Benchmarking Random Forests and XGBoost for Global Wheat Yield Prediction

Sebastian Dohne

CID: [01776389]

Internal Supervisor:
Samraat Pawar

Word Count: 5853

A Final Project report submitted for the MSc in:
Computational Methods in Ecology and Evolution

Department of Life Sciences
Imperial College London
2025

1 Declaration and Acknowledgements

For this thesis I declare the use datasets provided to me by CABI, by Anna Szyniszewska and Salar Mahmood. Further Preprocessing, cleaning and further supplementation of the final datasets was performed by me, details of which are discussed in the following text.

Abstract

Global, field-scale prediction of wheat yields remains difficult under realistic distribution shifts. Using 22,709 yield observations from 228 locations across 29 countries (2005–2020) spanning 27 agro-ecological zones, multi-source climate, soil, vegetation, management, and socioeconomic co-variates were assembled and Tree-based ensemble models fit. Spatial dependence was controlled via a semivariogram-informed threshold to form spatial clusters for group k-fold cross-validation, and generalization was tested on a spatiotemporal holdout. Random Forest and XGBoost fit to the training data but showed limited transfer to new space–time domains (RF: R^{2} = 0.333, MAE = 1.142 t ha^{1}, RMSE = 1.764 t ha^{1}; XGB: R^{2} = 0.325, MAE = 1.433 t ha^{1}, RMSE = 1.775 t ha^{1}). SHAP analyses was used to rank feature importance and consistently elevated select management intensity (nitrogen rate, pesticide use) and socioeconomic context (GDP per capita) features as the strongest global signals, while climate, soil, and vegetation indices were weaker and less transferable. Overall, standard ensemble methods with rich co-variates did not yield globally generalizable wheat predictions, motivating phenology-aware data curation and models that explicitly handle domain shift or hybridize with process knowledge.

2 Introduction

Wheat *(Triticum aestivum and Triticum durum)*, throughout humanities agricultural history until the modern day, has been one of human kinds most important and culturally significant crops. It is one that has seen successfully cultivation on a global scale, with 2018 estimate stating a total land usage of 217 M ha and total crop production of 760 M mt in 2020 [Igrejas and Branlard, 2020] [Dadrasi et al., 2023] making it the most widely grown crop in the world. Factors pertaining to its environmental stability such as its adaptability to temperature variations, thriving in tropical and subtropical regions aswell as northern regions above 60°N, ability to grow at altitude (3000m above sea-level) [Enghiad et al., 2017] can be attributed to a genetic diversity present across strains and species that have genetic markers that contribute to resistances of abiotic stresses such as drought, salinity and temperature [Mao et al., 2023] as a result of natural selection and selective breeding on a global scale [Sun et al., 2020]. Currently, wheat alone supplies an estimated fifth of global food calories and protein, playing a crucial role in maintaining global food security [Erenstein et al., 2022].

Broadly, the yield intensification of wheat is primarily influenced by factors such as water stress, nutrient deficiency (such as N availability, SOC, Phosphorus etc) and growing media composition , solar radiation, growing season temperatures, management practices (equipment usage and pesticide/fungicide application, sowing date, planting density etc.) and the effects of weeds, pests and diseases [Raza and Bebber, 2022] [Lobell et al., 2009] [Zhou et al., 2019] [Ghimire et al., 2017]. Where these variables are optimised, a value for the so-called yield gap (Y_g) is generated against actual yield obtained by farmers in practice (Y_a). Digging deeper and observing climatic variables, wheat has been shown to have an optimum temperature of between 15 - 20°C [Food and Agriculture Organization of the United Nations (FAO) and International Institute for Applied Systems Analysis (IIASA), 2024] during the reproductive stage with every degree above optimum decreasing yield by 6%. In controlled environments, temperature stress has been shown to decrease overall yield, particularly when wheat is in different reproductive stages, reducing yields by up to 29-44%, with the timing of stress determining the observed mechanism of yield loss - anthesis stage stress can reduce grain number through impaired fertility, while stress during the grain-filling stage reduces individual grain weight through decreased photosynthetic capacity [Djanaguiraman et al., 2020]. In global models, the yield difference between attainable rainfed and irrigated yield in wheat variants was found to be 34 ± 9% [Wang et al., 2021].

As for soil variables, nitrogen fertilizer has been observed to have a consistent increase in grain yield from 2.10 ± 0.36 to 3.27 ± 0.51 t ha^{-1} at 0 and 300 kg N ha^{-1} respectively, with diminishing economic and growth returns being found at the 150 kg N ha^{-1} [Sapkota et al., 2020], acidification has also been observed in other studies at this level of nitrogen-based fertiliser application, which sees a critical yield threshold at ¡ pH 5 [Ghimire et al., 2017]. Globally, microbial, as well as fungal and viral pathogens can contribute to an annual reduction in wheat yield approaching 22% with fungal diseases such as rusts, powdery mildew and mosaic viruses being major contributors [Gupta et al., 2022]. Pests and weeds have also been observed to have a global effect on wheat yield of of 7.7 and 7.9 % respectively (CI 5-10 and 3-13 %) [Oerke, 2006]. A recent expansion on this concept posits the idea of a 'genetic yield gap' (Y_{ig}), which is defined by the difference between a genotypically optimized cultivar grown under irrigated (Y_{ip}) or rainfed (Y_{iw}) conditions to the strain used in practice. As such it is of importance to minimise both Y_g and Y_{iw} where possible [Senapati et al., 2022].

Crop yield prediction has been extensively studied across diverse geographical contexts and with a range of methods having been applied. Literature reviews revealed that Neural Networks, decision trees, Random Forest - both due to their ability to identify crucial patterns within the data - and gradient boosted trees, linear regression and Support Vector Machines are the most frequently applied algorithms. The with the most commonly used features being solar, soil, nutrient and humidity and water availability information. While traditional approaches have shown promise, recent research has increasingly adopted deep learning methods, with Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Deep Neural Networks being the most prevalent [Van Klompenburg et al., 2020] [Jabed and Murad, 2024]. Tree-based models have however have been a leading choice in the pursuit of predicting crop yield via regression, wheat yield, where in particular Random forest and XGBoost are considered the most field tested. They are useful in part as they require minimal data preprocessing while excelling at feature selection and are able to naturally handle collinearity, while while requiring limited computational resources, with this especially being the case for XGBoost. However, Random Forests have been prone to overfitting and biased towards categorical variables. Despite output interpretability being limited in biological context due to their "black box" nature than traditional regression, but are more likely to capture complex non-linear relationships [Jabed and Murad, 2024] [Cravero et al., 2022].

Knowledge gaps remain however, while numerous studies have compared statistical and machine learning approaches for crop yield prediction, these comparisons have been largely confined to regional studies or specific climatic contexts [Zampieri et al., 2018], Uncertainties in yield prediction can arise across space from limited knowledge of biophysical processes, variable accuracy of meteorological and field data and unknown local agronomical practices, with these variables typically usually only being available at the field scale. While global analyses have attempted to systematically evaluate or predict crop production, as is the case with the conceptualization of AEZ (Agro-ecological Zones) which define crop yield suitability zones by by different climatic and edaphic variables according to the the Food and agriculture organization (FAO) Framework [Fischer et al., 2021]. Historically a typical limitation, in finding an optimal predictive approach is that generalisable may vary between different environmental contexts. As such supervised approaches using global high resolution yield and associated co-variate data have been minimally leveraged [Wang et al., 2018] [Qin et al., 2024], especially, where additional globally available features with strong predictive accuracy become easier to access, such as vegetation indices [Muruganantham et al., 2022]. what works well in temperate regions may perform poorly in arid or tropical zones. This gap is particularly important given growing concerns about model interpretability and the "black box" nature of machine learning approaches in agricultural applications.

Via the use of a comprehensive historical global dataset of annual crop yields with diverse climatic, soil, management, and biotic co-variates, I will employ supervised gradient boosting and bagging ensemble methods to develop predictive models across heterogeneous global agricultural systems. I will evaluate model performance using standard metrics and analyze feature importance to identify key drivers of yield variation globally. Subsequently, I will assess the model's generalization capability by testing its predictive accuracy on unseen regions in space and time, leveraging powerfully globally available satellite-derived and environmental features to assess broad applicability.

3 Materials and Methods

3.1 Data Collection

In order to build a global predictive model, the following project outline was followed.

