by the CNN ($R^2 = 0.66$ and RMSE = 5.69) and the CNN-LSTM ($R^2 = 0.61$ and RMSE = 6.14). Comparing the application of RScv on ML algorithms, which was carried out in this study (with a highest R^2 of 0.901 and an average R^2 of 0.86, Table 5), with the above mentioned literatures, as well as other researches such as Nosratabadi et al. (2020) and Prasad et al. (2021), it should be highlighted the robustness of ML algorithms and the optimisation power of RScv improving the estimation of different crop types. Likewise, although last years the RScv has been coupled with ML algorithms in certain fields of research as econometric assessments or studies of gastric diseases (Priscilla and Prabha, 2020; Sharma et al., 2023; Vishnu et al., 2023), from the author's knowledge currently exist a great lack of scientific studies covering ML and RScv methods on the cultivars yield topic.

Additionally, the ANN, KNN and SVR are the two or three main competitors of the ensemble ML algorithms as have been included in most of the related literatures (Ahmad et al., 2018; Keerthana et al., 2021). To have a more robust evaluation of the proposed RScv-AB, the same dataset was used as the input of these three classic AI algorithms which were also optimised by the RScv approach. These three models were also employed considering the SKlearn machine learning packages. Table 6 tabulated the evaluation of the proposed RScv-AB to these counterparts.

As Table 6 shows, while KNN depicts significant superiority to ANN and SVR, the RScv-AB outperforms them all based on R^2 and RMSE metrics which once again confirms the hybrid models robustness. Furthermore, among the four utilized predictors, Ada-Boost was identified as the best predictive algorithm in our study. Similar trustworthy results were obtained in Wang et al. (2020) where authors also confirmed the power of the Ada-Boost algorithm compare to various number of predictive algorithms such as several linear models, SVR, RF and deep neural network (DNN). Keerthana et al. (2021) also showed that AB is the supreme predictive method compare to DT, RF, GB, K-Nearest Neighbour (KNN) and Bagging regressors.

The represented approach in this study integrates remote sensing and machine learning techniques, which obtained good prediction results and therefore can be utilized as a powerful alternative to the existing crop yield simulation models (Xevi et al., 1996; Van Wart et al., 2013; Zhang et al., 2020). Thus, in line with Ryan (2022), the proposed method could imply positive impacts on the agricultural sector within various aspects, from which the improvement of crop yield estimation at global scale could be the most important. Accurate prediction of crop yields over both spatial and regional concepts is also a principal component for devising a viable large-scale food security strategy (Jeong et al., 2016; Spanaki et al., 2022). Furthermore, the developed model can help the policymakers and managers to enhance the current allocation of resources such as water and fertilizers which will

