Predicting the Yield of CIMMYT Wheat Genotypes in Novel Locations using Remote Sensing and Machine Learning

By

Aaron M Scherf

A thesis submitted in partial satisfaction of the

Requirements for the degree of

Master of Development Practice

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor David Zilberman, Chair

Professor Alain de Janvry

Professor David Roland Holst

Professor Siamak Khorram

Spring 2020

Abstract

Predicting the Yield of CIMMYT Wheat Genotypes in Novel Locations using Remote Sensing and Machine Learning

by

Aaron M Scherf

Master of Development Practice

University of California, Berkeley

Professor David Zilberman, Chair

Predicting crop yields based on the genotype of the seed variety is a critical field of research for climate adaptation. Existing crop breeding research focuses on maximizing yield gains by determining favorable crosses or genetic alterations for new varieties but offers fewer resources for determining the optimal locations for existing genotypes. Given the limited budgets of many regions which will be most affected by climate change, there is an urgent need for accessible, low-cost research tools which can help model the potential benefits of new seed varieties. Remote sensing data and common machine learning frameworks offer an alternative to the expensive, technically challenging methods currently favored by crop breeding researchers, but their applicability at scale is still uncertain.

This project focuses on assessing the potential of novel data sources, predictive models, and open-source programming tools for crop breeding research. The primary research question is whether remotely sensed environmental data can compare with existing sources in terms of the accuracy of wheat yield predictions, particularly for novel environments without prior experimental data. Another line of inquiry is whether non-parametric machine learning methods can compare with a standard genotype-by-environment interaction model using a Bayesian generalized linear regression. Assessing the viability of each method in terms of data accessibility and computational efficiency informs a tertiary research question, which seeks to determine if crop breeding research can be conducted in a fully open-source process.

The results support the use of machine learning models as more accurate than linear mixed models, though at the cost of interpretability. The inclusion of remote sensing data provided mixed results: improved accuracy for linear models but decreased scores for machine learning models. It is unclear if the results are due to a theoretical superiority of manually collected data, imprecisely measured remote sensing data, or poorly specified statistical models. The machine learning models offered a much more reproducible workflow than the linear models, largely due to the computational intensity of the matrix calculations required by the random effect distributions. Access to pedigree information for CIMMYT wheat varieties also prevents the analysis from full

reproducibility. Future research on improved remote sensing data, machine learning models, and more open data access seems to offer promising alternatives to current crop breeding research methods.

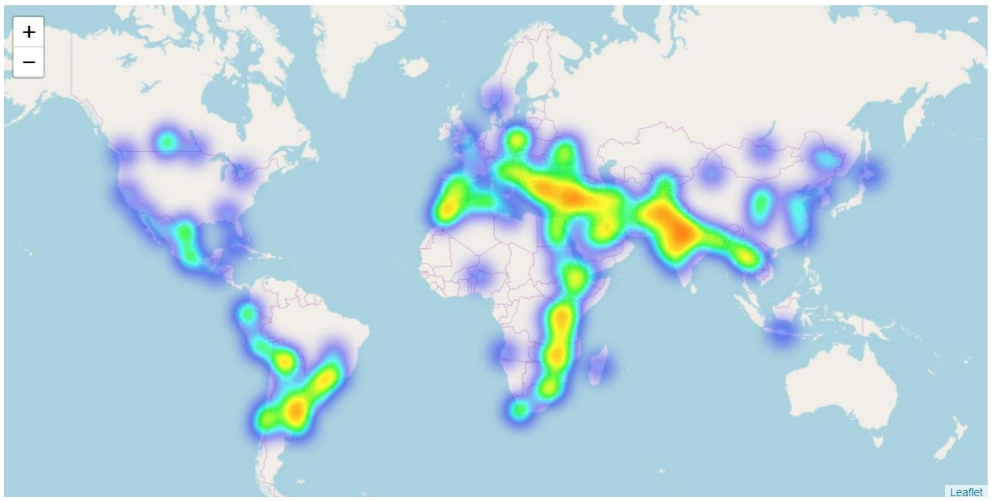
Table of Contents

Introduction

Agricultural researchers face an acute challenge to develop more productive and resilient crop varieties that can help farmers adapt to a rapidly changing climate. Effectively sharing knowledge with agricultural decision makers on the geographic suitability of new crop breeds for their regions is an essential aspect of distribution, but most crop breeding literature fails to adequately address the question of "which variety wins where". Modern farm trial techniques provide a framework for evaluating genotype-environment interactions that can predict yield effects for new crop varieties with some in-sample validity, but extension to new and untested environments remains a barrier to more rapid adoption in low-income countries without extensive testing facilities (Sukumaran et al, 2017).

Advances in remote sensing, big data storage, and machine learning applications offer novel solutions to the problem of out-of-sample prediction using satellite data, but these techniques are limited to highly sophisticated users (Gilbert et al, 2019; Cai et al, 2019; Jain et al, 2019). Google Earth Engine and associated machine learning models hosted in Google Colab offer an open-source, easy to use, and low-cost alternative to support the extension of cutting-edge models for more practical usage but have thus far been limited mostly to land classification (Shelestov et al, 2017). This project aims to create a reproducible and open-source framework for making out-of-sample predictions using multi-environment field trial data, using the collected wheat yield trials from the Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT), the International Centre for the Improvement of Maize and Wheat. While the individual models and technology used are not novel, the combination of CIMMYT's open-source crop and genotype data with Earth Engine's remote sensing catalogue offers a major advancement in the accessibility of an otherwise highly specialized field.

Figure 1: Heatmap of a random subsample of site locations for yield trials



(Source: CIMMYT Wheat Yield Trials; Map by Author)

The initial results from the comparative analysis support the use of machine learning frameworks as a more accurate, computationally less expensive method of predicting grain yields in a genotype-by-environment interaction framework. Not only were prediction accuracies for the machine learning models higher than for the linear mixed models, they were far more accessible in terms of software availability and reproducibility. An important caveat, however, lies in the interpretability of the models; while linear mixed models provide coefficient weights with p-values, only the random forest and gradient boosting models provide a list of feature importance, without any indication of the magnitude or direction of the effect. Machine learning can thus serve as a complement to linear models, particularly for yield predictions or simulations to determine optimal genotype varieties but may be less directly useful for identifying specific genetic attributes, environmental conditions, or interactions that can help researchers make breeding decisions.

The inclusion of remote sensing data led to mixed results in terms of predictive accuracy: linear mixed models were more accurate compared to those using manually collected data, but machine learning models were less accurate. It is unclear whether these results are due to a theoretical advantage of manual observation, such as the inclusion of management practices like row separation, or a methodological limitation. The remote sensing data employed by this study relied on inaccurate location data and considerable missing temporal information. The parameters for the machine learning models may also be sub-optimal. Given the advantages of remote sensing data for out-of-sample predictions, including its wider availability and lower cost, further research on the applicability of Earth Engine datasets for crop breeding seems warranted.

The reproducibility and accessibility of crop breeding research methods were likewise mixed. CIMMYT and Earth Engine data were freely available online but genotype information on pedigree coefficients were not. Linear mixed models depend on computationally intensive matrix calculations that exceed the capacity of most computing environments; a virtual machine with 1.2 terabytes of temporary memory was necessary. Machine learning models, on the other hand, were easily calculated on common hardware. The opportunities for increased transparency and reproducibility from open-science methods do not seem to impose binding constraints on high accuracy yield prediction modelling.

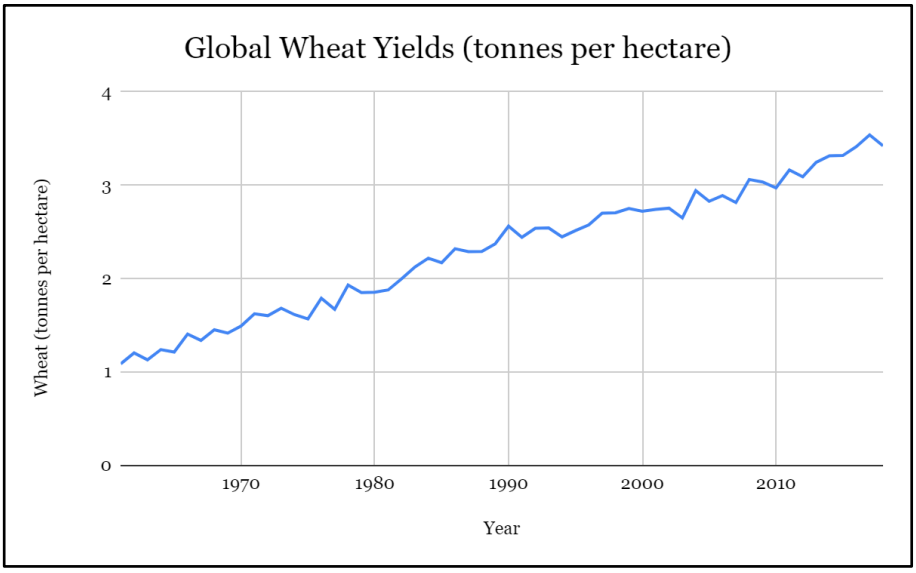
Context and Literature Review

Global food security in the next few decades will depend on the development and dissemination of crop varieties which can increase the yields, resilience, and nutritious value of agricultural goods in historically less productive regions. The combined effects of population growth and climate change are exacerbating the challenge facing crop breeders to develop better varieties in a shorter time frame. The gradual decline of yield productivity improvements from genetic gains demands new approaches to meet the need for new crop varieties. Equally vital is the targeted distribution of these varieties to geographic regions with climatic and management systems conducive to their optimal growth. To meet these challenges, agricultural researchers across the world have made use of a constantly evolving set of crop selection methods, with an increasing focus on the capacity of statistical modelling and computer simulations to help develop and distribute new varieties. Technological adaptation has played a pivotal role in increasing the capacity of the world's agricultural system and promises to mitigate future food security challenges, but only if it can be distributed efficiently to the places where it is needed most.

Trends in Global Wheat Production and Yields

Wheat has always been central to the field of crop selection. Since the Green Revolution of the 1960's in South Asia, improved varieties of wheat have helped to avert Malthusian forecasts of mass hunger by multiplying the productive capacity of farmers. Norman Borlaug, a researcher at CIMMYT, is credited with the development of disease-resistant semi-dwarf wheat varieties which led to the doubling of grain yields in the period between 1965 and 1970 (Curtis and Halford, 2014). The then Director of the US Agency for International Development (USAID), William Gaud, coined the phrase "Green Revolution" to describe the dramatic increase in agricultural productivity that began in India and spread throughout the world. The World Food Conference of 1974 was convened to determine how advances in crop science could be disseminated, leading to the establishment of organizations such as the International Food Policy Research Institute (IFPRI), International Fund for Agricultural Development (IFAD), and the Consultative Group on International Agricultural Research (CGIAR). The collective work of these and similar initiatives drove global wheat production to increase by 273% between 1961 and 2007, an average annual rate of 5.93% (USAID, 2016).

Figure 2: Global Wheat Yields (1961 - 2018)



(Source: Ritchie and Roser, 2013; Chart by Author)

Wheat yields have continued to increase through much of the past two decades, albeit at a much slower pace than in the Green Revolution. By 2020, production of wheat grain globally reached 759 million tons, but yield growth rates in some regions have stagnated. The divergence in yield trends between highly productive countries such as China and lower yielding regions such as Kazakhstan, Morocco, and Iraq have introduced many challenges to international trade networks but has also provided many opportunities to increase productivity for poorer regions through technological leapfrogging. Leveraging drought and heat resistant varieties to increase productivity in these regions can help to improve global supply and diversify production regions without necessarily cultivating more land area, thus preserving more habitat and avoiding the harmful emissions caused by land clearance (Curtis and Halford, 2014).