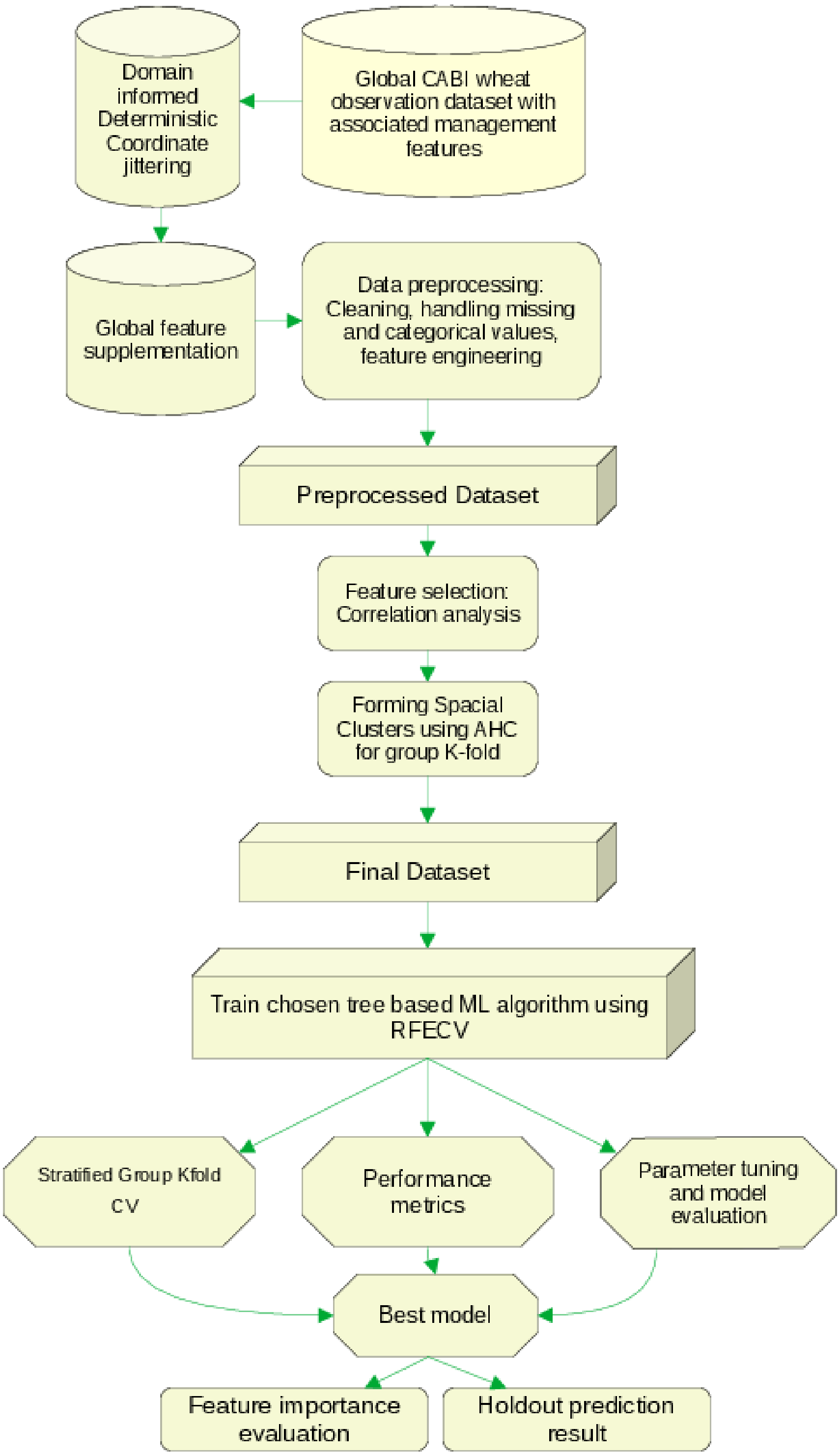

Domain informed Deterministic Coordinate jittering
Global CABI wheat observation dataset with associated management features
Global feature supplementation
Data preprocessing: Cleaning, handling missing and categorical values, feature engineering
Preprocessed Dataset
Feature selection: Correlation analysis
Forming Spacial Clusters using AHC for group K-fold
Final Dataset
Train chosen tree based ML algorithm using RFECV
Stratified Group Kfold CV
Performance metrics
Parameter tuning and model evaluation
Best model
Feature importance evaluation
Holdout prediction result

Figure 1: Model Architecture Outline

3.2 Data preprocessing and collection

The global wheat yield dataset utilized in this project was provided by CABI and spans a total global period of between 2005-2020, encompassing 228 unique locations across 29 countries worldwide with each having individual temporal ranges. Yield values are measured in tons per hectare (t ha^{-1}) and represented values associated with the final harvest number. The dataset includes diverse site types and their associated coordinates, ranging from commercial farms to institutional research fields, distributed across 27 distinct agroecological zones. with zones being defined globally and representing regions with varying levels of aridity, temperature regimes, and associated climatic factors that otherwise influence influence wheat yield production. Alongside these values was a comprehensive feature-set, where associated management variables were recorded per observation.

Before analysis of the data could begin, preprocessing was required, Given that the yield values obtained were from many different types of institutions, These had to be individually combed through to analyze which final values would hurt the model or not. extreme yield locations such as vertical farms or greenhouses were excluded as they would hurt Generalisability, and as such a max yield value of less than 30 ha. acres was settled on. Most columns had to be cleaned and high dimensionality columns had to be categorized using domain knowledge. Repeated rows were removed and extremely over-represented areas down-sampled where possible with observations, by way of preserving rows which contained the most variability and allowed for global generalisability. Where associated coordinates were unreliable (i.e in the ocean), these were critically assessed for reliability and excluded from the final dataset if they were deemed unfixable. This left the final dataset with 22709 individual yield values.

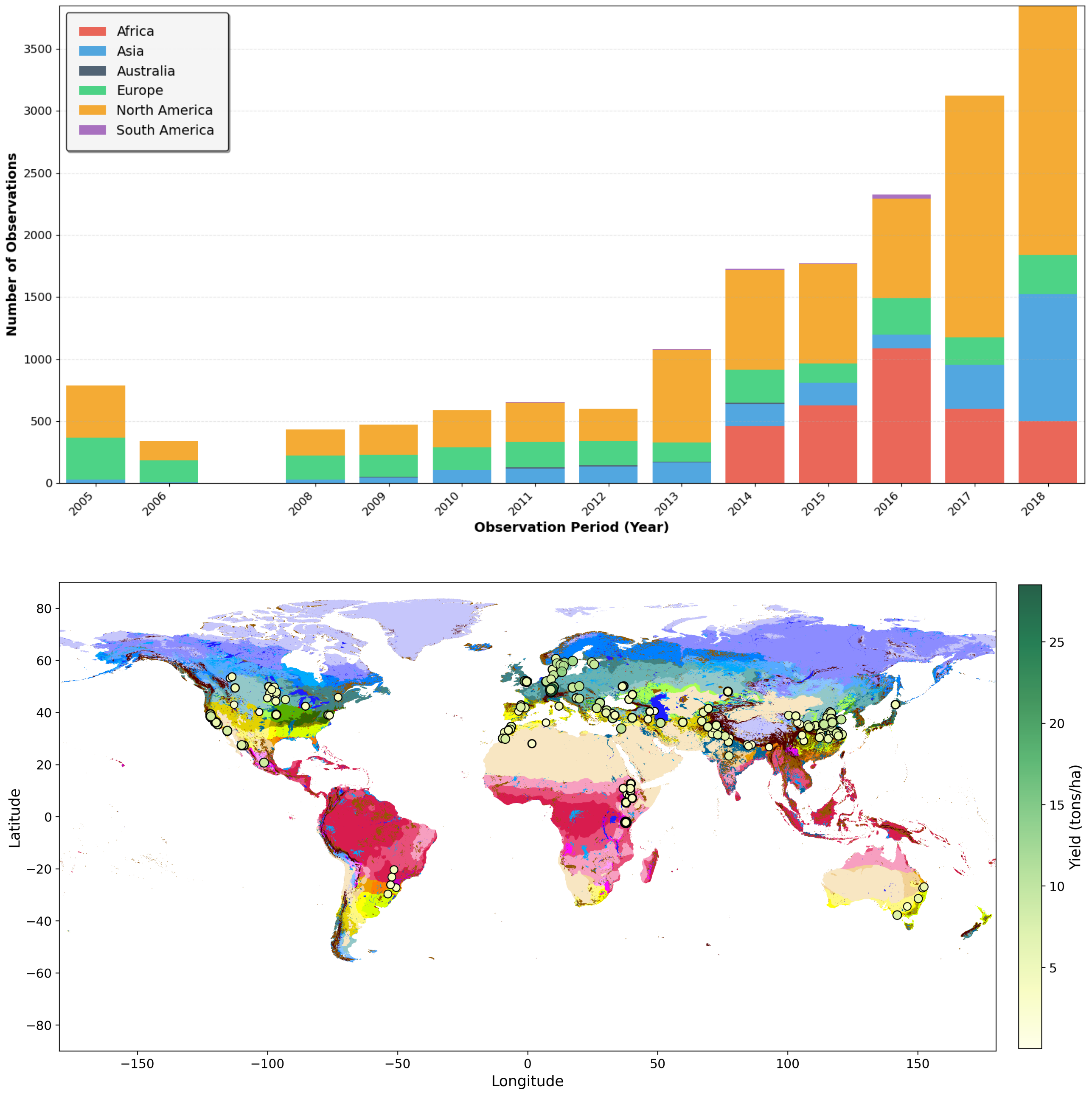


Figure 2: Plot showing frequency of obsrvations by continent and point data of all observation locations in the dataset overlayed on a world-map of AEZ zones, colours refer to individual zonal categories which were defined by different climatic and edaphic variables according to the Food and agriculture organization (FAO) Framework [Fischer et al., 2021].

A decision was also made, in order to better capture potential real-life micro-variability within individual sites, to implement a jiterring algorithm to the coordinate values which scaled with observation count per site per year during preprocessing as coordinate values associated with observation location were to the second decimal place and thus were not accurate to the field level. For each observation, the algorithm calculates an adaptive sampling radius using the equation: $r_{\text{jitter}} =$

$\min(r_{\max}, r_{\min} + \alpha \cdot \ln(n_{\text{obs}}))$ using a min of 0.5km , max of 10km and α of 0.7 and per site, so locations with more observations received larger sampling areas to explore greater spatial heterogeneity. Multiple coordinate samples are generated by randomly sampling displacement distances $d \sim \mathcal{U}(0, r_{\text{jitter}})$ and bearing angles $\theta \sim \mathcal{U}(0, 360)$, followed by a geodesic transformation to calculate new sampling positions as $\text{New Point} = \text{destination}(\text{Original Point}, d, \theta)$. The collision avoidance mechanism ensures that samples maintain a minimum separation distance ($d_{\min} = 100$ m) to prevent redundant sampling of identical micro-climates, using iterative rejection sampling with up to 50 attempts per point. The end result is that each observation now had a unique jittered coordinate value inside the defined radius. As such this approach attempted to capture spatial variability in environmental conditions that exist around reported farm locations, providing a more representative characterization of the growing environment by location.

Global features were assembled in accordance to their direct relevance to wheat biology, and by their observed efficacy in ML modeling approaches [Van Klompenburg et al., 2020]. For crops in general, factors like temperature and precipitation where extremes and stresses can impede growth by reducing carbon assimilation by compromising chloroplast proteins and membranes, as well trigger survival mechanisms such as the triggering stomatal closure in order to prevent waterloss, thus reducing photosynthesis. Other factors include the impeding of seedling and flower development through the reduction of leaf expansion rates. drought stresses are known to to reduce photosynthetic capability and damage hydraulic tissues as deficits persist. Excessive precipitation can lead to excessive soil-moisture and as anoxia and impeded gas exchange in the rhizophere, and root-tip damage inhibiting uptake among other factors. Heat moisture interactions can be particularly bad as they can cause crop to keep stomata closed to conserve water while heat damage increases [Lesk et al., 2022], with the effect of these stressors being particularly pronounced at critical growth stages like anthesis or grain-filling [Djanaguiraman et al., 2020].

