Projecting Crop Yields based on ESM Emulation

Nilay Menon nimenon@ucsd.edu

Gina Roberg groberg@ucsd.edu

Charles Wang czwang@ucsd.edu

Qilong Zou q1zou@ucsd.edu

Duncan Watson-Parris dwatsonparris@ucsd.edu

Abstract

The world continues to face the impact of climate change, posing great challenges, particularly with global agriculture. Policymakers and researchers need efficient tools to assess crop productivity under certain climate scenarios. Traditional crop modeling approaches offer high accuracy, but are often computationally intensive and difficult to scale. In response, this study aims to develop a machine learning-based emulator to predict crop yields, specifically for maize, rice, wheat, and soybeans using climate and environmental data from Earth System Model simulations. By leveraging models like Random Forest Regressor and Gaussian Process trained on historical crop yield and climate data from the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP), we efficiently predict annual crop yields based on key climate variables such as solar radiation, temperature, and precipitation. Our approach significantly reduces costs while maintaining accuracy, providing an easily accessible framework to measure the impacts of climate change on global food security. Our method aims to improve the ability of policymakers to anticipate agricultural challenges and ensure that food production is sustainable in a changing climate.

Website: https://github.com/Ginaroberg/Global-Crop-Yield-Climate-Simulator/tree/main
Code: https://github.com/Ginaroberg/DSC-180B-Capstone-B13

1 Introduction

The Norwegian Earth System Model (NorESM 2) can develop complex models that can project global climate conditions, by simulating various climate scenarios (Watson-Parris et al. 2022). Given a set of conditions, like increase in various greenhouse gas emissions, or change in temperature, the model can create predictions of how various climate variables (surface temperature, pre- cipitation, etc.) could change across global regions. These projections can provide insight into Earth's climate trajectory and ultimately aid policy makers to make informed decisions. However, presently, the information that can be extracted from these projections does not provide context into how this will impact society and what specific action must be taken. Specifically, we examine the impact on agriculture and crop yields. By utilizing climate data and agricultural data from (Lange and Büchner 2020), we are able to train the emulators render various climate scenarios that can simulate how crop yields for staple crops, like maize, wheat, barley, and soybeans, will change in a range from 50 to 100 years in the future. These projections can be crucial in developing effective policies in response to global climate change, which can ultimately impact the agricultural sector, farmers, landowners, and overall consumers.

There have been previous research on climate driven crop yield prediction, many of which have largely relied on statistical approaches. For example, models like the "Lund-Potsdam-Jena managed Land" simulate crop growth based on changing environmental conditions and detailed mathematical equations (Potsdam Institute for Climate Impact Research (n.d.)). The problem with these models is that they are computationally expensive. Statistical models have helped solve this issue as these models rely on historical climate and crop yield data to establish relationships. As a result, however, they often struggle to capture complex and nonlinear interactions. There have also been recent studies that have explored machine learning techniques including the use of neural networks to enhance prediction accuracy while reducing computational costs. For example, convolutional neural networks combined with satellite image data have been utilized for crop yield prediction (Russello and Shang (2018)). Our project aims to build on prior work by integrating machine learning models with the Earth System Model outputs to provide a scalable and efficient framework for predicting future crop yields under various climate conditions.

In this project, we utilized climate and crop yield data to train emulators for crop yield predictions. The input data consists of climate variables such as precipitation (pr), downward longwave radiation at the surface (rlds), downward shortwave radiation at the surface (rsds), surface wind speed (sfcwind), near-surface air temperature (tas), daily maximum near-surface air temperature (tasmax), and daily minimum near-surface air temperature (tasmin). All input data is sourced from the Earth System Model (ESM) simulations and more specifically from the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP). The output data represents historical and projected crop yields for maize, wheat, rice, and soybean. These crop yields are obtained from the LPJml global crop model. To process both input and output data, we first extracted the relevant climate variables and crop yield data for specified regions and time regions. Most missing values were handled by filling out NaN's or by averaging the spatial dimensions (Latitude and Longitude) to obtain global and regional means. This preprocessing of the data allowed us gather the proper climate and crop variables to run our models.

