**Interpretable rice yield prediction in India using autoML and dynamic visualization for enhanced early warning systems**

Djavan De Clercq1, Adam Mahdi1

1. Oxford Internet Institute, University of Oxford

**Abstract**

Yield forecasting, the science of predicting agricultural productivity in advance, helps a wide range of stakeholders make better decisions around agricultural planning in response to early warning of crop yield outcomes. This study aims to build yield prediction models capable of accurately predicting Kharif season rice yields at the district level in India several months before the rice harvest takes place. The methodology involved training learning models such as CatBoost, LightGBM, Orthogonal Matching Pursuit, and Extremely Randomized Trees on 20 years of climate, satellite, and rice yield data across 247 of India’s rice-producing districts. In addition to model-building, a dynamic dashboard was built to both visualize model outputs and diagnose model performance. The results demonstrated out-of-sample R2, MAE, and MAPE performance of up to 0.82, 0.29, and 0.16 respectively, outperforming test set performance reported in related literature on rice yield modeling. In addition SHAP value analysis was conducted to infer both the importance and directional impact of the climate and remote sensing variables included in the model. Important features driving rice yields included XX, XX, and XX. SHAP dependence plots showed that Kharif season rice yields typically increase with. The visual dashboard allows users to clearly see which districts may experience a rise or fall in yield relative to the previous year. In addition, the dashboard tool showed that the model performed well in certain regions (for instance XX% in region X and YY% in region Y). In order to build resilience to meteorological shocks in agricultural populations, we recommend that subsequent research on crop yield prediction focuses on the following areas: (a)

**1. Introduction**

**1.1. The societal implications of accurate crop yield forecasting in India**

Yield forecasting is the science of predicting agricultural productivity as measured by crop yield – the ratio of the total mass of the harvested product (such as rice) to the area used to cultivate the crop1 – before the harvest takes place, typically a few months in advance.

Pre-harvest prediction of crop yields is important in helping a wide range of stakeholders make better decisions around agricultural planning. For farmers, accurate crop yield forecasts can facilitate decision-making around what to grow and when to grow it2. In addition, near real-time monitoring of crop growth can inform the use of preventive measures such as irrigation and fertilization to boost agricultural productivity where needed3. For governments, yield prediction is relevant to the formulation of policies related to national food security, such as pricing policies for domestic markets, and policy decisions on the import and export of different crops4.

Accurate crop yield forecasting may also enable better design of insurance products that mitigate climate risks and stabilize farmer incomes5. Weather-based crop insurance, for instance, uses a weather index such as total precipitation to determine payments to farmers, meaning that insurance companies do not need to visit farmers to assess damages and arbitrate claims. Rather, if the weather reaches a certain threshold, rapid automatic payments are distributed to farmers, who avoid the need to sell assets to survive due to adverse climate events6.

The need for accurate information on crop yields is particularly important in countries like India, where the agricultural sector provides livelihoods for hundreds of millions of farmers, with 70% of rural households depending on agriculture for their main source of income7.One of India’s major staple crops is rice, which contributes to 30% of calories consumed in India and is a key export commodity for the country8. India cultivates rice on about 45 million hectares of land, with a total production of 178 million metric tonnes in 20209. In addition, the distribution of monsoon rainfall, which is a major source of water for rice cultivation, has become erratic in recent years due to climate variability10,11. In such contexts, crop yield predictions may be able to supplement agricultural early warning systems that give advanced notice of potential risks to crop productivity.

**1.2. Overview of approaches and variables that have been used to model crop yields**

Crop yield prediction is a challenging problem in precision agriculture, as final yields depend on a variety of factors such as weather, climate, soil, seed type, and agronomic practices such as irrigation and fertilizer use12.

This complexity is evident from the variety of variables included and methods applied in the growing body of literature on crop yield forecasting. For example, recent examples in literature involving deep learning approaches include corn and soybean yield forecasting in the US based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs)13, soybean yield forecasting in Argentina based on deep transfer learning14, and vineyard grape yield estimations based on CNNs. Recent examples based on machine learning approaches include sugarcane yield prediction using random forests15, prediction of wheat, barley, and canola yields in Western Australia using random forest16, yield forecasting of spring maize in Pakistan based on LASSO regression and support vector machine17, and Jojoba yield prediction in Israel based on gradient boosted regression trees18. Other examples of the machine learning approaches that have been applied to yield prediction have been summarized in a systematic literature review, which also analyzed the variables most frequently included in crop yield prediction studies. Across 50 studies between 2008 and 2019, features used as predictors of yield have included temperature, soil type, rainfall, humidity, pH-value, NDVI, wind speed, and more2.

For studies specific to rice, the staple crop of over half the world’s population19, a number of approaches have been applied to yield forecasting in recent years. Recent examples include rice yield prediction for 81 counties in southern China based on recurrent neural networks20; application of the ecological distance algorithm to model rice yields21; field and county-level rice yield prediction based on synthetic aperture radar (SAR), optical and meteorological data22; random forest yield prediction based on high-resolution imagery collected from unmanned ariel vehicles (UAVs)23; simulation of yields using the Cropping System Model-CERES-RICE24; pixel-scale rice yield prediction in South Korea based on a combination of deep learning and crop models25; rice yield estimation at 500m spatial resolution based on gradient boosted regression and vegetation indices derived from the Moderate Resolution Imaging Spectroradiometer (MODIS)26; and rice paddy yield prediction using sentinel-based optical and SAR data in India based on random forest27.

**1.3. Research contributions**

The present work aims to build on previous efforts in the crop yield prediction literature by building models capable of predicting rice yields at the district level for 362 districts across India. To this end, the present study brings several innovations over the existing literature. Firstly, an automated machine learning (AutoML) approach is applied in order to test a wide range of models, whereas previous literature on crop yield prediction in India has largely focused on using a narrower range of algorithms such as random forest or support vector machine2. This approach allows for rapid experimentation to identify models and feature engineering combinations most capable of forecasting rice yield prior to harvesting.Secondly, a novel combination of data sources is used to predict Indian rice yields. These include data from ERA5, a climate re-analysis product developed by the European Centre for Medium Range Weather Forecasts (ECMWF), which combines observations with modelled data to provide hourly data on atmospheric, land-surface, and sea-state parameters globally28. Vegetation data was also derived from the MODIS sensors on-board NASA’s TERRA and AQUA satellites. Third, a dashboard tool was developed to both dynamically visualize the outputs of the rice yield prediction learning model and diagnose model performance.

**2. Methodology**

**2.1. Study area**

In this study, climate and remote sensing data will be used as predictors to model rice yields for the kharif season (wet summer monsoon season) from 2001 to 2020 at the district level in India (India consists of 36 states and 684 districts). In India, more than half of the annual rice crop is grown during kharif29, a season which is characterized by high temperature, high humidity, and medium to high rainfall 30. Kharif season rice is typically sowed between the start of June to the end of August and harvested between the end of September to early January, depending on the region 31. During the 2019-2020 season, harvesting of Kharif rice was completed in February 2020.

Indian rice typically develops across a number of phenological stages. Past case studies have shown that approximate number of days taken for each stage: the sowing to tillering phase (P1) can range from 30 to 60 days, and the rate of tillering tends to increase under higher temperatures; the tillering to panicle initiation phase (P2) can range between 42 to 49 days; the panicle initiation to flowering phase (P3) can range from 12 to 28 days; the flowering to mil phase (P4) can range from 7 to 20 days; and the mil to physiological maturity phase (P5) can range from 17 to 31 days32. One study showed reported a linear relationship between the days taken from sowing to flowering and average air temperature.33

