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://ginaroberg.github.io/DSC-180B-Capstone-B13/
Code: https://github.com/Ginaroberg/DSC-180B-Capstone-B13
Proposal: https://tinyurl.com/yv5yv5t8

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, carbon dioxide, 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 emulators to 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),daily minimum near-surface air temperature (tasmin), carbon dioxide (CO2), Methane (CH4), sulfur dixoide (SO2), and black carbon (BC). All input data is sourced from the Earth System Model (ESM) simulations and more specifically from the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) and the Coupled Model Intercomparison Project (CMIP). 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 ex-

tracted 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 to 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 emission 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 $\mathbf{x}$ 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, 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 the input data consists of four emission variables-global cumulative sum of CO2 and CH4, BC, SO2-each emission variable 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]) +  # CO2
    gpflow.kernels.Matern32(active_dims=[1]) +  # CH4
    gpflow.kernels.Matern32(active_dims=[2, 3, 4, 5, 6], lengthscales=[1.0]*5) +  # SO2
    gpflow.kernels.Matern32(active_dims=[7, 8, 9, 10, 11], lengthscales=[1.0]*5) +  # BC
)
```

Here, each Matern32 kernel operates over a specific set of active dimensions corresponding to one of the emission variables, with an initial lengthscale of 1.0 for each dimension. During training, the lengthscale and variance parameter of each kernel component are optimized by maximizing the log marginal likelihood. To manage the high-dimensional SO2 and BC data, they are processed using Empirical Orthogonal Functions (EOFs) to reduce dimensionality. Then the GPR is applied 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 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 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. In addition to the climate variables, aerosol variables were used from CMIP6 model to make a model to predict crop yields from emission data. 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. Before predictions were made, to handle missing data for the random forest model, an average was taken for each crop over the latitude and longitude values. This average also helps reduce dimensionality and helps with sensitivity issues due to variations in data when running the model.

For each crop, the random forest model was trained to predict yield based on the given climate variable inputs such as CO2, SO2, CH4, and BC. 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 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 Gaussian Process Regression

Table 1: Performance of the GP model on predicted variables (Root Mean Square Error, RMSE)

Variable	RMSE at 2050	RMSE at 2100	RMSE 2045-2055	RMSE 2090-2100	RMSE 2050-2100	RMSE Avg Last 20y
mai	0.56496	0.61513	0.55263	0.53758	0.55217	0.16077
ri1	0.51429	0.62238	0.51356	0.54758	0.54599	0.20426
ri2	0.38867	0.49715	0.39267	0.43840	0.42400	0.13573
soy	0.83446	1.06647	0.87165	0.94446	0.93954	0.28871
wwh	0.58599	0.81055	0.60733	0.68442	0.66739	0.27302
ssh	0.65266	0.91003	0.67516	0.79187	0.74753	0.28467

Looking at the table, performance of the Gaussian Process Regression differs greatly across variables, with prediction strength changing depending on factors like variability and patterns. RMSE or the root mean squared error was used to help calculate error between the actual and predicted crop yield. When we look at RMSE, a lower value means less error in predictions.

RMSE values vary significantly across crops with soybeans and wheat showing the highest prediction errors. This indicates a greater sensitivity to drastic climate changes and

emissions. In contrast, rice seems to show the lowest RMSE values over time, which could indicate a more stable relationship between climate variables and crop yields. Overall the model showcases an increase in RMSE over time for crops which could suggest that as climate variability intensifies, the accuracy of yield predictions will become more challenging. Crops that are influenced by climate variables like precipitation such as soybeans, will be harder to predict due to the non linear relationship between these variables and crop yields.

