



International Journal of Research in Agronomy

E-ISSN: 2618-0618
P-ISSN: 2618-060X
© Agronomy
NAAS Rating (2026): 5.20
www.agronomyjournals.com
2026; 9(5): 112-119
Received: 22-02-2026
Accepted: 25-03-2026

Haripriya Ngairangbam
Department of Agronomy, School
of Agriculture, Lovely Professional
University, Punjab, India

Tanya Singh
Department of Agronomy, School
of Agriculture, Lovely Professional
University, Punjab, India

Manzoor A Khanday
Department of Statistics, Lovely
Professional University, Punjab,
India

Statistical techniques for crop yield prediction under climate variability: A review

Haripriya Ngairangbam, Tanya Singh and Manzoor A Khanday

DOI: <https://www.doi.org/10.33545/2618060X.2026.v9.i5b.5486>

Abstract

Agriculture still depends a lot on weather conditions, which makes growing crops very sensitive to changes in the climate. Predicting how much crops will yield is crucial for making sure we have food globally using resources wisely and helping farmers make good decisions. This review looks at all the statistical methods used to predict crop yields when the climate is changing. It checks out both new methods. The review looks at 30 relevant studies that use different approaches like regression, machine learning, deep learning and mixed models. The main findings show that new machine learning methods, like Random Forest and Support Vector Machines and deep learning methods like Long Short-Term Memory networks are better at predicting yields than statistical methods. However, there are still some problems. These include not having data not being able to understand how the models work the complexity of the calculations and the need for models that can work in different places and over time. Using data from sources. Like satellite images, weather observations and soil information. Seems to be very important for making predictions better. In the future we might see use of mixed modelling methods, artificial intelligence that can be explained integration of big data and better modelling frameworks that can handle how complex and unpredictable the interactions between climate and crop growth are. The use of machine learning, in crop yield prediction is very promising. Machine learning can help farmers and decision-makers. Crop yield prediction using machine learning is an area of research.

Keywords: Crop yield prediction, climate variability, climate change adaptation, statistical modelling, food security

Introduction

Agriculture's fundamental dependence on climatic conditions renders crop production systems inherently vulnerable to climate variability and change. The variation in temperature, rainfall, weather extremes and seasonal variations have a direct impact on the growth, development and end result yield of crops. With the world climate patterns becoming more and more unpredictable, the agricultural sector is experiencing increasing challenges in ensuring the stability of food production to satisfy the needs of an increasing world population that is estimated to hit 9.7 billion in the year 2050. This is a critical vulnerability especially in areas where smallholder farmers rely on rainfed agriculture, and there is low adaptive capacity to cushion against climate shocks (Sarr and Sultan, 2022) ^[40].

Projecting crop yield in climatic variability is one of the most important issues in the contemporary agricultural sector, in which farmers and scientists have to consider the more and more unpredictable weather conditions, which directly influence the food production. Climate variability comprises of fluctuations in temperature, rainfall, humidity and extreme weather patterns that may have great impacts on the growth of crops and final yields. The statistical techniques offer the necessary instruments of comprehending these complex interactions between climate factors and agricultural performance, to make more precise forecasts that can be used as a tool in agricultural decision making, agricultural food security planning, and policy formulation. The prediction models enable the stakeholders to predict future crop failures, to maximize allocation of resources and to come up with adaptive measures so that food production can remain constant regardless of environmental changes. The discipline integrates conventional statistical techniques with current computational technologies to address the complexity of agricultural systems where a variety of climate variables are used together with

Corresponding Author:
Haripriya Ngairangbam
Department of Agronomy, School
of Agriculture, Lovely Professional
University, Punjab, India

soil conditions, crop genetics and management practices to yield final yields.

Statistical methods have been the foundation of crop yield forecast systems since early days of simple empirical relationships to elaborate data-based models that can reflect the complicated, nonlinear interplays of climate variables and crop responses. The last 20 years have seen the paradigm shift in the approach to yield prediction, due to the growth in computational capabilities, the spread of remote sensing technology, the access to high-resolution climate observations, and the development of machine learning and artificial intelligence (Salami *et al.*, 2025) [39]. These advances have facilitated the shift in research and practice towards more flexible, high dimensional modeling frameworks that are capable of taking into consideration a variety of data and explain the spatial and temporal heterogeneity in crop-climate relations. (Kheir *et al.*, 2024; Osibo *et al.*, 2023) [19, 34].

Traditional regression-based approaches

The basis of crop yield prediction is traditional regression techniques based on linear relationships between climate variables and yields (Parviz *et al.*, 2017) [35]. These methods involve simple linear regression, multiple regression and stepwise selection methods which aid in determining the most significant climate factors influencing production in crops.

The origin of statistical crop yield prediction was simple regression models that were used to analyse the interrelation between crop yields and time series or the correlation between crop yields and single weather factors. Statistical methods that averaged the yields on large regions over long periods of time and compared them to time were the first large-scale yield simulation models to provide insight into overall tendencies in crop productivity. The most popular statistical methodology to study crop-yield-weather modelling became multiple linear regression (MLR), and organizations such as the FAO started to use multivariate regression models in their operational crop predictions in developing nations (Parviz *et al.*, 2017) [35].