Table 1: Wheat Production, Area, and Yields per Country (2010)

Select Wheat Yields by Country (2010)			
Country	**Production (1000 tonnes)**	**Area (million ha)**	**Yield (tonnes per ha)**
European Union	132 251	25.50	5.19
China	120 600	24.14	5.00
India	94 880	29.69	3.20
United States	61 755	19.82	3.12
Russian Federation	37 717	21.30	1.77
Pakistan	23 300	8.66	2.69

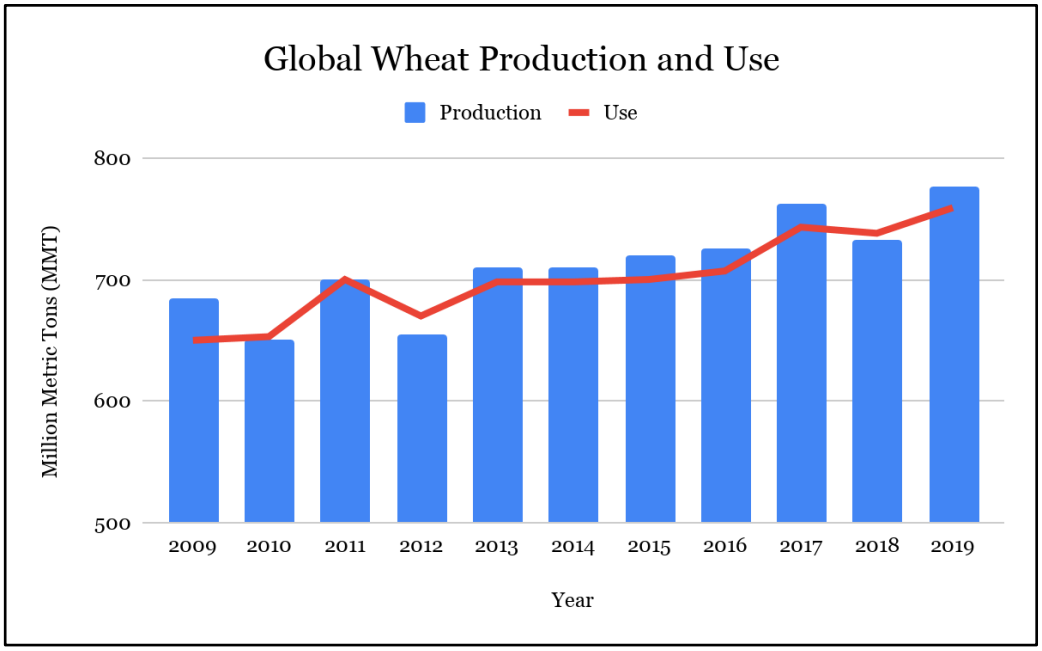
Turkey	15 500	7.80	1.99
Iran	14 000	7.00	2.00
Kazakhstan	11 000	12.40	0.89
Egypt	8 500	1.35	6.30
Afghanistan	4 150	2.51	1.65
Morocco	3 400	3.14	1.08
Mexico	3 230	0.57	5.67
Ethiopia	3 100	1.50	2.07
Iraq	2 100	1.25	1.68
Global	651 000	215.92	3.03

(Source: Curtis and Halford, 2014)

By 2050, global demand for wheat is expected to reach over 1.3 billion tons, doubling the levels from 2010 (Godfray, 2011). To meet this demand, production will have to increase by 553 million tons, a 71.3% increase from the 777 million tons produced in 2020. Over the next thirty years, that represents an average increase of 1.8% per year. By 2050, wheat supply will have to increase by 23.7 million tons per year just to meet demand. These figures were calculated assuming stable dietary patterns; if residents of lower income countries adopt rich-world, meat-heavy diets as their incomes increase, the total demand for wheat could increase by approximately 500 million additional tons, doubling the pace at which yields would have to increase (Tilman et al, 2011; calculations by author).

Some of this demand may be met by expanding available farmland or increasing productivity in rich countries, though the former option has high environmental and economic costs and the latter seems increasingly unlikely as yield improvements in highly productive regions slow. The most cost-effective strategy for meeting global demand for wheat, therefore, is to improve yields in low-productivity countries, thereby increasing production on the intensive margin in areas where yields are currently weaker. Tilman et al. (2011) also calculated that, by focusing on intensification of existing croplands in low-productivity regions, 800 million hectares of land could be saved from clearing for agriculture and 2 gigatons of CO_2 emissions could be avoided.

Figure 3: Global Wheat Production and Use (2009 - 2019)



(Source: U.S. Wheat Associates, 2020; Chart by Author)

Preserving our environment and mitigating the effects of climate change caused by greenhouse gas emissions is vital for our global food security. Compounding the case for a focus on yield improvements in lower income, low-productivity regions is the potential effects of climate change on agriculture in Africa and South Asia. Knox et al. (2012) conducted a meta-analysis of climate forecasting literature and calculated that the average change in yield across eight major crops by 2050 is -8% for both Africa and South Asia. Across the continent of Africa, wheat yields are expected to fall by -17% due to changing environmental conditions. Not only would such declines devastate the agriculturally dependent economies of many countries, further impoverishing farmers, it would increase the continent's dependency on food imports.

Losses in crop yields across equatorial regions are expected to be caused primarily by drought and water scarcity. Tnka et al. (2019) projected that, without significant mitigation of climate change, up to 60% of agricultural areas currently dedicated to wheat globally will face simultaneous water scarcity, compared to current rates of 15%. Even with full implementation of the Paris Climate Agreement, water scarcity would still reach double the current levels, causing more frequent and severe shocks to the supply of wheat, inevitably increasing prices and exacerbating the food insecurity of poorer countries.

Understanding the Role of Crop Breeding for Yield Improvements

In recent years, average genetic gains--the increase in yield attributable to improved genotypic varieties--have declined to less than 1% annually (Reynolds et al, 2012). Crespo-Herrera et al. (2017) found genetic gains between 2006 and 2015 of 1 to 2.7% for CIMMYT's Elite Spring Wheat Yield Trials (ESWYT), a collection of the most

promising varieties currently in development. A similar study found grain yield increases of 1.6 to 1.8% for the Semi-Arid Wheat Yield Trials between 2002 to 2014 (Crespo-Herrera, 2018). These gains were achieved in experimental conditions, however, not on-farm trials. Even if similar yield improvements could be sustained at scale, they would still fall far short of the expected 2-3.5% increase in annual demand. Without faster yield improvements, the production gap will most likely be filled by clearing more land for extensive increases in crop production, potentially exacerbating climate change processes and the destruction of natural habitat.

Researchers are in a race against rising populations and higher calorie diets to increase grain yields. Fortunately, new technologies have provided unprecedented insights into the genetic sources of crop traits. Large-scale genomics research projects have provided promising data resources, including genotype-phenotype maps and trait-associated markers (Juliana et al, 2019). A meta-analysis of 154 studies of agricultural innovation adoption by Ogundari and Bolarinwa (2018) indicated that recent literature has focused largely on high-yielding crop varieties, suggesting that the issue is not lack of academic attention on crop productivity research, but rather a failure to deploy these productive varieties to the areas where they are most needed.

The gap between highly productive varieties and their implementation in low-income regions is believed to stem largely from financial and legal barriers due to high costs for research, intellectual protection of varieties by corporations, or restrictive regulations regarding seed adoption. Challinor et al. (2016) found that the average time to develop and deliver new varieties of maize in sub-Saharan Africa is 30 years, over three times the time required for higher income areas, mostly due to the lack of technological access and funding gaps. If distribution of improved seeds is left entirely to the profit incentive, subsistence farmers in poor areas will continue to be passed over in favor of helping the already productive regions increase their output, even though the net increase in marginal productivity is far lower.

Scientific advances in the development of new wheat varieties using ever more sophisticated technology may lead to an accelerated increase in yields, but it is difficult to imagine genetic gains alone can achieve the growth rates necessary to keep pace with demographic and agricultural demand forecasts. If crop selection research is to contribute to global food security, it is necessary to invest more in the targeted dissemination of seed varieties to specific regions based on predicted climate models. Atlin, Cairns, and Das (2017) argue that just to keep pace with changing environmental conditions, climate-vulnerable areas such as sub-Saharan Africa will need access to more elite varieties of crop germplasm, they will have to develop seeds modified to their local conditions, and they will have to deploy them to farmers rapidly. The publicly available varieties developed by CIMMYT and CGIAR, as opposed to privately protected genotypes, offer the most promising solutions for low-income countries to combat the negative effects of climate change on crop yields. Connecting researchers, regulators, and agricultural extension officers in these regions will therefore require low-cost, transparent, and easily communicated models justifying why particular varieties need to be adopted in geographically specific locations, both to speed delivery and regulatory approval of new cultivars.

Finally, while the impact of different approaches to agriculture on biodiversity are the subject of debate (Butler et al, 2007; Scherr and McNeely, 2008; Tscharntke et al, 2012), improved seed varieties can be employed in both commercial and smallholder farms, with monocropping or agroecological practices. The process of crop breeding may develop specific varieties better suited to certain fertilizer or management regimes, but the endeavour of selecting genotypes for further development is beneficial for any option. Ultimately, this means that crop breeding research can be employed in any agricultural framework and is thus necessary for future climate adaptation regardless of the farming approach. The potential benefits for climate mitigation are significant; Burney, Davis, and Loebell (2010) estimated that for every dollar invested in agricultural yields since 1961, 68 kilograms of CO_2 emissions were prevented, saving the atmosphere from a total of 3.6 gigatons of CO_2 per year.

The Role of Remotely Sensed Climate Data and Machine Learning

In crop breeding and the selection of useful genotypes, the accuracy of statistical models is paramount. A biased model or one with low predictive power can lead to the inefficient allocation of resources into varieties that do not provide the benefits expected, either due to misleading results or lack of transferability of these results to the intended environments. Lobell (2013) estimates that failure to account for the measurement error intrinsic to many current sources of precipitation data can cause impact calculations to be underestimated by a factor of two, significantly affecting research results that inform funding, policy, and production decisions. Methods of improving predictive accuracy, therefore, are central to the field. Two areas of promising research for increasing the accuracy and efficiency of crop breeding models are the integration of remotely sensed satellite data and advances in supervised machine learning.

Ornella, Kruseman, and Crossa (2019) survey methods and applications of satellite data and supervised learning for drought prediction and mitigation, citing the potential for both technologies to help researchers combat the effects of climate change. They specifically mention the potential of Google Earth Engine (GEE) and its multiple petabytes of satellite images and geospatial data, combined with distributed high-performance computing, as a resource that could alter how crop breeding trials are conducted. While CGIAR has developed and deployed improved climate data resources as well (Ramirez-Villegas et al, 2020), their scope and level of detail are exceeded by several datasets available on Earth Engine. Not to mention, Earth Engine's API allows data to be queried using over 800 different functions for filtering and transforming images prior to their download, saving the user computer processing time and memory storage, neither of which are trivial when there are hundreds of variables and hundreds of thousands of observations to query. Earth Engine data can also be used to power interactive web applications without the need for advanced HTML coding expertise (Gorelick et al, 2017), further improving the transparency and accessibility of climate and agricultural information.

Supervised machine learning methods were specifically developed to solve prediction problems involving complex, high-dimensional datasets. They have been employed by crop breeding researchers for a variety of tasks, most revolving around

genomics for complex trait selection such as rust resistance in wheat (González-Camacho et al, 2018) or physiological and agronomic trait associations in maize (Shekoofa et al, 2014). Ramstein, Jensen, and Buckler (2019) describe the computational challenges of high-dimensional genetic data for crop breeding and the potential of machine learning to help identify causal loci in genes. For a comprehensive overview of machine learning for genomic selection in crop breeding, the author recommends Ornella, Cervigni, & Tapia's (2013) chapter in Crop Stress and Its Management: Perspectives and Strategies. Esposito et al (2020) also provide a summary of more recent advances in machine learning applications for genomics. The objective of this project is not necessarily to improve upon any of these models--the use of pedigree information rather than full genomic data makes that all but impossible--but rather to demonstrate their applicability to crop yield modelling in a simplified genotype-by-environment framework and test if machine learning can provide more accurate or more efficient alternatives to traditional mixed models in the prediction of optimal genotype varieties for new locations.

Research Questions and Hypotheses

This project focuses on assessing the potential of novel data sources, predictive models, and open-source programming tools for crop breeding research. The primary research question is whether remotely sensed environmental data can affect the accuracy of wheat yield predictions, particularly for novel environments without prior experimental data. Another line of inquiry is whether non-parametric machine learning methods can compare with a standard genotype-by-environment interaction model using a Bayesian generalized linear regression. An ancillary field of interest is whether all the necessary data manipulation and analysis can be conducted in an open-source, fully reproducible programming environment.

Efficacy of Remote Sensing Data for Crop Selection

Crop selection research typically focuses on two broad categories of prediction questions: choosing the most promising genotypes or genes for further trials and estimating which existing varieties will perform best in each location. The first question has benefited over the past decades from significant advances in gene sequencing, big data methods, and simulation techniques. These improvements in gene selection are expanding several important frontiers, from a focus on resilience to disease or extreme weather as well as more nutrient rich food crops. The optimal application of new varieties on actual farms, however, depends on the prediction of crop performance in novel locations and realistic field conditions. From a methodological perspective, this is similar to an out-of-sample prediction, which is fundamentally a question of the external validity of a particular variety as a "treatment" intervention. The extensibility of crop varieties to new locations depends on the interaction of the genotype with its environmental factors, often sub-divided into endogenous management decisions and exogenous climatic conditions. The quality of exogenous environmental data, therefore, is essential to accurate prediction modelling. The availability of comparable environmental data between trial locations and commercial agricultural fields is also

critical for extending recommendations on optimal varieties from test sites to actual farms.

This project investigates whether the out-of-sample predictive ability of typical genotype-by-environment interaction models can be improved with the inclusion of remotely sensed environmental data. We first compare the environmental data provided by CIMMYT in its wheat trials with similar variables from Google Earth Engine, by simple descriptive statistics and correlations. The quality of each environmental data set is then compared, with an attention to missing data problems, internal consistency, and level of spatial and temporal detail. Each dataset is then used as a feature set for a battery of predictive models, including standard Bayesian generalized linear regressions with mixed effects and several common machine learning approaches. The models are compared on a holdout set designed to simulate externally valid predictions to approximate their utility in a decision framework for regulators and agricultural extension officers.

The null hypothesis of this test would be if predictive accuracy for all models is statistically equivalent between the CIMMYT-provided and Earth Engine datasets, as measured by mean squared error and r-squared values. There is theoretical uncertainty between which alternative hypothesis is more likely; the CIMMYT data should have a higher predictive capacity given its greater spatial resolution and specificity, but the Earth Engine data can capture greater temporal detail. Beyond statistical accuracy, however, remotely sensed data offers three key advantages to manually collected variables: it is much more cost efficient to collect, rarely suffers from measurement bias or missing data issues, and can be more readily extended to locations outside of the experimental site.