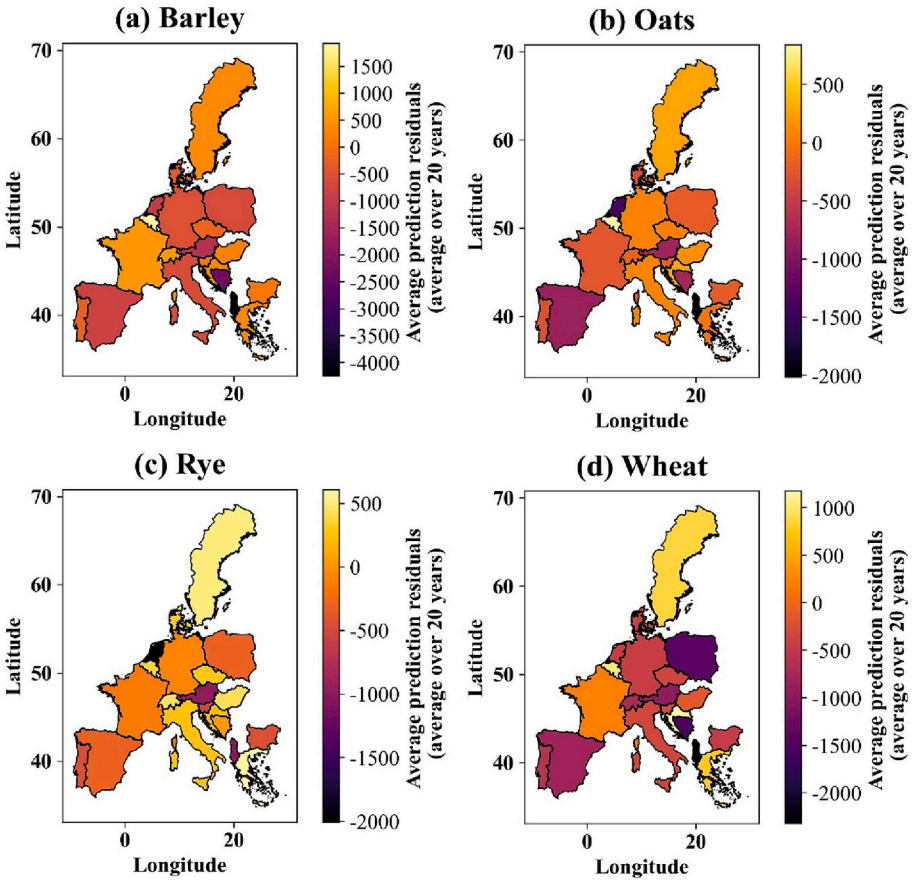


Fig. 7. Spatial distribution of R^2 across the studied European countries.

subsequently increase the overall crop growth rates and can aid to maintain sustainable agricultural practices (Konefal et al., 2023; Leo et al., 2023). Finally, application of satellite-based climate variables could be beneficiary to handle climate resilience in case of inevitable change in weather patterns because of climate change.

5. Conclusion

Ensuring food security, a critical aspect of UN Sustainable Development Goals, requires doubling crop yields by 2030. Developing accurate crop yield prediction tools is essential for devising effective food strategies and aiding decision-makers worldwide. Failure to address these challenges could lead to socioeconomic crises, particularly in underdeveloped nations. Thus, there is an urgent need for strategic planning to enhance agricultural productivity and to mitigate the impact of climate change on crop yields. In this study, the effectiveness of RScv as a novel optimisation technique on machine learning algorithms was evaluated for predicting crop yield from four major cultivation patterns in Europe. The four crops examined were Barley, Oats, Rye and Wheat, widely cultivated in the region. In order to have the highest spatial and temporal coverage, satellite products from NASA's GPCP and GLDAS missions were considered as inputs of this study. Prediction comparison between the classic AB, GB, RF and ET, as four popular machine learning algorithms, as well as their optimised versions using RScv was performed using the satellite-based input dataset. By comparing the algorithms before and after the application of RScv, it was found that the Ada-Boost (AB) algorithm produced the most accurate predictions of crop yield. The outcomes of the study showed that the RScv-AB algorithm achieved a prediction accuracy of over 90% when forecasting crop yield. Spatial analysis of the prediction error distribution for RScv-AB was also performed for the studied 20 countries in Europe and over four crop types. While the overall fluctuation of residuals was around zero, the model depicts more tendency toward underestimation. Crop-based uncertainty examination shows that the machine learning model produces more uncertainty in Barley, while Wheat, which also has the highest prediction accuracy, had the least associated uncertainty. While the proposed predictive approach has proven its robustness, certain limitations were found, which will be considered in future research. For instance, whereas this study tries to include the major climate parameters, utilizing other weather variables or vegetation indices, which also could be obtained from other remote sensing sources, could increase the prediction of crop yield accuracy. While, this study only measured the impacts of climate-related variables, accompanying socio-economical and land cover/use parameters such as cropping areas and management methods could also be addressed in future studies. Furthermore, this study was regionally limited to 20 countries in the Europe continent, but applying the same method to the global scale may lead to higher geospatial diversity and more applicability of suggested endeavour. Application of remote sensing particularity enables this method to be applied in regions which lack the proper recording of historical or real time crop yields. Open accessorily to satellite-based soil and weather variables in certain areas can aid policy planners to have better approximations of crops productivity. Similarly, by using a combination of AI models and satellite data, farmers can obtain correct predictions of crop yield and make informed decisions about how to proceed.