Recent regression methods embrace systematic processes to deal with the complexity in climate-yield relationships. The standard statistical crop yield model is a regression of panel data of crop yields on other weather data, where both climate and crop yield data may need to be detrended to remove linear trends and technological advances (Ray *et al.*, 2015) [37] (Ortiz-Bobea, 2021) [32]. To estimate the impacts of various climate variables on the economy, researchers typically run multiple regression equations with a wide range of combinations of climate variables and model selection criteria such as the Akaike Information Criterion (AIC) to determine what models perform the best (Ray *et al.*, 2015) [37].

The use of stepwise regression methods has become the norm when it comes to the identification of pertinent climate predictors among a large pool of potential variables. These techniques stepwise include or exclude climate indices by the criteria of their effect on model performance, such as R-squared values to decide on the best combination of predictors (Zhu *et al.*, 2019) [53]. The adjusted R-squared, AIC and Bayesian Information Criterion (BIC) are usually used to compare the models with varying numbers of parameters to determine the best regression formulation (Shammi *et al.*, 2021) [42].

Researchers may prefer to use nonlinear terms like a piece-wise linear relationship or a form of the climate variables that is a power of a few variables instead of linear terms in order to represent nonlinear responses of crops to climate conditions (Sidhu *et al.*, 2023) [44] (Fishman, 2016) [11]. Bootstrap

resampling methods are commonly used to test uncertainty in regression estimates and generate several different datasets to test the stability of a model and provide confidence intervals on predictions (Leng *et al.*, 2017) [23] (Coffel *et al.*, 2022) [6]. The median R-squared values of these traditional regression methods are 0.37-0.49 with major crops, which implies that the variables of temperature and precipitation can explain a significant part of interannual variations in yields (Coffel *et al.*, 2022) [6].

Machine learning and advanced statistical methods

Random Forests, neural networks, and boosted regression tree are all machine learning techniques that have become effective alternatives to a more traditional regression in crop yield prediction. These methods have the ability to model complicated nonlinear relationships and interactions among climate variables that are often not represented in linear models.

It is the weaknesses of the conventional linear regression techniques that have propelled scientists to machine learning and more sophisticated statistical techniques that are able to more effectively reflect the complex, nonlinear interactions between climate variables and crop yields. Although deterministic biophysical crop models can be useful in understanding the mechanisms, statistical models tend to be more effective at predicting over larger spatial scales (Crane-Droesch, 2018) [8]. Machine learning methods have a tremendous benefit in working with big data as well as modelling nonlinearities that are hard to model using the traditional methods.

The capability to work with nonlinearities and interact with climate variables in a complex manner has made Random Forests one of the most popular machine learning algorithms in predicting crop yields (Vogel *et al.*, 2019) [48]. This is an ensemble approach that employs several decision trees, and can be used in both regression and classification problems, in many cases surpassing traditional linear methods when compared to other algorithms such as Support Vector Machines, Generalized Additive Models and the K-Nearest Neighbours (Vogel *et al.*, 2019) [48] (Chen *et al.*, 2024) [5]. Random Forests have been demonstrated to be less sensitive to aggregation methods of climate data compared to linear regression models and, in general, offers a better predictive performance (Chen *et al.*, 2024) [5].

More sophisticated methods of regularization such as LASSO (Least Absolute Shrinkage and Selection Operator) have also been useful in automated selection of variables, especially when there are large numbers of potentially collinear climate predictors. LASSO regression is able to identify automatically drivers of extreme effect of compounds and can deal with cases where the ordinary least squares estimation is afflicted with overfitting (Zscheischler *et al.*, 2020) [54].

Neural networks and their variations are deep learning methods that have demonstrated impressive capabilities in prediction of crop yields. Semi parametric neural networks which are parametric structure coupled with deep learning features can be used to perform better than classical statistical models as well as fully-nonparametric neural networks (Crane-Droesch, 2018) [8]. Long Short-term Memory (LSTM) networks have been found to be particularly effective at capturing cumulative and nonlinear relationships between yields and environmental conditions with researchers finding that LSTM models explain 76% of yield changes because of meteorological indices as opposed to 39% in models solely explained by meteorological indices (Jiang *et al.*, 2019) [18] (Cao *et al.*, 2021) [4]. Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN) are also becoming popular, with CNN being the most popular deep learning

algorithm used in crop yield research (Klompenburg *et al.*, 2020) [47].

Gradient boosting algorithms, such as XGBoost, LightGBM, and CatBoost, have become one of the most effective models in prediction of yields (Sidhu *et al.*, 2023) [44] (Yavrum *et al.*, 2025) [50] (Mahesh *et al.*, 2024) [29]. Compared to simple least squares methods, boosted regression trees are more capable of simulating physiological mechanisms, and performing better in models, especially when the climate variables are in a form of polynomials, and the relationships are piecewise (Sidhu *et al.*, 2023) [44]. The recent research has demonstrated CatBoost with prediction accuracy of more than 99% in predicting rice yields (Mahesh *et al.*, 2024) [29].