2 Methods

2.1 Gaussian Process Regression

Gaussian Process Regression (GPR) is a non-parametric, Bayesian approach to regression that provides not only point predictions but also a principled quantification of uncertainty. In this study, GPR is used to emulate the relationship between climate variables and crop yields, offering a flexible framework to capture nonlinear dependencies inherent in ESM outputs. At its core, GPR assumes that the underlying function $f(\mathbf{x})$ relating the input features $\mathbf{x}$ (e.g. climate variables such as pr, rsds, tas) to the target crop yields is drawn from a Gaussian process:

$$f(\mathbf{x}) \sim \mathbf{GPR}\big(m(x), k(x, x')\big),$$

where $m(x)$ is the mean function (often set to zero without loss of generality) and $k(x, x')$ is the covariance kernel. In our implementation, we incorporated a composite kernel built from several Automatic Relevance Determination (ARD) Matern32 kernels. ARD allows the model to learn a separate lengthscale for each input dimension, thereby identifying the relative importance of individual features. Given that our input data is partitioned into seven groups corresponding to key climate variables—pr, rlds, rsds, sfcwind, tas, tasmax, and tasmin—each group is modeled by its own Matern32 kernel

$$k(r) = \sigma^2\left(1 + \sqrt{3}r\right)\exp\left\{-\sqrt{3}r\right\},$$

where r is the scaled Euclidean distance between input points (with scaling provided by feature-specific lengthscales). This formulation allows each kernel component to capture the moderately smooth yet complex behavior of a specific group of climate variables. The composite kernel is defined as follows:

```
kernel = (
    gpflow.kernels.Matern32(active_dims=[0, 1, 2, 3, 4], lengthscales=[1.0]*5) +  # pr
    gpflow.kernels.Matern32(active_dims=[5, 6, 7, 8, 9], lengthscales=[1.0]*5) +  # rlds
    gpflow.kernels.Matern32(active_dims=[10, 11, 12, 13, 14], lengthscales=[1.0]*5) +  # rsds
    gpflow.kernels.Matern32(active_dims=[15, 16, 17, 18, 19], lengthscales=[1.0]*5) +  # sfcwind
    gpflow.kernels.Matern32(active_dims=[20, 21, 22, 23, 24], lengthscales=[1.0]*5) +  # tas
    gpflow.kernels.Matern32(active_dims=[25, 26, 27, 28, 29], lengthscales=[1.0]*5) +  # tasmax
    gpflow.kernels.Matern32(active_dims=[30, 31, 32, 33, 34], lengthscales=[1.0]*5)    # tasmin
)
```

Here, each Matern32 kernel operates over a specific set of active dimensions corresponding to one of the climate variables, with an initial lengthscale of 1.0 for each dimension. During training, the hyperparameters (including the lengthscales and variance parameters) of each kernel component are optimized by maximizing the log marginal likelihood. This process automatically determines the relevance of each feature group, allowing the model to focus on the most influential climate variables. To manage the high-dimensional climate input data—originally processed using Empirical Orthogonal Functions (EOFs) to reduce dimensionality—we then applied GPR separately for each crop yield target (maize, wheat, rice, and soy). This preprocessing step ensured that the model focuses on the most significant modes of variability in the climate data, which improves both the interpretability and computational efficiency of the GP model. A key advantage of GPR is its ability to provide predictive distributions:

$$p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mu^*, \sigma^{*^2}),$$

where μ_* and σ_*^2 denote the predictive mean and variance at a new input $\mathbf{x}_*$. This probabilistic output is particularly valuable in the context of climate impact assessment, as it enables policymakers to understand the confidence bounds of crop yield projections and to identify regions of high uncertainty that may require further investigation or data collection. In summary, our GPR model successfully leverages the flexibility and uncertainty quantification inherent in Gaussian processes to emulate crop yield predictions under various climate scenarios. Although the computational cost is higher compared to other deterministic methods, the additional insights provided by the uncertainty estimates make GPR an appealing tool for risk assessment and decision-making in the context of global agriculture under climate change.

2.2 Random Forest Regressor

The random forest model is an ensemble machine learning approach that incorporates the construction of multiple decision trees and combines their predictions to improve overall accuracy, increase robustness, and reduce overfitting. An important advantage of the random forest model is that it can handle complex non linear relationships. This is useful as climate data is non linear and not easily usable with standard regression techniques. Given this model and the goal of reproducing climate model data, a random forest regressor model can be something to look at.