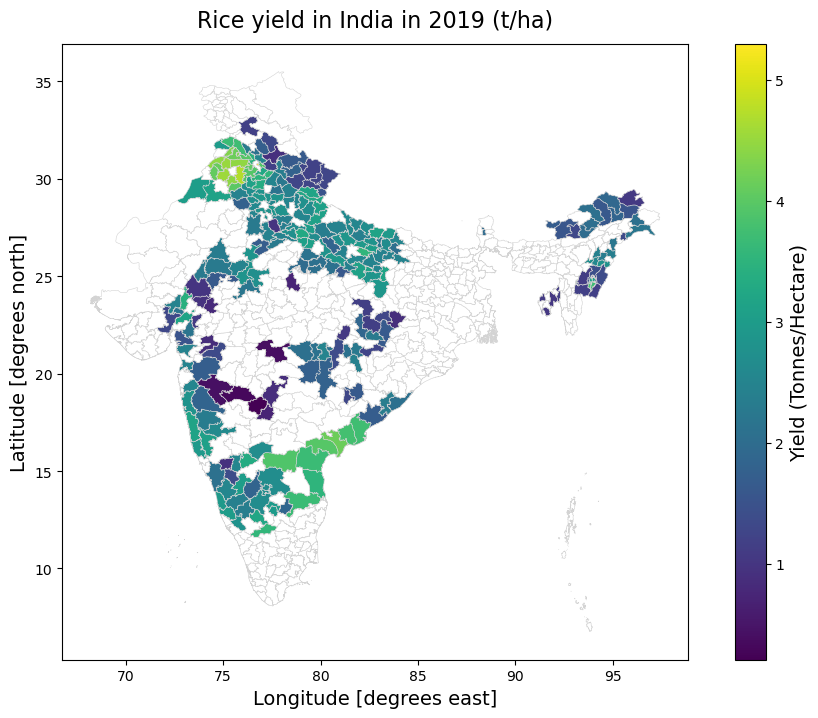
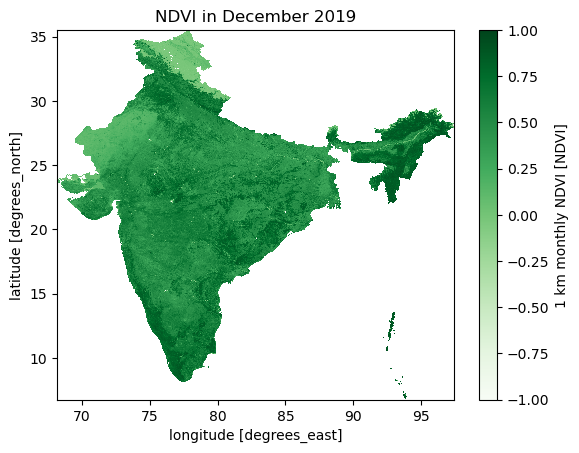
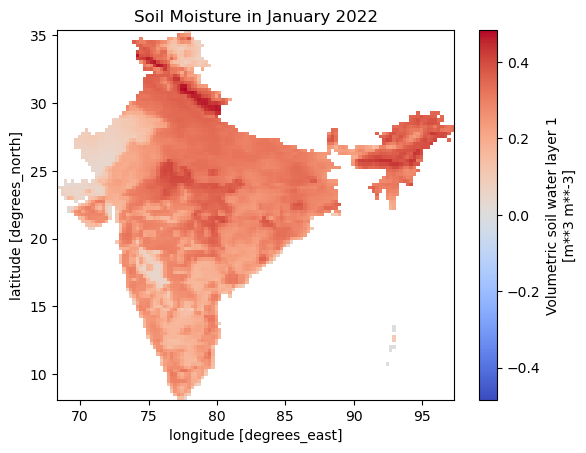
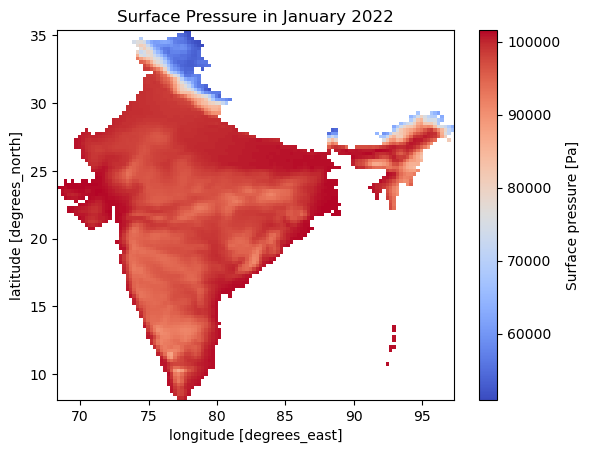
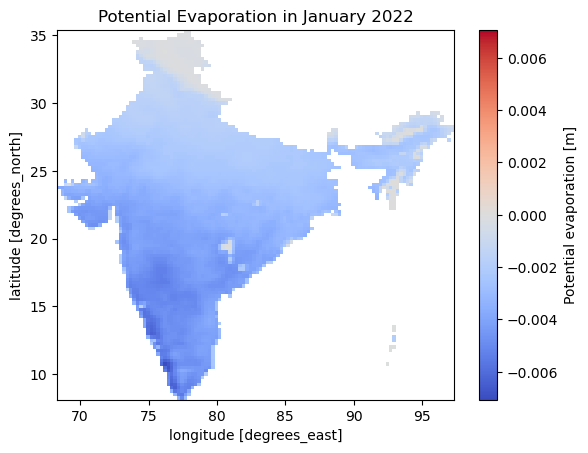
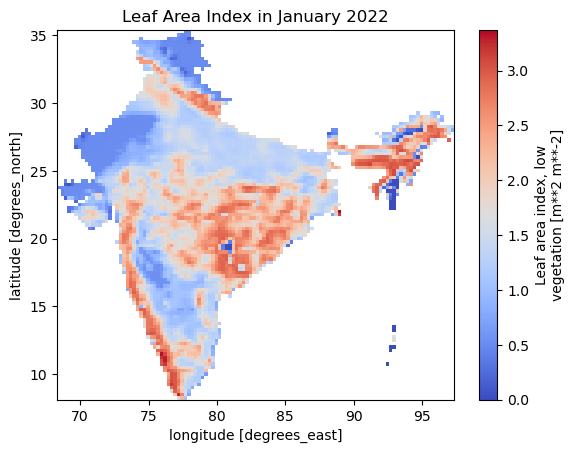
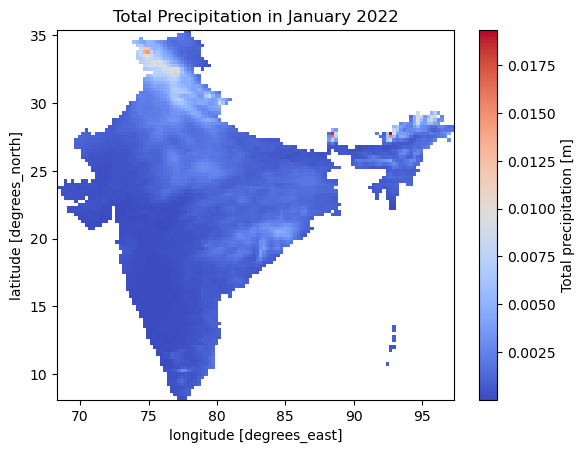
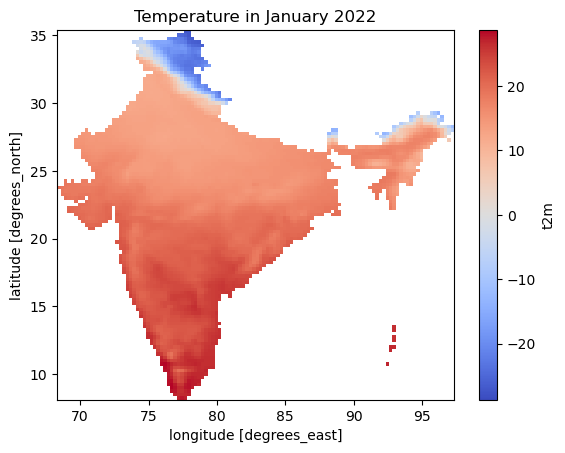
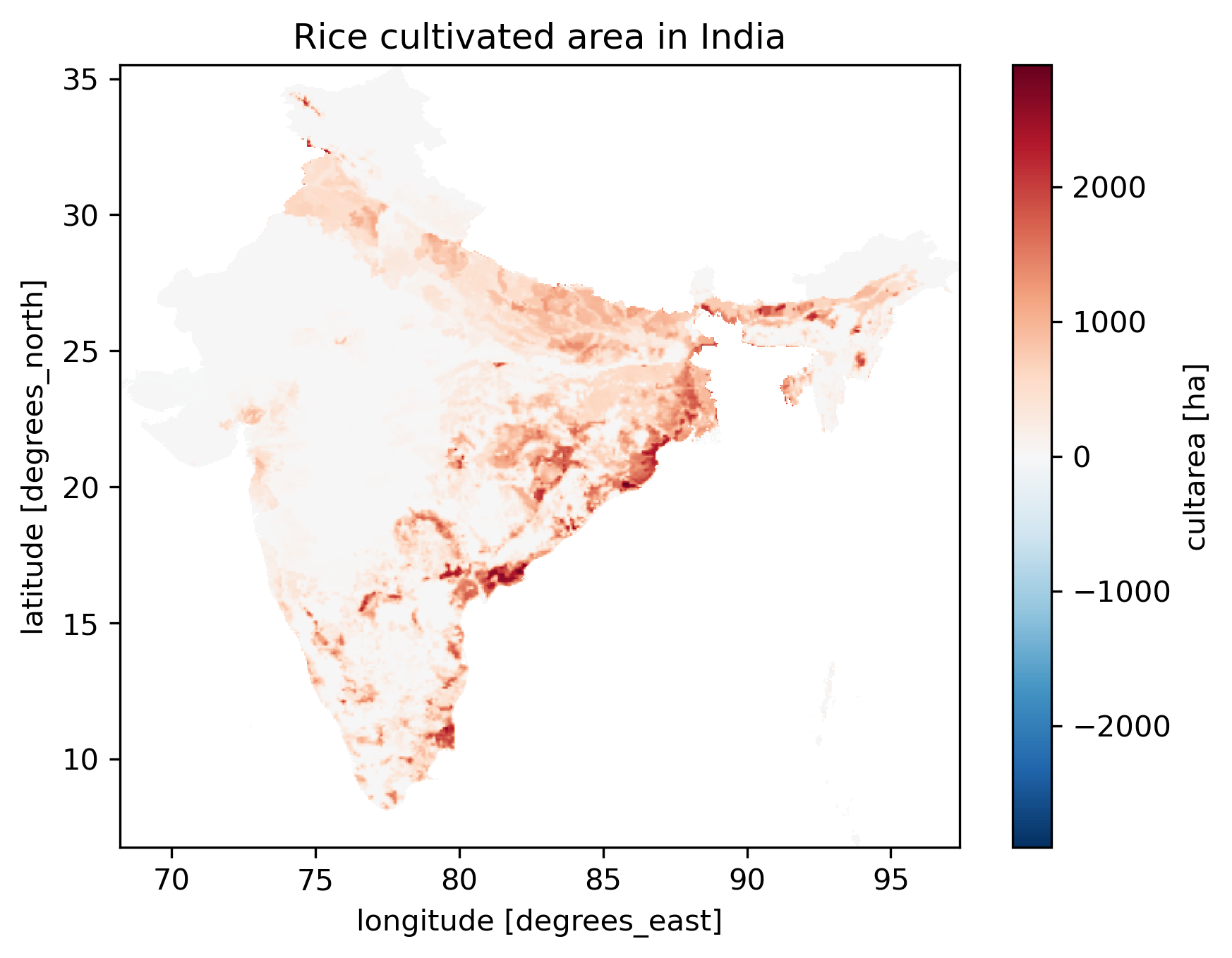
**and overview of methodology**

The methods that will be applied in this study can be summarized as follows. Firstly, 20 years of data on climate, vegetation, and rice yields will be collected from various sources. Next, climate and vegetation data will be pre-processed and aggregated to the same level as historical rice yields prior to development of machine learning models. Lastly, a range of models including Bayesian ridge regression and LightGBM will be trained, evaluated, and interpreted, after which the implications of the modelling results will be discussed with regards to their inclusion within early warning systems.

**2.2. Data Collection**

**Climate reanalysis data.** Daily climate reanalysis data on temperature, potential evaporation, surface pressure, leaf area index, total precipitation, and soil water content was obtained from ERA5 data from the European Center for Medium-Range Weather Forecasts (ECMWF), which provides global estimates of surface and atmospheric parameters since 1950 at a resolution of approximately 30\*30 km34. Climate reanalysis data, which are often freely available, provide temporally and spatially homogenous data35, which makes them suitable for applications such as crop yield prediction in contexts where in-situ weather station measurements are inadequate or incomplete. In addition, weather stations vary in their accuracy and generally record a limited number of variables, such as rainfall, temperature, pressure, and wind speed; variables that are more technically demanding to measure, such as humidity and solar radiation, may be lacking36.

The climate variables used in this study were selected due to their influence on rice yields. An extensive body of research has shown that rice growth is affected by factors such as soil water content37, temperature38, potential evaporation (as a proxy for transpirational demand)39, surface pressure40, and precipitation41.



**a**

**b**

**c**

**c**

**d**

**e**

**f**

**g**

**h**

**Figure 1.** Overview of geospatial data used in this research. Panels **a-h** show cultivated rice area in India; temperature, total precipitation, potential evaporation, surface pressure, soil moisture, and leaf area index from the ECMFW, NDVI data from NASA’s MODIS; and district-level rice yield data from India’s Ministry of Agriculture and Farmers Welfare.

**Remote sensing data**. Normalized Difference Vegetation Index (NDVI) is a dimensionless index that describes the difference between visible and near-infrared reflectance of vegetation cover, and can be used to estimate the density of green on an area of land42. To determine the density of green on a patch of land, the wavelengths of visible and near-infrared sunlight reflected by the plants are observed. NDVI values range from −1 to +1; higher values of NDVI imply healthy and dense vegetation, whereas lower NDVI values indicate sparser vegetation. NDVI data was obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) onboard NASA’s Terra and Aqua satellites, due to their wide coverage and temporal resolution. There are several examples in academic literature of using NDVI to investigate the progress of crops, such as for wheat in Argentina43, cereals in Europe44, and rice in Vietnam45. NDVI data was masked using CROPGRIDS, a global, geo-referenced dataset providing information on areas for 173 crops circa the year 2020, at a resolution of 0.05° (~5.55 km at the equator).46

**Yield data.** District-level rice production and yield data from 1995 to 2021 for 367 districts were obtained from the APY dataset of the Directorate of Economics and Statistics in India’s Ministry of Agriculture and Farmers Welfare47. In this dataset, the year denotes the year in which the crop was harvested. For kharif season rice, the sowing is in the previous calendar year48.