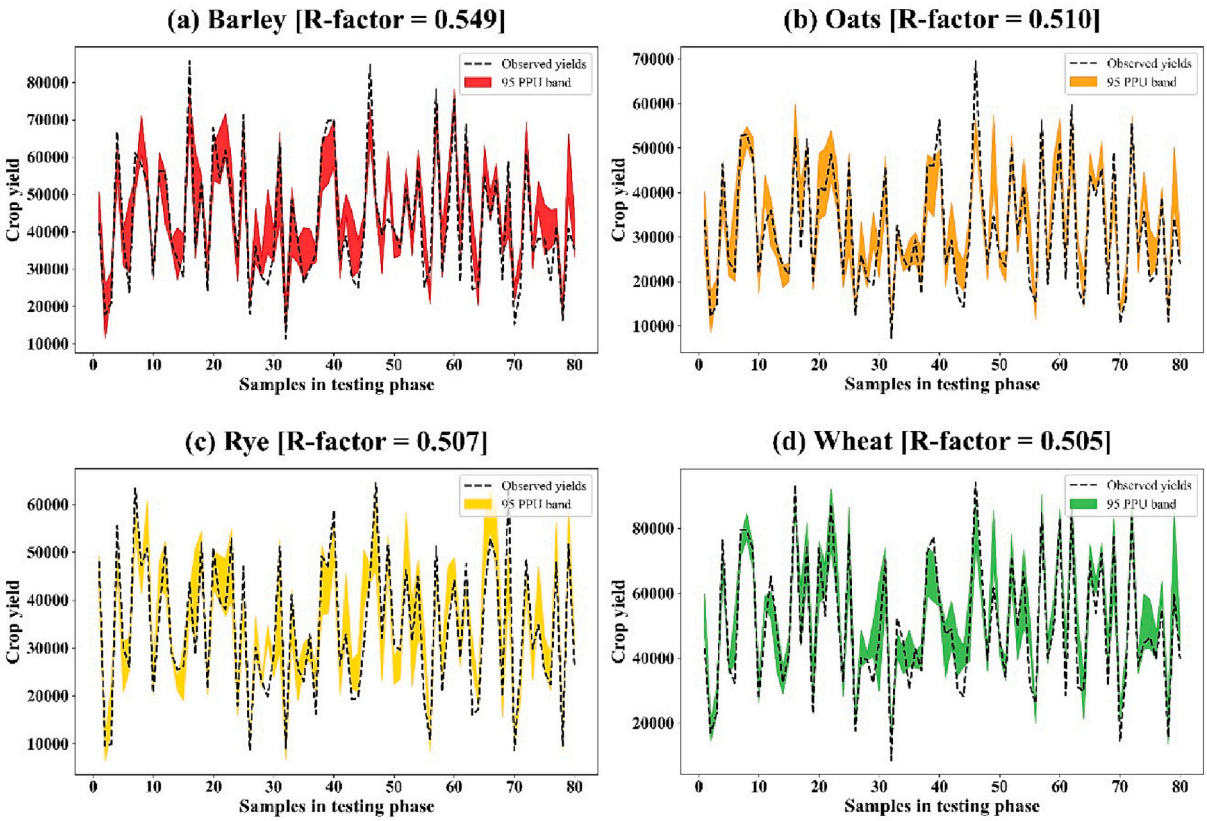


Fig. 8. Crop-based uncertainty analysis of optimised machine learning predictions using Monte-Carlo simulation and R-factor criteria.

Table 6
Comparison between RScv-AB and the optimised KNN, ANN and SVR algorithms in each studied crop for 20 European countries. Note: RMSE values are expressed in tons/ha.

Metrics	Algorithms	Barley	Oats	Rye	Wheat
R^2	RScv-AB	0.837	0.875	0.846	0.901
	RScv-KNN	0.772	0.837	0.756	0.844
	RScv-ANN	0.584	0.674	0.531	0.662
	RScv-SVR	0.510	0.657	0.608	0.622
RMSE	RScv-AB	7285	5196	5838	6816
	RScv-KNN	8699	5876	7480	8747
	RScv-ANN	11,252	8043	10,128	12,185
	RScv-SVR	12,138	8297	9264	12,945