Applicability of Machine Learning Models to GxE Predictions

The second set of research questions examined by this project focus on the statistical model used to make predictions from the data. A common approach in genotype-by-environment interaction literature has been the linear mixed model (LMM), estimated using a Bayesian generalized linear regression (BGLR). LMMs have proven robust to high dimensional datasets, often incorporating thousands of genotype varieties, environmental variables, and location effects in complex interactions. Their use of Bayesian approaches to estimate the distributions of random effects in these interactions approximates the complexity of agricultural processes much better than a purely fixed effects model. Calculating these complex models is computationally intensive, however, requiring the creation of several intermediate variance-covariance matrices which can reach hundreds of gigabytes even for relatively simple sets of pedigree data. These limitations make it difficult to replicate existing crop breeding models outside of well-resourced research centers--a dilemma both for the credibility of the field and the applicability of its models for agricultural stakeholders, such as government regulators or development institutions.

An alternative approach to modelling genotype-by-environment interactions has made use of machine learning frameworks. Non-parametric models such as random forests, gradient boosted decision trees, and neural networks offer similar theoretical

advantages as an LMM in comparison to linear regression, in that they can incorporate spatially or temporally dependent observations in hierarchical experimental structures without violating statistical assumptions and biasing predictions (Grinberg et al, 2020). Machine learning approaches also offer three key advantages over a BGLR: they are commonly available in several open-source programming environments, they are often less computationally intensive while providing more flexible hyperparameters, and they are being developed more intensively than methods specific to genomic selection research (Heslot et al, 2012; Raschka and Mirjalili, 2019). Advances in the interpretability of machine learning models, an often-cited limitation, has also made them a more attractive alternative in recent years (Ribeiro et al, 2016; Rudin, 2019).

This project investigates three common machine learning methods for the prediction of continuous outcomes: random forest regression, gradient boosted regression trees, and artificial neural networks. All three are readily available functions in the Python package Scikit-Learn, which provides a robust set of supporting methods for cross-validation, feature engineering, feature selection, and model evaluation. The BGLR package in R is used to implement the comparison LMM, adjusted to improve computational efficiency using the OpenBLAS system. The prediction results for each method will be calculated on a holdout set of 15% of the available data and assessed according to their root mean squared error (RMSE) and R-squared values.

The second null hypothesis of this study would be finding no statistical difference between the predictive accuracy of the LMM regression and the machine learning models. Given the historical preference for LMM in the crop breeding literature and their closer theoretical approximation of genotype-by-environment interactions, we expect them to outperform the more general machine learning models, though perhaps with higher computational requirements.

If the null hypothesis is refuted, however, and at least one of the machine learning models outperforms the LMM, it may provide motivation for further research into the assumptions of the LMM. Since BGLR approaches require a prior distribution for the variance of their random effects, it is possible that previous methods imposed too strict of assumptions on the normality or independence of their model terms. Previous approaches have also frequently incorporated the main effect of locations in their models, a term only available for fields which have previously conducted yield trials. For out-of-sample predictions, statistical models can only make use of genotype effects, environmental covariates, and their interactions, providing a theoretical advantage to machine learning methods which do not require random location effects to ensure unbiased residuals.

Accessibility of Crop Selection Modelling using Open-Source Platforms

The final line of inquiry pursued by this project focuses on the availability of data and capacity for replication of crop selection models. Recent trends in open-science have pushed research to adopt more replicable methods, using open-source statistical packages like R or Python, publishing data sets and coding scripts online, and relying only on commonly available computing resources. Crop breeding and genomic selection are known as computationally intensive fields which rely on specific technological

resources, not all of which are openly available (Wu et al, 2011). Genetic information is at the forefront of "big data" techniques which promise significant advances in crop breeding, but potentially at the cost of replicability. Given ethical arguments surrounding the public nature of genetic information--that crop breeding represents an international common good, especially when funded and conducted by organizations like CGIAR and the FAO (Gardner and Lesser, 2003; Sally, 2011)--there is sufficient motivation to ask whether a shift towards open-science in crop selection is possible, and if so, whether it involves any trade-offs in terms of predictive accuracy.

This project therefore seeks to replicate common genotype-by-environment interaction models using only publicly available data and methods. The only data which was not openly accessible online was the genotype pedigree matrix, though CGIAR has committed to an open-data policy (CGIAR, "Open Access and Open Data"). All programming was conducted in either R or Python using a Google Colab or JupyterLab interface, which provides server-side computational resources and removes the burden of processing from users. The code itself was written to be as comprehensible as possible even for non-technical users and will be distributed openly online through a GitHub repository. A detailed readme specifying the steps for complete replication will also be included.

While there is no statistical measure of reproducibility, if the project can successfully replicate existing models given the above limitations and provide scientifically useful results, it seems reasonable that other research projects can follow suit and adopt an open-science framework. Furthermore, given that the project itself was conducted over the course of just three months, with limited support from technical specialists, there is reason to believe that better resourced institutions can improve their reproducibility without incurring significant costs in adjusting workflows or retraining researchers.

The potential benefits of an open-science approach to crop selection are difficult to assess. It seems reasonable, however, that a project which can be easily replicated or adapted by non-academic stakeholders may lead to improved communication and novel applications, thus helping the research impact a larger set of farmers, particularly in less resourced regions.

Data

Following the Open Data and Open Access policy of CGIAR, this project seeks to use only publicly available data sources. The three primary data sources include: 109 CIMMYT Wheat Yield Trial results, 4 Google Earth Engine datasets, and a coefficient of parentage matrix for 3,443 wheat varieties from the International Crop Information System (ICIS). Both CIMMYT and Google have made their data freely available online, though CIMMYT's dataverse is still organized as a series of separate downloads through their website, as opposed to Earth Engine's convenient access through a Python-enabled API. ICIS seems to have a publicly accessible client for downloading crop genotype data, but the supporting web infrastructure is defunct. The data was instead requested directly from Dr. Jose Crossa, a scientist at CIMMYT.

CIMMYT Wheat Trial Yield, Environment, and Location Data

The wheat yield trials used in this study were aggregated from four long-running CIMMYT projects: the Elite Selection Wheat Yield Trials (iterations 1-39), the Semi-Arid Wheat Yield Trials (iterations 1-26), the High Rainfall Wheat Yield Trials (iterations 1-26), and the High Temperature Wheat Yield Trials (iterations 1-17). Each trial contains information on the grain yield per hectare, genotype, location, and environmental conditions for every field plot, including data on local checks--control varieties sourced from nearby farms--which allows for a causally valid estimate of the potential effects of new wheat varieties and their interaction with environmental factors over the entire growing cycle. The locations, years, varieties, and environmental conditions for each iteration of each trial type differ significantly. CIMMYT has used several data structures and variable naming conventions over time, though there is enough similarity to allow for comparisons over the entire 40-year data span.

After accounting for the differences in file formats and input structure, the aggregated dataset consists of 189,418 unique plots of wheat in 2,759 locations from 1980 through 2019. The combined data included 130 environmental features, ranging from extrinsic factors like total precipitation per month to intrinsic management decisions like fertilizer application. Many of these environment covariates were not common to all trials, however, resulting in a significant missing data problem. 80 environmental factors contain values for less than 1% of the aggregate wheat plots across all trials. After removing rows with missing data in key variables such as grain yield, harvest finishing date, and genotype, the processed dataset consisted of 169,529 observations.

Table 2: CIMMYT Wheat Yield Trial Summary Statistics

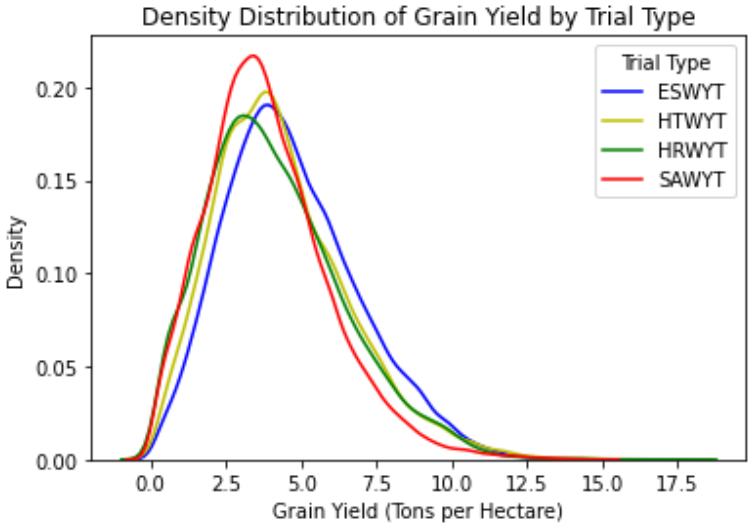
Trial Type	Number of Trials	Total Fields (N)	Grain Yield (Tons per Hectare)			Unique Varieties	Unique Locations
			Mean	Std. Dev.	Range		
ESWYT	39	70,365	4.697	2.216	15.423	1082	1530
HRWYT	26	27,694	4.151	2.273	17.823	997	493
HTWYT	17	24,825	4.357	2.228	16.141	536	441
SAWYT	26	46,645	3.886	2.044	14.844	942	853
Full Data	108	169,529	4.345	2.207	17.823	3443	2759

(Source: Aggregated CIMMYT Wheat Trial Data; Calculations by Author)

The dependent or target variable of the predictive models throughout this project is the grain yield of each field plot, measured in tons per hectare. Each of the four CIMMYT trial types demonstrated a similar distribution of grain yields, centered around 4.345 tons per hectare, though the ESWYT and High-Temperature Wheat Yield Trials

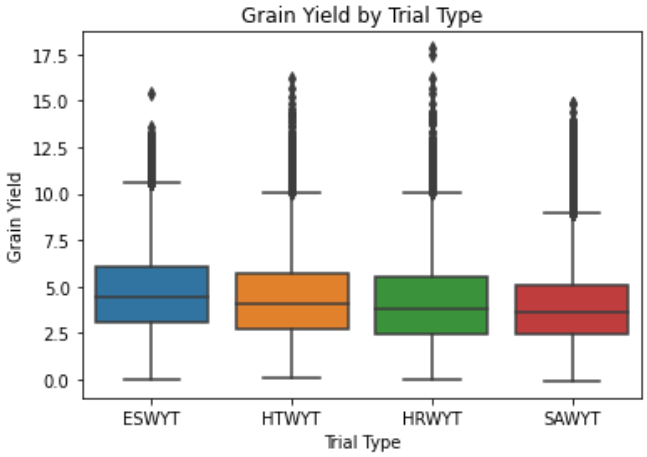
(HTWYT) trials had a slightly higher yield on average. Yield varied from 0 to 17.82 tons per hectare, with an average standard deviation of 2.21, as described in Table 1. The density and box plots below indicate the distributions and central tendencies of yield values per trial.

Figure 4: Density Distribution of Grain Yield by Trial Type



(Source: Aggregated CIMMYT Wheat Trial Data; Calculations by Author)

Figure 5: Boxplot of Grain Yield by Trial Type

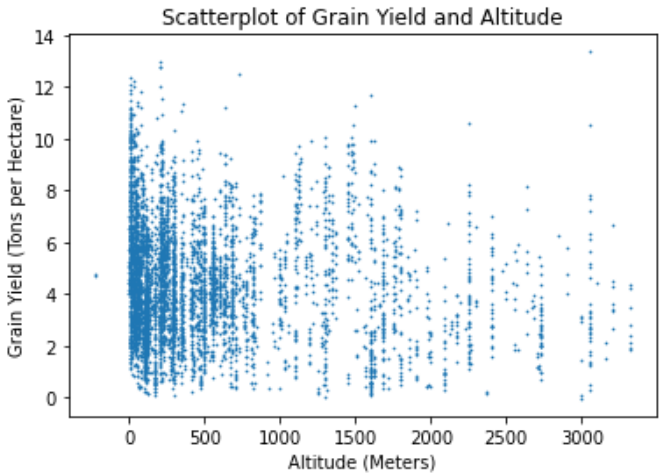


(Source: Aggregated CIMMYT Wheat Trial Data; Calculations by Author)

Thirty-four environmental variables from the aggregated data were selected for inclusion in further modelling, 20 of which were continuous numeric and 14 were categorical or binary. These 34 were chosen for their relative completeness and theoretical relevance; a partial list of the 40 most relevant variables in the CIMMYT data can be found in Appendix 5, and a correlation matrix describing the variables can be found in Figure 9. No independent variables in the CIMMYT environmental data

displayed any clear linear relationship with grain yield, indicating that a simple parametric estimation of yield values through direct linear regression would likely not be highly predictive. Altitude, selected because it is a stable variable with little room for measurement error, was plotted against grain yield below for a random subsample of the CIMMYT observations, to demonstrate the dispersion of the data.