Vegetation indices, are popular predictive features because they utilize spectral signatures emitted from plants, and thus are essentially a direct approximation of plant health and photosynthetic activity. NDVI(normalized vegetation index) is one of the most widely used and is a normalized measurement of the visible and near-infrared (NIR) spectral bands that analyses the contrast between the two, where crops absorb red-light when they are photosynthetically active more effectively. EVI (Enhanced vegetation index), was developed to address problems in regards to signal saturation by the inclusion of blue light bands and as such is considered to be more effective in cases to vegetation vitality than pure NDVI [SKINNER et al., 1994]. LAI and FPAR are further proxies of Vegetation health. The former measures leaf density directly in relation to the floor below and FPAR is a measurement of the photosynthetically active radiation absorbed by plants (between 400-700nm) within a vegetation canopy.

Given the temporal and geographic scope of the dataset, along with limited availability and reliability of direct pathogenic and management data for specific locations, we utilized proxies to infer potential pathogen, pest, and weed presence. Annual herbicide, pesticide, and bactericide application rates can indicate outbreak patterns or responses to climate-driven changes in pest pressure. Additionally, regional GDP serves as a proxy for management intensity, reflecting farmers' access to resources such as irrigation systems and farming equipment [Najafi et al., 2018].

3.3 Feature matching and engineering

In order to match predictive features to the dataset, The approximate start and end dates to the growing seasons were defined by the MIRCA2000 global irrigated/rained cropping calendar, which defines the proximal months in which different growing seasons are observed by region globally via the use of climatic variables yearly. These months were matched to each observation by harvest date/sowing year and coordinates. wether or not cropland was considered to be irrigated or not was defined by AEZ classification and proximal level of irrigation quantified. Using this framework, competitive features could be joined to the dataset by jittered coordinate/observation year and monthly time series features across growing seasons could be associated with yield values via the use of the defined growing start date. A comprehensive feature set was compiled encompassing climatic, biotic, management, and socioeconomic variables via the use o a multi-source data integration approach. Raw netCDF files were obtained from primary literature, APIS and web-scraping techniques, while vegetation indices (EVI, LAI, and FPAR) were acquired using the MODIS Coordinate Extraction Tool APPears to extract point-specific data across the complete temporal range [AppEEARS Team, 2025] and aligned with the existing dataframe structure. To ensure spatial and temporal consistency, variables not available in monthly, 10km resolution format were temporally averaged to monthly intervals and spatially coarsened to 10km resolution where technically feasible. MODIS pixels affected by cloud cover or atmospheric interference were retained as missing values. Visual inspection of missing data patterns across space and time suggested no spacial or temporal bias, and missing values were imputed using only when required by modeling algorithms.

Finally, these features were further engineered to create domain-informed indices that capture comprehensive temporal patterns where individual time steps might miss critical information. This was considered particularly important for time series features, where individual monthly values (e.g., NDVI-1 through NDVI-9) provide only monthly snapshots of dynamic processes. To address this limitation, combinatorial metrics such as area under the curve (AUC), seasonal averages, and variability measures and indexes were created to complement the granular monthly data. This approach provides both detailed temporal information at the monthly scale and integrated seasonal patterns that capture the full growing cycle dynamics relevant to crop yield prediction to provide the constructed model with a full comprehensive image of the feature space, and have demonstrated efficacy in increasing predictive accuracy in crop models [Johnson et al., 2021].

Variable Type	Variables	Description	Source
Remote Sensing Data			
NDVI	NDVI1-9, unitless	Normalized Difference Vegetation Index: 9-period time series. $NDVI = \frac{NIR-Red}{NIR+Red}$	Copernicus [Copernicus Global Land Service, 2021], 10km
EVI	EVI1-9, unitless	Enhanced Vegetation Index: 9-period time series. $EVI = 2.5 \times \frac{NIR-Red}{NIR+6Red-7.5Blue+1}$	MODIS/Terra [Running et al., 2021], 500m
LAI	lai_month_1-7, m^2/m^2	Leaf Area Index: 7-month time series	MODIS/Terra [Myneni et al., 2021], 500m
FPAR	fpar_month_1-7, fraction	Fraction of photosynthetically available radiation: 7-month time series	MODIS/Terra [Myneni et al., 2021], 500m
Climate Data			

Continued on next page

Continued from previous page

Variable Type	Variables	Description	Source
Temperature	temp1-9, °C	Monthly temperature time series with seasonal statistics	CABI, 10km
Precipitation	prc1-9, mm	Monthly precipitation time series with seasonal totals	CABI, 10km
Evaporation	Evap1-9, mm	Total evaporation: monthly time series	GLEAM [Miralles et al., 2025], 10km
Evaporative Stress	EvapS1-9, unitless	Evaporative stress factor (evaporation/potential evaporation)	GLEAM [Miralles et al., 2025] [Hulsman et al., 2023], 10km
Soil Moisture	SoilM1-9, m^3/m^3	Root zone soil moisture time series with seasonal statistics	GLEAM [Miralles et al., 2025] [Hulsman et al., 2023], 10km
Transpiration	Trans1-9, mm/day	Plant transpiration time series	GLEAM [Miralles et al., 2025], 10km
Soil Data			
Soil Texture	Clay, Sand, Silt (%)	Soil texture percentages (0-30cm depth)	SoilGrids [Poggio et al., 2021], 10km
Soil Chemistry	pH, SOC, Soil_N	Soil pH, organic carbon (cg/kg), available nitrogen (cg/kg)	SoilGrids [Poggio et al., 2021], 10km
Topography	Elevation (m), AEZ	Elevation above sea level and agro-ecological zone	CABI/FAO [Fischer et al., 2021]
Management Data			
Crop Variety	Crop.variety	Specific wheat variety/cultivar	CABI

Continued on next page

Continued from previous page

Variable Type	Variables	Description	Source
Tillage	Tillage.type	Tillage practice used	CABI
Irrigation	% irrigated (0-1)	Irrigation percentage by area	MICRA2000
Nitrogen	N.rate (kg/ha), treatment type	Nitrogen application rate and treatment type	CABI/CROPGRIDS [Tang et al., 2023]
Pest data	Pest presence(yes/no)	Pest presence and detection	CABI
Pathogen and pest management	Use per area (kg/ha), (tonnes)	Fungicide, herbicide and bacteriocide use	FAOSTAT [FAO, 2024]
Socioeconomic Data			
GDP	GDP per capita (USD)	GDP per capita by region	Gridded dataset [Kummu et al., 2025]
Year	Observation period	Harvest year (technological progress proxy)	CABI
Target Variable			
Grain Yield	Grain.yield (t/ha)	Wheat grain yield per hectare	CABI/CROPGRIDS [Tang et al., 2023]

Table 1: Data Variables and Sources for Crop Yield Prediction Model

For missing values in the dataset an imputation strategy using the mice package in R with a random forest–based approach was implemented [Van Buuren and Groothuis-Oudshoorn, 2011]. All variables with more than 50% missingness were removed prior to imputation and from the final dataset. For the remaining variables, a single-pass random forest imputation (method=”rf”) to generate a completed dataset. To reduce computational burden and avoid unstable imputations, the predictor matrix was pruned using quick-pred with minimum correlation and usable-case thresholds of 0.1 and 0.25 respectively. In the case of high-cardinality categorical variables (=>100 levels), these were restricted to being imputed only from spatial co-variates. In order to percent target leakage into the full dataset the outcome variable (yield) was excluded from the predictor set.

3.4 Feature selection

When refining the feature selection for continuous variables, in order to prevent overfitting, a combinatorial approach of using pearsons correlation matrices using a high threshold of 0.95 then choosing features more correlated with the target yield variable, aswell as Recursive feature selection was opted for when training the global model in favor of other traditional feature selection methods such as VIF(features with zero variation were removed however) or specifically LASSO regression as is commonly observed in agricultural modeling. This is because in the feature set there are many unseen interaction effects and non-linear relationships that the model can learn from. Both these approaches were deemed the best option as the high multi-collinearity threshold of 0.95 percent ensured essentially redundant features would be removed, and recursive feature selection works backwards when training the model, beginning with the whole feature-set, in theory capturing all interaction effects and working backwards to remove features that were not observed to interact with others [Ganati and Sitote, 2025]. Categorical features were either one-hot encoded if there were less than 20 categories, or binary encoded if there were more, in order to minimize the curse of multidimensionality.

3.5 Spacial and temporal autocorrelation

For the model to generalize globally, without overfitting, the potential presence of spacial and temporal autocorrelation needed to be addressed which is the idea the places closer to each other in space and time tend to be more similar to each other [Roberts et al., 2017], and is a pattern found most datasets that have any kind of spacial-temporal structure, and it is typically not possible to capture all of these similarities within a feature-set. In predictive crop yield modeling, this could exemplify itself via similarities within the soil micro-biome, soil compaction on farms from heavy machinery use, unrecorded persistence of pathogens in the soil or past crop residue among many other potential factors which are difficult to capture within a feature-set. For these sites within close points in time. These factors change slowly over time, requiring years to vary sufficiently for effective model generalization across temporal scales.

In spacial modelling, Autocorrelation is typically mitigated via a the use of an appropriate group K-Fold strategy. For this dataset, it was decided that an agglomerative hierarchical clustering (AHC) approach should be used in order to cluster locations together into spacial groups based on a haversine distance threshold informed via the use of a semivariogram, which quantifies how the similarity between observations decreases with increasing distance. Spatial autocorrelation was assessed using semivariogram using mean yield across years per location, with this quantifying how the similarity between observations decreased with increasing distance in the dataset [Wang et al., 2023].