**Table 1: Data overview.** A range of agronomically-relevant datasets were used as predictors of the target variable (district-level rice yield in India); rice areas masks were used to filter NDVI data by rice-growing area46,47,49,50

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data type | Parameter | Description | Unit | Source |
| Climate reanalysis | Potential evaporation (pev) | A measure of the extent to which near-surface atmospheric conditions are conducive to the process of evaporation. | m | ECMFW (ERA5) |
| Climate reanalysis | 2m-temperature | The temperature of air at 2m above the surface of land, sea or inland waters. 2m temperature is calculated by interpolating between the lowest model level and the Earth's surface, taking account of the atmospheric conditions. | K | ECMFW (ERA5) |
| Climate reanalysis | Total precipitation | The accumulated liquid and frozen water, comprising rain and snow, that falls to the Earth's surface. It is the sum of large-scale precipitation and convective precipitation. | m | ECMFW (ERA5) |
| Climate reanalysis | Leaf area index, low vegetation | The surface area of one side of all the leaves found over an area of land for vegetation classified as 'low'. 'Low vegetation' consists of crops and mixed farming, irrigated crops, short grass, and more.51 | m2m-2 | ECMFW (ERA5) |
| Climate reanalysis | Total precipitation | The accumulated liquid and frozen water, comprising rain and snow, that falls to the Earth's surface. It is the sum of large-scale precipitation and convective precipitation. | m | ECMFW (ERA5) |
| Climate reanalysis | Volumetric soil water | The volume of water in soil layer 1 (0 - 7cm, the surface is at 0cm) | m3m-3 | ECMFW (ERA5) |
| Remote Sensing | Normalized Difference Vegetation Index (NDVI) | A dimensionless index that describes the difference between visible and near-infrared reflectance of vegetation cover, and can be used to estimate the density of green on an area of land | (-) | NASA EOSDIS (AQUA MODIS) |
| Crop mask | Rice crop mask | A comprehensive global, geo-referenced dataset providing information on areas for 173 crops circa the year 2020, at a resolution of 0.05° (~5.55 km at the equator). | ha | CROPGRIDS |
| Rice production | District-level yield | District-level rice production and yield data from 1995 to 2021 for 367 districts of India. | t/ha | APY |

Data used in this analysis were programmatically downloaded via API and automated Python scripts. ERA5 data was ingested via the CDS API, while NDVI data was ingested via USGS’ AρρEEARS API.

**2.3. Data pre-processing**

Climate data from ERA5 in NetCDF format over a bounded area comprising India was clipped to the Indian country boundary. NDVI data from NASA’s AQUA MODIS satellite in NetCDF format were clipped to the Indian country boundary and masked with a rice cropland layer.

Next, the climate variables and NDVI were aggregated to the district level based on zonal statistics. The vector geometry data for India’s ADM2 (district-level) boundaries which raster pixels were aggregated to were obtained from the Database of Global Administrative Areas (GADM)52. District-level yield data from APY was then merged to the climate and remote sensing data aggregated at the district level to produce a spatially consistent geodataframe. Yield outliers beyond three standard deviations were removed as they were assumed not achievable at the district level in India.26

Feature engineering was conducted to produce monthly averages for the climate and NDVI parameters for every month between May and November, corresponding to the full sowing and growing period for kharif rice.26 This process was repeated for all variables to produce a set of 52 features used as input for the modelling. The months selected for climate and NDVI feature aggregation were chosen to reflect the full range of rice growth stages, including the grain filing, vegetative, and reproductive stages.53

**2.4. Model development & interpretation**

This study developed and tested the performance of multiple rice yield prediction models based on a variety of machine learning models. These included LightGBM54, an efficient and distributed gradient boosting framework that uses tree-based learning, Bayesian ridge regression55–57, which has been recognized for its ability to deal with hierarchical data structures58, gradient boosting regression59, random forest60, Huber regression61, decision tree regression62, elastic net regression63, AdaBoost64, orthogonal matching pursuit65, and extremely randomized trees66.

The models above were trained on district-level data for 2001 to 2018 (4,606 observations), and validated on out-of-sample test data for 2019 and 2020 (502 observations). The data was split in a manner that reflects how yield prediction models may be used in practice, avoiding random splits in favor of chronological splits to help ensure the model’s robustness to future, unseen data. This out-of-sample approach to testing regression models with temporal dependency has been shown to be more robust than cross-validation approaches tailored to time series problems.67

The top-performing models were evaluated based on three out-of-sample performance measures including R2, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). Also reported were Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Logarithmic Error (RMSLE). In addition, model results were evaluated based on prediction error plots, residual plots, and spatio-temporal plots of prediction error to evaluate potential model bias (for example, better model performance for certain rice-growing regions of India). Lastly, Shapley Additive exPlanations (SHAP) were used to explore the impact of features on model output.68,69 SHAP values are a model-independent methodology utilized for quantifying the significance of features in predictive modeling

The SHAP of feature of for observation is defined as:

where is the feature evaluated, the total number of features, a subset of the full feature set {1,  …, } that does not include the feature , a subset of features in , and the model's prediction function.70

**2.5. Computation**

Data ingestion, pre-processing, and modelling was conducted in a conda-based python environment with a diverse set of python libraries. Data processing and geospatial operations were carried out using python libraries such as numpy, xarray, pandas, rasterio, rasterstat, and geopandas. Modelling and visualization was conducted using python libraries such as scikit-learn, pycaret, matplotlib, and seaborn.

**2.6. Interactive visualization for model evaluation and decision-making**

Visual dashboards can help stakeholders more easily understand the outputs of analytical models, by translating predictive analytics into accessible visual formats. This may enable users from diverse backgrounds – such as farmers, policymakers, and researchers – to interact with and understand the data effectively. Tools such as Streamlit or PowerBI have been used for building visual web apps or dashboards related to bioinformatics, bacterial testing, financial auditing, twitter sentiment analysis, credit card fraud detection, drug target prioritization, and pharmaceutical sales forecasting71–76.

In this study, we leverage PowerBI to visualize model results (e.g., what are the predicted yields in each Indian district), while also providing model diagnostic information (e.g., does the model have varied predictive power in different regions).

**3. Results & discussion**

**3.1. Overview of out-of-sample model performance**

Table 2 summarizes the out-of-sample (validation set) performance across the models tested: R2 and MAPE values of up to 0.82 and 0.16 respectively are achieved. Compared to *out-of-sample results* reported in previous literature on rice yield prediction in different parts of the world, the models perform well. For instance, one study which developed rice yield prediction models for China based on support vector machine regression, neural networks, and random forest, achieved R2 values ranging from 0.24 to 0.31 and MAE values ranging from 0.58 to 0/66 t/ha.77 Another study estimating rice yields in Vietnam’s Mekong Delta reported out-of-sample MAE values ranging from 0.46 to 0.55 t/ha for Winter and Summer rice models.78 A study on county-level rice yield prediction in China’s Jiangsu province reported out-of-sample R2 values of 0.39 to 0.59 on an independent holdout set.22 A study on pixel-scale rice yield prediction in South Korea reported test-set R2 values of 0.80.25 One image-driven yield prediction study reported test-set R2 values of 0.65.79 A study using multi-temporal UAV-based multispectral vegetation indices reported test set R2 values of up to 0.80.80