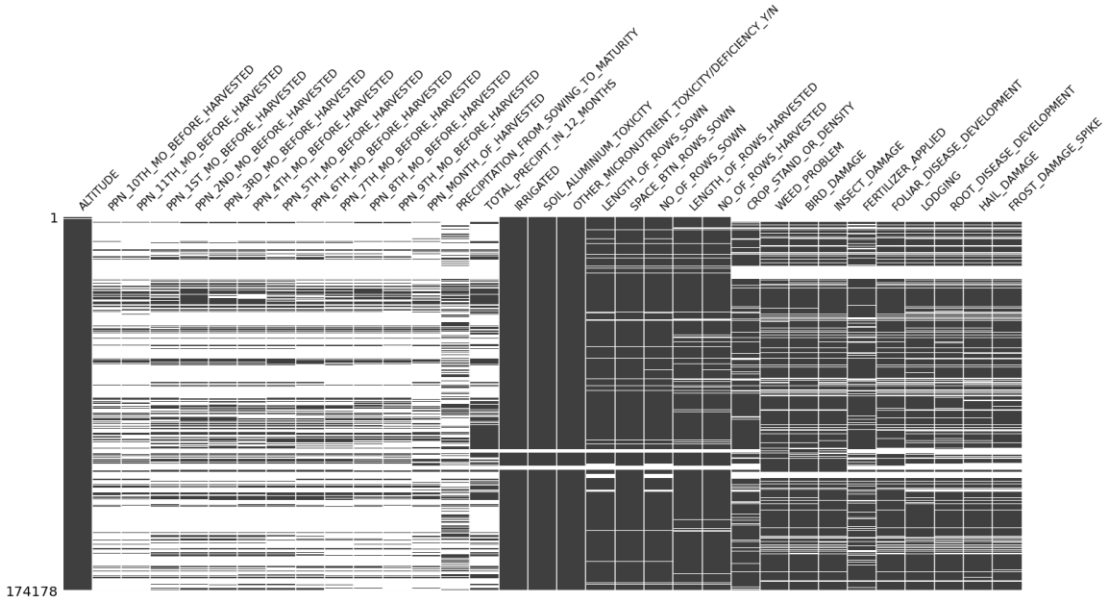
Figure 6: Scatterplot of Grain Yield per Hectare and Altitude



(Source: Aggregated CIMMYT Wheat Trial Data; Calculations by Author)

The primary drawback of the CIMMYT environmental data is the proportion of trial observations with incomplete information. Missing data in the aggregated data set may be the result of missed measurements or shifting covariate sets across different studies; in either case, the result is an incomplete set of predictors. The matrix graph below indicates the degree of missingness in the CIMMYT data, with each column representing an environmental variable and each row an observation. The white spaces indicate missing data, while gray stands for complete information. Some variables like altitude or irrigation status have relatively little missing data (0.1% and 1%, respectively) while others, like most precipitation measures, are barely represented, with completeness rates ranging from 20% to 40%. There are also several overlapping precipitation measures, such as total precipitation in the twelve months preceding harvest as well as precipitation from sowing to maturity. In many trials both variables seem to have been recorded simultaneously, indicating that they are not necessarily substitutes.

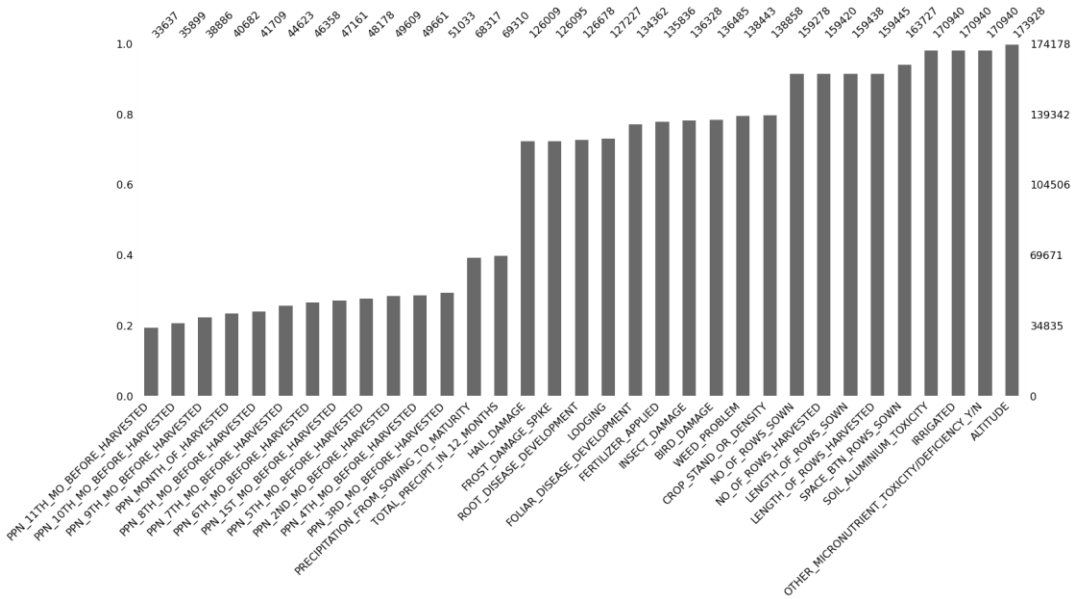
Figure 7: Missing Data Matrix of CIMMYT Environmental Variables



(Source: Aggregated CIMMYT Wheat Trial Data; Calculations by Author)

Figure 7 demonstrates another method of visualizing the missingness of the CIMMYT data, with each variable represented as a bar plot. The y-axis is the total number of observations containing data, so again white space represents missing information. Most CIMMYT variables of interest contain data for less than 80% of observations; most precipitation measures have less than 30% completeness.

Figure 8: Bar Graph of Missing Data by Variable for CIMMYT Trials



(Source: Aggregated CIMMYT Wheat Trial Data; Calculations by Author)

Another drawback of the aggregated wheat trial data, as it is available on the public dataverse, is its lack of genotype identification values. The CIMMYT trials contained information on the cross identification (CID) and selection identification (SID) numbers, which together can uniquely identify a genotype variety, but not the genotype identification (GID) which is necessary for obtaining pedigree information from ICIS. Without a GID, it is impossible to link the wheat trial data to genotypic information, whether a pedigree or genome, which is essential to predicting which variety will be optimal under different conditions. Fortunately, the CIMMYT wheat germplasm bank (http://wgb.cimmyt.org/gringlobal/search.aspx) does have complete records of CIDs, SIDs, and GIDs for all varieties. A simple web-scraping algorithm was able to recover these paired values so that the CIMMYT wheat trial data was able to be genotyped properly. Not all potential users of CIMMYTs data should be presumed to have web-scraping knowledge, however, so the lack of GID values in the original dataverse data sets represents a significant limitation for accessibility and reproducibility.

Google Earth Engine Environmental Data

Google Earth Engine hosts over 20 petabytes of remote sensing data in its catalog, including constantly updating sets of satellite imagery from the LANDSAT, MODIS, and SENTINEL missions. This project made use of 27 variables drawn from 4 datasets, detailed in Table 3 below, mostly consisting of post-processed raster data such as the Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System. Each dataset spans the entire Earth, though their temporal range, spatial resolution, and update frequency differ. The datasets and variables were selected based on their theoretical relevance to agriculture and comparability to the CIMMYT environmental covariates.

Table 3: Google Earth Engine Raster Datasets and Variables

Spatial Resolution	Dataset Name	Date range	Example Variables
30 meters (0.016 arc minutes)	NASA SRTM Digital Elevation Data	2000	• Altitude
2.5 arc minutes (4.63km)	TerraClimate: Monthly Climate and Climatic Water Balance for Global Terrestrial Surfaces, University of Idaho	1958 to present	• Actual evapotranspiration • Climate water deficit • Palmer Drought Severity Index • Precipitation accumulation • Soil moisture • Vapor pressure
0.1 arc degrees (6 arc	FLDAS: Famine Early Warning Systems Network	1982 - present	• Evapotranspiration • Surface pressure • Total precipitation rate • Snowfall rate

minutes, 11.1km)	(FEWS NET) Land Data Assimilation System		• Baseflow-groundwater runoff • Near-surface air temperature • Soil heat flux • Soil temperature
0.25 arc degrees (15 arc minutes, 27.78km)	ERA5 Daily aggregates - Latest climate reanalysis produced by ECMWF / Copernicus Climate Change Service	1979 - present	• Maximum air temperature at 2m • Average air temperature at 2m • Minimum air temperature at 2m • Dewpoint temperature at 2m • Total precipitation • Surface pressure

The NASA Shuttle Radar Topography Mission (SRTM) data for elevation was collected during an 11-day mission of the Space Shuttle Endeavor in February of the year 2000 (Farr et al, 2007). The shuttle was equipped with a synthetic aperture radar system that collected the elevation of tile images of the Earth's surface at a 30-meter pixel resolution. The data is only available from 2000 but is assumed to remain relevant for all years considering the slow rate of altitude changes for any given agricultural plot. The data was released to the public in 2015, providing a tenfold improvement in spatial resolution for elevation data compared to previous sources.

The TerraClimate dataset, produced by Abatzoglou et al (2017), provides monthly climate aggregates for a range of environmental variables at the 2.5 arc minute spatial resolution. The data is available from 1958 to the present, as GEE consistently updates its datasets as new information is released. TerraClimate combines WorldClim data from the Food and Agriculture Organization (FAO) of the United Nations with the Climatic Research Unit gridded Time Series version 4 (CRU Ts4.0) (Harris et al, 2020) and Japanese 55-Year Reanalysis (JRA55) (Ebita et al, 2011) using climatically aided interpolation.

The Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System (FLDAS) provides detailed environmental data at the 6-arc minute resolution, available from 1982 to the present. The FLDAS data was designed by McNally et al (2017) to help estimate agricultural production and predict famine around the world, making it particularly relevant to crop breeding applications. The data incorporates the Noah version 3.6.1 surface model with CHIRPS-6 hourly rainfall, downscaled using the NASA Land Surface Data Toolkit.

The European Centre for Medium-Range Weather Forecasts (ECMRWF) Atmospheric Reanalysis Generation 5 (ERA5) assimilates data from weather centers around the world to produce climatic estimates at the 15-arc minute resolution, available from 1979 to the present. The data collection and analysis were conducted by the European Union's Copernicus program and was published in 2018.

The datasets were queried using the GEE Python API, which can specify the coordinate points and date ranges of interest for each observation and aggregate data in time and space according to various functions. When possible, the variables were collected for each of the 13 months preceding the harvest finishing date of the associated

field plot, then aggregated in time by collapsing all data points within the month range through either a sum total (for variables which are considered agriculturally relevant in their aggregate stock value, such as precipitation) or an average (for variables which affect crop growth as a flow, such as temperature). The resulting raster images for each location in each month were then spatially reduced to a single average scalar value for all pixels. Google Earth Engine's dynamic scaling feature allowed the same query function to be passed to each dataset despite the varying spatial resolution of the raster data. The spatial radius used for cropping the raster image values around the coordinate point and later spatially reducing the pixels was specified as a parameter in the function, allowing for flexible comparison of datasets with different initial spatial resolutions.

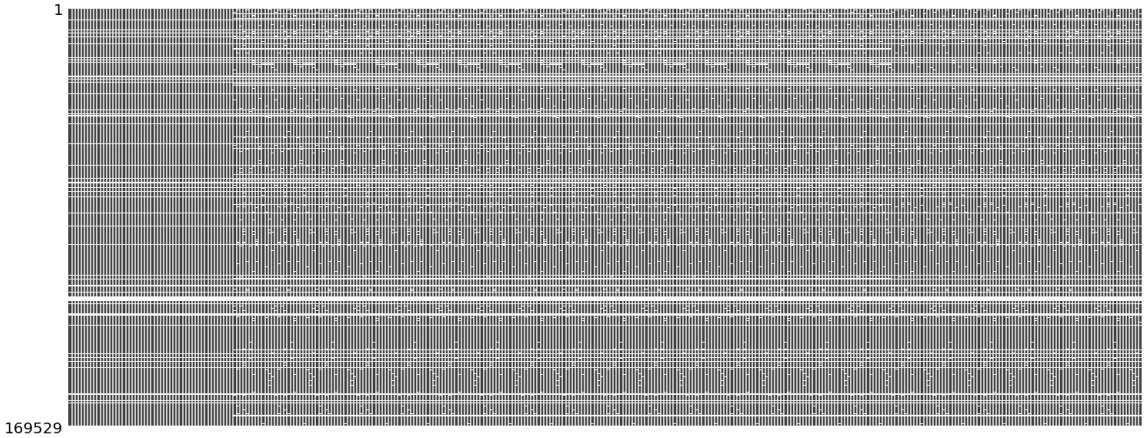
The query process resulted in a series of 351 columns, representing each variable-month combination for every field plot in the CIMMYT data. There is some overlap in the variables between datasets for common measures, such as evapotranspiration or soil temperature, though there is enough variation in the collection methodology and resolution of the data that there is little danger of collinearity. The Earth Engine data thus represents an environmental covariate source that is orders of magnitude more detailed than the factors included in the original CIMMYT data, which contained fewer than 50 relevant variables. The Earth Engine data does not include managerial factors such as irrigation or fertilization, though these could be feasibly combined with the environmental variables.