Nomenclature

AI	Artificial Intelligence
AB	Ada-Boost
ANFIS	Adaptive neuro-fuzzy inference systems
ANN	Artificial neural network
CD	Coefficient of Determination
CHIRPS	Climate Hazards Group Infrared Precipitation with Station data
CNN	Convolutional Neural Network
DT	Decision Tree
ECMWF	European centre for medium weather forecast
ET	Extra Tree
GBR	Gradient Boost Regression
GDD	Growing degree days
GDAP	Data and Assessment Panel
GEWEX	Global Water and Energy Experiment
GLDAS	Global Land Data Assimilation System
GPCP	Global Precipitation Climatology Project
GPROF	Goddard Profiling Algorithm
KNN	K-Nearest Neighbour
LAI	Leaf Area Index
LST	Land Surface Temperature
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
MLR	Multiple linear regressions
MODIS	Moderate Resolution Imaging Spectroradiometer
MSE	Mean Squared Error
NDVI	Normalized difference vegetation index
NN	Neural Networks
NSE	Nash Sutcliffe Efficiency
RF	Random Forest
RMSE	Root Mean Squared Error
RS	Remote Sensing
SSMI/SSMIS	Special Sensor Microwave Imager/Sounder.
SPI	Standardised precipitation index.
SVM	Support Vector Machine
TRMM	Tropical Rainfall Measuring Mission

CRediT authorship contribution statement

Seyed Babak Haji Seyed Asadollah: Software, Methodology, Data curation, Conceptualization. **Antonio Jodar-Abellan:** Writing – review & editing, Writing – original draft, Validation, Supervision. **Miguel Ángel Pardo:** Writing – review & editing, Writing – original draft, Visualization.

Declaration of generative AI and AI-assisted technologies in the writing process

The authors declare that no generative AI of any kind have been employed in the process of writing this paper.

Declaration of competing interest

The authors declare no competing interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Antonio Jodar-Abellan acknowledges financial support received form the *Margarita Salas* Postdoc Spanish Program.

References

Abbaspour-Gilandeh, Y., Aghabara, A., Davari, M., Maja, J.M., 2022. Feasibility of using computer vision and artificial intelligence techniques in detection of some apple pests and diseases. Appl. Sci. 12 (2), 906. MDPI.

Adler, R.F., Gu, G., Huffman, G.J., Sapiano, M.R.P., Wang, J.J., 2020. GPCP and the global characteristics of precipitation. In: Satellite Precipitation Measurement. Springer, pp. 677–697.

AghaKouchak, A., Pan, B., Mazdiyasni, O., Sadegh, M., Jiwa, S., Zhang, W., Love, C.A., et al., 2022. Status and prospects for drought forecasting: opportunities in artificial intelligence and hybrid physical–statistical forecasting. Phil. Trans. R. Soc. A 380 (2238), 20210288. The Royal Society.

Ahmad, M.W., Reynolds, J., Rezgui, Y., 2018. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. J. Clean. Prod. 203, 810–821. Elsevier.

Alexandridis, N., Feit, B., Kihara, J., Luttermoser, T., May, W., Midega, C., Öborn, I., et al., 2023. Climate change and ecological intensification of agriculture in sub-Saharan Africa–A systems approach to predict maize yield under push-pull technology. Agric. Ecosyst. Environ. **352**, 108511. Elsevier.

Asadollah, S.B.H.S., Sharafati, A., Motta, D., Jodar-Abellan, A., Pardo, M.A., 2023. Satellite-based prediction of surface dust mass concentration in southeastern Iran using an intelligent approach. Stoch. Env. Res. Risk A. 37, 3731–3745. SpringerNature.

Beaudoing, H., Rodell, M., 2016. GLDAS Noah Land Surface Model L4 Monthly 0.25 x 0.25 Degree V2. 1. Greenbelt, Maryland.

Beaudoing, H., Rodell, M., 2020. NASA/GSFC. HSL, GLDAS Noah Land Surface Model L4 monthly 0.25× 0.25 degree **2**.

Behrangi, A., Song, Y., 2020. A new estimate for oceanic precipitation amount and distribution using complementary precipitation observations from space and comparison with GPCP. Environ. Res. Lett. 15 (12), 124042. IOP Publishing.

Belloni, A., Chernozhukov, V., 2013. Least squares after model selection in high-dimensional sparse models. Bernoulli 19 (2), 521–547. https://doi.org/10.3150/11-BEJ410.

Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80 (6), 2369–2429. Wiley Online Library.

Bi, L., Hu, G., 2021. A genetic algorithm-assisted deep learning approach for crop yield prediction. Soft. Comput. **25** (16), 10617–10628. Springer.

Boix-Fayos, C., de Vente, J., 2023. Challenges and potential pathways towards sustainable agriculture within the European Green Deal. Agric. Syst. **207**, 103634. Elsevier. https://doi.org/10.1016/j.agsy.2023.103634.