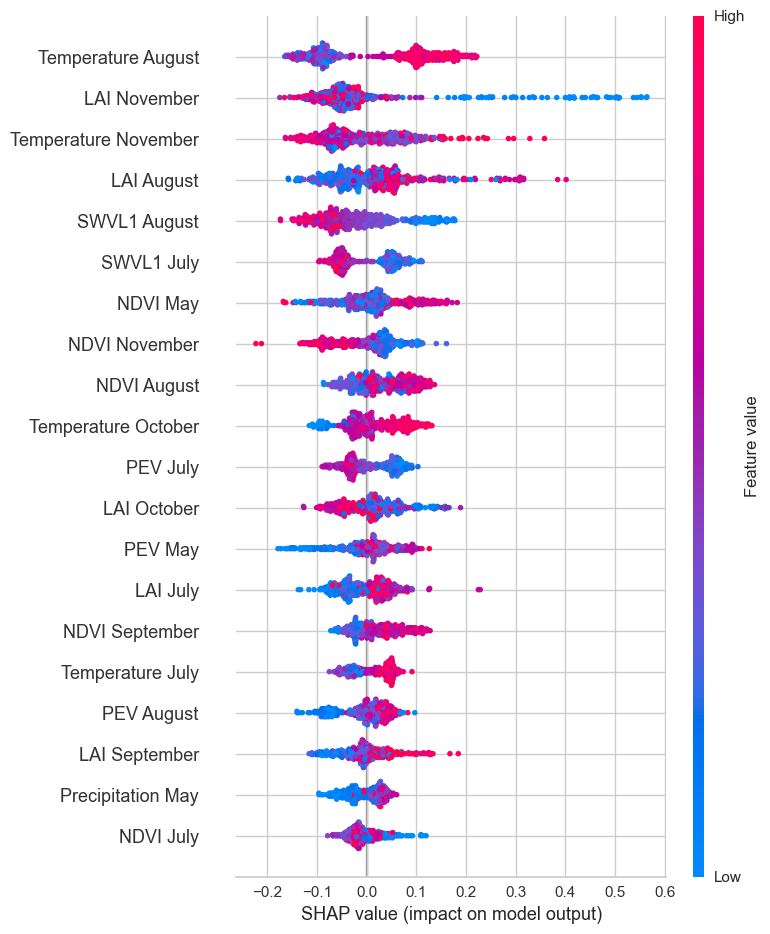
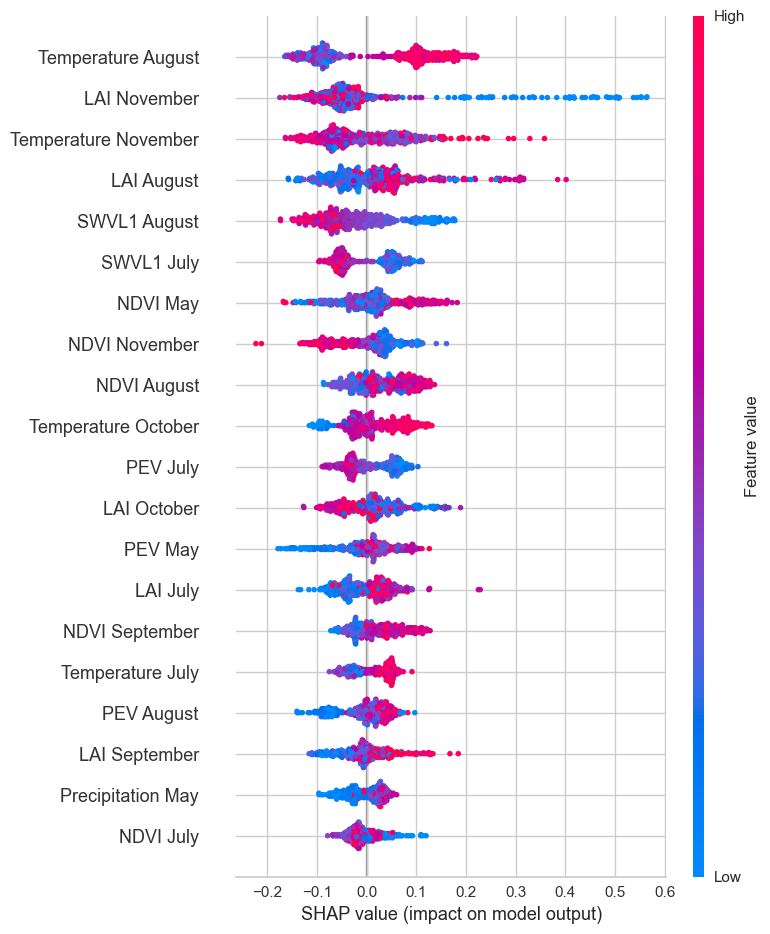
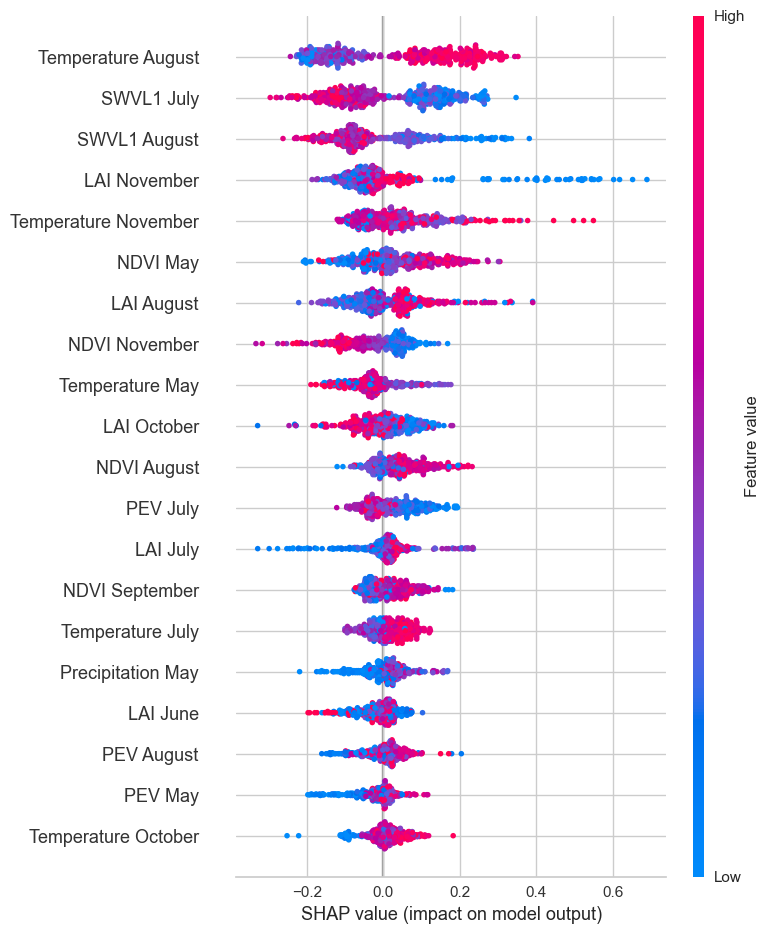
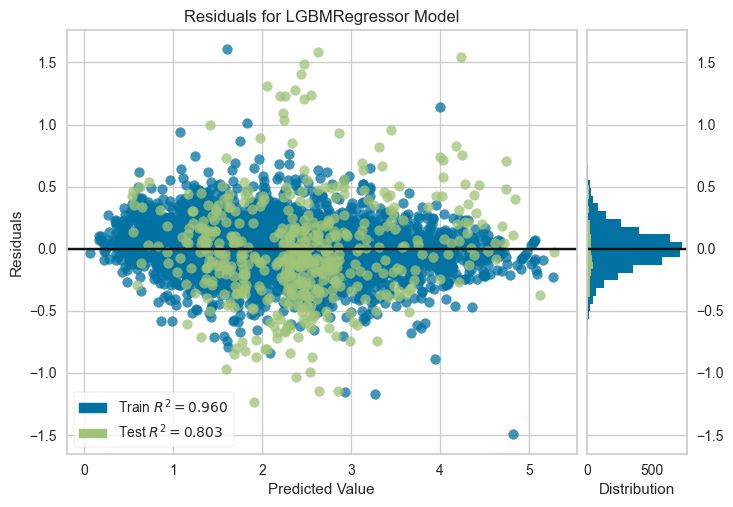
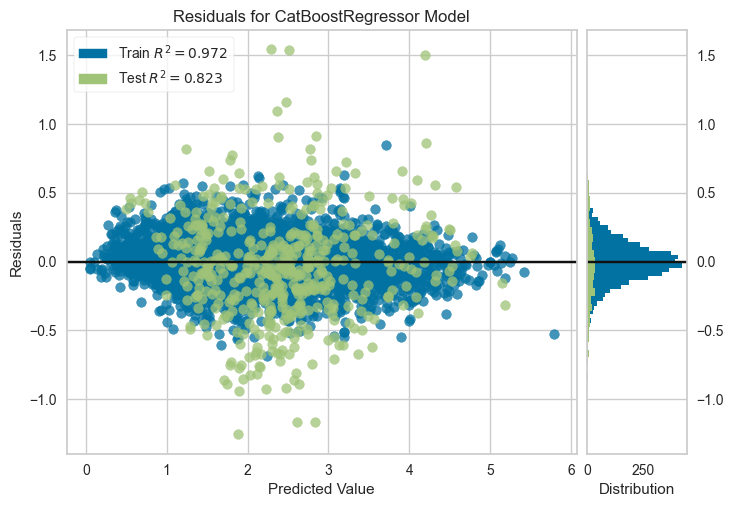
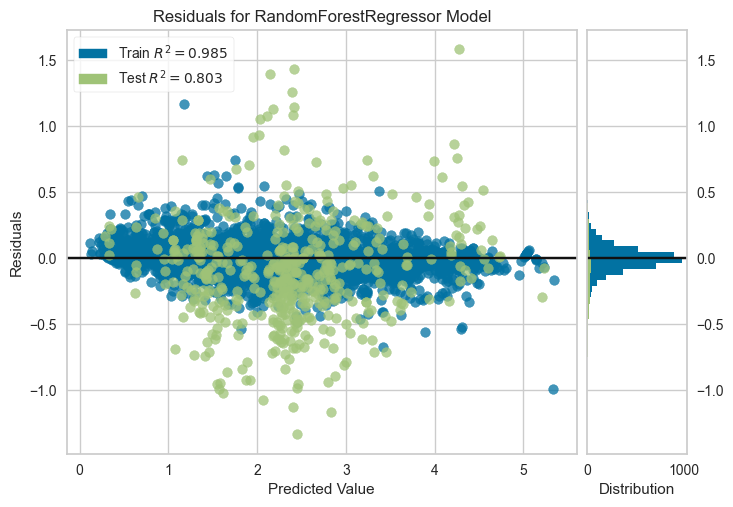
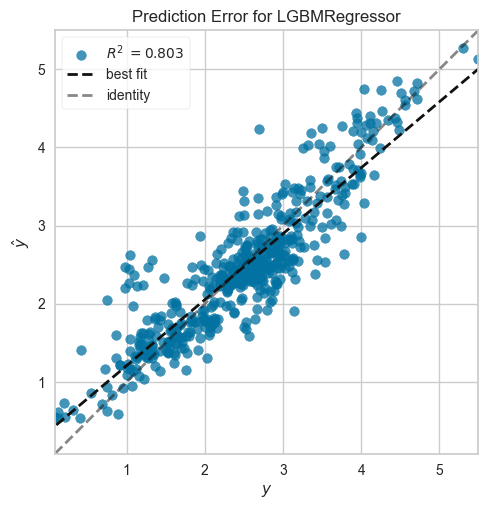
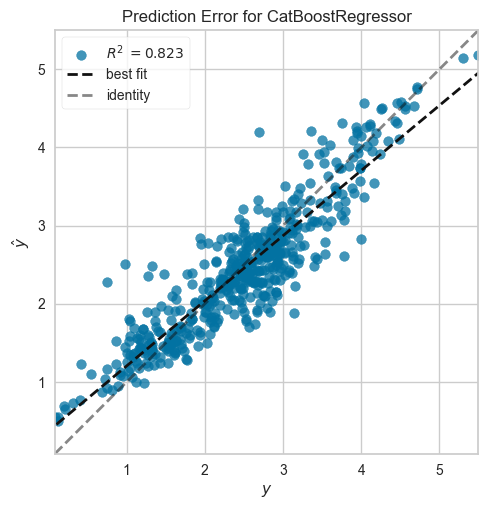
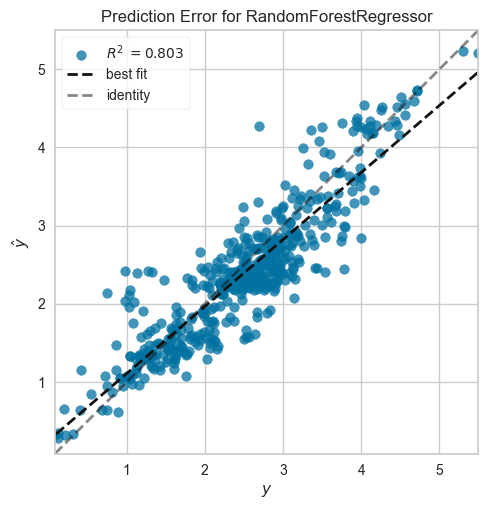
**Table 2.** Model performance on out-of-sample test data shows that the top three (based on R2 and MAPE) models include Random Forest, CatBoost, and Light Gradient Boosting. **Experiment 1** (“all features”) shows results for model runs including the “Year” and “District” features. **Experiment 2** shows results for model runs trained exclusively using climate and satellite data observations (“EO features only”).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Experiment 1 – all features* | | | | *Experiment 2 – EO features only* | | | |
| **Model** | **MAE** | **RMSE** | **R2** | **MAPE** | **MAE** | **RMSE** | **R2** | **MAPE** |
| Random Forest Regressor | 0.31 | 0.41 | 0.80 | 0.16 | 0.45 | 0.56 | 0.63 | 0.25 |
| CatBoost Regressor | 0.29 | 0.39 | 0.82 | 0.18 | 0.43 | 0.53 | 0.67 | 0.24 |
| Light Gradient Boosting Machine | 0.31 | 0.41 | 0.80 | 0.19 | 0.44 | 0.56 | 0.64 | 0.25 |
| Extreme Gradient Boosting | 0.33 | 0.43 | 0.78 | 0.19 | 0.44 | 0.58 | 0.61 | 0.23 |
| Orthogonal Matching Pursuit | 0.33 | 0.46 | 0.76 | 0.20 | 0.76 | 0.96 | -0.08 | 0.49 |
| Decision Tree Regressor | 0.41 | 0.56 | 0.63 | 0.20 | 0.58 | 0.81 | 0.24 | 0.33 |
| Bayesian Ridge | 0.33 | 0.46 | 0.76 | 0.21 | 7.51 | 11.81 |  | 4.15 |
| Gradient Boosting Regressor | 0.32 | 0.41 | 0.80 | 0.21 | 0.50 | 0.62 | 0.55 | 0.28 |
| Ridge Regression | 0.34 | 0.47 | 0.75 | 0.21 | 0.61 | 0.78 | 0.30 | 0.37 |
| Huber Regressor | 0.33 | 0.46 | 0.75 | 0.21 | 0.75 | 0.95 | -0.06 | 0.49 |
| K Neighbors Regressor | 0.39 | 0.51 | 0.70 | 0.21 | 0.55 | 0.70 | 0.44 | 0.31 |
| Linear Regression | 0.36 | 0.48 | 0.73 | 0.21 | 0.61 | 0.78 | 0.29 | 0.36 |
| AdaBoost Regressor | 0.45 | 0.55 | 0.65 | 0.27 | 0.63 | 0.75 | 0.34 | 0.40 |
| Passive Aggressive Regressor | 0.48 | 0.62 | 0.56 | 0.31 | 0.70 | 0.91 | 0.04 | 0.51 |
| Elastic Net | 0.66 | 0.81 | 0.24 | 0.41 | 0.74 | 0.94 | -0.04 | 0.49 |
| Lasso Regression | 0.80 | 0.99 | -0.13 | 0.53 | 0.74 | 0.94 | -0.03 | 0.49 |
| Lasso Least Angle Regression | 0.80 | 0.99 | -0.13 | 0.53 | 0.76 | 0.96 | -0.07 | 0.48 |
| Dummy Regressor | 0.80 | 0.99 | -0.13 | 0.53 | 0.80 | 0.99 | -0.13 | 0.53 |
| Least Angle Regression | 1.90 | 2.35 | -5.42 | 0.99 | 2.33 | 2.98 | -9.34 | 1.21 |