The key advantage presented by remotely sensed data over manually collected environmental information, such as that included in the CIMMYT data, is its completeness. There was some missing data, due to cloud masking or observations outside the date-range of a dataset, but the GEE environmental data only had 3.5% of data values missing, compared to 37% missing for the CIMMYT data. Remotely sensed data is also available for novel locations outside of the CIMMYT study areas, making it more appropriate for out-of-sample predictions which try to capture the extensibility of the effects of improved wheat varieties.

The disadvantage of remote sensing data is of course its relatively lower spatial resolution compared to on-farm sensors, though the degree to which such a generalization over space affects prediction accuracies for large sets of agriculture data is, yet, unclear. This spatial inaccuracy is compounded by the level of detail in the CIMMYT wheat trial locations; geographic coordinates were provided only at the arc minute level--just two decimal points of specificity after the degree, which means the actual location could be anywhere in an approximately 1.85km radius (at the equator). The lack of geographic specificity can obviously introduce measurement error into the variables, compromising their relevance to prediction.

The missing data matrix plot below follows the same format as above; each variable represents a column along the horizontal axis and each observation a row along the vertical axis. Missing data occurs only rarely, due to a harvest season out of the temporal range of the variable or the application of a cloud or snow mask for data quality. Some rows seem to be missing for all variables, indicating that there may be undetected errors in the location or date range of those values.

Figure 9: Missing Data Matrix of Google Earth Engine Data



(Source: Google Earth Engine Datasets; Calculations by Author)

ICIS Coefficient of Parentage Matrix

The final data input for the project is a matrix of genotype relatedness for the 3,495 CIMMYT wheat varieties contained in the initial trial data. The matrix contains the coefficient of parentage (COP) values calculated between each variety, measuring the associations between genotypes based on their pedigree. The structure of the data is essentially a variance-covariance matrix; each row and each represents one genotype variety, so that the diagonal of the matrix measures the internal COP for each wheat type and the remaining values provide information on how similar varieties are to each other. This structure is commonly referred to in genotype-by-environmental modelling as the additive relationship matrix, or A-matrix, and is used to estimate the variance of the effects for each line.

The COP matrix was obtained from the International Crop Information System (ICIS), which declares its data to be openly available to any interested scientist. The ICIS website (https://cropforge.github.io/iciswiki/index.html) contains links to a download center and data source, including instructions for the process of querying crop genotype data through a file transfer protocol, though neither the links nor the transfer process seem to function. The data was instead requested directly from a scientist at CIMMYT, Dr. Jose Crossa, who provided the tabular file for the COP matrix. The ICIS data request system, it should be noted, organized genotypes by their genotype identification number (GID), while the wheat yield trials code only the CID and SID numbers. In order to match the wheat yield data to the coefficient of parentage matrix, the corresponding GID values for each CID-SID pair had to be retrieved from the CIMMYT wheat germplasm databank through a web-scraping algorithm.

Methodology

Following the acquisition of the CIMMYT yield trial data, Earth Engine variables, and pedigree matrix from ICIS, the methodology for this project can be divided into three major steps: processing, analysis, and prediction. The data processing followed a standard data science workflow, including cleaning, imputing, and scaling the data before splitting it along a cross-validation scheme. The analysis phase can be subdivided into linear mixed model and machine learning approaches; the former utilized the Bayesian generalized linear regression (BGLR) package in R while the latter was conducted primarily using Scikit Learn and XGBoost in Python. The best parameter configurations for each model were then used to estimate grain yields for a holdout set of wheat trials containing at least 4 observations from each genotype variety, to approximate the decision space of agricultural stakeholders when selecting which seeds to distribute to what locations.

Data Processing

In order to make a proper comparison of the various statistical models and datasets being investigated, the data had to first be properly cleaned and standardized. While there are several well-researched trade-offs involved in decisions on whether to remove observations due to missingness, to impute the remaining data, and to standardize data using a scaler, this project sought to follow standard machine learning approaches to improve compatibility and model convergence. Each version of the processed data was then split into training sets, test sets, and holdout sets for assessing the ultimate model accuracy. Exploratory data analysis (EDA) was then employed to compare relevant variables between the CIMMYT and GEE datasets.

Cleaning, Imputing and Scaling the Data

The process of cleaning and imputing revealed several key differences between the CIMMYT and Earth Engine datasets, highlighted here both to explicate the full research methodology and demonstrate the potential advantages and disadvantages of each environmental data source. The degree of missingness in the CIMMYT environmental factors, documented in the data section, required a significant reduction in the number of observations to minimize missing data bias: from 189,697 rows in the initial aggregated data to 169,529 rows in the full processed dataset. Observations were dropped if they were missing values for grain yield, genotype, or if all relevant date variables were missing. While it is impossible to test with certainty whether these values were missing at random, the observations that were dropped were compared with the remaining sample to compare their averages and standard deviations for their altitude, latitude, and longitude. A paired t-test was then performed to test the statistical significance of the difference in distribution between the observations which were dropped and those remaining. The observations with missing values for grain yield had a higher altitude than the non-missing values and were closer to the equator and prime meridian. These differences were highly statistically significant, indicating there may be a pattern to the missingness of grain yields in the CIMMYT data. Fortunately, only 3,059 observations were missing the target variable, just 1.6% of the initial dataset.

The observations missing a harvest finishing date were more numerous: 34,648 rows (18.3% of the original CIMMYT data). The difference in mean altitude between the missing and non-missing observations was not statistically significant but their locations were, though not to the same degree as grain yields. Fortunately, around 17,539 observations missing a value for their harvest finishing date still had a value for either their harvest start date or sowing date. By taking the average duration of the period between harvest start and finish as well as sowing and harvest finish for the remaining observations, an estimated harvest finish date was imputed for those fields. The remaining 17,109 observations, for which there were no dates reported, were dropped from the data, leaving 169,529 observations for the remaining analysis.

Table 4: Statistical Testing for the Randomness of Missing Yield and Harvest Date Observations in the CIMMYT Wheat Yield Trial Data

Variable	Missing (%)	Altitude			Latitude			Longitude		
		Present Mean (Std Dev)	Missing Mean (Std Dev)	T-Test Stat (P-Value)	Present Mean (Std Dev)	Missing Mean (Std Dev)	T-Test Stat (P-Value)	Present Mean (Std Dev)	Missing Mean (Std Dev)	T-Test Stat (P-Value)
Grain Yield	3,059 (1.6%)	526.2 (658.4)	692.9 (895.9)	-13.79 (0) ***	24.53 (61.99)	0.59 (65.97)	21.16 (0) ***	20.61 (23.15)	7.24 (31.62)	31.47 (0) ***
Harvest Finish Date	34,648 (18.3%)	528.91(661.9)	529.21(669.2)	-0.07 (0.9)	20.18 (23.66)	21.34 (22.04)	-8.33 (0) ***	25.08 (62.14)	20.05 (61.97)	13.60 (0) ***

(Source: Aggregated CIMMYT Wheat Trial Data; Calculations by Author)

After removing observations with missing data in grain yield and harvest finishing date, the remaining CIMMYT environmental data still had significant missing portions, as described in the data section above. To avoid losing further samples, these values were instead estimated using an iterative random forest imputation, implemented through the missingpy library in Python. The MissForest function predicts values for missing observations based on the remaining feature set (excluding grain yield), starting with the variable containing the fewest missing values. The function proceeds through each column containing missing values, utilizing either a random forest regressor in the case of continuous numeric data or a random forest classifier for categorical data, to predict the most likely value for the missing observation. The algorithm iterates this process multiple times, utilizing the non-parametric and random nature of the imputation function to make multiple estimates for each missing value and comparing their rate of change to determine its stopping criterion.

After cleaning and imputing the dataset, the remaining numerical environmental variables as well as grain yield were standardized column-wise to ensure their comparability. Most machine learning methods require some form of standardization to ensure balanced variable weights, as columns with drastically different units are difficult to compare. CIMMYT precipitation data, for example, was measured in millimeters while Earth Engine's ERA5 Climate data uses meters for the same measure. Rather than try to convert all units for every variable--a time intensive and difficult to replicate

process--the data was simply standardized by its mean and standard deviation. Not only does this process help improve the comparability of the datasets, it decreases the time required for later calculations to converge to an optimal parameter by reducing the potential range of values to search through. Factor variables were left as binary dummies to preserve their potential interaction effects.

Cross-Validation Framework

Cross validation is a common technique to approximate the accuracy of a predictive model for observations in the population outside of the data sample. By withholding a subset of the data from the initial inputs used to train the data, cross validation allows a model to be assessed against a novel set of cases. As this project's research question of interest centers around the prediction of grain yields for a location or time frame outside of the scope of available training data, the cross-validation framework attempts to simulate such a situation by withholding 15% of the grain yield observations as a validation holdout set. From the remaining model data, another 15% (12.75% of the initial processed data) is sampled as a test set for assessing the regularization and optimization of hyper-parameters. The remaining 72.25% of observations are the training data, split further into five folds for cross-validation of the machine learning models, which tunes the hyperparameters on available data in sequential order in order to produce an optimized predictive model. The percentage values for these splits were chosen arbitrarily based on typical ranges for validation data; further adjustments with optimal split sizes could be an area of future research.

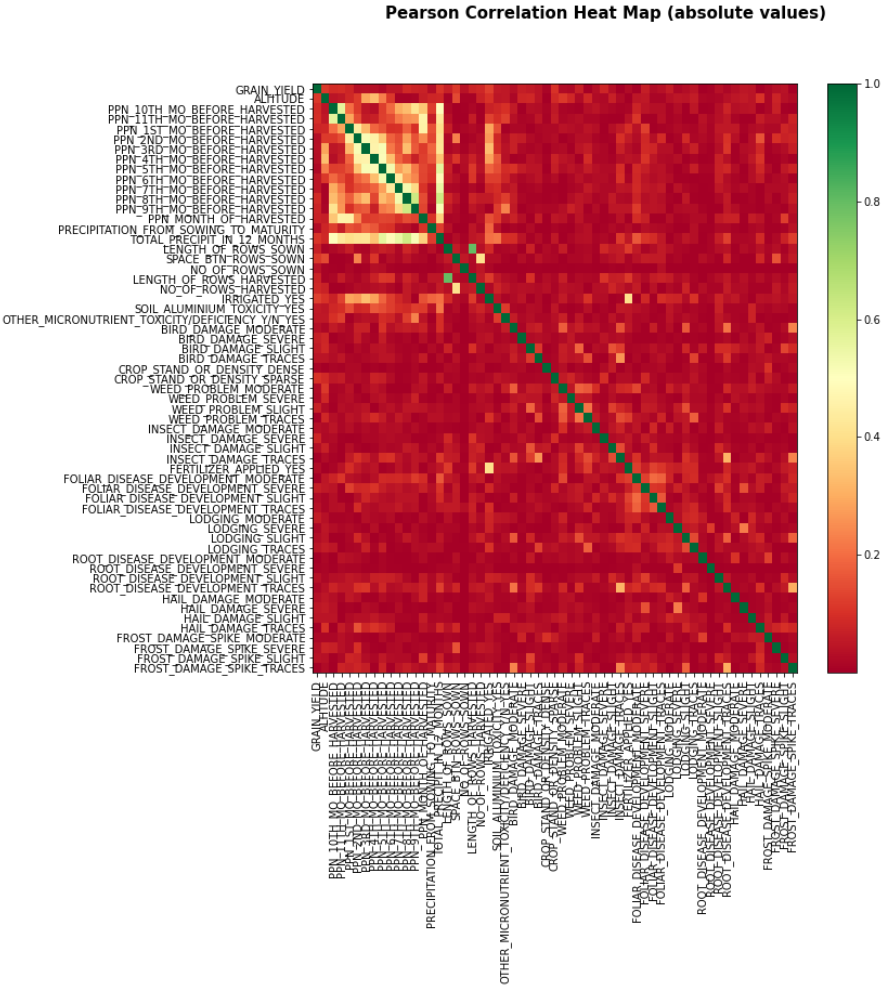
Table 5: Cross-Validation Scheme for Model Fitting and Evaluation

<table>
<tr><td colspan="3">Parameter and Model Selection Cross-Validation Scheme</td><td>N per Fold</td></tr>
<tr><td rowspan="6">Model Selection 7-Folds
(144,100)</td><td rowspan="5">Regularization Tuning 6-Folds
(122,485)</td><td>Training Fold 1 (14.45%)
(Calculates Coefficients, Tests Initial Hyper-Parameters)</td><td>24,497</td></tr>
<tr><td>Training Fold 2 (14.45%)
(Calculates Coefficients, Tests Initial Hyper-Parameters)</td><td>24,497</td></tr>
<tr><td>Training Fold 3 (14.45%)
(Calculates Coefficients, Tests Initial Hyper-Parameters)</td><td>24,497</td></tr>
<tr><td>Training Fold 4 (14.45%)
(Calculates Coefficients, Tests Initial Hyper-Parameters)</td><td>24,497</td></tr>
<tr><td>Training Fold 5 (14.45%)
(Calculates Coefficients, Tests Initial Hyper-Parameters)</td><td>24,497</td></tr>
<tr><td colspan="2">Regularization Test Set (12.75%)
(Test Set for Selecting Final Hyper-Parameters)
(Uses novel locations to simulate out-of-sample predictions)</td><td>21,615</td></tr>
<tr><td colspan="3">Model Comparison Holdout Test Set (15%)
(Tests Final Accuracy of Each Feature and Hyper-Parameter Optimized Model)
(Use novel locations and most recent data to simulate near-term forecasting)</td><td>25,429</td></tr>
</table>

Exploratory Data Analysis

As an initial assessment of the comparability of the Earth Engine data with the environmental data provided by CIMMYT, exploratory data analysis was employed to visualize relevant variables. First, the author investigated the correlations between grain yield and the numeric environmental variables provided by CIMMYT, to determine which variables might be most relevant for prediction. The correlation heatmap below plots each variable on both the x and y axes, with warmer colors indicating a low correlation and coloring closer to green indicating a high correlation. The diagonal demonstrates perfect correlation, since the variables are measured against themselves. Grain yield is on the left-most and top-most position and does not seem linearly correlated with any variables. Precipitation by month variables are correlated with each other and total precipitation, as would be expected.