Breiman, L., 2001a. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., 2001b. Statistical modeling: the two cultures. Stat. Sci. 16 (3), 199–215. https://doi.org/10.1214/ss/1009213726.

Cao, J., Wang, H., Li, J., Tian, Q., Niyogi, D., 2022. Improving the forecasting of winter wheat yields in Northern China with machine learning–dynamical hybrid subseasonal-to-seasonal ensemble prediction. Remote Sens. **14** (7), 1707. MDPI.

Cavan, N., Omon, B., Dubois, S., Toqué, C., Van Inghelandt, B., Queyrel, W., Colbach, N., Angevin, F., 2023. Model-based evaluation in terms of weed management and overall sustainability of cropping systems designed with three different approaches. Agric. Syst. 208, 103637 https://doi.org/10.1016/j.agsy.2023.103637.

Challinor, A.J., Watson, J., Lobell, D.B., Howden, S.M., Smith, D.R., Chhetri, N., 2014. A meta-analysis of crop yield under climate change and adaptation. Nat. Clim. Chang. 4 (4), 287–291. Nature Publishing Group UK London.

Dang, C., Liu, Y., Yue, H., Qian, J., Zhu, R., 2021. Autumn crop yield prediction using data-driven approaches:-support vector machines, random forest, and deep neural network methods. Can. J. Remote. Sens. 47 (2), 162–181. Taylor & Francis.

Davenport, T.H., 2019. The AI Advantage: How to Put the Artificial Intelligence Revolution to Work. Mit Press, Publisher, p. 248. ISBN: 9780262538008.

Derdour, A., Jodar-Abellan, A., Melián-Navarro, A., Bailey, R.T., 2023. Assessment of land degradation and droughts in an arid area using drought indices, the modified soil-adjusted vegetation index, and landsat remote sensing data. Cuadernos de Investigacion Geografica 49 (2), 65–81. https://doi.org/10.18172/cig.5523.

Dourado-Neto, D., Teruel, D.A., Reichardt, K., Nielsen, D.R., Frizzone, J.A., Bacchi, O.O. S., 1998. Principles of crop modeling and simulation: I. Uses of mathematical models in agricultural science. Sci. Agric. 55, 46–50. SciELO Brasil.

FAOSTAT, 2023. Food and Agriculture Organization of the United Nations/ Statistics Division. Available at: https://www.fao.org/faostat/en/#home. Accessed on: 24 April 2023.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Ann. Stat. 1189–1232. JSTOR.

Fukase, E., Martin, W., 2020. Economic growth, convergence, and world food demand and supply. World Dev. 132, 104954. Elsevier.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63 (1), 3–42. Springer.

Gopal, P.S.M., Bhargavi, R., 2019. A novel approach for efficient crop yield prediction. Comput. Electron. Agric. 165, 104968. Elsevier.

Han, G., Niles, M.T., 2023. An adoption spectrum for sustainable agriculture practices: A new framework applied to cover crop adoption. Agric. Syst. 212, 103771. Elsevier. https://doi.org/10.1016/j.agsy.2023.103771.

Hanke, M., Dijkstra, L., Foraita, R., Didelez, V., 2023. Variable selection in linear regression models: choosing the best subset is not always the best choice. Biom. J. 66 (1), 2200209. Wiley Online Library.

Hastie, T., Tibshirani, R., Tibshirani, R., 2020. Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. Stat. Sci. 35 (4), 602–608. https://doi.org/10.1214/19-STS733.

Hochmuth, G.J., Cantliffe, D.J., Soundy, P., 1998. A comparison of three mathematical models of response to applied nitrogen: A case study using lettuce. HortScience 33, 5.

Hu, X., Ren, H., Tansey, K., Zheng, Y., Ghent, D., Liu, X., Yan, L., 2019. Agricultural drought monitoring using European Space Agency sentinel 3A land surface temperature and normalized difference vegetation index imageries. Agric. For. Meteorol. 279, 107707. Elsevier.

Huber, F., Yushchenko, A., Stratmann, B., Steinhage, V., 2022. Extreme gradient boosting for yield estimation compared with deep learning approaches. Comput. Electron. Agric. 202, 107346. Elsevier.