Support Vector Machines (SVM) have been shown to be very effective in sequential climate data and some studies have demonstrated that SVM has R² of 84-90% and root mean square error of low when used to predict crop yields in climate variability. It has been found in multiple comparative studies that machine learning models perform better than traditional statistical regression models, and that Random Forest and Gradient Boosting regressors typically perform better on predictive tasks compared to traditional regression methods, in terms of standard predictive metrics such as RMSE, MAE, and R² (Akwabeng *et al.*, 2026) [2].

Recent advances in frameworks are trending towards the integration of various machine learning techniques via ensemble methods whereby various base models such as Random Forest, Gradient Boosting and LSTM networks are stacked dynamically (Tomar *et al.*, 2025) [46]. Such adaptive ensemble strategies have the capacity of integrating real-time information on the soil, climate, and weather stress environment to offer a stronger forecast of yields in different environmental conditions (Tomar *et al.*, 2025) [46].

Probabilistic and Bayesian approaches

Probabilistic and Bayesian approaches give complex ways to predict crop yields, which quantitatively measure uncertainty and deal with complex climate-yield interactions. These methods apply probability distributions, Monte Carlo sampling and Bayesian inference to come up with more resilient predictions with a confidence interval.

Probabilistic and Bayesian statistics is an important step forward in predicting crop yields in the face of climate variability, providing advanced techniques to measure uncertainty and to include a variety of sources of variability in a prediction model. Such techniques go beyond deterministic forecasting to give probability distributions of the possible outcomes, which is especially useful in assessing risk and making decisions in agricultural systems.

Bayesian models have particularly been successful in uniting a variety of data and in taking into consideration several forms of uncertainty when making predictions of crop yield. Among the notable applications are multivariate statistical models which calculate bias and uncertainty of predicted yield through Markov-Chain Monte Carlo (MCMC) simulation with machine learning methods (Newlands *et al.*, 2014) [31]. These combined strategies have the potential to decrease the forecast uncertainty of model-based crop yield forecasts to 1-4 percent by integrating satellite remote-sensing and agro-climate data on a probabilistic framework that integrates both data and model structural uncertainty.

Particularly promising have been generative probabilistic models whose parameters are estimated using Bayesian inference to model crop yield as a function of temperature and

precipitation. These methodologies tend to use a two-dimensional Gaussian function to represent crop growth response to climate variables, which capture non-linear crop growth response and interactions of meteorological factors (Shirley *et al.*, 2020) [43] (Cottrell *et al.*, 2025) [7].

The use of Monte Carlo simulation techniques is important in probabilistic yield prediction as it allows the stochastic sampling of model parameters, and produces distributions of possible outcomes. In Monte Carlo studies, stepwise regression is used by researchers to identify climate predictors then regionally pool regression coefficients and fit probability distributions so that during stochastic sampling, 10,000 simulations can be created to cover the entire spectrum of potential yield responses (Zhu *et al.*, 2019) [53].

Copula functions have been developed as effective in modelling joint dependence between yield variations and climate conditions with little input data, giving probabilistic results to evaluate the risk of crop yield loss due to extreme climate events (Wang *et al.*, 2024) [49]. These techniques are unlike the conventional linear or nonlinear fitting techniques as they estimate the probabilities of yield failures in varying climatic conditions (Wang *et al.*, 2024) [49]. The copula-based studies have shown moderate and exceptional droughts can elevate the risk of crop losses by 32-56% with respect to crops and regions (Leng, 2021) [21] (Leng *et al.*, 2019) [22].

Spatial and temporal dependencies are progressively being considered in more sophisticated Bayesian modelling models by use of hierarchical models and dynamic statistical models. These are spatial basis function methods which involve Bayesian representations of structural uncertainty on both parametric and climate model structures, and Bayesian Structural Time Series methods which break time series down into trend, seasonality, and covariate components and also permit quantifications of uncertainty. Other frameworks incorporate dynamic Generalized Extreme Value distributions whereby the parameters can change with time together with covariates in response to the non-stationarity of climate effects (Michaelides *et al.*, 2025) [30].

Incorporating modern Bayesian hierarchical modelling schemes offers a holistic solution to crop yield prediction, model choice and uncertainty prediction across various future conditions of greenhouse gas emissions (Li *et al.*, 2025) [24]. These frameworks systematically break down predictive uncertainty to the contributions of climate models, emission scenarios, and crop model parameters, and hybrid methods such as machine learning paired with global gridded crop models cut uncertainties by 33-78% of major crops (Li *et al.*, 2023) [25].

Hybrid and ensemble methods

Hybrid and ensemble techniques use various modelling techniques in order to address individual weaknesses and enhance the accuracy of crop yield prediction. These approaches combine physics-based models with machine learning procedures and apply ensemble strategies to minimize uncertainty and model complicated climate-yield interactions.