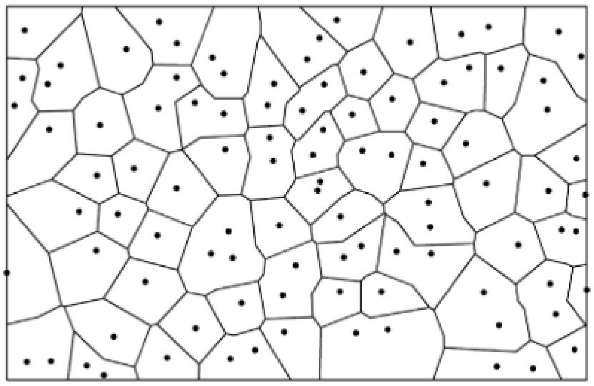
The spatial semivariogram (semi-variance) quantifies average dissimilarity with distance,

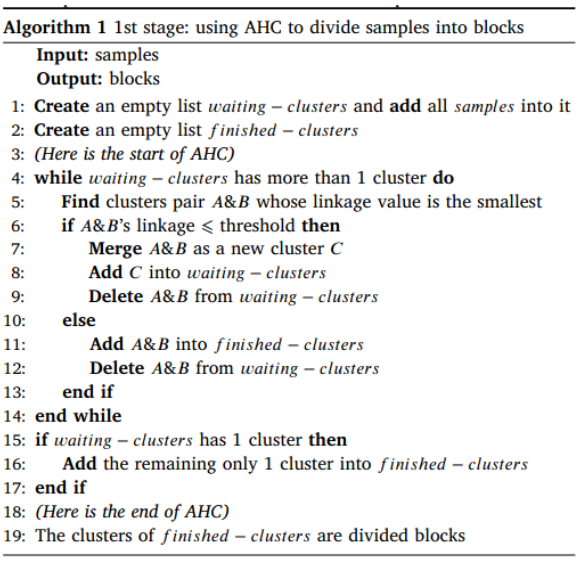
$$\gamma(h) = \tfrac{1}{2}\, E\big[(Z(x) - Z(x+h))^2\big]\,,$$

where h is the lag (separation) distance, $Z(x)$ the observed yield at x, and $E[\cdot]$ the expectation over all pairs at distance h. Its empirical form is

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{(i,j):\, d(x_i,x_j)\approx h} [Z(x_i) - Z(x_j)]^2,$$

with $N(h)$ the number of such pairs [Wang et al., 2023, Gasch et al., 2015]. We used the estimated *range* (plateau distance) to set a great-circle (Haversine) cutoff for agglomerative hierarchical clustering (AHC), yielding 179 spatial clusters. To handle unequal cluster sizes and prevent spatial leakage, we applied a greedy Group K-Fold that kept clusters intact and created balanced 10 folds.





```
Algorithm 1 1st stage: using AHC to divide samples into blocks
    Input: samples
    Output: blocks
Create an empty list $waiting-clusters$ and add all $samples$ into it
Create an empty list $finished-clusters$
(Here is the start of AHC)
while $waiting-clusters$ has more than 1 cluster do
    Find clusters pair $A\&B$ whose linkage value is the smallest
    if $A\&B$'s linkage $\leqslant$ threshold then
        Merge $A\&B$ as a new cluster $C$
        Add $C$ into $waiting-clusters$
        Delete $A\&B$ from $waiting-clusters$
    else
        Add $A\&B$ into $finished-clusters$
        Delete $A\&B$ from $waiting-clusters$
    end if
end while
if $waiting-clusters$ has 1 cluster then
    Add the remaining only 1 cluster into $finished-clusters$
end if
(Here is the end of AHC)
The clusters of $finished-clusters$ are divided blocks
```

Figure 3: The AHC algorithm as detailed in the Spacial+ paper aswell as a visual of the final formed spacial groups informed by the semi-variance threshold [Wang et al., 2023]

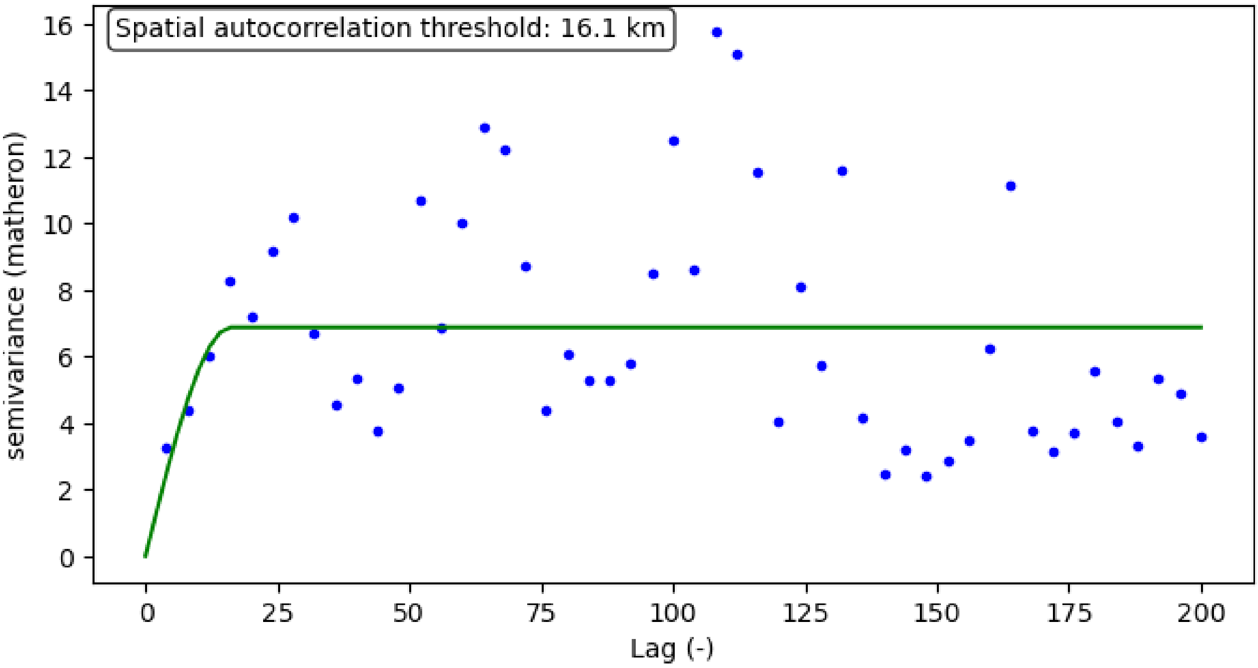


Figure 4: The above semivariogram reveals spatial structure through three key parameters: the range (distance where semivariance plateaus, indicating spatial independence), the sill (maximum semivariance value at the plateau), and the nugget (semivariance at zero distance, representing measurement error and micro-scale variation). A spacial autocorrelation threshold was identified at 16.1km.

While spacial independence was ensured during testing, the model was able to train on the entire and past location data between 2005-2018. While techniques to avoid temporal leakage such as forward chaining would have liked to have been employed, this was not possible given the dataset structure and varying temporal ranges for locations. So for our testing set, in order to address the temporal autocorrelation aspect, an independent holdout testing group was created between the years 2019-2020, in order to validate the models ability to generalize to unseen locations in space and time, as is recommended when this underlying pattern is present [Roberts et al., 2017]. The holdout set comprises 1,100 observations (5% of the total dataset) drawn from 17 unique locations, validated by the spacial autocorrelation threshold across 14 agroecological zones, spanning Asia, Europe, North America, and Africa, was used to assess the model's ability to generalize.

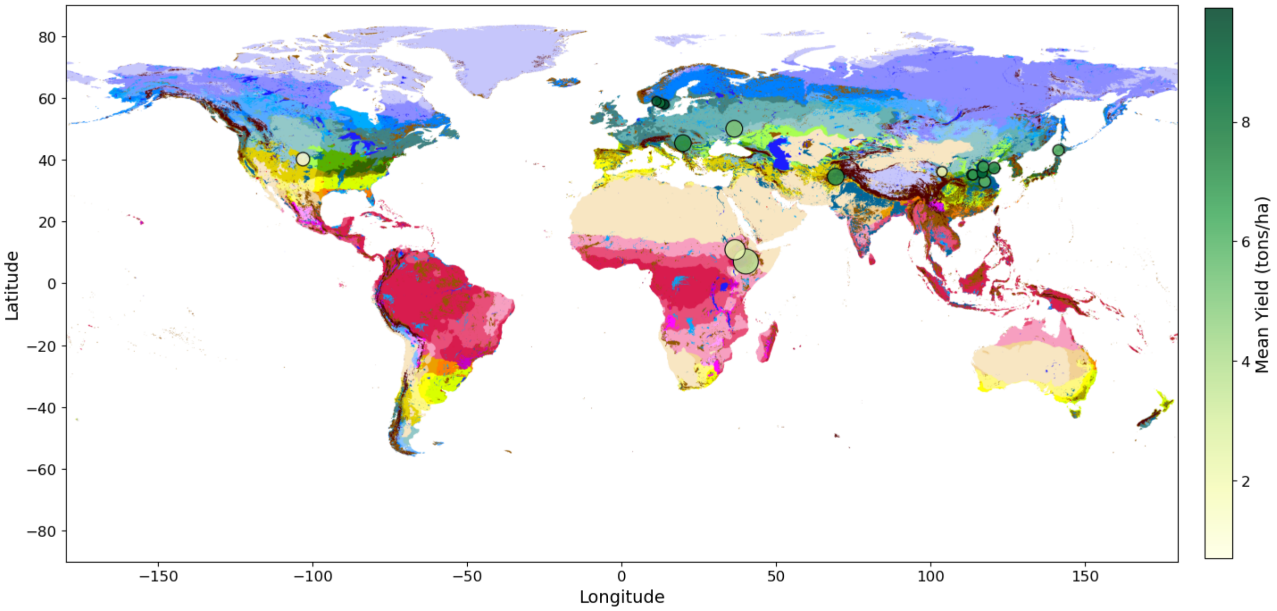


Figure 5: 2019-2020 Spatiotemporal Holdout between set locations overlayed over AEZ zone map. Bubble size refers to observation count per location

4 Models

4.1 Random Forest Regressor

Random Forest is a supervised learning algorithm that operates by constructing an ensemble of decision trees during training and outputting the average prediction of the individual trees for regression tasks [Breiman, 2001]. Each tree is trained on a different bootstrap sample of the data, and at each split, only a random subset of features are considered, helping to reduce overfitting and de-correlate the trees. This ensemble approach enhances generalization performance by reducing variance without increasing bias. Random Forest is particularly robust to noise, can naturally model complex nonlinear relationships, performing well with both numerical and categorical input and requires minimal hyperparameter tuning.