Hunink, J.E., Eekhout, J.P.C., Vente, J.D., Contreras, S., Droogers, P., Baille, A., 2017. Hydrological modelling using satellite-based crop coefficients: A comparison of methods at the basin scale. Remote Sens. 9, 174. MDPI.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., et al., 2016. Random forests for global and regional crop yield predictions. PLoS One 11 (6), e0156571. Public Library of Science San Francisco, CA USA.

John, V., Liu, Z., Guo, C., Mita, S., Kidono, K., 2016. Real-time lane estimation using deep features and extra trees regression. Image Video Technol. 721–733. Springer.

Ju, S., Lim, H., Ma, J.W., Kim, S., Lee, K., Zhao, S., Heo, J., 2021. Optimal county-level crop yield prediction using MODIS-based variables and weather data: A comparative study on machine learning models. Agric. For. Meteorol. 307, 108530. Elsevier.

Kang, Y., Khan, S., Ma, X., 2009. Climate change impacts on crop yield, crop water productivity and food security–A review. Prog. Nat. Sci. 19 (12), 1665–1674. Elsevier.

Keerthana, M., Meghana, K.J.M., Pravallika, S., Kavitha, M., 2021. An ensemble algorithm for crop yield prediction. In: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV). IEEE, pp. 963–970.

Khairunniza-Bejo, S., Mustaffha, S., Ismail, W.I.W., 2014. Application of artificial neural network in predicting crop yield: A review. J. Food Sci. Eng. 4 (1), 1. David Publishing Company, Inc.

Konefal, J., de Olde, E.M., Hatanaka, M., Oosterveer, P.J.M., 2023. Signs of agricultural sustainability: A global assessment of sustainability governance initiatives and their indicators in crop farming. Agric. Syst. 208, 103658. Elsevier. https://doi.org/10.1016/j.agsy.2023.103658.

Kotsias, G., Lolis, C.J., Hatzianastassiou, N., Levizzani, V., Bartzokas, A., 2020. On the connection between large-scale atmospheric circulation and winter GPCP precipitation over the Mediterranean region for the period 1980-2017. Atmos. Res. 233, 104714. Elsevier.

Kumar, N., Kumar, U., 2022. Artificial intelligence for classification and regression tree based feature selection method for network intrusion detection system in various telecommunication technologies. Comput. Intell. 40, 1–23. Wiley Online Library.

Lencastre, P., Gjersdal, M., Gorjão, L.R., Yazidi, A., Lind, P.G., 2023. Modern AI versus century-old mathematical models: How far can we go with generative adversarial networks to reproduce stochastic processes? Phys. D: Nonlinear Phenomena 453, 133831. Elsevier.

Leo, S., Migliorati, M.D.A., Nguyen, T.H., Grace, P.R., 2023. Combining remote sensing-derived management zones and an auto-calibrated crop simulation model to determine optimal nitrogen fertilizer rates. Agric. Syst. 205, 103559. Elsevier. https://doi.org/10.1016/j.agsy.2022.103559.

Li, L., Wang, B., Feng, P., Wang, H., He, Q., Wang, Y., Li Liu, D., et al., 2021. Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across China. Agric. For. Meteorol. 308, 108558. Elsevier.

Lin, H.T., Liang, T.J., Chen, S.M., 2012. Estimation of battery state of health using probabilistic neural network. IEEE Trans. Indust. Inform. 9 (2), 679–685. IEEE.

Lu, Y., 2019. Artificial intelligence: a survey on evolution, models, applications and future trends. J. Manag. Anal. 6 (1), 1–29. Taylor & Francis.

Luo, L., Sun, S., Xue, J., Gao, Z., Zhao, J., Yin, Y., Gao, F., et al., 2023. Crop yield estimation based on assimilation of crop models and remote sensing data: A systematic evaluation. Agric. Syst. 210, 103711. Elsevier. https://doi.org/10.1016/j.agsy.2023.103711.

Mahato, J.K., Gupta, S.K., 2022. Exploring applicability of artificial intelligence and multivariate linear regression model for prediction of trihalomethanes in drinking water. Int. J. Environ. Sci. Technol. 19 (6), 5275–5288. Springer.

Mazumder, R., 2020. Discussion of "best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons". Stat. Sci. 35 (4), 579–592. https://doi.org/10.1214/20-STS807.