3.2 Random Forest Regressor

Table 2: Performance of the RF model on predicted variables (Root Mean Square Error, RMSE)

Variable	RMSE at 2050	RMSE at 2100	RMSE 2045-2055	RMSE 2090-2100	RMSE 2050-2100	RMSE Avg Last 20y
mai	0.60445	0.76512	0.57226	0.66501	0.62661	0.40600
ri1	0.58795	0.90276	0.56869	0.82899	0.70535	0.60385
ri2	0.46413	0.70047	0.43300	0.66851	0.54798	0.47902
soy	0.92437	1.39343	0.91536	1.21869	1.09213	0.77653
wwh	0.64245	1.23159	0.68146	1.05876	0.89180	0.80462
ssh	0.69616	1.23707	0.72671	1.04736	0.89009	0.69779

Examining the RMSE values for the Random Forest Model, we can see that prediction error tends to be high for all the crops. This could be due to multiple reasons. When looking at crops such as soybeans that are very dependent on climate variables like temperature and precipitation, the predictions tend to be worse as these variables are often non linear with relation to time. Precipitation is affected not just by climate but a lot of factors like aerosol gases and oceanic processes, which could lead to bad predictions. Crops like rice and wheat, which have multiple growing seasons, might have higher variability due to temperature differences and overall changing climate conditions throughout the seasons. This could make it difficult to capture a proper relationship, resulting in predictions that aren't as accurate. We can also examine regions of crop growth as certain crops like maize and rice cannot be grown in every region. This, combined with the fact that maize growth is heavily dependent on temperature, could make predictions very difficult, resulting in a higher error.

4 Findings and Future Work

The comparison between the Gaussian Process model and the Random Forest model shows key differences in prediction accuracy for crop yields. The GP model produced lower RMSE values throughout all crop yields, indicating better performance in capturing climate to yield production relationships. However, its sensitivity to missing data may limit its scalability for large scale agricultural predictions. In contrast, the RF model, while computationally efficient, showed higher RMSE values in all crops, suggesting a greater difficulty in modeling yield variability with correct accuracy. The variability in RMSE across different crops highlights how some crops, such as soybeans and wheat, are harder to predict due to

their complex relationship with climate factors like precipitation, temperature, and aerosol gases like carbon dioxide, unlike the crop rice, which showed the lowest RMSE generally with both models.

For farmers, these findings provide valuable insight into how different crops might respond to changing climate conditions. The higher RMSE for soybeans suggests that its yield is more sensitive to climate variability, making soybeans a riskier choice in regions that experience extreme weather conditions. Maize and rice show significantly lower RMSE values, which means that their yields can be predicted with more confidence. This information can guide crop selection, schedules for crop planting, and irrigation planning which can help farmers reduce the risks that come with climate change. To add on, policymakers and scientists can use these insights to develop agricultural policies and possibly a early warning systems for farmers in high-risk areas.

In the future, improving such predictive models will be crucial for increasing food security and farming practices. Refinement of the GP model by balancing accuracy and computational efficiency could enhance its applicability in the real world. In addition, incorporating additional data sources such as soil quality and extreme weather events could further improve predictions, particularly for crops with higher RMSE values. Understanding which crops are most affected by climate variability allows for the correct adaptation strategies, such as developing drought resistant crop variations, improving irrigation systems, and/or adjusting planting schedules. Using these insights into policy frameworks and agricultural planning, farmers and decision makers can make more informed decisions, ensuring flexibility against climate-related challenges in global food production.

5 Conclusion

The growing issue of climate change places even more importance on accessible computational models that can emulate crop yields to further help policymakers and farmers with food security.By comparing Gaussian Process and Random Forest models, we found that the GP model generally showed lower prediction error and the RF model was computationally efficient but less accurate. The variability in prediction error across different crops highlights the challenges of predicting yields, especially for crops like soybeans. Farmers and policymakers can leverage these results to make actionable choice with food security in a future with changing climate.

6 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 using climate variables and aerosol variables like carbon dioxide. Attempted implementation of Random Forest model Dashboard. Worked on final report, poster, and code.

Gina Roberg: Researched possible data sources to use for predictions, looked into CNN model for crop yield predictions. Developed dashboard implementing Gaussian Process Model. Worked on final report, poster, and dashboard.

Charles Wang: Researched possible data sources to use for predictions,looked into CNN model for crop yield predictions. Developed static website. Worked on final report, poster, dashboard, code.

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 using climate variables and aerosol variables like carbon dioxide. Worked on final report, poster, dashboard, and code.