The results above also perform well compared to studies which only reported *in-sample performance metrics*. One study on rice yield modelling in Bangladesh reported in-sample R2 values ranging from 0.44 to 0.91; out-of-sample performance was not reported.81 Another study on rice yield in China report in-sample R2 values of 0.77, lower than the out-of-sample R2 performance achieved in this study of 0.82 (CatBoost regressor).82 For rice yield prediction in the Philippines, one study reported an in-sample RMSE of 0.46 t/ha.83 Another study using drones reported in-sample R2 values of 0.60 to 0.81 for rice yield prediction in Japan based on NDVI.84,85

**3.2. Errors, residuals, and SHAP value analysis**

Observed and simulated yields show a high level of agreement for some of the top-performing models including the random forest, CatBoost, and LightGBM regressors. In addition, the residual plot residual plots show that the majority of both training set and test set observations are randomly dispersed along the horizontal axis, indicating a reasonable low level of bias and homoscedasticity. The distribution of residuals here is roughly centred around zero but with some skewness, indicating the potential presence of outliers.



**a1**

**a2**

**a3**

**b1**

**b2**

**b3**

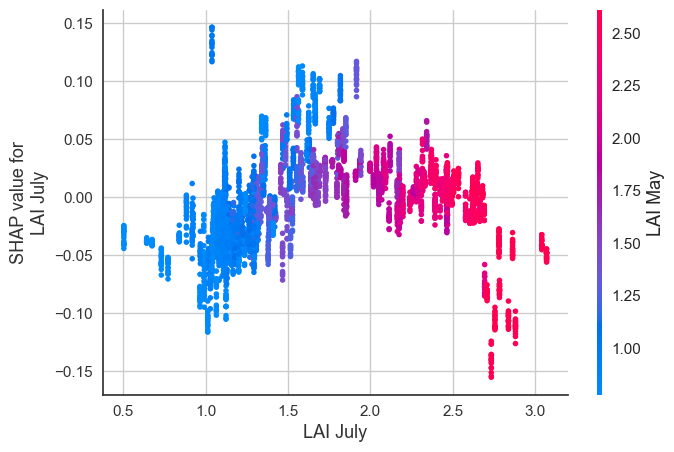
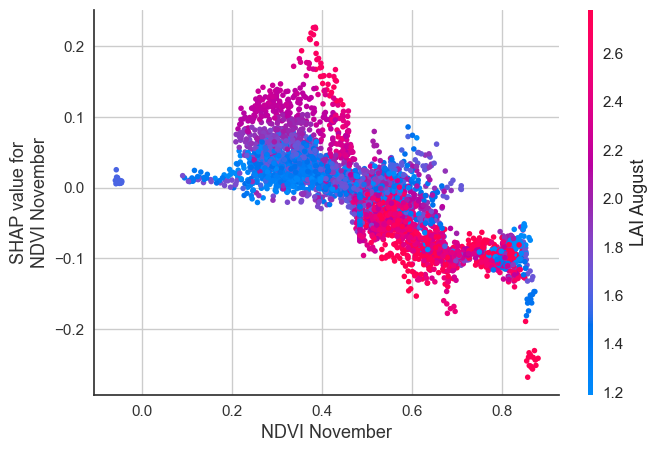
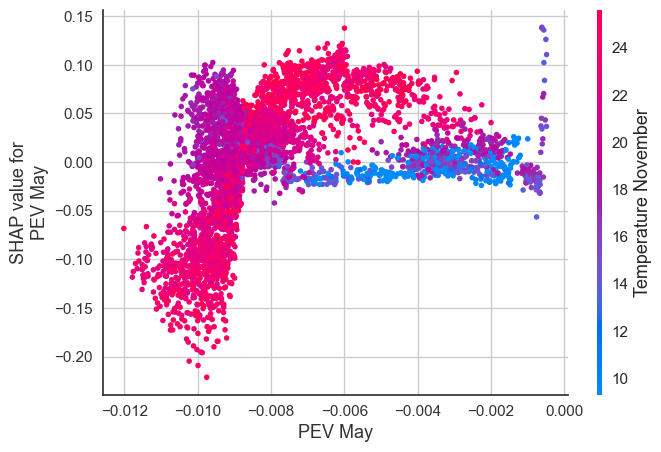
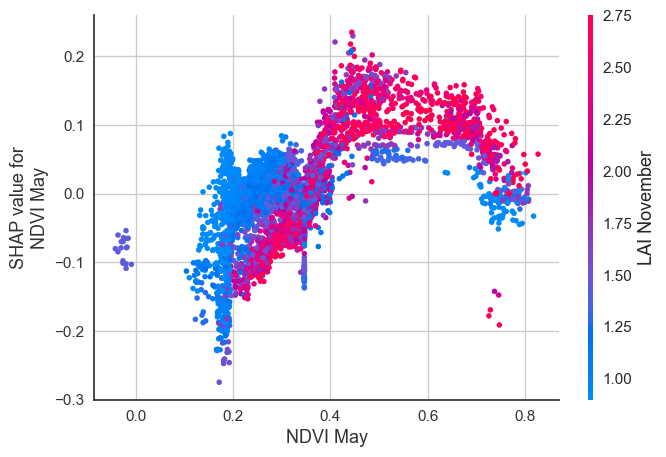
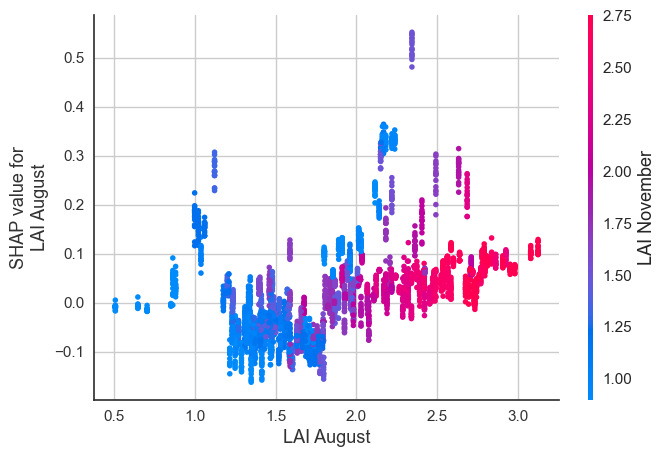
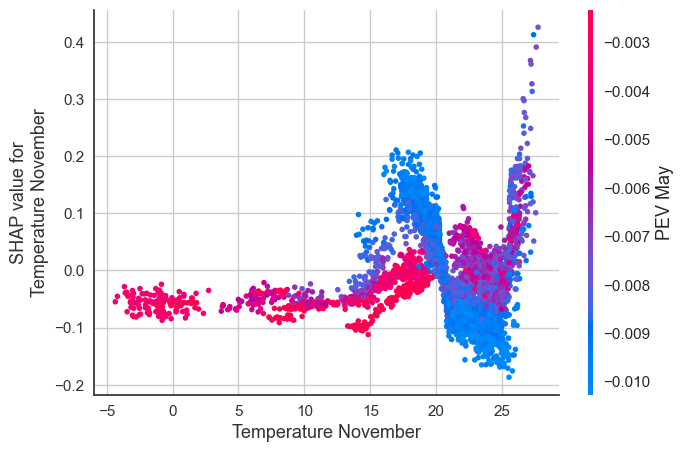
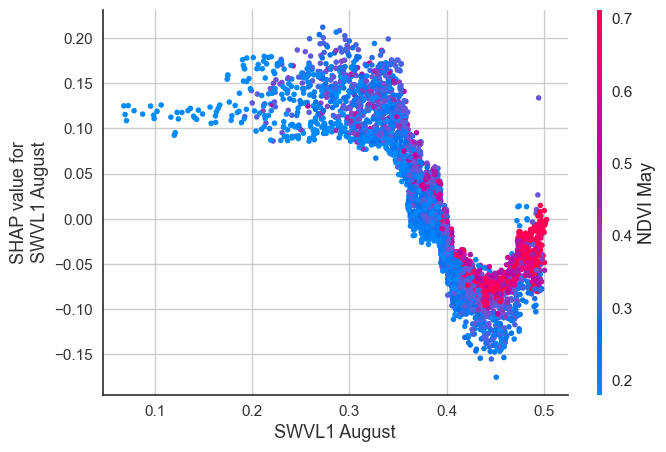
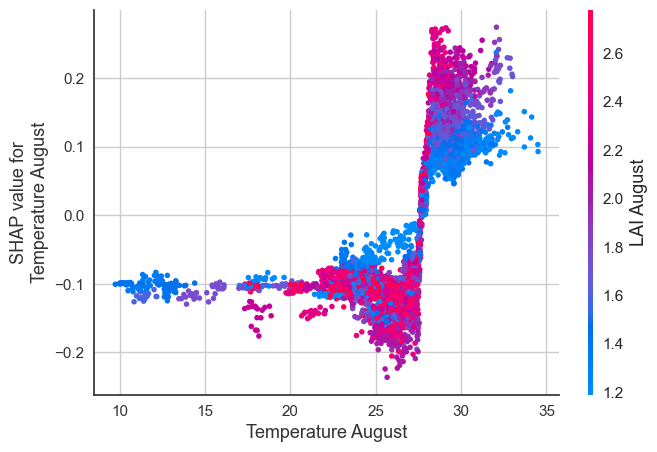
**c1**

**c2**

**c3**

**Figure 2.** A visual representation of prediction errors, residuals, and SHAP-based feature contributions using random forest, CatBoost, and LightGBM regressors. Two years of observations (502 observations in total) where used for the out-of-sample validation data. **a1** to **a3** show theerrors, residuals, and SHAP summary plot for the random forest model, **b1 to b3** for the CatBoost regressor, and **c1 to c3** for the LightGBM regressor. Out-of-sample R2 ranges from 0.80 to 0.82 across these models. The SHAP summary plots from experiment 2 (only climate and satellite used for features) in **a3**, **b3**, and **c3** indicates that important variables included temperature, soil water volume, NDVI, and LAI.