Hybrid and ensemble methods are another important advance in crop yield prediction, which tries to overcome limitations of individual modelling methods by integrating their advantages to gain high predictive powers. These models combine mechanisms of mechanistic knowledge and complex recognition of patterns by combining physics-based crop models with statistical and machine learning in a systematic way (Li *et al.*, 2023) [25].

Process-based crop models, which are enhanced by machine learning algorithms, have emerged as the most successful hybrid

methods in the field of crop forecasting, to enhance accuracy and uncertainty quantification. Some of the most notable implementations are the combination of APSIM model outputs with Random Forest based on indicators of extreme climate events that produced 81% of the yield changes with a 33% enhanced performance compared to APSIM alone. In the same way, hybrid models that integrated MCWLA-Wheat model with the Random Forest minimized the root mean squared error to 365 kg/ha as opposed to the conventional models.

New ensemble systems show impressive gains in prediction quality and diminishing uncertainty. The combination of Dynamic Linear Models and Random Forest and nine Global Gridded Crop Models in studies resulted in significant performance improvement, raising correlation coefficients between 0.15 and 0.77 with maize, and between 0.37 and 0.70 with soybean and reducing uncertainties by 33-78% and 56-68% respectively (Li *et al.*, 2023) ^[25]. Observational constraints methods based on machine learning and random forest can minimize uncertainties (62% yield average and 52% yield variability) in crop yield projections in the future, compared to traditional linear regression constraint, by far (Yin *et al.*, 2022) ^[51].

Multi-base model Advanced ensemble architectures have mechanisms of dynamic stacking of multiple base models. Modern models integrate Random Forest, gradient boosting and LSTM networks with adaptive ensemble frameworks that have the ability to incorporate real-time data regarding the soil, climate and weather stress factors such as drought severity and erratic rainfall patterns (Tomar *et al.*, 2025) ^[46]. These mixed approaches effectively model complicated yield reaction to severe climate conditions, pests and illnesses and recreates stressors that were not previously modelled models (Li *et al.*, 2025) ^[24] (Heinicke *et al.*, 2022) ^[15].

The success of hybrid methods lies in their capability to reduce the predictive uncertainty systematically to reflect the contribution of climate models, emission scenarios, and crop model parameters (Li *et al.*, 2025) ^[24]. Current Bayesian hierarchical models offer holistic solutions that do not make highly constraining homogeneity assumptions and allow quantification of uncertainty to be transparently expressed at a variety of future greenhouse gas emission levels. These models are the latest edge in crop yield forecasting, and provide powerful approaches that integrate mechanistic knowledge with data-driven pattern identification to respond to the intricate issues of forecasting agricultural results in the face of climate change.

Model evaluation and performance assessment

Crop yield prediction Model evaluation involves statistical measures such as R-squared, AIC and cross-validation to determine the accuracy and the most appropriate models. The variation in performance generally depends on type of crop, climate variable, and spatial scale where median R-squared is about 0.37-0.49 among major crops.

To effectively evaluate crop yield prediction models, it is necessary to systematically assess the model through a number of performance measures and validation methods to guarantee that the prediction can be reliable in the face of climate variability. Standardized assessment protocols are vital in comparing other statistical methods due to the fact that the evaluation methodology adopted immensely influences the selection of models, and interpretation of findings.

Cross-validation methods are the gold standard of evaluating models in crop yield prediction as they allow researchers to

evaluate the out-of-sample performance and prevent overfitting. Random Forest models often rely on cross-validated predictions to compute R² values that approximate the amount of variance in yield anomalies that can be attributed to climate predictors (Vogel *et al.*, 2019) ^[48]. More complex cross-validation techniques based on machine learning are being used in contexts of causal inference, which entails three steps: eliminating variation due to non-parametric controls, cross-validation of residualized data to select weather measures, and adjusting models to pick climate dimensions useful in adaptation (Hultgren *et al.*, 2025) ^[16].

The test of statistical significance is important in establishing the validity of the model and in this context ANOVA or F-tests are generally performed at the $p = 0.10$ level to establish that the model chosen is statistically significantly better than random climate models. Reporting on the extent to which the variability in yield can be attributed to that in climate is only done using models that satisfy statistical significance levels (Ray *et al.*, 2015) ^[37].

There are several criteria of information that are needed to compare the models with varying number of parameters and level of complexity. The adjusted R-squared values are used to compare models with varying number of explanatory variables by determining the best model by minimizing variability of dependent variables (Shammi *et al.*, 2021) ^[42] (Li *et al.*, 2019) ^[26]. The Akaike Information Criterion (AIC) is used to estimate the difference between models that are given and those that are true, with the best model often being the one which has the lowest value of the Akaike Information Criterion. Bayesian Information Criterion (BIC) decreases the maximum likelihood choice and punishes the model complexity with numerous parameters (Shammi *et al.*, 2021) ^[42].

The performance benchmarks have a wide difference between crops and modelling methods. The median R² values in traditional statistical models are about 0.49 in the case of maize and 0.37 in the case of soybean, meaning that the variables of temperature and precipitation have a significant predictive power of interannual variability in yield (Coffel *et al.*, 2022) ^[6]. To compare traditional statistical regression with machine learning methods such as Random Forest and Gradient Boosting regressors, standard regression measures such as root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R²) are compared (Akwabeng *et al.*, 2026) ^[2].