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]

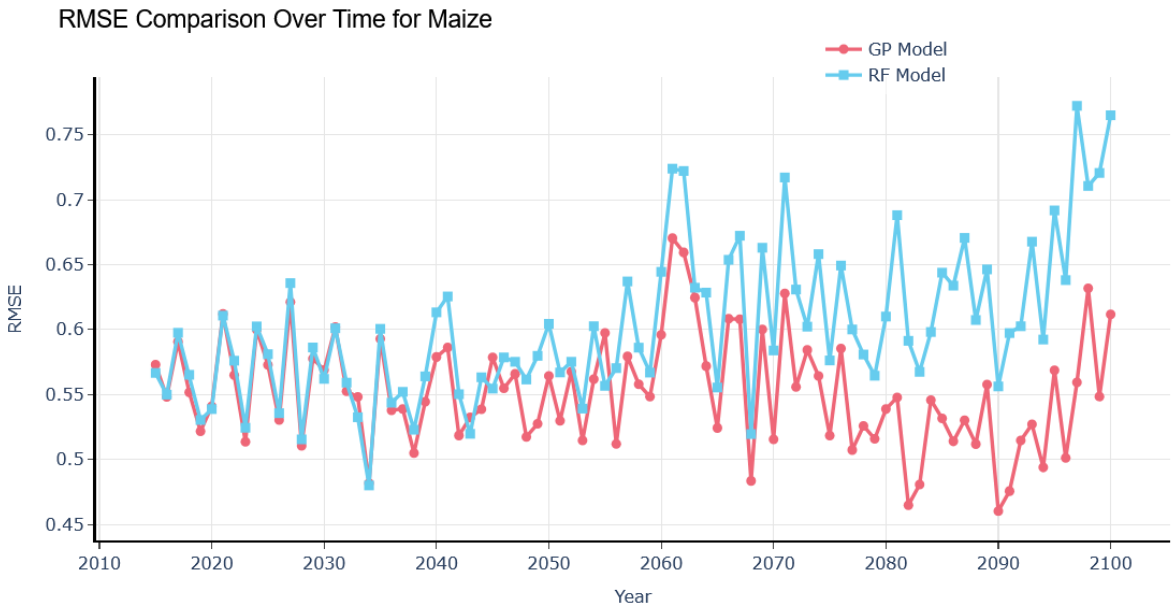
Russello, H., and Wenling Shang. 2018. "Convolutional Neural Networks for Crop Yield Prediction using Satellite Images." Online. [Link]

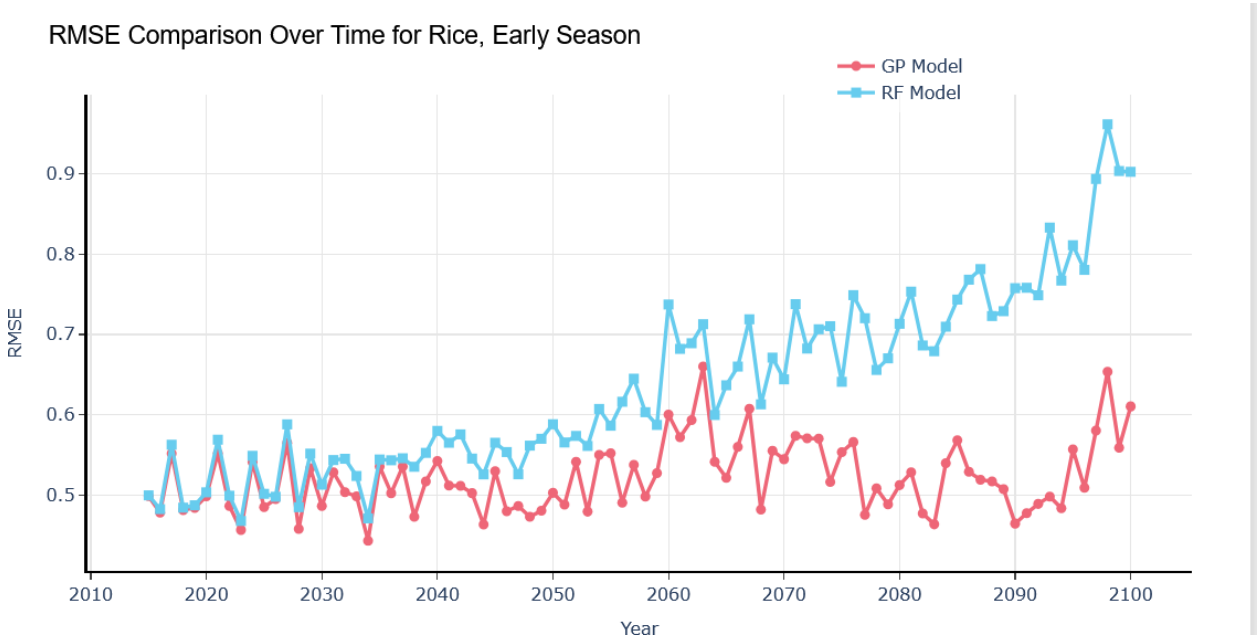
Watson-Parris, D., Y. Rao, D. Olivié, Ø. Seland, P. Nowack, G. Camps