In addition to the errors and residuals, the SHAP summary plots in figure 2 concisely display the magnitude, prevalence and direction of a variable’s effect on final rice yield. The plots reveal that important variables include temperature, soil water volume, NDVI, and LAI in selected months. The importance ranking of these variables corroborates previous findings that that factors such as soil water content, temperature, and NDVI are important factors in estimating rice growth.37,38



**a**

**b**

**c**

**d**

**e**

**f**

**g**

**h**

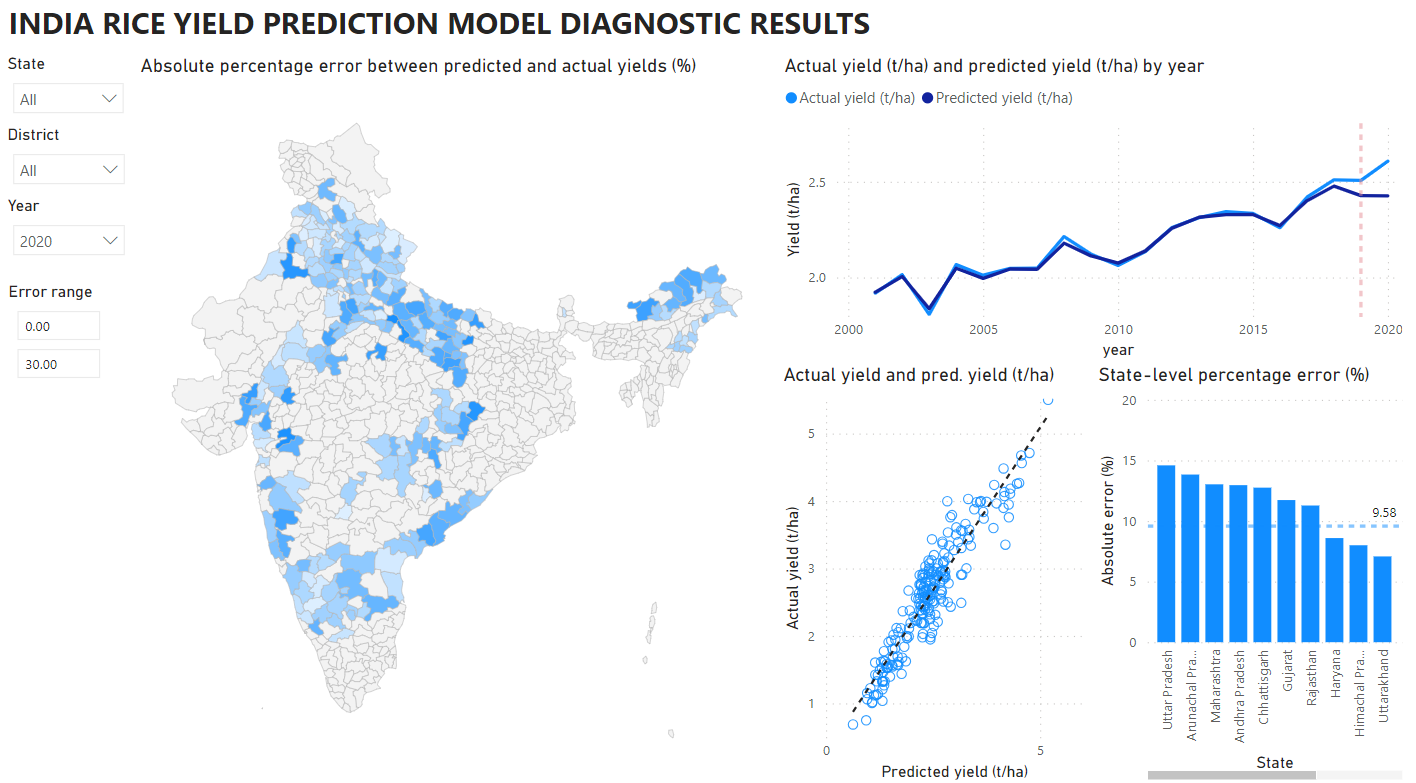
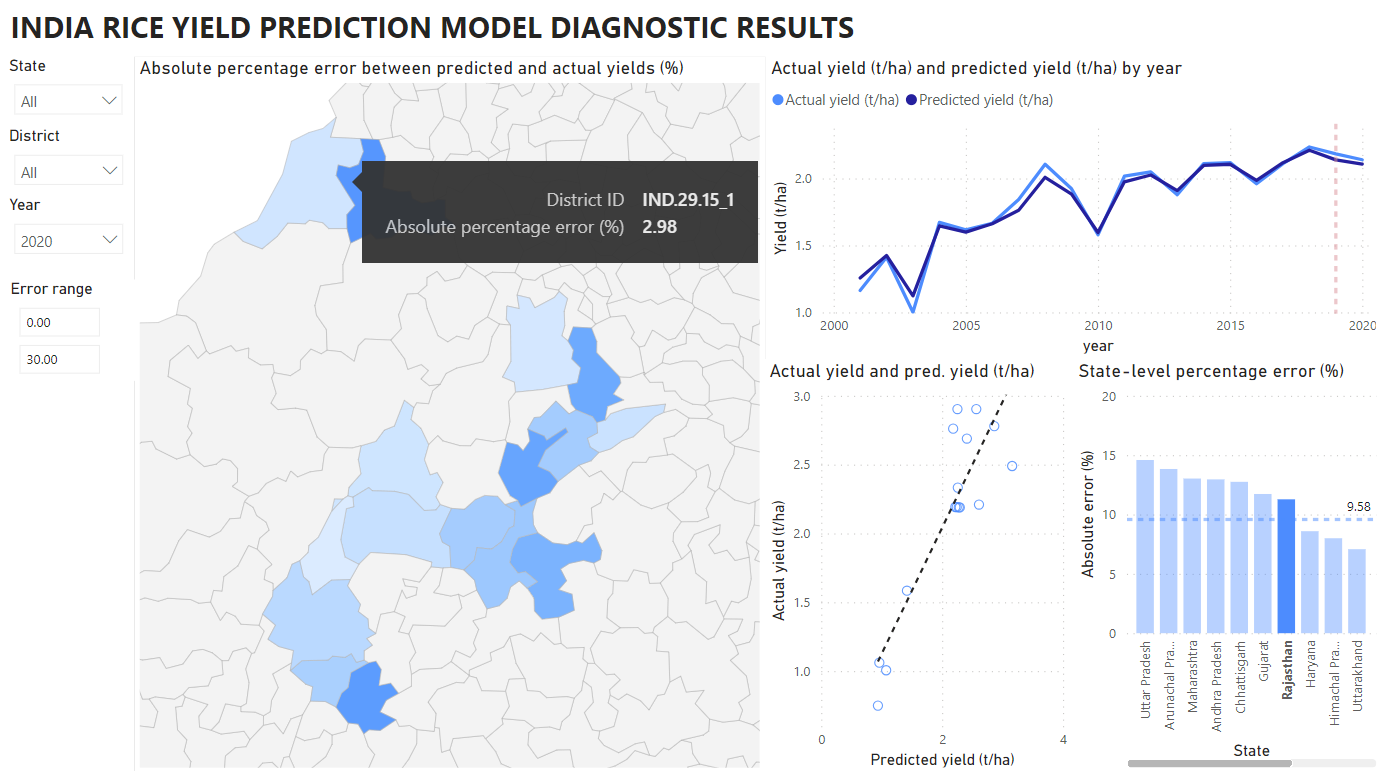
**i**

**Figure 3**. SHAP dependence plots for selected features versus their SHAP values in the rice yield prediction model. The colored bar represents interaction effects of the selected features with other features for which SHAP value correlations were highest. **a.** The dependence plot indicates that higher temperatures in August correlate with increased SHAP values for rice yield, particularly when the leaf area index (LAI) in August is also high, as shown by the red color intensity, signifying a strong interaction between temperature and LAI. **b.** As soil water volume (SWVL) in July decreases, SHAP values for yield prediction increase, especially when the temperature in May is high, as indicated by the red color intensity. Similar interactions are visualized in panels **c** to **f** for soil water volume, temperature, LAI, NDVI, and PEV in various months.

A closer analysis of the specific impacts of features on rice yield is shown in figure 3. Increases in temperature in August, which coincides with the sowing to panicle initiation phases of rice growth, are associated with higher yields, corroborating previous findings in India that above average yields may be associated with higher maximum temperatures32.

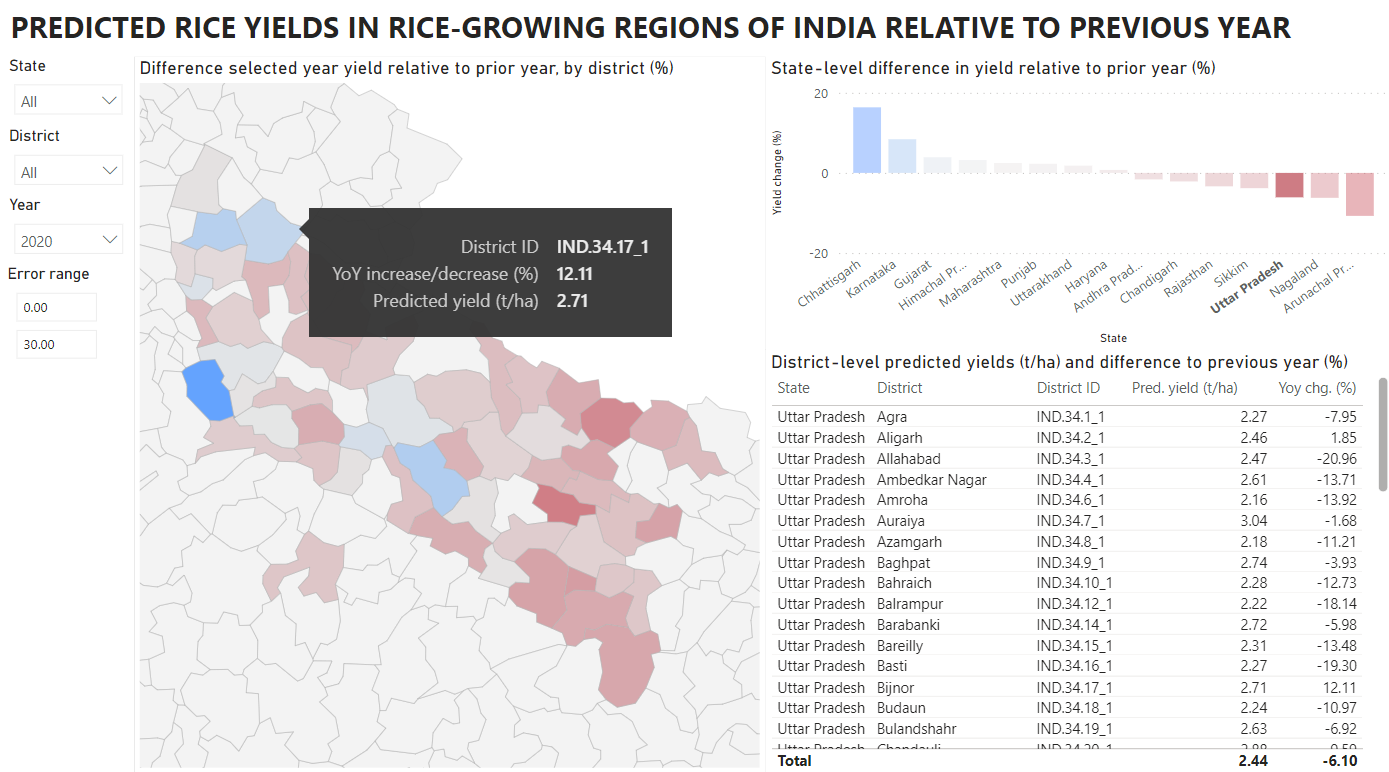
**3.3. Interactive visualization tool**

In addition to the static plots of model performance above, two visual dashboards were developed to both (a) provide an easy-to-understand, spatially explicit summary of model predictions, and (b) to help to identify potential biases in model performance.