Mo, C., Jiang, C., Lei, X., Lai, S., Deng, Y., Cen, W., Sun, G., et al., 2022. Combining standard artificial intelligence models, pre-processing techniques, and post-processing methods to improve the accuracy of monthly runoff predictions in karst-area watersheds. Appl. Sci. 13 (1), 88. MDPI.

Mohamadou, Y., Halidou, A., Kapen, P.T., 2020. A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19. Appl. Intell. 50 (11), 3913–3925. Springer.

Naderloo, L., Alimardani, R., Omid, M., Sarmadian, F., Javadikia, P., Torabi, M.Y., Alimardani, F., 2012. Application of ANFIS to predict crop yield based on different energy inputs. Measurement 45 (6), 1406–1413. Elsevier.

Naimaee, R., Kiani, A., Jarahizadeh, S., Haji Seyed Asadollah, S.B., Melgarejo, P., Jodar-Abellan, A., 2024. Long-term water quality monitoring: using satellite images for temporal and spatial monitoring of thermal pollution in water resources. Sustainability 16, 646. https://doi.org/10.3390/su16020646.

NASA, 2023. NASA Goddard Earth Sciences (GES) Data and Information Services Center (DISC). Available at: https://disc.gsfc.nasa.gov/. Accessed on: 12 May 2023.

Nosratabadi, S., Szell, K., Beszedes, B., Imre, F., Ardabili, S., Mosavi, A., 2020. Comparative analysis of ANN-ICA and ANN-GWO for crop yield prediction. In: 2020 RIVF International Conference on Computing and Communication Technologies (RIVF). IEEE, pp. 1–5.

O'brien, R.M., 2007. A caution regarding rules of thumb for variance inflation factors. Qual. Quant. 41, 673–690. Springer.

Padhee, S.K., Dutta, S., 2020. Spatiotemporal reconstruction of MODIS land surface temperature with the help of GLDAS product using kernel-based nonparametric data assimilation. J. Appl. Remote. Sens. 14 (1), 14520. SPIE.

Pede, T., Mountrakis, G., 2022. Towards daily maximum heat index estimation across the conterminous United States using satellite-derived products. Int. J. Remote Sens. 43 (8), 2861–2884. Taylor & Francis.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al., 2011. Scikit-learn: machine learning in Python. J. Machine Learn. Res. 12, 2825–2830. JMLR. org.

Poornappriya, T.S., Gopinath, R., 2022. Rice plant disease identification using artificial intelligence approaches. Int. J. Elect. Eng. Technol. 11 (10), 392–402.

Prasad, N.R., Patel, N.R., Danodia, A., 2021. Crop yield prediction in cotton for regional level using random forest approach. Spat. Inf. Res. 29 (2), 195–206. Springer.

Priscilla, C.V., Prabha, D.P., 2020. Influence of optimizing XGBoost to handle class imbalance in credit card fraud detection. In: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, pp. 1309–1315.

Ramirez-Villegas, J., Koehler, A.-K., Challinor, A.J., 2017. Assessing uncertainty and complexity in regional-scale crop model simulations. Eur. J. Agron. 88, 84–95. Elsevier.

Roser, M., Ritchie, H., Ortiz-Ospina, E., Rodés-Guirao, L., 2013. World population growth. Our world in data. https://www.scirp.org/reference/referencespapers?referenceid=2979311.

Ryan, M., 2022. The social and ethical impacts of artificial intelligence in agriculture: Mapping the agricultural AI literature. In: AI & SOCIETY. Springer, pp. 1–13.

Sadeghi, M., Akbari Asanjan, A., Faridzad, M., Afzali Gorooh, V., Nguyen, P., Hsu, K., Sorooshian, S., et al., 2019. Evaluation of PERSIANN-CDR constructed using GPCP V2. 2 and V2. 3 and a comparison with TRMM 3B42 V7 and CPC unified gauge-based analysis in global scale. Remote Sens. 11 (23), 2755. MDPI.

Sakamoto, T., 2020. Incorporating environmental variables into a MODIS-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm. ISPRS J. Photogramm. Remote Sens. 160, 208–228. Elsevier.

Sarker, I.H., 2022. AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. SN Comp. Sci. 3 (2), 158. Springer.

Shanmugasundar, G., Vanitha, M., Čep, R., Kumar, V., Kalita, K., Ramachandran, M., 2021. A comparative study of linear, random forest and AdaBoost regressions for modeling non-traditional machining. Processes 9 (11), 2015. MDPI.