A random forest regressor is a specific version of the random forest model which predicts a numeric value instead of classifying. To detail how training works on a RF regressor, the model trains multiple decision tress on a random subset of the data and each of these tress predicts a value for the target variable. For predictions, each new input is passed through the regressor and each tree in the regressor gives a prediction value. The final prediction would then be an average of all the tree predictions.

The model used in this project is from the Earth System Emulator (ESM) library. The data preparation for this model involved extracting input climate variables sourced from the ISIMIP repository such as temperature, radiation, precipitation, and wind speed. To address any dimensionality in these input variables, Empirical Orthogonal Functions (EOFs) were applied to the input variables such as precipitation (pr), downward longwave radiation at the surface (rlds), downward shortwave radiation at the surface (rsds), surface wind speed (sfcwind), near-surface air temperature (tas), daily maximum near-surface air temperature (tasmax), and daily minimum near-surface air temperature (tasmin). This reduces dimensionality in the variables while preserving variability patterns. The target variables in our study were crop yields for maize (mai), rice (ri1,ri2), wheat (swh, wwh), and soy which were obtained from the LPJml crop model simulations. The target data for this model covered the time period 2015-2100.

For each crop, the random forest model was trained to predict yield based on the given climate variable inputs. The model operated on key hyperparameters including n_estimators (controls the number of trees), min_samples_split (minimum number of samples required to split an internal node), min_samples_leaf (minimum number of samples in leaf nodes), max_depth (tree depth limit), and max_features (choice for RF model to reduce correlation among trees). The values for these hyperparameters were initially set based on previous knowledge on the domain but to further optimize model performance, a small grid search was performed. Even after the grid search however, the performance of the models were very similar.

The trained random forest models were then evaluated by comparing their predictions to the LPJml crop yield outputs. In general, the models tend to slightly under-predict crop yields which could be attributed to data limitations, simplifications in the overall model structure, or even the inability to fully capture the complex interactions between climate variables and crop productivity. Despite these challenges, the emulator provides an efficient way to estimate crop yields under varying climate scenarios.

3 Results

3.1 Random Forest

4 Conclusion

5 Contributions

Nilay Menon: Researched possible data sources to use for predictions, looked into ways to reduce size of input data, built and refined Random Forest Model for crop yield predictions, looked into ideas of how to structure website for our deliverable.

Gina Roberg: Researched possible data sources to use for predictions, looked into CNN model for crop yield predictions, started implementation of website and built overall structure of website, and attempted to link prediction data into interactive graphs on the website.

Charles Wang: Researched possible data sources to use for predictions,looked into CNN model for crop yield predictions,looked into ideas of how to structure website for our deliverable.

Qilong Zou: Researched possible data sources to use for predictions, processed input data into usable size, built Gaussian Process Regression model for crop yield predictions, started implementation of website and deliverable. Built one structure of what our website could look like.

References

Lange, Stefan, and Matthias Büchner. 2020. "ISIMIP3b bias-adjusted atmospheric climate input data." [Link]

Potsdam Institute for Climate Impact Research. n.d.. "LPJML - Lund-Potsdam-Jena managed land." [Link]

Russello, H., and Wenling Shang. 2018. "Convolutional Neural Networks for Crop Yield Prediction using Satellite Images." Online. [Link]

Watson-Parris, D., Y. Rao, D. Olivié, Ø. Seland, P. Nowack, G. Camps-Valls, P. Stier, S. Bouabid, M. Dewey, E. Fons, J. Gonzalez, P. Harder, K. Jeggle, J. Lenhardt, P. Manshausen, M. Novitasari, L. Ricard, and C. Roesch. 2022. "ClimateBench v1.0: A Benchmark for Data-Driven Climate Projections." *Journal of Advances in Modeling Earth Systems* 14 (10), p. e2021MS002954. [Link]

Appendices

Link to Project Proposal: https://www.overleaf.com/read/kbxgnxfchzdte62882

A.1 Training Details

A.2 Additional Figures

A.3 Additional Tables