Hyperparameters

Key Random Forest hyperparameters include:

- **n_estimators**: Number of trees in the forest.
- **max_depth**: Maximum depth of each tree.
- **min_samples_split**: Minimum number of samples required to split an internal node.
- **min_samples_leaf**: Minimum number of samples required to be at a leaf node.
- **max_features**: Number of features to consider when looking for the best split.
- **bootstrap**: Whether bootstrap samples are used when building trees.

4.2 XGBoost Algorithm

XGBoost is a supervised learning algorithm that leverages gradient boosting to make accurate predictions for a target variable by amalgamating an ensemble of estimates from a group of simpler and weaker models. The algorithm functions by adding weak learners, in the form of decision trees, to the ensemble iteratively in order to minimize gradient loss, with each successive model trying to correct the errors of the previous models. Like random forest it can handle both numerical and categorical data and can detect non-linear trends in data as well as handle missing values. Apart from its flexibility, XGBoost is scalable, parallel, distributed, out-of-core, and uses cache-aware computing features and comes with an extensive range of hyper-parameters to fine-tune performance. This enables it to handle large datasets while limiting computing time, thus making it particularly suitable for yield regression [Chen and Guestrin, 2016].

Hyperparameters

Key XGBoost hyperparameters include:

- **max_depth**: Maximum depth of each tree ($q(x)$).
- **eta (learning rate)**: Shrinks the contribution of each tree.

- **subsample**: Fraction of training data sampled per tree.
- **colsample_bytree**: Fraction of features sampled per tree.
- **lambda** (λ): L_2 regularization on leaf weights.
- **alpha** (α): L_1 regularization on leaf weights.
- **gamma** (γ): Minimum loss reduction required to make a split.
- **min_child_weight**: Minimum sum of instance weights (h_i) in a child node; larger values make the algorithm more conservative.
- **n_estimators**: Number of boosting rounds (K).

5 Accuracy measurements for modelling

The following criteria were selected to be evaluated when observing our regression model output as they represent the standard when evaluating supervised regression models. where t_i is the true yield value and y_i is the predicted yield value.

Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(t_i - y_i)^2}{\sum_{i=1}^{n}(t_i - \bar{t})^2} \tag{1}$$

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - y_i)^2} \tag{2}$$

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|t_i - y_i| \tag{3}$$

For measuring feature importance in the models, the SHAP package was used due to its integration into the Scikit-learn and XGBoost ecosystem. SHAP is based on the Shapley value which for i is defined as:

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

where N is the set of all features, S is a subset not containing i, and $v(S)$ is the model output using features in S. The four axioms SHAP satisfies are: (i) **Efficiency** – the contributions of all features sum to the model output, (ii) **Symmetry** – if two features contribute equally they receive the same value, (iii) **Dummy** – features with no effect receive zero value, and (iv) **Additivity** – contributions remain consistent when models are combined as is the case in the ensemble models utilized. Thereby providing a mathematically justified approach to interpreting model predictions and quantify the relative importance of the final feature-set used in explaining our yield variability. Since our the models used were tree-based algorithms, tree-explainer was used to calculate SHAP values for speed and accuracy.

Random Forest and XGBoost regressor were cross-validated using a 10-group Kfold approach and parameters were identified via a randomized grid search, which has been shown in previous yield modelling studies to improve computational efficiency without reducing accuracy. The optimal parameters obtained were then applied to recursive feature elimination with cross-validation (RFECV), which reduced the feature space in steps of five to a minimum of 50 features, using mean absolute error (MAE) as the scoring metric. Redundant features were removed, and a second randomized search was conducted on the reduced feature set to produce the final models. For XGBoost, the final parameters were: 600 estimators, maximum tree depth of 5, learning rate of 0.00267, subsample and gamma values of 1.0, $\alpha = 0.0167$, $\lambda = 10.0$, minimum child weight = 1, and column sampling by tree (`colsample_bytree`) = 0.6. Runtime was 38 minutes, with 120 features selected from 130. For Random Forest, the final parameters were: 600 trees, maximum depth of 18, and `log2` feature sampling at each split. Regularization was applied via 10 samples per leaf, a minimum impurity decrease of 1×10^{-7}, and cost-complexity pruning with $\alpha = 0.004$. The model was trained without bootstrapping (`bootstrap = False`). Runtime was 138 minutes, with 95 features selected from 125.

5.1 Computing Tools

All computations were performed on a Linux system using a 12th Gen Intel® Core™ i7-1265U processor with 12 threads (6 cores, 2 threads per core). with CPU frequencies ranging from 400 MHz to 4.8 GHz. Analysis was conducted in jupyter notebooks using python version 3.12.7 using libraries such as XGboost and scikit-learn. A full dependency list for the vitual environment used aswell as other packages mention in the paper will be listed in the appendices.

6 Results

Using the trained models on the holdout testset yielded the following results:

Model	$\mathbf{R^2}$	**MAE(Tons/ha)**	**RMSE(Tons/ha)**
XGBoost	0.325	1.433	1.775
Random Forest	0.333	1.142	1.764

Table 2: Performance metrics for ensemble models on the holdout test set.

Both XGBoost and RF were unable to predict yield with high accuracy on the spatially independent future test set, and showed similar performances on the validation sets. The models achieved prediction accuracy within approximately 1.4 tons/ha on average (MAE), with RMSE values of 1.77 tons/ha indicating some larger prediction errors when observing outliers in the dataset. this represents roughly 25% predictive error relative to the actual yields recorded. While performance was similar between the two models, Random forest had a slight edge in all three categories compared to its counterpart, with an MAE 0.291 lower.

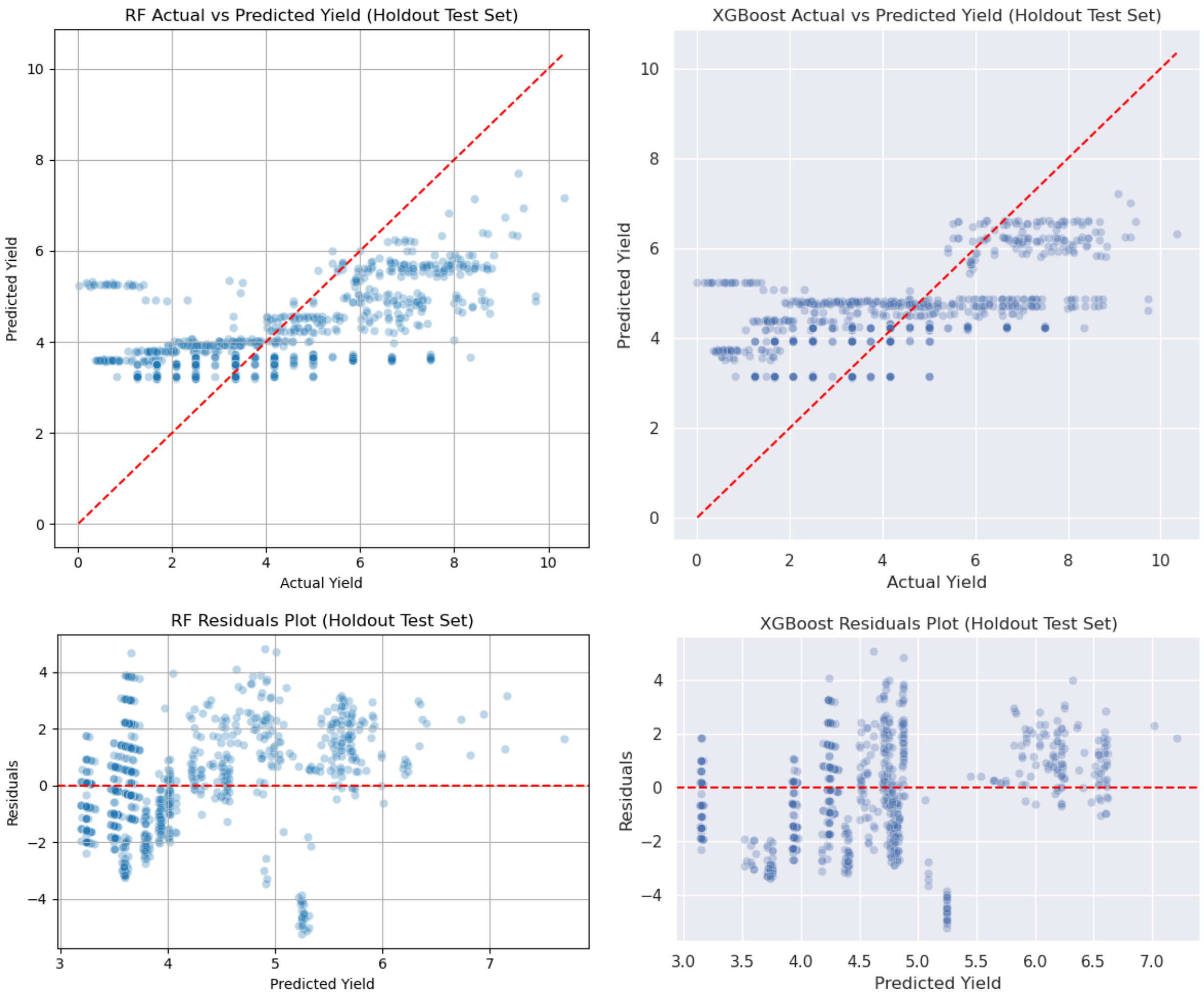


Figure 6: Model performance evaluation on spatiotemporal holdout test set. Top panels show predicted vs. actual wheat yields for Random Forest (left) and XGBoost (right) models. Bottom panels display corresponding residual plots.

Observing the final predictive patterns in the test data shows evidence of poor generalization to the holdout dataset and strong overfitting to the training data. Vertical clustering in the residual plots indicate that both ensemble models defaulted to certain predicted values regardless of final yield variability. This pattern suggests the models learned representations that were too specific to the training data distribution rather than capturing generalizable relationships across the full feature space. Both Models performed particularly poorly when it came to predicting yields at extreme ranges. Where 95% of actual yield values in the holdout fell between 0.6 and 8.43, for Random Forest 95% of the predicted values range was between 3.24 and 5.90 and for XGBoost a similar pattern occurred where predicted yield had a range of 3.15 to 6.56.