Figure 10: Correlation Heatmap of CIMMYT Environmental Variables

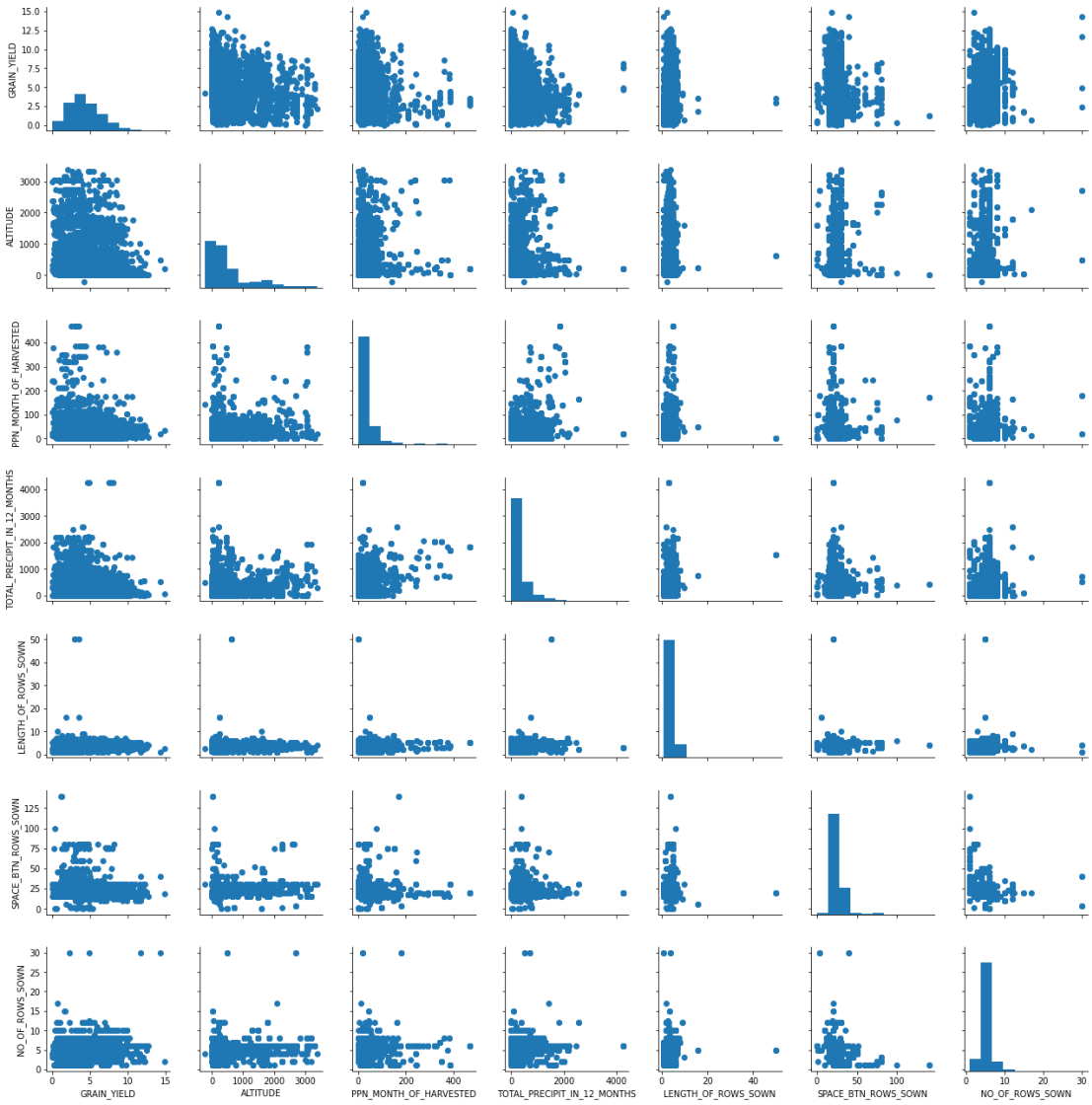


(Source: Aggregated CIMMYT Wheat Trial Data; Calculations by Author)

Six environmental variables with theoretical linkages to grain yield were inspected by scatterplot for further investigation: altitude, precipitation in the month of harvest, total precipitation in the twelve months prior to harvest, length of rows sown,

space between rows sown, and number of rows sown per field. Altitude and precipitation seem to have some negative relationship with grain yield, but the correlation is not strong. Some of the additional plots in the pairwise grid below indicate potential outliers, particularly in the sowing dimensions.

Figure 11: Scatterplots of Select CIMMYT Environmental Variables with Grain Yield

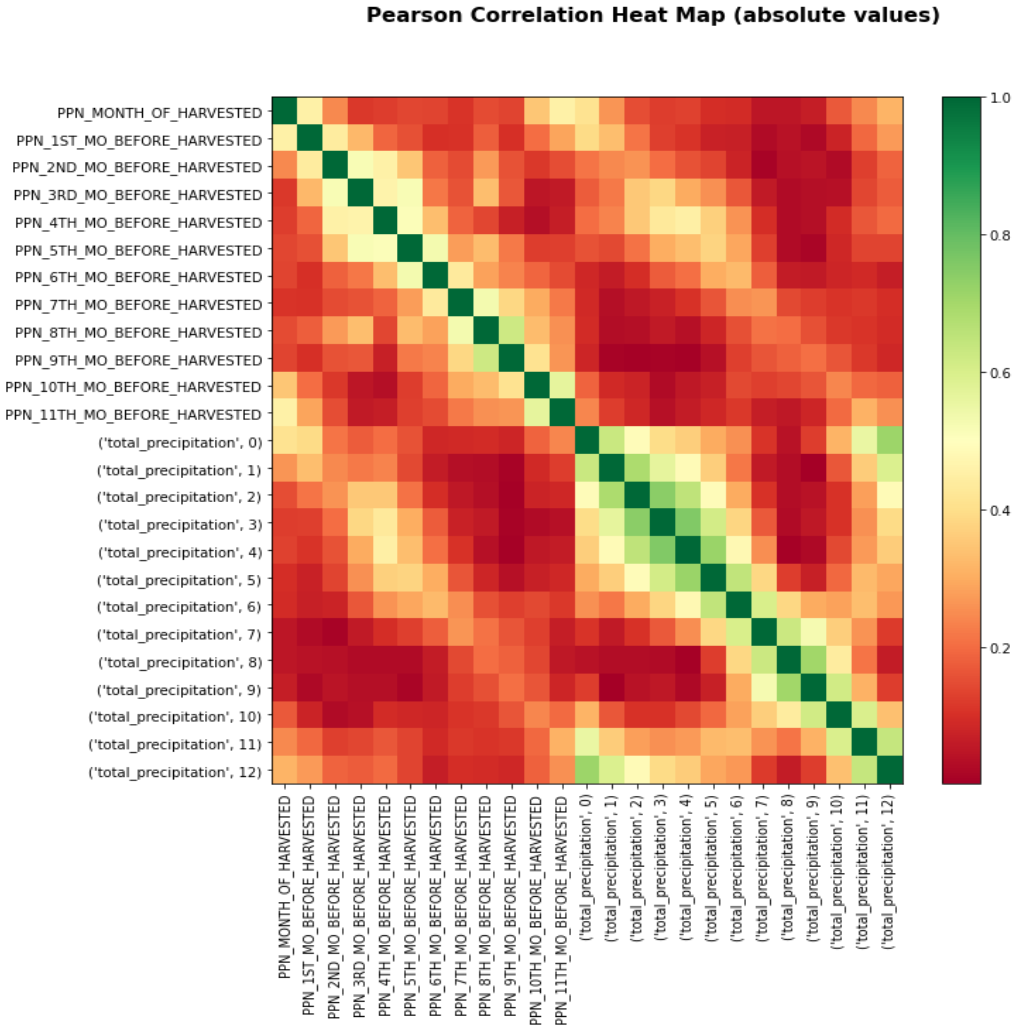


(Source: Aggregated CIMMYT Wheat Trial Data; Calculations by Author)

The correlation heatmap below tests the similarity of the CIMMYT Wheat Trial data--which is presumably collected at a higher fidelity spatial resolution--with the Earth Engine data for precipitation over the 12 months prior to harvest. The lower-left and upper-right quadrants of the plot indicate the correlation coefficient between CIMMY and Earth Engine data, with the diagonal providing insights on the correlation between variables measuring precipitation for the same month. The correlations range from 0.25 to 0.55, with stronger association in the months closer to harvest. The

positive correlation values indicate that the remotely sensed data is in fact comparable to the manually collected CIMMYT data.

Figure 12: Correlation Heatmap of CIMMYT and GEE Precipitation Variables



(Source: Aggregated CIMMYT Wheat Trial Data and Google Earth Engine ERA5 Daily Climate Data; Calculations by Author)

What is strange, however, is that the correlation between CIMMYT and GEE data is not consistent over the months of precipitation. If both datasets are attempting to measure the same precipitation values in the same areas at the same times, it seems unusual that their correlation scores would vary over different months. There are two plausible explanations for the difference in levels of association over time: months further in advance of harvest (7 months prior or earlier) experience greater variance in the spatial distribution of weather patterns, making correlations harder to detect, or data collection methods differ for one of the two datasets.

Since the data encompasses most regions of Earth over a 40-year period, it seems unlikely that a pattern of spatial weather variance would persist over the entire dataset.

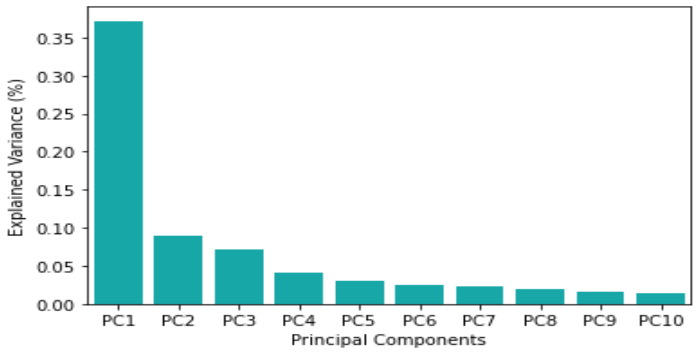
Temporal seasonality in precipitation would be expected, as would spatial patterns, but both datasets should detect these trends equivalently. Unless there were changes in the methods used to collect or compile satellite information for the ERA5 precipitation estimates, the most likely explanation is that CIMMYT weather data is inconsistent for at least some studies. This theory, while difficult to test directly, is also supported by the stronger degree of correlation between neighboring months' precipitation values in the Earth Engine variables (in the bottom-right quadrant) compared to the CIMMYT variables (in the top-left quadrant). Again, there does not seem to be a plausible explanation for precipitation to correlate more closely between adjacent months in the remote sensing data as compared to the CIMMYT data for the same locations during the same years except for a measurement bias in one of the datasets. There is at least reasonable suspicion that at least some of the CIMMYT data may suffer from measurement bias, or that there was a mistake made during data processing.

Principal Component Analysis of the Coefficient of Parentage Values

The pedigree data contained within the coefficient of parentage matrix is a theoretically inefficient way of representing the relatedness of genotype varieties. This is because the matrix is sparse, containing many zero values, and repetitive, since it is by definition symmetric. To take advantage of the genotype data in a more computationally efficient way, the dimensionality of the relatedness matrix can be reduced through principal component analysis. The resulting principal components describe the dimensions of greatest variance in the multi-dimensional cloud, meaning that each provides a measure of the position of each genotype in the dispersion of relatedness values. The trade-offs for reducing the dimensions of the dataset are, of course, a loss in the overall detail of the information and potential challenges in interpreting any analysis conducted with the components. Given that the principal components can be returned to their original dimensions through multiplication with the transpose of the eigenvectors, and the computational benefits from reducing the dimensions from 3,443 to 100 are significant, PCA seems appropriate.

Before conducting PCA, the coefficient of parentage matrix containing the variance and covariance of genotypes was merged with the yield observations to assign relatedness values for each observation's genotype to all others in the dataset. Since this results in 3,433 columns of mostly sparse data added to the dataset, methods of dimensionality reduction were considered to improve future analytical model performance. Following the research of Yao and Ochoa (2019), principal component analysis was used to decompose the coefficient of parentage genotype columns into a set of principal components. While the optimal number of components to include in future models is the subject of active research, for this project 100 principal components were selected, which together accounted for 88.9% of the total variance of the original columns. The first 10 components are plotted below by the percentage of variance explained. Component 1 seems to capture over a third of the total variance, with subsequent components declining from 9% to 2% by component 10.