**a1**

**b1**



**a2**

**b2**

**Figure 4**. Panels **a1** and **a2** illustrate the model results. Panel **a1** shows panel shows a map of India with different districts highlighted, indicating the predicted rice yields relative to the previous year. The color coding represents the difference in predicted yield relative to the prior year, with a corresponding table listing the districts and their respective yield differences. **a2** shows a similar view, zoomed in to districts within the state of Uttar Pradesh. Panels **b1** and **b2** provide model diagnostic information. **b1** shows a map with the average percentage error by region; a scatter plot comparing the predicted yield and the actual yield; and a line graph showing the actual yield and predicted yield each year. **b2** shows a similar view, zoomed in to districts within the state of Rajasthan.

**3.4. Implications for future research on the design of early warning systems in India**

How good do predictions need to be? One study showed that model accuracy is “very good” if MAPE values are less than 10%, and “good” if the values range between 10% to 20%.86

**3.5. Future research directions**

**Modeling and feature engineering.**

Increased predictive power of rice yields model may potentially be achieved by incorporating a wider array of agronomically relevant climatological variables. Prior investigations have highlighted the influence of thermal extremes, precipitation, diurnal humidity variations, and solar radiation on rice yields.32 In addition, agricultural yield models – particularly those that are grounded in time series analysis – may stand to gain from incorporating auto-regressive elements, rolling averages, or cumulative indices.87 Nevertheless, despite the potential benefits of expanding the feature set, rice yield models with expanded variables may be exposed to more overfitting, feature redundancy, challenges in model interpretability. For models where interpretability is important, reduced feature sets may contribute to more effective interpretation.

However, the augmentation of the feature spectrum must be approached with caution. An expanded variable set within rice yield models can inadvertently lead to overfitting, introduce feature redundancy, and complicate model interpretability. In scenarios where model clarity and understandability are paramount, a more conservative approach with a reduced set of features might be advisable. This focused feature set could streamline the interpretive process, allowing for clearer insights and more straightforward decision-making.

Agronomist information (pairing with LLMs)

Integrating crop yield prediction models into early warning systems

Clarifying how the end users of these tools can directly or independently benefit from such systems

Monitoring important variables at key stages of crop growth

Include broader sets of variables (e.g., agronomic practices such as fertilizer production etc.)

Hierarchical models

**User-friendly interfaces for both farmers and policymakers.** Additional research is needed on access to climate and yield anomaly information via user-friendly tools can help mitigate farmers’ losses. Studies have already shown that poor access to access to climate information in South Asia has been observed as a factors driving perceived losses88.

**Early warning systems with greater context-specificity.** For example, Kharif rice in India can be grown under both irrigated and non-irrigated conditions. Moreover, strategies that aim to diffuse early warning systems may require tailoring depending on the socio-economic status of farmers, the size of their farms, the level of internet connectivity, household income, and other factors which may affect how farmers interact with digital technologies. Past research has shown that differences in farmers backgrounds can significantly affect adoption of agricultural technologies89.

**Research on how government support mechanisms can most optimally enhance the diffusion of yield prediction-enabled early warning systems.** Indian government agencies may be able to boost farmers’ climate resilience by boosting connectivity in regions most affected by weather anomalies. For example, to counteract the negative effects of climate change on agriculture, Bangladesh has boosted its support services for farmers, including enhanced information access and extension services. The government's 'Info Sarkar' project aims to link government offices nationwide and has established over 4,500 internet-equipped Union Digital Centres (UDCs) to aid rural communities. Studies indicate these centers are being effectively utilized by educated youths, suggesting a potential for these individuals to lead in disseminating climate adaptation strategies to farmers88.

**Anticipatory allocation of aid.** Studies thatfocus on mechanisms to embed crop yield predictions in anticipatory aid schemes such as crop index insurance or anticipatory allocation of critical inputs such as fertilizer or irrigation could help mitigate the risks of crop failure from natural events like droughts and floods. This is particularly important for protecting small and marginal farmers from the financial impacts of crop loss.

Diffusion of models for uptake to

**4. Conclusion**

**References**

1. Weiss, M., Jacob, F. & Duveiller, G. Remote sensing for agricultural applications: A meta-review. *Remote Sens. Environ.* **236**, 111402 (2020).

2. van Klompenburg, T., Kassahun, A. & Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **177**, 105709 (2020).

3. Sagan, V. *et al.* Field-scale crop yield prediction using multi-temporal WorldView-3 and PlanetScope satellite data and deep learning. *ISPRS J. Photogramm. Remote Sens.* **174**, 265–281 (2021).

4. Li, A., Liang, S., Wang, A. & Qin, J. Estimating Crop Yield from Multi-temporal Satellite Data Using Multivariate Regression and Neural Network Techniques. *Photogramm. Eng. Remote Sens.* **73**, 1149–1157 (2007).

5. Wang, R., Rejesus, R. M. & Aglasan, S. Warming Temperatures, Yield Risk and Crop Insurance Participation. *Eur. Rev. Agric. Econ.* **48**, 1109–1131 (2021).

6. Wong Jing Wen, Y., Rajeswari Ponnusamy, R. & Ming Kang, H. Application of weather index-based insurance for paddy yield: The case of Malaysia. *Int. J. Adv. Appl. Sci.* **6**, 51–59 (2019).

7. FAO. India at a glance. *Food and Agriculture Organization of the United Nations* (2018).

8. Nayak, H. S. *et al.* Rice yield gaps and nitrogen-use efficiency in the Northwestern Indo-Gangetic Plains of India: Evidence based insights from heterogeneous farmers’ practices. *Field Crops Res.* **275**, 108328 (2022).

9. FAOSTAT. Crops and livestock products. *Food and Agriculture Organization* (2021).

10. Ishfaq, M. *et al.* Alternate wetting and drying: A water-saving and ecofriendly rice production system. *Agric. Water Manag.* **241**, 106363 (2020).

11. Sattar, A. & Srivastava, R. C. Modelling climate smart rice-wheat production system in the middle Gangetic plains of India. *Theor. Appl. Climatol.* **144**, 77–91 (2021).

12. Xu, X. *et al.* Design of an integrated climatic assessment indicator (ICAI) for wheat production: A case study in Jiangsu Province, China. *Ecol. Indic.* **101**, 943–953 (2019).

13. Khaki, S., Wang, L. & Archontoulis, S. V. A CNN-RNN Framework for Crop Yield Prediction. *Front. Plant Sci.* **10**, (2020).

14. Wang, A. X., Tran, C., Desai, N., Lobell, D. & Ermon, S. Deep Transfer Learning for Crop Yield Prediction with Remote Sensing Data. in *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies* (ACM, 2018). doi:10.1145/3209811.3212707.

15. Charoen-Ung, P. & Mittrapiyanuruk, P. Sugarcane Yield Grade Prediction Using Random Forest with Forward Feature Selection and Hyper-parameter Tuning. in *Recent Advances in Information and Communication Technology 2018* (eds. Unger, H., Sodsee, S. & Meesad, P.) 33–42 (Springer International Publishing, 2019).

16. Filippi, P. *et al.* An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precis. Agric.* **20**, 1015–1029 (2019).

17. Ahmad, I. *et al.* Yield Forecasting of Spring Maize Using Remote Sensing and Crop Modeling in Faisalabad-Punjab Pakistan. *J. Indian Soc. Remote Sens.* **46**, (2018).

18. Goldstein, A. *et al.* Applying machine learning on sensor data for irrigation recommendations: revealing the agronomist’s tacit knowledge. *Precis. Agric.* **19**, 421–444 (2018).

19. Chaurasiya, A. *et al.* Layering smart management practices to sustainably maintain rice yields and improve water use efficiency in eastern India. *Field Crops Res.* **275**, 108341 (2022).

20. Chu, Z. & Yu, J. An end-to-end model for rice yield prediction using deep learning fusion. *Comput. Electron. Agric.* **174**, 105471 (2020).

21. Tian, L., Wang, C., Li, H. & Sun, H. Yield prediction model of rice and wheat crops based on ecological distance algorithm. *Environ. Technol. Innov.* **20**, 101132 (2020).

22. Yu, W. *et al.* Improved prediction of rice yield at field and county levels by synergistic use of SAR, optical and meteorological data. *Agric. For. Meteorol.* **342**, 109729 (2023).

23. Wan, L. *et al.* Grain yield prediction of rice using multi-temporal UAV-based RGB and multispectral images and model transfer – a case study of small farmlands in the South of China. *Agric. For. Meteorol.* **291**, 108096 (2020).

24. Jha, P. K. *et al.* Using daily data from seasonal forecasts in dynamic crop models for yield prediction: A case study for rice in Nepal’s Terai. *Agric. For. Meteorol.* **265**, 349–358 (2019).

25. Jeong, S., Ko, J. & Yeom, J.-M. Predicting rice yield at pixel scale through synthetic use of crop and deep learning models with satellite data in South and North Korea. *Sci. Total Environ.* **802**, 149726 (2022).

26. Arumugam, P., Chemura, A., Schauberger, B. & Gornott, C. Remote Sensing Based Yield Estimation of Rice (Oryza Sativa L.) Using Gradient Boosted Regression in India. *Remote Sens.* **13**, 2379 (2021).

27. Ranjan, A. K. & Parida, B. R. Paddy acreage mapping and yield prediction using sentinel-based optical and SAR data in Sahibganj district, Jharkhand (India). *Spat. Inf. Res.* **27**, (2019).

28. Gómez, D., Salvador, P., Sanz, J. & Casanova, J. L. Regional estimation of garlic yield using crop, satellite and climate data in Mexico. *Comput. Electron. Agric.* **181**, 105943 (2021).