Using SHAP to compare feature importance recorded in the global models and comparing this to the holdout detailed patterns showed the following results:

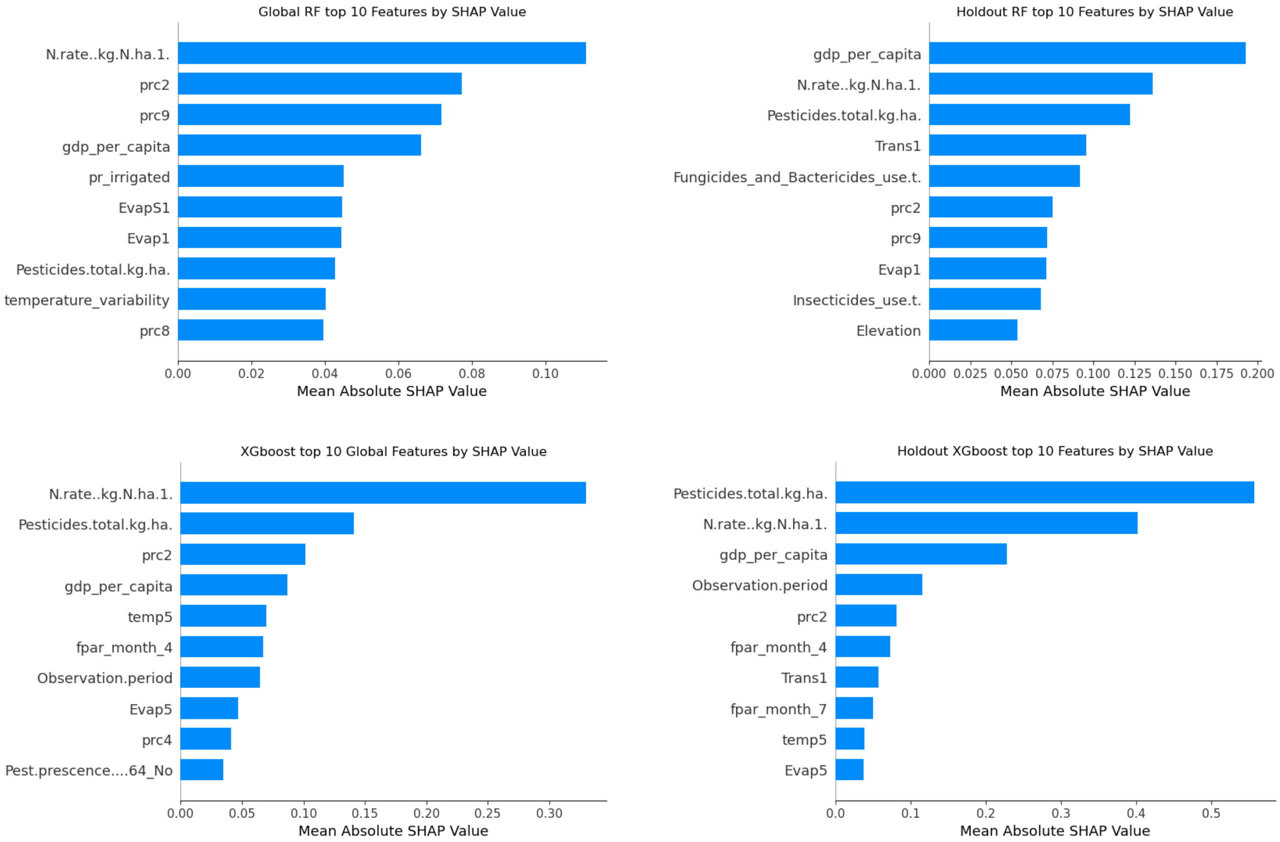


Figure 7: Mean absolute SHAP values comparing training and holdout performance. Top row: Random Forest (RF) models. Bottom row: XGBoost (XGB) models. Left column shows training data, right column shows holdout data.

General Analysis of Absolute SHAP values reveals substantial differences in feature importance between training and holdout datasets, providing evidence of poor model generalization. However Across both Random Forest and XGBoost models, nitrogen application, pesticide use per hectare, and GDP per capita consistently ranked among the most important features, indicating that management variables and economic factors strongly influence yield predictions. Among climatic variables, precipitation in the second month (prc2) appeared consistently important across models, which broadly aligns with agronomic knowledge, aligning broadly with the tillering life-stage for wheat for most crop varieties [Acevedo et al., 2006].

However, feature importance rankings diverged significantly between training and holdout datasets beyond these core variables. This inconsistency indicates that the models learned feature relationships specific to the training regions and failed to transfer these to new geographic regions. Particularly concerning is the high importance of 'Observation.period' in the XGBoost holdout predictions, suggesting the model defaulted to temporal patterns rather than learning meaningful climate-yield relationships when encountering unfamiliar conditions.

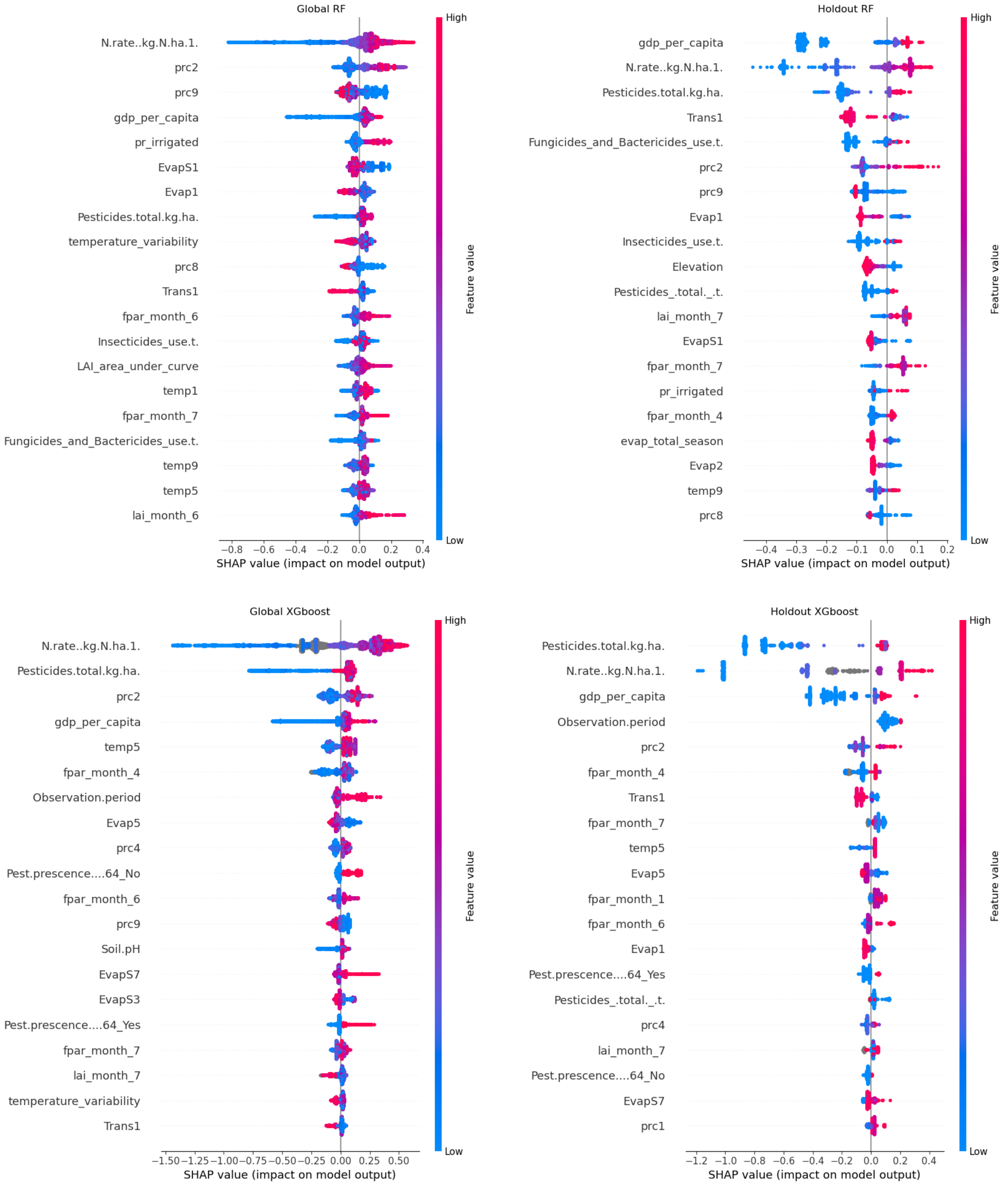


Figure 8: SHAP beeswarm plots showing feature value distributions and impacts of top 20 features per model. Top row: Random Forest (RF) models. Bottom row: XGBoost (XGB) models. Left column shows training data, right column shows holdout data.

Observing the The Bee-swarm provide more a clearer picture in how the top features as rated by the model provide more insight in how the most important features observed influenced the model output and how. Likely the clearest pattern observed across models was how Nitrogen application overwhelmingly increased final yield output, a pattern which was not only observed in the testing set, but also in the holdout with SHAP values covering the largest range and showing the clearest pattern with increases in it translating to increases in yield according the the models. A similar pattern was observed with the other management features observed, mainly pesticide application aswell as GDP per capita. Reaffirming that management and technological access above all leads to direct increase in yield. While model final model accuracy was low, the testing and holdout data showed evidence of monthly and feature engineered climatic variables and vegetation indices making small yet meaningful impact given the number of final features within both the XGBoost and RF global models. Features relating to soil biochemistry, specific treatment categories and crop.variety were notably absent from the final model, indicating their potential lack of predictive ability on a global scale.

While interaction effects can't explicitly be quantified due to the nature of the models used to the 'black box' nature of the models used. Evidence of interactions can be captured observed using SHAP. An interaction matrix between features for their respective SHAP values per observation was generated.