Figure 13: Explained Variance Captured by First Ten Principal Components of Coefficient of Parentage Values



(Source: ICIS Coefficient of Parentage Matrix; Calculations by Author)

Comparative Analysis of Regression Models

The two primary hypotheses for the project were tested by comparing the predictive accuracy of a series of regression models. The linear mixed model utilizing Bayesian generalized linear regression (BGLR) is taken as a baseline, since it has been commonly employed in crop breeding research for assessing genotype- by-environment interactions and thus serves as a familiar approach for other researchers. The other models employed are common machine learning methods available in most statistical analysis software, including random forest, gradient boosting, and multi-layer perceptrons. By comparing the relative performance of various datasets across different models, the project aims to provide insights on which combinations may be most promising for predicting optimal seed varieties for novel locations based on environmental conditions and previous experimental evidence. By testing all the methods with both the CIMMYT and GEE data, the project hopes to assess the viability of remote sensing data as a replacement for manually collected factors, particularly for making out-of-sample predictions in novel locations.

Linear Mixed Model with BGLR

The linear mixed model is an extension of the ordinary least squares regression which allows for both fixed and random effects. Fixed effects in linear modelling describe the average relationship between the independent variables and the dependent outcome (or target) variable. The calculation of these fixed effects requires statistical assumptions on the distribution of the variables and their covariance with the outcome, notably that residuals are independently and identically distributed (iid). This iid assumption does not allow the model to capture variance that arises from the dependence of observations within the model from anything except for the random sampling of observations from a population. In other words, only the distribution of the full sample of observations is accounted for, not potential distributions within sub-samples. Since agricultural data is often collected from multiple locations or genotypes, it is reasonable to expect that the data within these sub-groups may be related, so that fields in the same location are more similar to each other than fields in different locations, or related genetic varieties have more similar yields than unrelated ones.

The hierarchical structure of environmental trials like CIMMYT's, therefore, necessarily violates the assumption of independence between observations, leading to biased results in fixed effect models. It is not enough to simply introduce dummy variables for each category, as it would assume that these effects extend to observations in the population outside of the sample, leading to overfitting and inflated error rates for prediction. Fixed effect dummies also do not account for potential dependence between categories, such as spatial dependence between sites or genetic relatedness between varieties. In the case of multi-environment crop breeding trials, it would also increase the dimensionality of the input (or feature) data significantly. The aggregated CIMMYT data employed by this project, for example, has over 3,443 unique genotypes and 2,773 unique locations, compared with just 34 environmental variables in a base model of just environmental factors. Furthermore, when the question of interest is the prediction of genotype performance in a novel location, utilizing previous location fixed effects would not add predictive power to the model since the test sample, by definition, does not have previous location data.

The addition of random effects, on the other hand, allow for the inclusion of variance in the model from hierarchical designs without violating independence assumptions. Introducing a random term for locations or genotypes, for example, lets the model account for the random intercept of these groupings by assuming the distribution of the variable in the population a priori. By setting these distributions in advance, the only variance left for the model to estimate is the effect of sampling from a population, satisfying the independence assumptions and calculating unbiased coefficient values for the main effects. The drawback of such a technique, of course, is that it introduces an assumed distribution for variables which may or may not be known in advance. These distributions are often assumed to be standard normal Gaussian with a mean of zero and variance estimated from available data. In the case of genotypes, for example, the distribution of the effects for each variety can be calculated in a matrix using the pedigree of selection history. The estimation of these distributions incorporating variance-covariance matrices can be calculated using best linear unbiased predictors (BLUPs), as described by Henderson (1975).

By predicting the BLUP for the random effects of genotypes in a mixed model, the regression analysis can incorporate pedigree information as stored in the coefficient of parentage (COP) matrix to explicitly model genetic similarity between varieties. The effect of genetic covariance between species is typically modelled as an additive matrix, calculated as twice the coefficient of parentage value multiplied by the population genetic variance. Bernando (2019) gives a more comprehensive explanation of how genetic effects can be modelled for different crop species.

Linear mixed models (LMMs) have been common practice for studying genotype-by-environment trials for decades. According to Crossa et al (2006), the first use of random effects modelling for the assessment of genotype means on locations was by Gogel et al. (1995). Piepho (1997) extended this work through a factor analytic mixed model which incorporated the variance-covariance structure of genotypes. For a more comprehensive history of the applications of mixed models in crop evaluation trials, see Smith et al. (2005), Malosetti et al. (2013) and Pérez-Rodríguez et al. (2017).

Application of the LMM using the BGLR Package

The standard form of the linear mixed model equation is like a fixed effects linear regression, in that there is a target variable (y) that is estimated using a set of input features. The input features are multiplied by a set of coefficient parameters, estimated by the model, and added to calculate the predicted outcome. The general form of an LMM for genotype-by-environment interactions, therefore, can be written as follows:

Equation 1: Linear Mixed Model with Genotype-by-Environment Interaction

$$y_{ij} = 1\mu + X_i\beta_1 + X_j\beta_2 + (X_iX_j)\beta_3 + \varepsilon$$

Where $y_{ij}$is the target phenotype value, in this case grain yields, for genotype i in location j; μ is an intercept term; X_i is a design matrix of genotype values identifying the variety of each observation; $\beta_1$is the vector of random effects from genotypes estimated using a Bayesian ridge regression (BRR) model fit to the additive variance-covariance pedigree matrix; X_j is the design matrix of environmental covariate values; $\beta_2$is the vector of fixed effects associated with each environmental variable; $X_iX_j$is the interaction matrix of the genotype and environmental covariate matrices; $\beta_3$is the vector of random effects of these interactions, also modeled with a BRR; and ε is an error term of residuals. Reduced forms of the model, containing either the environmental or pedigree effects only, are constructed by removing the interaction term and complementary set of design matrix and effects vector.

Estimating the fixed and random effects of the linear mixed model requires several steps to calculate the appropriate matrices and regression coefficients. In applications with high dimensionality due to the number of genotypes, sites, and interactions considered, it can be computationally difficult to estimate all the necessary unknown values. The Bayesian Generalized Linear Regression package for R, developed by Paulino Pérez-Rodríguez, addresses these problems using Bayesian approaches to shrinkage and variable selection for estimating the model terms. For a more complete description of the BGLR package, the CRAN page contains both a reference manual describing the functions included and a vignette with example code and data (https://cran.r-project.org/web/packages/BGLR/index.html).

The full genotype-by-environment interaction LMM was built in 3 stages: first with just the random effects of seed variety lines using the pedigree matrix, then with the addition of the environmental covariates as fixed effects, and finally with the introduction of the interaction terms between the pedigree coefficients and the environmental variables as random effects.

In order to fit an LMM with random genotype effects, the distribution of the pedigree matrix must first be calculated. The default method for parametrizing the additive relationship matrix (**A**) is to divide the matrix by the average of its diagonal and then compute the Cholesky decomposition of the resulting values, then multiply these by the incidence matrix of the varieties to apply a genotype value to each observation. The Cholesky decomposition converts the matrix into the product of a lower triangular matrix and its conjugate transpose. While this method is more computationally efficient,

it is unstable for large and sparse datasets, which is often the case when the number of genotypic varieties is high. An alternative is to replace the Cholesky decomposition with the eigen-decomposition of the A matrix into its eigenvectors and eigenvalues. Multiplying the genotypic incidence matrix by the eigenvectors and the square root of the eigenvalues was shown by Meyer (2009) to be equivalent to the calculation including the Cholesky decomposed **A** matrix. The resulting product (**Z***) can then be included in the BGLR model through a Bayesian ridge regression model to fit the main effects of varieties on yields using a Markov Chain Monte Carlo method.

The main environmental covariate effects can be added to form the second LMM simply by dividing each value by the square root of the number of variables, then fit similarly with a Bayesian ridge regression. The final interaction effects model is calculated similarly to the pedigree main effects. First, we take the cross product of the transpose of the **Z*** matrix (**ZAZ**), then the cross product of the transpose of the environmental covariates (**W***). We then multiply these two cross product matrices to calculate the intermediate interaction matrix (**K**). This matrix would also normally be Cholesky decomposed, but again we employed eigen-decomposition to avoid the instability of the Cholesky requirements. Multiplying the eigenvectors of K by the square root of its eigenvalues results in the final interaction matrix (**L2**), which is then included in the BGLR model along with the main effects of pedigree and environments, also using Bayesian ridge regression model.

The input features for the BGLR model still include several large, potentially sparse variance-covariance matrices, however, which can approach several hundred gigabytes in size--too large for the temporary random-access memory (RAM) of most computers. In calculating the covariance matrices for this project, a virtual computing environment provided by Google Compute Engine on the Google AI Platform with 1.4 terabytes of RAM was necessary to complete all calculations. The average laptop available to most consumers has around 16Gb of RAM, around 1.14% of the memory necessary to calculate the BGLR models used in this analysis. At the time of writing, Google provided $300 USD of credit for Compute Engine applications to new users, allowing others to replicate the same environment, though clearly this requirement is not ideal for scientific replication since it depends on a promotional offer. The computational requirements of the matrix calculations, therefore, represent a significant limitation of the LMM approach. Given that more recent crop breeding approaches make use of even more computationally intensive genomic models, computing resources are a clear limitation in the accessibility of any approach that requires the calculation of large matrices.

Various statistical and computational workarounds were attempted, including the bigmemory R package, developed by Emerson and Kane, which can be used to store the matrices in disc memory and call portions into RAM as needed. An adjusted version of the BGLR package was developed to accommodate these large matrices, as described in Pérez-Rodríguez et al. (2017) and made accessible at (http://genomics.cimmyt.org/wheat_50k/PG/), but when applied to the data for this project the adjusted package was not able to convert the bigmatrix objects successfully. Since the adjusted BGLR package is not published or maintained, it is also likely to become deprecated and stop working with newer versions of the bigmatrix package as

they are released. Further efforts to parallelize the matrices and make the calculations less dependent on large amounts of RAM were discontinued due to time constraints but may be investigated further in subsequent research.

As suggested by Crossa et al (2006), the OpenBLAS software for optimizing linear algebra calculations was also used to configure R for faster calculations of the necessary matrices (Xianyi et al, 2012; Wang et al, 2013). Changing the linear algebra solver in the R environment can help improve computational efficiency for large matrix equations.

Machine Learning Models with Scikit-learn and XGBoost

The broad field of machine learning encompasses all algorithms which improve their performance based on past iterations or experience. In statistical analysis, any process can be categorized as machine learning which employs a training set of data to make decision outputs given a test set of inputs (typically not included in the initial data). The specific mathematical function used to predict outputs is commonly referred to as the model or algorithm. While there are numerous fields and subfields of machine learning, its application for most predictive problems can be divided into supervised and unsupervised models.

Supervised models take a set of independent variables, also called features or inputs, and relate them to a dependent variable--the target or output. Once the model is "fitted" to the data, the statistical relationship between these and the output variable can be used to predict the value of the target for any set of input features. In the case of a linear regression, for example, this modelled relationship is given in the form of a simple matrix-based equation. Unsupervised models, on the other hand, do not provide a target variable and instead attempt to understand the underlying structure of a set of features using methods like clustering or principal component analysis. This project makes use of both methods, but the methodological focus is on supervised regression models for the prediction of continuous numeric output values, in this case grain yields.

Supervised machine learning models can be further subdivided into two categories of prediction: classification and regression. Classification models try to guess the category of an observation from a discrete set of options, with myriad applications for image recognition, land use mapping, assisted decision making, etc. Regression models estimate the numeric value of the target variable for an observation from a potentially infinite space of real numbers. Many statistical models have been adapted for use in both cases, such as applications of logistic regression for classification based on predicted probability. Similarly, models which are more commonly employed for classification, such as neural networks or decision trees, can also be extended to regression analysis. This project investigates three widely used categories of machine learning: random forests, gradient boosting (specifically the XGBoost algorithm), and neural networks (specifically a multi-layer perceptron using the Adam method of stochastic gradient-based optimization).

The random forest and multi-layer perceptron models were implemented using Scikit-Learn (Pedregosa et al, 2011), a popular package for machine learning and other data manipulations in Python. The XGBoost algorithm was developed by Tianqi Chen as

a software package which can be implemented in Python, R, and other statistical programming languages (Chen and Guestrin, 2016). Both follow a similar framework for machine learning, in which a regressor object is created and fit to the dataset in a cross-validation framework which tunes the parameters of the model to optimize accuracy over repeated sections of the data. The parameters for models differ by algorithm and may also differ in their optima by dataset, so the process of searching for the best values can be time and labor intensive. The GridSearchCV function in Scikit-Learn was used to help automate this process, using a 5-fold cross validation system across the training data and ultimately assessing accuracy against the test data, as described in the cross-validation section above.

Random Forest Regressor

Random forest models build upon two simple concepts: decision trees and bagging. The trees which make up the random forest are "grown" by making binary splits in a decision path according to the input features; if an observation has a value for precipitation higher than 10, as an arbitrary example, they are classified on one branch of the tree and if not then they fall on the other side. The random nature of the rules used for splitting branches of the tree allows the model to approximate nonlinear relationships and complex interactions without feature engineering or theoretical assumptions about the nature of variables' relationships. The trees assess the validity of each split according to their contribution to prediction accuracy of the target variable, making them powerful yet computationally inexpensive models. The downside of decision trees is that they tend to overfit the training data; their prediction parameters do not extend well from the sample they are given to the population of interest.