29. Auffhammer, M., Ramanathan, V. & Vincent, J. R. Climate change, the monsoon, and rice yield in India. *Clim. Change* **111**, 411–424 (2012).

30. Sharma, S. *et al.* Field-specific nutrient management using Rice Crop Manager decision support tool in Odisha, India. *Field Crops Res.* **241**, 107578 (2019).

31. FAO. GIEWS - Global Information and Early Warning System. *Food and Agriculture Organization of the United Nations* (2021).

32. Bal, S. K. *et al.* Critical weather limits for paddy rice under diverse ecosystems of India. *Front. Plant Sci.* **14**, (2023).

33. Alvarado, J. R. Influence of air temperature on rice population, length of period from sowing to flowering, and spikelet sterility. (2002).

34. Hersbach, H. *et al.* The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).

35. Urban, A. *et al.* Evaluation of the ERA5 reanalysis-based Universal Thermal Climate Index on mortality data in Europe. *Environ. Res.* **198**, 111227 (2021).

36. Colston, J. M. *et al.* Evaluating meteorological data from weather stations, and from satellites and global models for a multi-site epidemiological study. *Environ. Res.* **165**, 91–109 (2018).

37. Chakraborty, P. K., Banerjee, S., Nath, R. & Samanta, S. Assessing congenial soil temperature and its impact on root growth, grain yield of summer rice under varying water stress condition in Lower Gangetic Plain of India. *J. Saudi Soc. Agric. Sci.* (2021) doi:https://doi.org/10.1016/j.jssas.2021.07.001.

38. Jia, Y. *et al.* Effects of low water temperature during reproductive growth on photosynthetic production and nitrogen accumulation in rice. *Field Crops Res.* **242**, 107587 (2019).

39. Kuwagata, T. *et al.* Hydrometeorology for plant omics: potential evaporation as a key index for transcriptome in rice. *Environ. Exp. Bot.* 104724 (2021) doi:https://doi.org/10.1016/j.envexpbot.2021.104724.

40. Tang, Y. *et al.* Effects of long-term low atmospheric pressure on gas exchange and growth of lettuce. *Adv. Space Res.* **46**, 751–760 (2010).

41. Poddar, R., Acharjee, P. U., Bhattacharyya, K. & Patra, S. K. Effect of irrigation regime and varietal selection on the yield, water productivity, energy indices and economics of rice production in the lower Gangetic Plains of Eastern India. *Agric. Water Manag.* 107327 (2021) doi:https://doi.org/10.1016/j.agwat.2021.107327.

42. Schinasi, L. H., Benmarhnia, T. & De Roos, A. J. Modification of the association between high ambient temperature and health by urban microclimate indicators: A systematic review and meta-analysis. *Environ. Res.* **161**, 168–180 (2018).

43. Lopresti, M. F., Di Bella, C. M. & Degioanni, A. J. Relationship between MODIS-NDVI data and wheat yield: A case study in Northern Buenos Aires province, Argentina. *Inf. Process. Agric.* **2**, 73–84 (2015).

44. Panek, E. & Gozdowski, D. Analysis of relationship between cereal yield and NDVI for selected regions of Central Europe based on MODIS satellite data. *Remote Sens. Appl. Soc. Environ.* **17**, 100286 (2020).

45. Son, N. T., Chen, C. F., Chen, C. R., Minh, V. Q. & Trung, N. H. A comparative analysis of multitemporal MODIS EVI and NDVI data for large-scale rice yield estimation. *Agric. For. Meteorol.* **197**, 52–64 (2014).

46. Tang, F. H. M. *et al.* CROPGRIDS: A global geo-referenced dataset of 173 crops circa 2020. *Earth Syst. Sci. Data Discuss.* 1–22 (2023) doi:10.5194/essd-2023-130.

47. MOA. Crop Production Statistics Information System. *Ministry of Agriculture and Farmers Welfare* (2021).

48. Sonkar, G. *et al.* Vulnerability of Indian wheat against rising temperature and aerosols. *Environ. Pollut.* **254**, 112946 (2019).

49. Hersbach, H. *et al.* ERA5 hourly data on pressure levels from 1940 to present, Tech. Rep. (2023).

50. Didan, K. MODIS/Aqua Vegetation Indices Monthly L3 Global 0.05Deg CMG V061. (2021) doi:10.5067/MODIS/MYD13C2.061.

51. ECMWF | Parameter details. https://codes.ecmwf.int/grib/param-db/?id=66.

52. Areas, G. A. GADM database of Global Administrative Areas, version 2.0. (2012).

53. Longfei, Z. *et al.* Improved Yield Prediction of Ratoon Rice Using Unmanned Aerial Vehicle-Based Multi-Temporal Feature Method. *Rice Sci.* **30**, 247–256 (2023).

54. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).

55. Shi, Q., Abdel-Aty, M. & Lee, J. A Bayesian ridge regression analysis of congestion’s impact on urban expressway safety. *Accid. Anal. Prev.* **88**, 124–137 (2016).

56. Tipping, M. E. Sparse Bayesian Learning and the Relevance Vector Machine. *J Mach Learn Res* **1**, 211–244 (2001).

57. MacKay, D. J. C. Bayesian Interpolation. in *Maximum Entropy and Bayesian Methods: Seattle, 1991* (eds. Smith, C. R., Erickson, G. J. & Neudorfer, P. O.) 39–66 (Springer Netherlands, 1992). doi:10.1007/978-94-017-2219-3\_3.

58. Huang, H. & Abdel-Aty, M. Multilevel data and Bayesian analysis in traffic safety. *Accid. Anal. Prev.* **42**, 1556–1565 (2010).

59. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).

60. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

61. Huber, P. & Ronchetti, E. *Robust Statistics*. (Wiley, 2009).

62. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*. (Springer New York, 2009). doi:10.1007/978-0-387-84858-7.

63. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).

64. Hastie, T., Rosset, S., Zhu, J. & Zou, H. Multi-class AdaBoost. *Stat. Interface* **2**, 349–360 (2009).

65. Rubinstein, R., Zibulevsky, M. & Elad, M. Efficient Implementation of the K-SVD Algorithm using Batch Orthogonal Matching Pursuit. in (2008).

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

67. Cerqueira, V., Torgo, L. & Mozetič, I. Evaluating time series forecasting models: an empirical study on performance estimation methods. *Mach. Learn.* **109**, 1997–2028 (2020).

68. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 4768–4777 (Curran Associates Inc., 2017).

69. Molnar, C. *Interpretable Machine learning. A Guide for Making Black Box Models Explainable*. (lulu.com, 2019).

70. Lenaers, I. & De Moor, L. Exploring XAI techniques for enhancing model transparency and interpretability in real estate rent prediction: A comparative study. *Finance Res. Lett.* **58**, 104306 (2023).

71. Jain, V., Kavitha, H. & Mohana Kumar, S. Credit Card Fraud Detection Web Application using Streamlit and Machine Learning. in *2022 IEEE International Conference on Data Science and Information System (ICDSIS)* 1–5 (2022). doi:10.1109/ICDSIS55133.2022.9915901.

72. Nantasenamat, C., Biswas, A., Nápoles-Duarte, J. M., Parker, M. I. & Dunbrack, R. L. Chapter 27 - Building bioinformatics web applications with Streamlit. in *Cheminformatics, QSAR and Machine Learning Applications for Novel Drug Development* (ed. Roy, K.) 679–699 (Academic Press, 2023). doi:10.1016/B978-0-443-18638-7.00001-3.

73. Patil, S. & Lokesha, V. Live Twitter Sentiment Analysis Using Streamlit Framework. SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.4119949 (2022).

74. Belghith, M., Ben Ammar, H., Elloumi, A. & Hachicha, W. A new rolling forecasting framework using Microsoft Power BI for data visualization: A case study in a pharmaceutical industry. *Ann. Pharm. Fr.* (2023) doi:10.1016/j.pharma.2023.10.013.

75. Abusager, K., Baldwin, M. & Hsu, V. Using Power BI to Inform Clostridioides difficile Ordering Practices at an Acute Care Hospital in Central Florida. *Am. J. Infect. Control* **48**, S57–S58 (2020).

76. Nickell, E. B., Schwebke, J. & Goldwater, P. An introductory audit data analytics case study: Using Microsoft Power BI and Benford’s Law to detect accounting irregularities. *J. Account. Educ.* **64**, 100855 (2023).

77. Guo, Y. *et al.* Integrated phenology and climate in rice yields prediction using machine learning methods. *Ecol. Indic.* **120**, 106935 (2021).

78. Clauss, K., Ottinger, M., Leinenkugel, P. & Kuenzer, C. Estimating rice production in the Mekong Delta, Vietnam, utilizing time series of Sentinel-1 SAR data. *Int. J. Appl. Earth Obs. Geoinformation* **73**, 574–585 (2018).

79. Han, J. *et al.* Rice yield estimation using a CNN-based image-driven data assimilation framework. *Field Crops Res.* **288**, 108693 (2022).

80. Su, X. *et al.* Grain yield prediction using multi-temporal UAV-based multispectral vegetation indices and endmember abundance in rice. *Field Crops Res.* **299**, 108992 (2023).

81. Islam, Md. M., Matsushita, S., Noguchi, R. & Ahamed, T. Development of remote sensing-based yield prediction models at the maturity stage of boro rice using parametric and nonparametric approaches. *Remote Sens. Appl. Soc. Environ.* **22**, 100494 (2021).

82. Zhou, X. *et al.* Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery. *ISPRS J. Photogramm. Remote Sens.* **130**, 246–255 (2017).

83. Setiyono, T. D. *et al.* Spatial Rice Yield Estimation Based on MODIS and Sentinel-1 SAR Data and ORYZA Crop Growth Model. *Remote Sens.* **10**, (2018).

84. Guan, S. *et al.* Assessing Correlation of High-Resolution NDVI with Fertilizer Application Level and Yield of Rice and Wheat Crops Using Small UAVs. *Remote Sens.* **11**, (2019).

85. Ajith, K., Geethalakshmi, V., Ragunath, K. P., Pazhanivelan, S. & Dheebakaran, G. Rice Yield Prediction Using MODIS - NDVI (MOD13Q1) and Land Based Observations. *Int. J. Curr. Microbiol. Appl. Sci.* **6**, 2277–2293 (2017).

86. Son, N.-T. *et al.* Field-scale rice yield prediction from Sentinel-2 monthly image composites using machine learning algorithms. *Ecol. Inform.* **69**, 101618 (2022).

87. Cerqueira, V., Moniz, N. & Soares, C. VEST: automatic feature engineering for forecasting. *Mach. Learn.* (2021) doi:10.1007/s10994-021-05959-y.

88. Islam, Z., Alauddin, M. & Sarker, Md. A. R. Determinants and implications of crop production loss: An empirical exploration using ordered probit analysis. *Land Use Policy* **67**, 527–536 (2017).

89. Samal, P., Pandey, S., Kumar, G. A. K. & Barah, B. C. Rice Ecosystems and Factors Affecting Varietal Adoption in Rainfed Coastal Orissa: A Multivariate Probit Analysis. *Agric. Econ. Res. Rev.* **24**, (2011).