Model performance also depends on climate variable and spatial scale to a large extent providing significant considerations to evaluation strategies. Statistical models exhibit varying strengths when predicting different types of climate response, with panel and cross-sectional models being more effective in predicting temperature response and time-series models more effective in predicting precipitation change response (Islam *et al.*, 2018) ^[17] (Gyamerah *et al.*, 2024) ^[13] (Gyamerah *et al.*, 2020) ^[14].

Quantification of uncertainty has become part of model evaluation, and bootstrap resampling methods have frequently been applied in the evaluation of uncertainty in regression estimates. Researchers usually generate 1000 distinct datasets as a result of resampling and model different statistical models to each dataset to measure ranges of uncertainty (Leng *et al.*, 2017) ^[23]. The higher-order methods take samples of the coefficient estimates based on the t-distributions of students in mean coefficient estimates and standard errors to produce thousands of estimates to measure the uncertainty in prediction (Coffel *et al.*, 2022) ^[6].

A stress test on a variety of different statistical procedures can

be used to ensure that the findings are not reliant on one particular modelling decision. Random Forest models are often evaluated comparatively with Support Vector Machines, Generalized Additive Models, K-Nearest Neighbours, and multiple linear regression to ensure that the performance is not only better with the random forest model but also with other statistical learning models (Vogel *et al.*, 2019) ^[48].

Model Category	Specific Algorithm/Method	Key Performance Metric (R2 or Accuracy)	Advantage/Characteristic
Traditional Regression	Multiple Linear Regression (MLR)	R2: 0.37 – 0.49	Foundation for operational predictions in developing nations.
Machine Learning	Random Forest (RF)	High Predictive Power	Less sensitive to climate data aggregation compared to linear models.
Deep Learning	LSTM Networks	R2: 0.76 (76%)	Effective at capturing cumulative and non-linear relationships.
Deep Learning	CatBoost	Accuracy: > 99%	Specifically demonstrated exceptional results for rice yields.
Advanced Stats	Support Vector Machines (SVM)	R2: 0.84 – 0.90	Very effective with sequential climate data.
Hybrid Systems	APSIM + Random Forest	81% of yield changes explained	33% better performance than using the process-based model alone.
Hybrid Systems	Bayesian Hierarchical / Ensemble	Uncertainty reduction: 33% – 78%	Systematically breaks down predictive uncertainty from climate models and emission scenarios.

Challenges and limitations

Overfitting with large numbers of variables, poor performance in a complex climate and assumptions of stationarity that break down in climate change are some of the major challenges confronting statistical crop yield prediction. Such techniques also have the challenge of nonlinear relationships and spatial-temporal dependencies that are hard to model using conventional techniques. Although statistical approaches towards crop yield prediction under climate variability have been widely used, and have proven to be useful, a number of fundamental challenges and limitations impact their reliability and applicability. These constraints are more acute when climate patterns are becoming more volatile and more complex agricultural systems are being modelled with more sophisticated models. The overfitting issue that arises with a large number of predictor variables is one of the most problematic issues with statistical crop yield models. Although multiple linear regression models have been found to be effective when predictors are not too many in number, ordinary least squares estimation is usually affected by overfitting issues when many potential explanatory variables are taken into account and thus undermines the predictive ability of these methods (Bachmair *et al.*, 2018) ^[3]. The situation is the most acute when the number of variables is significant relative to the number of data points, or when there is a high degree of correlation among input variables. Another basic challenge of statistical methods is the nonlinear and complicated association between crop yield and climate variables. The effect of climate variability on crop production, due to complexity and non-linear correlation between crop production and climate factors, may be challenging to accurately measure and estimate using statistical methods (Gyamerah *et al.*, 2020) ^[14]. This complication is enhanced by the fact that the behaviour of statistical models varies with the climate variable and spatial scale under consideration with responses to temperature change being better forecasted using panel and cross-sectional models than responses to precipitation change (Gyamerah *et al.*, 2024) ^[14]. Conventional crop yield forecasting models tend to be simplistic in their use of statistical applications that do not consider the intricate interaction of climatic conditions, spatial variation, and crop behaviour (Michaelides *et al.*, 2025) ^[30]. Such models generally consider the situation as stationary and neglect the dynamic character of the climatic effects which are more and