Sharafati, A., Asadollah, S.B.H.S., Neshat, A., 2020. A new artificial intelligence strategy for predicting the groundwater level over the Rafsanjan aquifer in Iran. J. Hydrol. 591, 125468. Elsevier.

Sharma, N., Malviya, L., Jadhav, A., Lalwani, P., 2023. A hybrid deep neural net learning model for predicting coronary heart disease using randomized search cross-validation optimization. Decision Anal. J. 9, 100331. Elsevier.

Shastry, A., Sanjay, H.A., Hegde, M., 2015. A parameter based ANFIS model for crop yield prediction. In: 2015 IEEE International Advance Computing Conference (IACC). IEEE, pp. 253–257.

Shobha, K., Nickolas, S., 2018. Analysis of importance of pre-processing in prediction of hypertension. CSI Trans. ICT **6**, 209–214. Springer.

Spanaki, K., Karafili, E., Sivarajah, U., Despoudi, S., Irani, Z., 2022. Artificial intelligence and food security: swarm intelligence of AgriTech drones for smart AgriFood operations. Prod. Plan. Control 33 (16), 1498–1516. Taylor & Francis.

Syed, T.H., Famiglietti, J.S., Rodell, M., Chen, J., Wilson, C.R., 2008. Analysis of terrestrial water storage changes from GRACE and GLDAS. Water Resour. Res. 44 (2). Wiley Online Library.

Taufiqurrahman, A., Putrada, A.G., Dawani, F., 2020. Decision tree regression with AdaBoost ensemble learning for water temperature forecasting in aquaponic ecosystem. In: 2020 6th International Conference on Interactive Digital Media (ICIDM). IEEE, pp. 1–5.

Thompson, C.G., Kim, R.S., Aloe, A.M., Becker, B.J., 2017. Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. Basic Appl. Soc. Psychol. 39 (2), 81–90. Taylor & Francis.

Van Wart, J., Kersebaum, K.C., Peng, S., Milner, M., Cassman, K.G., 2013. Estimating crop yield potential at regional to national scales. Field Crop Res. 143, 34–43. Elsevier.

Vishnu, M.K., Rupak, V.R.V., Vedhapriyaa, S., Sangeetha, M., Manjuladevi, R., Sagana, C., 2023. Recurrent gastric cancer prediction using randomized search Cv optimizer. In: 2023 International Conference on Computer Communication and Informatics (ICCCI). IEEE, pp. 1–5.

Wang, Y., Zhang, Z., Feng, L., Du, Q., Runge, T., 2020. Combining multi-source data and machine learning approaches to predict winter wheat yield in the conterminous United States. Remote Sens. 12 (8), 1232. MDPI.

Weber, M., Engert, M., Schaffer, N., Weking, J., Krcmar, H., 2023. Organizational capabilities for ai implementation—coping with inscrutability and data dependency in ai. Inf. Syst. Front. 25 (4), 1549–1569. Springer.

Wu, Z., Feng, H., He, H., Zhou, J., Zhang, Y., 2021. Evaluation of soil moisture climatology and anomaly components derived from ERA5-land and GLDAS-2.1 in China. Water Resour. Manag. **35** (2), 629–643. Springer.

Xevi, E., Gilley, J., Feyen, J., 1996. Comparative study of two crop yield simulation models. Agric. Water Manag. 30 (2), 155–173. Elsevier.

Xiao, F., Wang, Y., He, L., Wang, H., Li, W., Liu, Z., 2019. Motion estimation from surface electromyogram using adaboost regression and average feature values. IEEE Access 7, 13121–13134. IEEE.

Yang, F., Wang, D., Xu, F., Huang, Z., Tsui, K.-L., 2020. Lifespan prediction of lithium-ion batteries based on various extracted features and gradient boosting regression tree model. J. Power Sources 476, 228654. Elsevier.

Ying, X., 2019. An overview of overfitting and its solutions. J. Phys. Conf. Ser. 1168, 22022. IOP Publishing.

Zhang, J., Chen, Y., Zhang, Z., 2020. A remote sensing-based scheme to improve regional crop model calibration at sub-model component level. Agric. Syst. **181**, 102814. Elsevier. https://doi.org/10.1016/j.agsy.2020.102814.