-Valls, P. Stier, S. Bouabid, M. Dewey, E. Fons, J. Gonzalez, P. Harder, K. Jeggle, J. Lenhardt, P. Manshausen, M. Novitasari, L. Ricard, and C. Roesch. 2022. "ClimateBench v1.0: A Benchmark for Data-Driven Climate Projections." *Journal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

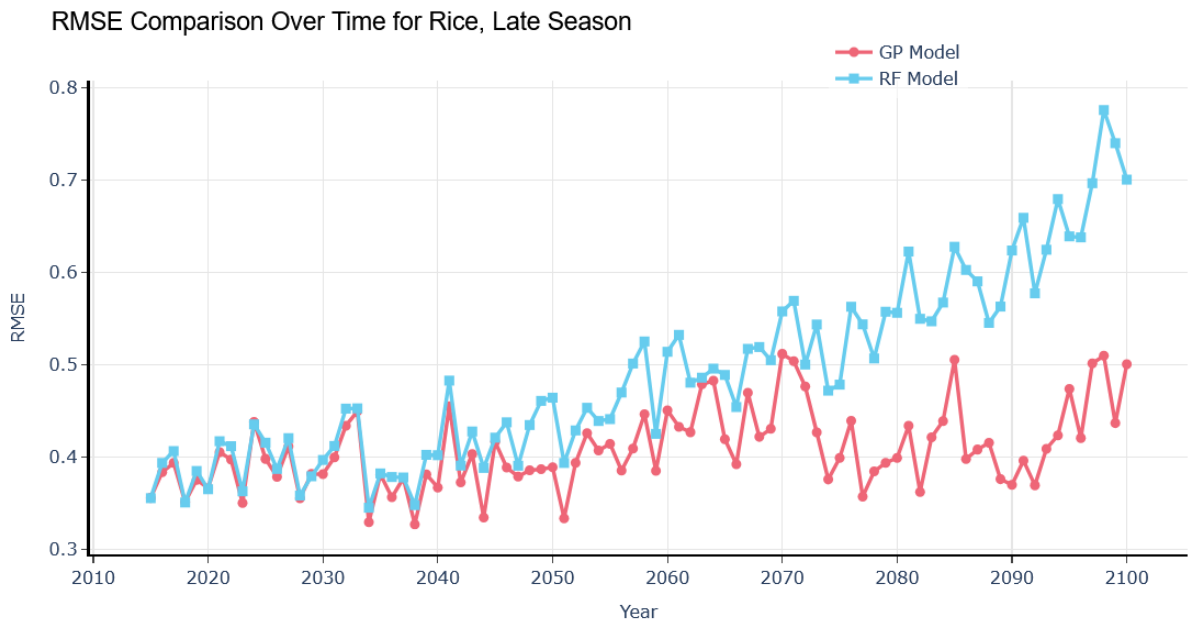
Appendices