more unstable due to the global climate change (Michaelides *et al.*, 2025) ^[30]. The stationarity assumption is especially troublesome to extreme value modelling, where classical methods might be inadequate to reflect how climate extremes vary over time. Statistical models indicate that there are significant performance constraints in some environments and agricultural systems. Findings are inaccurate in agricultural systems that are subjected to oceanic weather and high levels of inputs where agro-meteorological processes associated with excess water like nitrogen loss, diseases, anoxia and unfavourable harvest conditions are not sufficiently represented by existing modelling frameworks (Lecerf *et al.*, 2019) ^[20]. Such shortcomings underscore the importance of enhanced modelling of complex biological and physical mechanisms that influence crop yields in changing climatic conditions. The spatial and temporal inconsistency of predictor performance provides more problems to model reliability. Although crop model predictors are more predictable as they generalize the series of agro-meteorological conditions over the entire growing season, the success of meteorological predictors is spatially and temporally incoherent and varies depending on crop phenology (Lecerf *et al.*, 2019) ^[20]. This inconsistency complicates the ability to come up with robust models that can work consistently across various regions and time periods. The problem of data quality and availability also limits the usefulness of statistical methods. The fact that there is a difference in the performance of models that use reported yield data, and those that use remote sensing-obtained data, indicates that the quality of data is a paramount constraint, and remote sensing yield data exhibit worse performance than ground-based data. These data constraints are especially troublesome to the extent that they attempt to represent regional variability and to address local adaptation strategies that play a major role in determining yield outcomes (Schauberger *et al.*, 2017) ^[41].

Data preprocessing and feature engineering

The preprocessing of data used in crop yield prediction includes detrending climate and yield data to eliminate technological gains and linear tendencies, followed by feature engineering approaches such as aggregation of climate variables at various time scales and dimensionality reduction approaches to manage large volumes of correlated predictors. Successful preprocessing of data and feature engineering are

important processes in coming up with strong statistical models to predict crop yields in the face of climate fluctuation. These preprocessing tasks should deal with the complexity of agricultural data in which yield data have responses to not only climate variability, but also non-climate variables that include technological advancements, management transitions, and economic realities.

Most statistical models of crop yield modelling are based on the detrending process. A common preprocessing process consists of three steps conducted systematically, starting with the calculation of harvested-area-weighted average seasonal and annual average temperature and precipitation data at political entities that grow crops; then detrending climate and yield data; and finally developing regression models between detrended yield data and climate data (Ray *et al.*, 2015) [37]. This detrending is essential since the change in crop yield is not just due to climate change but other factors such as technological change that have to be distinguished by climate signals (Feng *et al.*, 2023) [10].

Several detrending techniques can be obtained to eliminate non-climate trends in yield data. Popular methods are differencing techniques, multi-year moving averages, linear regression and filtering analysis (Feng *et al.*, 2023) [10]. One of the most common methodologies that have been effectively used to separate contributions of climate and non-climate factors to observed crop yield evolution and to extract and analyse climate impacts on yield are the first-difference methods (El-Maayar *et al.*, 2013; Lobell *et al.*, 2007; Osborne *et al.*, 2013) [9, 28, 33].

Another important preprocessing problem is climate variable aggregation, and the temporal resolution and aggregation approach has a strong impact on model performance. Comparative studies of various methods have indicated that principal component analysis or an average of monthly data-based models have better predictive ability than those based on more complex dimension reduction methods. Random Forest models are less sensitive to climate aggregation schemes than linear regression models to predict yields, indicating that machine learning models can be more resilient to preprocessing decisions (Chen *et al.*, 2024) [5].

Improved feature engineering methods deal with the issue of large volumes of potentially correlated climate predictors. Principal component analysis and functional principal component analysis are becoming more popular to improve the reliability of the model and solve the problem of multicollinearity among weather-related variables (Yavrum *et al.*, 2025) [50]. In complicated modelling systems, a researcher will tend to come up with various versions of climate variables, with single and squared terms, many possible combinations of predictors, which must be systematically chosen with criteria such as the Akaike Information Criterion (Ray *et al.*, 2015) [37].

Different kinds of statistical models have been given special preprocessing methods. Linear mixed-effects models must be specified carefully to include fixed effects such as changes in temperature and precipitation, CO₂ concentration changes, and categorical variables such as adaptation measures, and random effects on the study areas, crops and climate models. To perform distribution-based analyses, relative distribution approaches are employed to represent variations in whole yield distributions across baseline and climate change conditions, breaking down the changes into location and shape terms with non-parametric frameworks (Stella *et al.*, 2021) [45].

The preprocessing workflow should also be able to support the combination of various data sources and modelling methods. The frameworks of meta-analysis involve the development of

standardized data set of several studies that vary in terms of crops, regions, scenarios, time periods and methods of analysis, which involves a careful harmonization of various data forms and various time scales (Zhang *et al.*, 2021; Liu *et al.*, 2019) [52, 27]. The detailed preprocessing of these steps makes it possible to have statistical models that are capable of capturing the climate-yield relationships and also control the confounding factors that would otherwise bias the outcome of prediction.