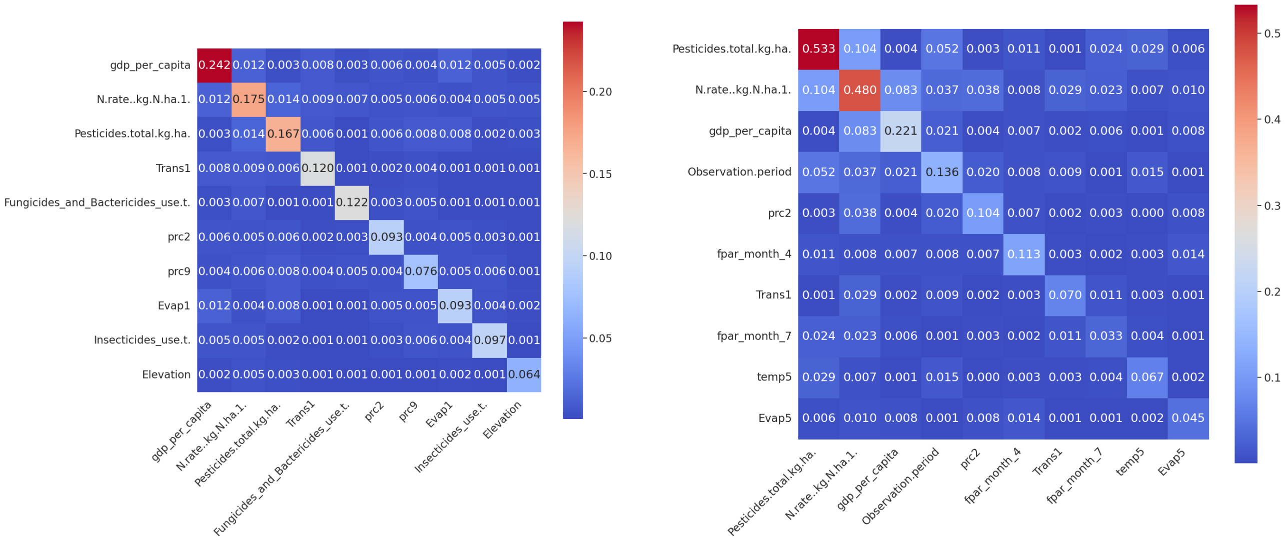


Figure 9: Matrix of mean absolute Tree-SHAP interaction values for the top–10 features on the holdout set (diagonal = main effects; off–diagonal = doubled pairwise interactions; units are model output. Left for RF, Right for XGboost).

The results of the pairwise interaction matrix of the top features suggest that across the holdout test set, while top predictors dominated the models with the model being mostly additive in the holdout.

Unfortunately, interation effects between climatic variables were not evident in the results and likely not able to be generalized by the model, and the only evidence of weak synergistic feature relationships being only observed between the top predictors representing socioeconomic and management variables, with this being the case across both models.

7 Discussion

The Results of the global model revealed that despite using a comprehensive global feature set covering the most important drivers pertaining to wheat yield across a significant temporal span, it remains difficult to build a an ensemble model while taking spacial and temporal autocorrelation into account, that is able to generalize with strong predictive accuracy to areas unseen in space and time, in a way that it sufficiently captures feature importance and interaction effects in a way that typical regional yield models are able to, and performed poorly to other global yield prediction models [Cao et al., 2025] [Blanc and Sultan, 2015]. The paper did however provide an insight into, was the importance of general certain management practices in the form of fertilizer and pesticide application as well as location based economics in regards to yield an the importance of these practices globally. As such the paper is a cautionary tale to other prospective crop yield modelers that may want to attempt this exercise, but it may provide some insight into important features that transfer to modelling regardless of location.

The dataset presented significant challenges, particularly spatial and temporal imbalances in observation counts. A key trade-off emerged: retain valuable management-related and biotic features from data-rich experimental sites in overrepresented locations, preserving important yield variability but risking overfitting and spatial bias; or discard many observations to reduce bias, but lose critical information and variability. Despite preprocessing efforts including strategic down-sampling of over-represented locations and coordinate jittering to inject spatial variation, this fundamental pattern was however never fully resolved and likely impeded model generalization. The sparse representation of southern hemisphere observations compared to northern locations further limited the model's ability to capture global environmental variation.

A further potential limitation was the reliance on the MICRA Crop Calendar to approximate wheat growing seasons due to missing precise planting and harvest dates in many locations. This affected the reliability of monthly time series feature attachment and may have reduced the potential for richer feature engineering, such as the incorporation of Growing Degree Days (GDDs). Additionally, the dataset did not distinguish between winter and spring wheat variants, which have different life cycles and growing seasons [Hyles et al., 2020]. This prevented targeted feature engineering based on cultivar-specific interactions, as cultivar identification was hindered by inconsistent naming conventions in the data. These limitations highlighted the challenges of balancing data richness against spatial representativeness.

Provided these issues are addressed, the methodology explored could open up many possibilities for future research. With the changing climate being a key question that could be addressed when training a global model to predict how yield may respond to future scenarios, further exploring range and frequency shifts in plant pests and pathogens in response to this [Singh et al., 2023]. The creation of a global model could also allow for more detailed analysis regarding feature interactions by region, using interpretable models like GAMS that are able to model non-linear relationships effectively. High resolution imaged based satellite data, which has seen success in regional modelling but was out of the scope of this question such as sentinel 2 could also provide an additional direction to explore in regards to finding a globally portable solution yield prediction [Zhao et al., 2020].

8 Conclusions

Global, data-driven ensembles (Random Forest, XGBoost) were not able to generalize to the spatiotemporal holdout despite rich co-variates and spatial–temporal controls, highlighting limits of current data and methods for worldwide generalized wheat-yield prediction. SHAP analyses consistently elevated management intensity (nitrogen rate, pesticide use) and socioeconomic context (GDP) as the strongest signals, while climate, soil, and vegetation features were weaker and less transferable. Methodological improvement will require better-balanced sampling, precise phenology/cultivar labels, harmonized management data, and models designed for domain shift or hybridization with process knowledge.

References

[Acevedo et al., 2006] Acevedo, E., Silva, P., and Silva, H. (2006). Growth and wheat physiology, development. *Laboratory of Soil-Plant-Water Relations. Faculty of Agronomy and Forestry Sciences. University of Chile. Casilla*, 1004.

[AppEEARS Team, 2025] AppEEARS Team (2025). Application for extracting and exploring analysis ready samples (AppEEARS), version 3.93.

[Blanc and Sultan, 2015] Blanc, E. and Sultan, B. (2015). Emulating maize yields from global gridded crop models using statistical estimates. *Agricultural and Forest Meteorology*, 214:134–147. Publisher: Elsevier.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32. Publisher: Springer.

[Cao et al., 2025] Cao, J., Zhang, Z., Luo, X., Luo, Y., Xu, J., Xie, J., Han, J., and Tao, F. (2025). Mapping global yields of four major crops at 5-minute resolution from 1982 to 2015 using multi-source data and machine learning. *Scientific Data*, 12(1):357. Publisher: Nature Publishing Group UK London.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

[Copernicus Global Land Service, 2021] Copernicus Global Land Service (2021). Copernicus global land service: NDVI 1km V3.0, 10-daily. tex.version: V3.0.

[Cravero et al., 2022] Cravero, A., Pardo, S., Sepúlveda, S., and Muñoz, L. (2022). Challenges to use machine learning in agricultural big data: a systematic literature review. *Agronomy*, 12(3):748. Publisher: MDPI.

[Dadrasi et al., 2023] Dadrasi, A., Chaichi, M., Nehbandani, A., Soltani, E., Nemati, A., Salmani, F., Heydari, M., and Yousefi, A. R. (2023). Global insight into understanding wheat yield and production through Agro-Ecological Zoning. *Scientific Reports*, 13(1):15898. Publisher: Nature Publishing Group UK London.

[Djanaguiraman et al., 2020] Djanaguiraman, M., Narayanan, S., Erdayani, E., and Prasad, P. V. (2020). Effects of high temperature stress during anthesis and grain filling periods on photosynthesis, lipids and grain yield in wheat. *BMC Plant Biology*, 20:1–12. Publisher: Springer.

[Enghiad et al., 2017] Enghiad, A., Ufer, D., Countryman, A. M., and Thilmany, D. D. (2017). An overview of global wheat market fundamentals in an era of climate concerns. *International Journal of Agronomy*, 2017(1):3931897. Publisher: Wiley Online Library.

[Erenstein et al., 2022] Erenstein, O., Jaleta, M., Mottaleb, K. A., Sonder, K., Donovan, J., and Braun, H.-J. (2022). Global trends in wheat production, consumption and trade. In *Wheat improvement: food security in a changing climate*, pages 47–66. Springer International Publishing Cham.

[FAO, 2024] FAO (2024). Pesticides use and trade – 1990–2022. Technical report, FAOSTAT Analytical Briefs, No. 89, Rome.

[Fischer et al., 2021] Fischer, G., Nachtergaele, F. O., van Velthuizen, H. T., Chiozza, F., Franceschini, G., Henry, M., Muchoney, D., and Tramberend, S. (2021). Global agro-ecological zones v4 – model documentation. Technical Report, FAO, Rome.

[Food and Agriculture Organization of the United Nations (FAO) and International Institute for Applied System Food and Agriculture Organization of the United Nations (FAO) and International Institute for Applied Systems Analysis (IIASA) (2024). THE GLOBAL AGRO-ECOLOGICAL ZONING version 4 Crop profile: Wheat.

[Ganati and Sitote, 2025] Ganati, B. A. and Sitote, T. M. (2025). Predicting land suitability for wheat and barley crops using machine learning techniques. *Scientific Reports*, 15(1):15879. Publisher: Nature Publishing Group UK London.

[Gasch et al., 2015] Gasch, C. K., Hengl, T., Gräler, B., Meyer, H., Magney, T. S., and Brown, D. J. (2015). Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D+ T: The Cook Agronomy Farm data set. *Spatial Statistics*, 14:70–90. Publisher: Elsevier.

[Ghimire et al., 2017] Ghimire, R., Machado, S., and Bista, P. (2017). Soil pH, soil organic matter, and crop yields in winter wheat–summer fallow systems. *Agronomy Journal*, 109(2):706–717. Publisher: Wiley Online Library.