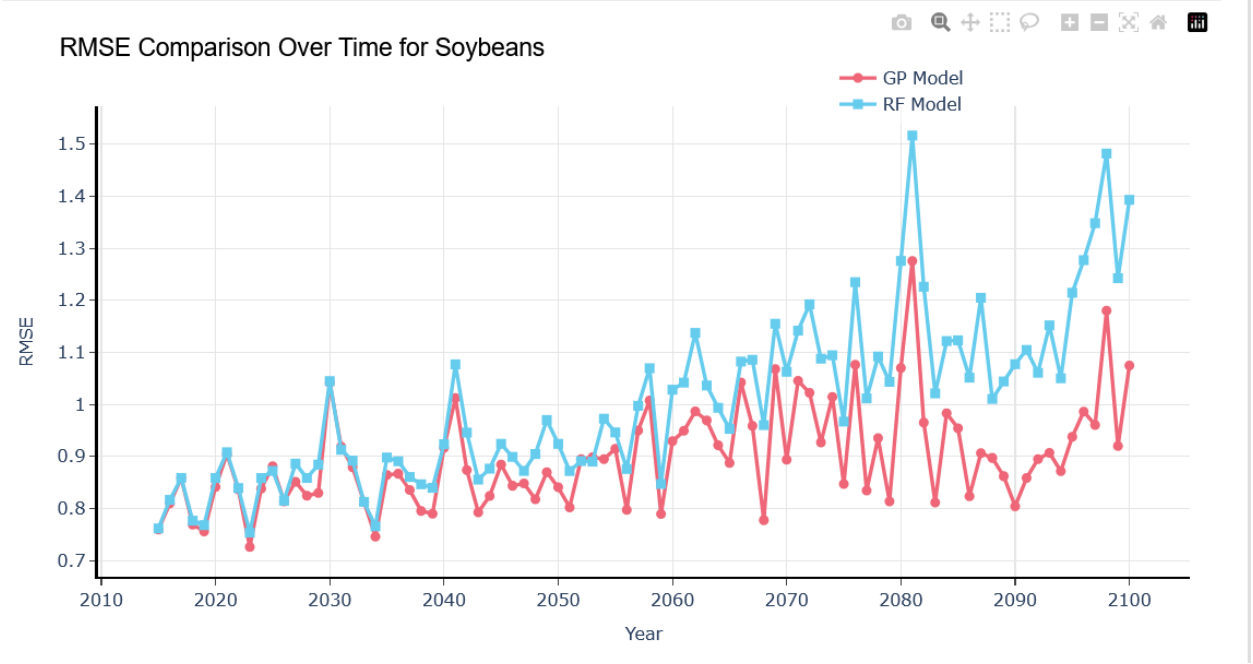
A.1 Additional Figures

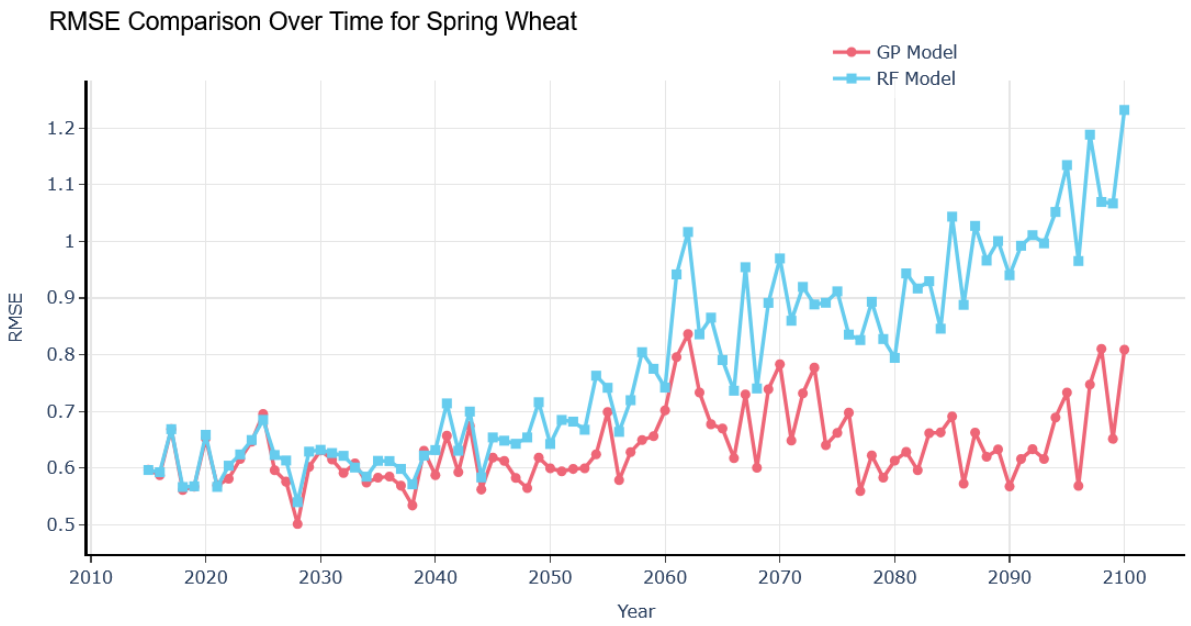
A.1.1 RMSE Comparisons For Crops Per Model

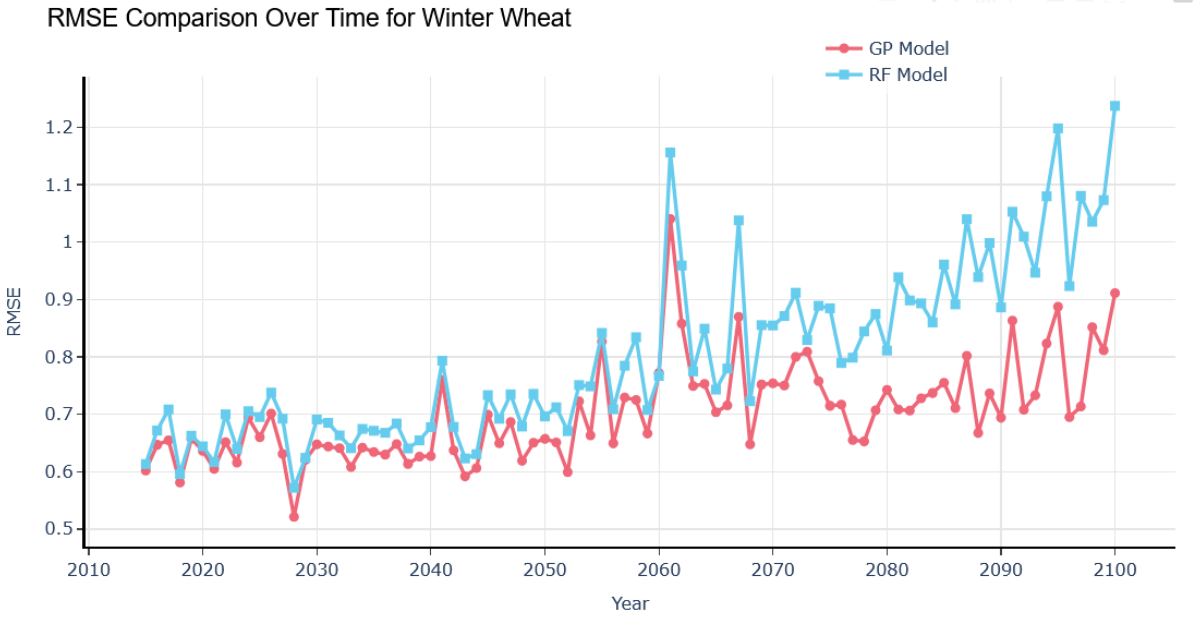




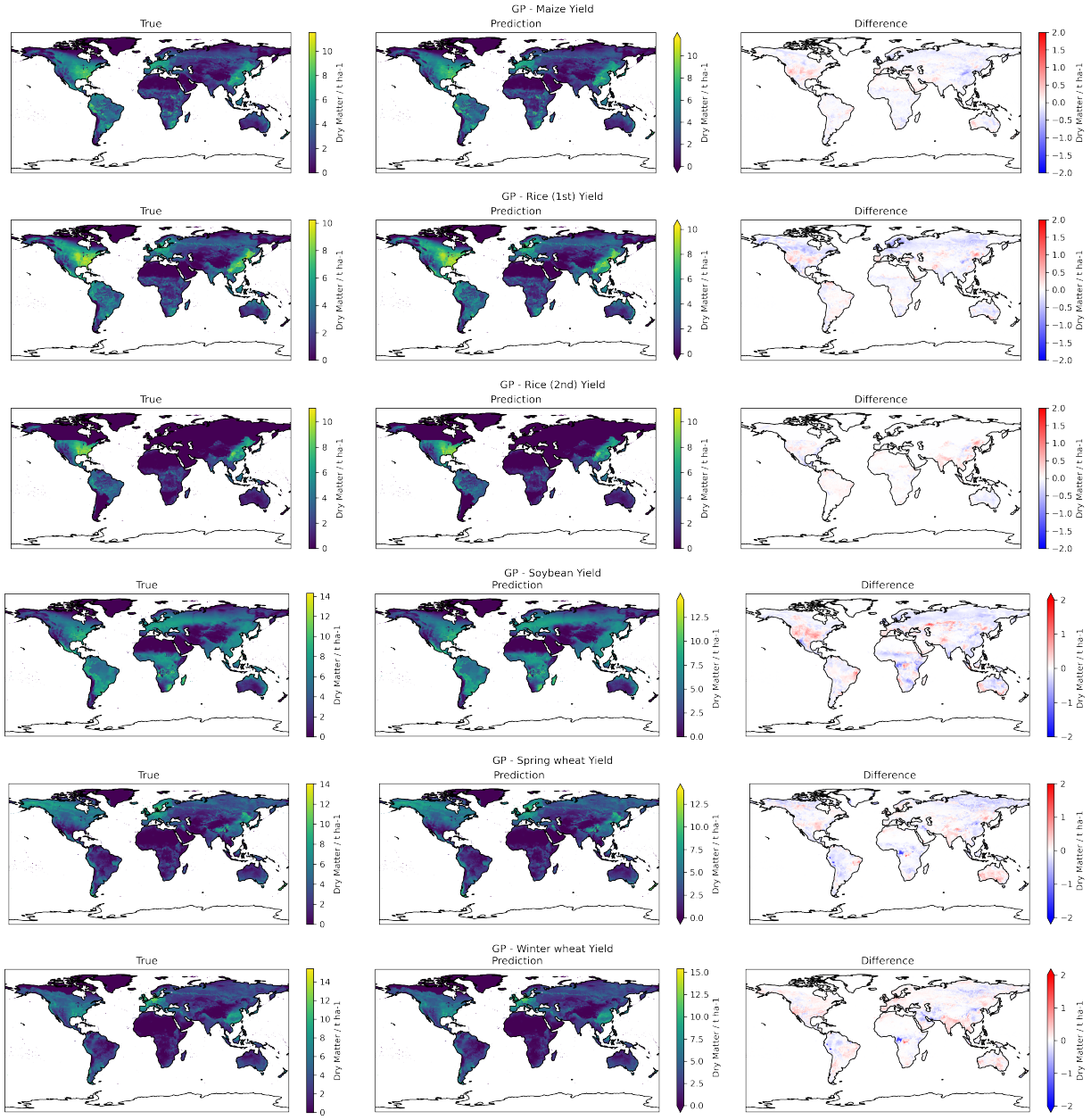

RMSE Comparison Over Time for Rice, Late Season
GP Model
RF Model
RMSE
0.8
0.7
0.6
0.5
0.4
0.3
2010
2020
2030
2040
2050
2060
2070
2080
2090
2100
Year


RMSE Comparison Over Time for Soybeans
GP Model
RF Model
RMSE
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
2010
2020
2030
2040
2050
2060
2070
2080
2090
2100
Year


RMSE Comparison Over Time for Spring Wheat
GP Model
RF Model
RMSE
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
2010
2020
2030
2040
2050
2060
2070
2080
2090
2100
Year


RMSE Comparison Over Time for Winter Wheat
GP Model
RF Model
RMSE
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
2010
2020
2030
2040
2050
2060
2070
2080
2090
2100
Year

A.1.2 Gaussian Process Regression Prediction Error Per Crop



A.1.3 Random Forest Regressor Prediction Error Per Crop