References

1. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 2003;19(6):716-723.
2. Akwabeng PM, Awele O, Jegede OA, Odili J. A data-driven assessment of crop yield variability and global food security under climate change. *World Journal of Advanced Research and Reviews*. 2026;29(1):295-304.
3. Bachmair S, Tanguy M, Hannaford J, Stahl K. How well do meteorological indicators represent agricultural and forest drought across Europe? *Environmental Research Letters*. 2018;13.
4. Cao J, Zhang Z, Luo Y, Zhang L, Zhang J, Li Z, *et al.* Wheat yield predictions at a county and field scale with deep learning, machine learning, and Google Earth Engine. *European Journal of Agronomy*. 2021;123:126204.
5. Chen M, Guilpart N, Makowski D. Comparison of methods to aggregate climate data to predict crop yield: an application to soybean. *Environmental Research Letters*. 2024;19(5):054049.
6. Coffel ED, Lesk C, Winter JM, Osterberg EC, Mankin JS. Crop-climate feedbacks boost US maize and soy yields. *Environmental Research Letters*. 2022;17(2):024012.
7. Cottrell A, Pope E, Davie J, Zhang P, Falloon P, Crocker T, *et al.* A data-driven model for quantifying how maize yields in North-East China respond to summer climate. *Environmental Research Communications*. 2025;7(3):035010.
8. Crane-Droesch A. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*. 2018;13(11):114003.
9. El-Maayar M, Lange MA. A methodology to infer crop yield response to climate variability and change using long-term observations. *Atmosphere*. 2013;4:365-382.
10. Feng X, Tian H, Cong J, Zhao C. A method review of the climate change impact on crop yield. *Frontiers in Forests and Global Change*. 2023;6:1198186.
11. Fishman R. More uneven distributions overturn benefits of higher precipitation for crop yields. *Environmental Research Letters*. 2016;11(2):024004.
12. García-León D, López-Lozano R, Toreti A, Zampieri M. Local-scale cereal yield forecasting in Italy: lessons from different statistical models and spatial aggregations. *Agronomy*. 2020;10(6):809.
13. Gyamerah SA, Asare C, Agbi-Kaesler HO, Baffour-Ata F. Assessing the impact of climate variability on maize yields in the different regions of Ghana: a machine learning perspective. *PLOS ONE*. 2024;19.
14. Gyamerah SA, Ngare P, Ikpe D. Probabilistic forecasting of crop yields via quantile random forest and Epanechnikov kernel function. *Agricultural and Forest Meteorology*. 2020;280:107808.
15. Heinicke S, Frieler K, Jägermeyr J, Mengel M. Global gridded crop models underestimate yield responses to