[Gupta et al., 2022] Gupta, V., Kumar, M., Singh, V., Chaudhary, L., Yashveer, S., Sheoran, R., Dalal, M. S., Nain, A., Lamba, K., Gangadharaiah, N., and others (2022). Genotype by environment interaction analysis for grain yield of wheat (Triticum aestivum (L.) em. Thell) genotypes. *Agriculture (Nitra, Slovakia)*, 12(7):1002. Publisher: MDPI.

[Hulsman et al., 2023] Hulsman, P., Keune, J., Koppa, A., Schellekens, J., and Miralles, D. (2023). Incorporating plant access to groundwater in existing global, satellite-based evaporation estimates. *Water Resources Research*, 59.

[Hyles et al., 2020] Hyles, J., Bloomfield, M. T., Hunt, J. R., Trethowan, R. M., and Trevaskis, B. (2020). Phenology and related traits for wheat adaptation. *Heredity*, 125(6):417–430. Publisher: Springer International Publishing Cham.

[Igrejas and Branlard, 2020] Igrejas, G. and Branlard, G. (2020). The importance of wheat. In *Wheat quality for improving processing and human health*, pages 1–7. Springer.

[Jabed and Murad, 2024] Jabed, M. A. and Murad, M. A. A. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon*. Publisher: Elsevier.

[Johnson et al., 2021] Johnson, D. M., Rosales, A., Mueller, R., Reynolds, C., Frantz, R., Anyamba, A., Pak, E., and Tucker, C. (2021). USA crop yield estimation with MODIS NDVI: Are remotely sensed models better than simple trend analyses? *Remote Sensing*, 13(21):4227. Publisher: MDPI.

[Kummu et al., 2025] Kummu, M., Kosonen, M., and Masoumzadeh Sayyar, S. (2025). Downscaled gridded global dataset for gross domestic product (GDP) per capita PPP over 1990–2022. *Scientific Data*, 12(1):178. Publisher: Nature Publishing Group UK London.

[Lesk et al., 2022] Lesk, C., Anderson, W., Rigden, A., Coast, O., Jägermeyr, J., McDermid, S., Davis, K. F., and Konar, M. (2022). Compound heat and moisture extreme impacts on global crop yields under climate change. *Nature Reviews Earth & Environment*, 3(12):872–889. Publisher: Nature Publishing Group UK London.

[Lobell et al., 2009] Lobell, D. B., Cassman, K. G., and Field, C. B. (2009). Crop yield gaps: their importance, magnitudes, and causes. *Annual review of environment and resources*, 34(1):179–204. Publisher: Annual Reviews.

[Mao et al., 2023] Mao, H., Jiang, C., Tang, C., Nie, X., Du, L., Liu, Y., Cheng, P., Wu, Y., Liu, H., Kang, Z., and others (2023). Wheat adaptation to environmental stresses under climate change: Molecular basis and genetic improvement. *Molecular Plant*, 16(10):1564–1589. Publisher: Elsevier.

[Miralles et al., 2025] Miralles, D. G., Bonte, O., Koppa, A., Baez-Villanueva, O. M., Tronquo, E., Zhong, F., Beck, H. E., Hulsman, P., Haghdoost, S., and Dorigo, W. A. (2025). GLEAM4: global evaporation and soil moisture datasets at 0.1° resolution from 1980 to near present. *Scientific Data*, 12(1). Publisher: Nature Publishing Group.

[Muruganantham et al., 2022] Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N. H., and Islam, N. (2022). A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sensing*, 14(9):1990. Publisher: MDPI.

[Myneni et al., 2021] Myneni, R., Knyazikhin, Y., and Park, T. (2021). MODIS/terra+aqua leaf area index/FPAR 8-day L4 global 500m SIN grid V061.

[Najafi et al., 2018] Najafi, E., Devineni, N., Khanbilvardi, R. M., and Kogan, F. (2018). Understanding the changes in global crop yields through changes in climate and technology. *Earth's Future*, 6(3):410–427. Publisher: Wiley Online Library.

[Oerke, 2006] Oerke, E.-C. (2006). Crop losses to pests. *The Journal of agricultural science*, 144(1):31–43. Publisher: Cambridge University Press.

[Poggio et al., 2021] Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D. (2021). SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL*, 7:217–240. Publisher: Copernicus GmbH.

[Qin et al., 2024] Qin, X., Wu, B., Zeng, H., Zhang, M., and Tian, F. (2024). Global gridded crop production dataset at 10 km resolution from 2010 to 2020. *Scientific Data*, 11(1):1377. Publisher: Nature Publishing Group UK London.

[Raza and Bebber, 2022] Raza, M. M. and Bebber, D. P. (2022). Climate change, biotic yield gaps and disease pressure in cereal crops. *bioRxiv : the preprint server for biology*, pages 2022–08. Publisher: Cold Spring Harbor Laboratory.

[Roberts et al., 2017] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., and others (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929. Publisher: Wiley Online Library.

[Running et al., 2021] Running, S., Mu, Q., and Zhao, M. (2021). MODIS/terra gross primary productivity 8-day L4 global 500m SIN grid V061.

[Sapkota et al., 2020] Sapkota, T. B., Singh, L. K., Yadav, A. K., Khatri-Chhetri, A., Jat, H. S., Sharma, P. C., Jat, M. L., and Stirling, C. M. (2020). Identifying optimum rates of fertilizer nitrogen application to maximize economic return and minimize nitrous oxide emission from rice–wheat systems in the Indo-Gangetic Plains of India. *Archives of Agronomy and Soil Science*, 66(14):2039–2054. Publisher: Taylor & Francis.

[Senapati et al., 2022] Senapati, N., Semenov, M. A., Halford, N. G., Hawkesford, M. J., Asseng, S., Cooper, M., Ewert, F., van Ittersum, M. K., Martre, P., Olesen, J. E., and others (2022). Global wheat production could benefit from closing the genetic yield gap. *Nature Food*, 3(7):532–541. Publisher: Nature Publishing Group UK London.

[Singh et al., 2023] Singh, B. K., Delgado-Baquerizo, M., Egidi, E., Guirado, E., Leach, J. E., Liu, H., and Trivedi, P. (2023). Climate change impacts on plant pathogens, food security and paths forward. *Nature Reviews Microbiology*, 21(10):640–656. Publisher: Nature Publishing Group UK London.

[SKINNER et al., 1994] SKINNER, G. E., LARKIN, J. W., and RHODEHAMEL, E. J. (1994). Mathematical modeling of microbial growth: a review. *Journal of food safety*, 14(3):175–217. Publisher: Wiley Online Library.

[Sun et al., 2020] Sun, H., Zhang, W., Wu, Y., Gao, L., Cui, F., Zhao, C., Guo, Z., and Jia, J. (2020). The circadian clock gene, TaPRR1, is associated with yield-related traits in wheat (Triticum aestivum L.). *Frontiers in Plant Science*, 11:285. Publisher: Frontiers Media SA.

[Tang et al., 2023] Tang, F. H. M., Nguyen, T. H., Conchedda, G., Casse, L., Tubiello, F. N., and Maggi, F. (2023). CROPGRIDS.

[Van Buuren and Groothuis-Oudshoorn, 2011] Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45:1–67.

[Van Klompenburg et al., 2020] Van Klompenburg, T., Kassahun, A., and Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and electronics in agriculture*, 177:105709. Publisher: Elsevier.

[Wang et al., 2018] Wang, H., Li, Q., Du, X., Zhao, L., and Wang, N. (2018). Evaluation of potential crop productivity based on remote sensing and agro-ecological zones around the world. *Geocarto International*, 33(7):713–722. Publisher: Taylor & Francis.

[Wang et al., 2021] Wang, X., Müller, C., Elliot, J., Mueller, N. D., Ciais, P., Jägermeyr, J., Gerber, J., Dumas, P., Wang, C., Yang, H., and others (2021). Global irrigation contribution to wheat and maize yield. *Nature Communications*, 12(1):1235. Publisher: Nature Publishing Group UK London.

[Wang et al., 2023] Wang, Y., Khodadadzadeh, M., and Zurita-Milla, R. (2023). Spatial+: A new cross-validation method to evaluate geospatial machine learning models. *International Journal of Applied Earth Observation and Geoinformation*, 121:103364. Publisher: Elsevier.

[Zampieri et al., 2018] Zampieri, M., Ceglar, A., Dentener, F., and Toreti, A. (2018). Understanding and reproducing regional diversity of climate impacts on wheat yields: current approaches, challenges and data driven limitations. *Environmental Research Letters*, 13(2). Publisher: IOP Publishing.

[Zhao et al., 2020] Zhao, Y., Potgieter, A. B., Zhang, M., Wu, B., and Hammer, G. L. (2020). Predicting wheat yield at the field scale by combining high-resolution Sentinel-2 satellite imagery and crop modelling. *Remote Sensing*, 12(6):1024. Publisher: MDPI.

[Zhou et al., 2019] Zhou, W., Han, G., Liu, M., and Li, X. (2019). Effects of soil pH and texture on soil carbon and nitrogen in soil profiles under different land uses in Mun River Basin, Northeast Thailand. *PeerJ*, 7:e7880. Publisher: PeerJ Inc.

A APPENDICES

Code used for the project can be found here:

https://github.com/airbreather2/CMEECourseWork/tree/main/FinalProject

Data used for Final project can be found here: (will be hosted for 1 month then taken down, email sed24@ic.ac.uk if it is required)

https://www.dropbox.com/scl/fo/86k7zs8jwcvkqfckb815a/ACuExK6ti1Zei3ldYM0

A.1 Packages used for both R and python

B Software Dependencies

B.1 Python Packages

B.1.1 Core Libraries

- os
- time
- numpy
- pandas
- matplotlib.pyplot
- seaborn
- tabulate
- sys
- joblib

B.1.2 Geospatial Analysis

- geopandas
- shapely.geometry
- haversine
- folium

B.1.3 Machine Learning

- sklearn
- xgboost
- category_encoders
- shap

B.1.4 Statistical Analysis

- scipy.stats
- scipy.cluster.hierarchy
- scipy.spatial.distance
- gstools
- skgstat
- statsmodels.graphics.tsaplots

B.2 R Packages

- missForest
- Tidyverse
- readr
- mice
- corrplot