- droughts and heatwaves. *Environmental Research Letters*. 2022;17(4):044026.
16. Hultgren A, Carleton TA, Delgado MT, Gergel DR, Greenstone M, Houser T, *et al.* Impacts of climate change on global agriculture accounting for adaptation. *Nature*. 2025;642:644-652.
 17. Islam T, Chisty TA, Chakrabarty A. A deep neural network approach for crop selection and yield prediction in Bangladesh. In: *Proceedings of the 2018 IEEE Region 10 Humanitarian Technology Conference*. 2018. p.1-6.
 18. Jiang H, Hu H, Zhong R, Xu J, Xu J, Huang J, *et al.* A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: a case study of the US Corn Belt at the county level. *Global Change Biology*. 2020;26(3):1754-1766.
 19. Kheir AM, Govind A, Nangia V, Devkota M, Elnashar A, Omar MED, *et al.* Developing automated machine learning approach for fast and robust crop yield prediction using a fusion of remote sensing, soil, and weather dataset. *Environmental Research Communications*. 2024;6(4):041005.
 20. Lecerf R, Ceglar A, López-Lozano R, Van der Velde M, Baruth B. Assessing the information in crop model and meteorological indicators to forecast crop yield over Europe. *Agricultural Systems*. 2019;168:191-202.
 21. Leng G. Maize yield loss risk under droughts in observations and crop models in the United States. *Environmental Research Letters*. 2021;16(2):024016.
 22. Leng G, Hall J. Crop yield sensitivity of global major agricultural countries to droughts and the projected changes in the future. *Science of the Total Environment*. 2019;654:811-821.
 23. Leng G, Huang M. Crop yield response to climate change varies with crop spatial distribution pattern. *Scientific Reports*. 2017;7(1):1463.
 24. Li D, Kitsios V, Newth D, O'Kane T. A Bayesian ensemble projection of climate change and technological impacts on future crop yields. 2025.
 25. Li L, Zhang Y, Wang B, Feng P, He Q, Shi Y, *et al.* Integrating machine learning and environmental variables to constrain uncertainty in crop yield change projections under climate change. *European Journal of Agronomy*. 2023;149:126917.
 26. Li T, Meng Q. Forest dynamics in relation to meteorology and soil in the Gulf Coast of Mexico. *Science of the Total Environment*. 2019;702:134913.
 27. Liu Y, Li N, Zhang Z, Huang C, Chen X, Wang F. The central trend in crop yields under climate change in China: a systematic review. *Science of the Total Environment*. 2019;135355.
 28. Lobell D, Field CB. Global-scale climate–crop yield relationships and the impacts of recent warming. *Environmental Research Letters*. 2007;2:014002.
 29. Mahesh P, Soundrapandian R. Yield prediction for crops by gradient-based algorithms. *PLOS ONE*. 2024;19(8):e0291928.
 30. Michaelides M, Mailhot M, Li Y. Probabilistic crop yield forecasts with spatio-temporal conditional copula using extreme weather covariates. 2025.
 31. Newlands NK, Zamar DS, Kouadio LA, Zhang Y, Chipanshi A, Potgieter A, *et al.* An integrated, probabilistic model for improved seasonal forecasting of agricultural crop yield under environmental uncertainty. *Frontiers in Environmental Science*. 2014;2:17.
 32. Ortiz-Bobea A. Climate, agriculture and food. *arXiv*. 2021;arXiv:2105.12044.
 33. Osborne TM, Wheeler TR. Evidence for a climate signal in trends of global crop yield variability over the past 50 years. *Environmental Research Letters*. 2013;8(2):024001.
 34. Osibo BK, Ma T, Wahab MMA, Jia L, Wenzheng Y, Bediako-Kyeremeh B, *et al.* Interpretable deep learning model for crop yield prediction: a case study of wheat yield prediction in Egypt. 2023.
 35. Parviz L, Paymai M. Comparison of the efficiency of classical and fuzzy regression models for crop yield forecasting with climatological aspect. *Poljoprivreda i Sumarstvo*. 2017;63(1):235.
 36. Qin M, Zheng E, Hou D, Meng X, Meng F, Gao Y, *et al.* Response of wheat, maize, and rice to changes in temperature, precipitation, CO₂ concentration, and uncertainty based on crop simulation approaches. *Plants*. 2023;12(14):2709.
 37. Ray DK, Gerber JS, MacDonald GK, West PC. Climate variation explains a third of global crop yield variability. *Nature Communications*. 2015;6(1):5989.
 38. Roth S, Lee BS, Nicholas RE, Keller K, Haran M. Bayesian spatial models for projecting corn yields. *Remote Sensing*. 2023;16(1):69.
 39. Salami ZA, Babamuratov B, Ayyappan V, Prabakaran N, Khudayberganov K, Singh C. Modelling crop yield prediction with random forest and remote sensing data. *Natural and Engineering Sciences*. 2025;10(2):67–78.
 40. Sarr AB, Sultan B. Predicting crop yields in Senegal using machine learning methods. *International Journal of Climatology*. 2022;43(4):1817–1838.
 41. Schauburger B, Gornott C, Wechsung F. Global evaluation of a semiempirical model for yield anomalies and application to within-season yield forecasting. *Global Change Biology*. 2017;23:4750–4764.
 42. Shammi SA, Meng Q. Modeling the impact of climate changes on crop yield: irrigated vs. non-irrigated zones in Mississippi. *Remote Sensing*. 2021;13(12):2249.
 43. Shirley R, Pope E, Bartlett M, Oliver S, Quadrianto N, Hurley P, *et al.* An empirical Bayesian approach to modelling the impact of weather on crop yield: maize in the US. *arXiv*. 2020;2001.02614.
 44. Sidhu BS, Mehrabi Z, Ramankutty N, Kandlikar M. How can machine learning help in understanding the impact of climate change on crop yields? *Environmental Research Letters*. 2023;18(2):024008.
 45. Stella T, Webber H, Olesen JE, Ruane AC, Fronzek S, Bregaglio S, *et al.* Methodology to assess the changing risk of yield failure due to heat and drought stress under climate change. *Environmental Research Letters*. 2021;16.
 46. Tomar V, Milind E. Climate-resilient smart agriculture: real-time IoT-driven crop yield prediction using adaptive ensemble machine learning. In: *Proceedings of the 1st International Conference on Intelligent Methods and Advanced Computer Scientific Innovations*; 2025.
 47. Van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: a systematic literature review. *Computers and Electronics in Agriculture*. 2020;177:105709.
 48. Vogel E, Donat MG, Alexander LV, Meinshausen M, Ray DK, Karoly D, *et al.* The effects of climate extremes on global agricultural yields. *Environmental Research Letters*. 2019;14(5):054010.
 49. Wang B, Li L, Feng P, Chen C, Luo JJ, Taschetto AS, *et al.*

- Probabilistic analysis of drought impact on wheat yield and climate change implications. *Weather and Climate Extremes*. 2024;45:100708.
50. Yavrum C, Selcuk-Kestel AS, Garrido J. Comparative analysis of weather-based indexes and the Actuaries Climate Index for crop yield prediction and weather-derivative pricing. 2025.
 51. Yin X, Leng G. Observational constraint of process crop models suggests higher risks for global maize yield under climate change. *Environmental Research Letters*. 2022;17.
 52. Zhang Y, Niu H, Yu Q. Impacts of climate change and increasing carbon dioxide levels on yield changes of major crops in suitable planting areas in China by the 2050s. *Ecological Indicators*. 2021;125:107588.
 53. Zhu X, Troy TJ, Devineni N. Stochastically modeling the projected impacts of climate change on rainfed and irrigated US crop yields. *Environmental Research Letters*. 2019;14(7):074021.
 54. Zscheischler J, Vogel J, Rivoire P, Deidda C, Rahimi L, Sauter C, *et al.* Identifying meteorological drivers of extreme impacts: an application to simulated crop yields. In: EGU General Assembly Conference Abstracts; 2021. p. EGU21-15524.