To overcome the limitation of overfitting, decision trees can be extended using a technique called bagging. The principle behind bagging is simply taking random subsamples with replacement of the input data and using them as training sets, allowing each to grow its own decision tree, and then comparing the combined predictions of all trees. Due to the random nature of the subsamples, the trees are independent of each other and thus in aggregate predict a less biased result than any would individually. The extension of decision trees into random forests is part of a collection of machine learning approaches known as ensemble methods, which use averages of multiple predictions to make a collective decision on the value or class of the target. In addition to reducing bias, the aggregation of decision trees can help avoid issues of high dimensional feature data, such as multicollinearity between input variables, which would violate the assumptions of a linear model.

The specific random forest function used for this project is the RandomForestRegressor from the Scikit-Learn Ensemble package. The documentation for the function can be found on the Scikit-Learn website (https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html).

Gradient Boosting Regressor with XGBoost

Gradient boosting is another type of ensemble method which uses a series of weak learning functions to converge to an optimal prediction through iterative gradient descent. In general, gradient boosting starts with a single prediction function, like a

least squares regression, which combines the modelled relationship as well as the residual errors. The next iteration attempts to capture further variance from the residuals, approaching the minimum of a generalizable loss function by descending according to the gradient of the initial prediction.

XGBoost, for extreme gradient boosting, is an algorithm built on tree-based gradient boosting methods. The functional approach can thus be thought of similarly to a random forest, in that it captures non-parametric relationships as well as interactions between features, but rather than simply comparing many independent trees the model attempts to optimize its predictions further by directing iterations using the gradient of the loss function. XGBoost adapts the general class of gradient boosted regression trees with computational improvements that help process sparse, high dimensional datasets. Not only does this improve runtimes for the algorithm, the resulting predictions have generally been more accurate since they can incorporate more data in a more "intelligent" manner than pure randomness.

The specific XGBoost function utilized in this project is the XGBRegressor, part of the Python API adapted to fit within the Scikit-Learn framework. The documentation for the function can be found on the XGBoost website (https://xgboost.readthedocs.io/en/latest/python/python_api.html#)

Multi-Layer Perceptron Regressor with Adam

Neural networks comprise one of the more diverse categories of machine learning prediction. Artificial neural networks (ANNs) are the most straightforward of many systems which attempt to replicate the connectivity of neurons in a biological brain. All ANNs function as a collection of nodes (or neurons) connected by networked linkages (or edges, similar to synapses). When a node receives a signal, in the form of some data value, it processes this signal according to some function and passes it on to descendent nodes. The functions employed by the neurons for processing may be non-linear and can take multiple input signals from different neurons; in this way it is also analogous to a decision tree, with each node approximating a decision rule, but in the case of neural networks the rules do not need to result in binary outcomes. The links or edges between neurons can be weighted to increase or decrease the relative importance of that connection to the entire model; allowing the network to be regularized in a fashion like a linear regression. The individual nodes are often aggregated into a set of layers, with each layer connected only to its surrounding layers of neurons. By adjusting the number of layers in the model, the complexity of the overall network can be regulated.

The multi-layer perceptron is a type of artificial neural network that only allows nodes to connect neurons in a single direction, known as a feed-forward network. The limitation of signals traveling only from the input data layer towards the output prediction layer restricts and simplifies the network, avoiding any convolutional or recurrent layers. While these models may improve prediction performance, they often do so at the cost of interpretability, as the final network could involve complex patterns of repetitive neural processing.

The processing methods implemented by the nodes in a neural network determine the optimal value at each level through an objective function. One of the

algorithms that has been introduced to solve these optimization problems is Adam, developed by Kingma and Ba (2017). Adam has become the default optimizer for multilayer perceptrons in Scikit-Learn due to its computational efficiency with large, high-dimensional datasets; these properties allow it to calculate the optimal parameters for tuning a neural network and achieve convergence--reach the minimum value for a cost function by iterating through the gradient space.

The specific multilayer perceptron function utilized in this project is the MLPRegressor in the Scikit-Learn Neural_Network package. The documentation for the function can be found on the Scikit-Learn website (https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html)

Prediction for Holdout Sample

Statistical prediction methods are only applicable to hypothesis testing if implemented within the correct research framework. Given that the motivating question for this project is the prediction of grain yields for different genotypes in novel environments, the cross-validation framework attempts to simulate this problem by holding out a sample of the data as a validation set. This validation set is composed of 15% of observations stratified by genotype variety, to proxy a realistic decision framework in which regulators or farmers are interested in which seed varieties are expected to produce the highest yields. Each prediction method--LMM, random forest, XGBoost, and MLP--was tuned to its optimal parameters using a 5-fold cross validation on the training data (72.25%), tested for each of the 9 dataset types on the testing data (12.25%), and finally compared on the validation holdout dataset (15%) using Google Earth Engine environmental data.

Results and Interpretation

The results of the analysis are split into the accuracy scores of the predictive models, scatter and residual plots of the LMM and random forest models, lists of the top ten features by importance for the random forest model for each dataset, and a discussion of the reproducibility of the methodology.

The primary results indicate that machine learning approaches can provide higher levels of accuracy in a less computationally intensive environment. Remote sensing data improves upon the accuracy of the manually collected data for the linear models but not in the machine learning models, perhaps due to poor specification of parameters or extraneous variable dimensions. The most important features differed by data type: altitude, the spacing and length of rows, and precipitation were most important in the CIMMYT data while the GEE data models depended more on soil moisture and vapor pressure. The inclusion of genetic information improved accuracy for the linear models but decreased accuracy when included as principal components in the machine learning models, though this may still be due to problems in the parametrization of the models. The LMM approach with BGLR was impossible in a typical computing environment due to memory constraints; the machine learning models were more accessible but had longer runtimes and were more sensitive to parameter tuning.

Accuracy Scores for Predictive Models

The predictions for each of the four statistical models were assessed against the observed values for the same holdout set of fields. The accuracy statistics used for comparison included the root mean squared error (RMSE), the mean absolute error (MAE), Pearson's R correlation, and the R-squared. The RMSE and R-squared for each model and each variation of the data are given in Table 6 below.

Table 6: RMSE and R-Squared for LMM and ML Models for All Environment and Genotype Dataset Combinations

Holdout RMSE (R-Squared)		BGLR LMM GxE Interaction	Random Forest	XGBoost	Multi-Layer Perceptron Network
CIMMYT Env Data	Environment Data Only	0.941 (0.113)	0.414 (0.829)	0.423 (0.822)	0.419 (0.824)
	Env & Pedigree	0.928 (0.139)	**0.401 (0.839)**	0.423 (0.821)	0.456 (0.792)
GEE Env Data	Environment Data Only	0.869 (0.243)	0.504 (0.746)	0.507 (0.743)	0.514 (0.736)
	Env & Pedigree	0.859 (0.261)	0.509 (0.741)	0.499 (0.751)	0.509 (0.741)

(Source: Aggregated CIMMYT Data and Google Earth Engine Data; Calculations by Author)

The results of the analysis are mixed with respect to the first hypothesis, whether remotely sensed data can improve upon manually collected environmental variables. The linear mixed models were both more accurate with GEE data instead of CIMMYT data, but the opposite was true for the machine learning models. Paradoxically, the inclusion of genetic information also improved accuracy in the LMM but decreased it in some machine learning models. Without further robustness tests on the parameters used for training the machine learning models, it is difficult to conclude if the results are indicative of a generalizable trend or simply a misspecification of the models. Given the other advantages of remote sensing data in terms of cost, availability, and reduction in measurement bias, the comparable accuracies of the two models may still offer sufficient motive to replace manually collected climate information with Earth Engine data, particularly if geographic locations can be specified with a high degree of accuracy.

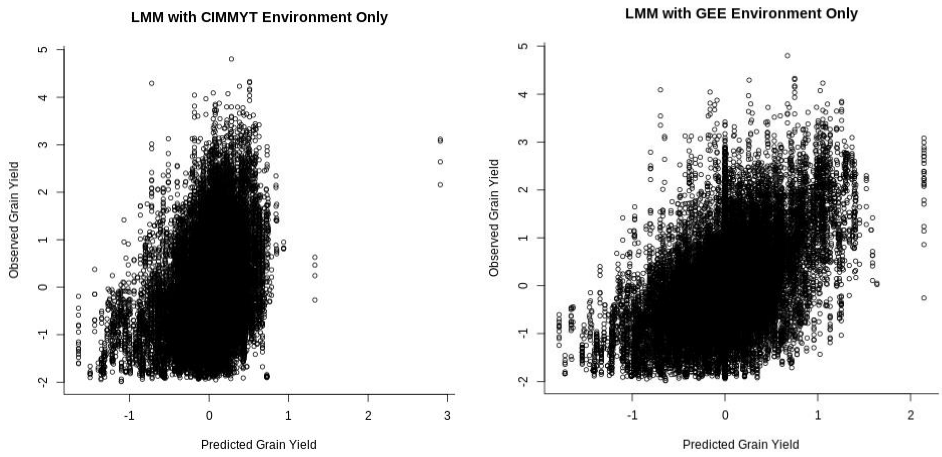
The results support a refutation of the second null hypothesis, however, since the machine learning models consistently outperformed the linear mixed models in terms of accuracy. The random forest model performed best on the holdout validation sample in all data cases except the GEE and pedigree data, which had the highest dimensionality. The random forest model also provides an easily interpretable list of feature importance values, which can be cross-referenced against variable coefficients in linear models to better understand which factors contribute to predictive modelling. Neural networks do not offer simple interpretation, however, so are of limited use for generalizing results.

Considering these factors, random forest models seem most appropriate for predicting out-of-sample wheat yields in a genotype-by-environment framework.

Diagnosis of Scatter and Residual Plots

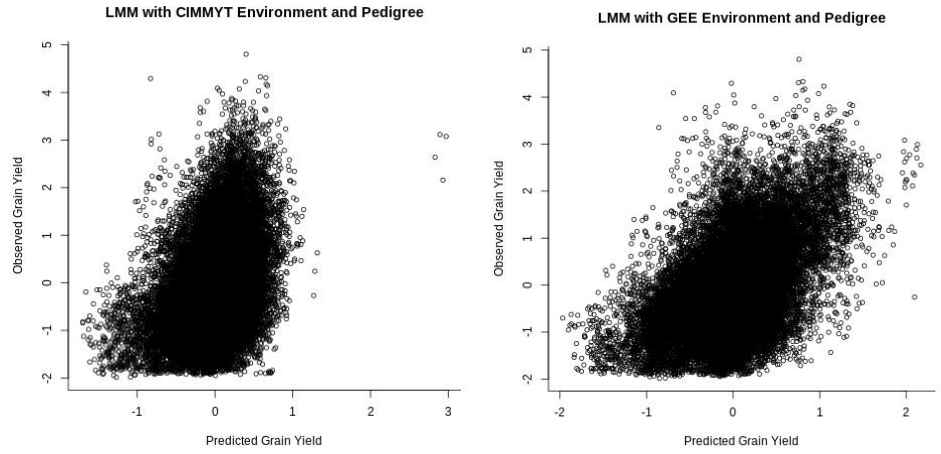
Examining the scatterplots of observed vs expected values for each model, the lower accuracy of the linear mixed models compared to the machine learning models is quickly evident from the wider distribution from the 45-degree line. The LMMs also seem to bound their prediction results at arbitrary thresholds, forcing most predicted values between -1 and 1 standard deviations from the mean. The predictions using Earth Engine data exhibit more linear patterns, corresponding to their higher accuracy scores.

Figure 14 and 15: Scatterplots of Actual vs Predicted Test Values using a BGLR LMM Model with Environmental Variables using CIMMYT (Left) and Earth Engine (Right)



(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Figure 16 and 17: Scatterplots of Test Values for Grain Yield using a BGLR LMM Model with GxE Interactions using CIMMYT Data (Left) and Earth Engine Data (Right)

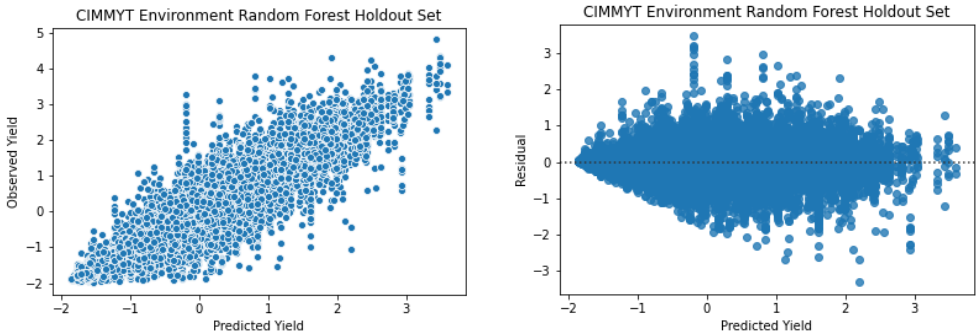


(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

The scatter plots and residual plots from the random forest models for each dataset demonstrate much tighter predictions around the 45-degree line, reflecting their higher accuracy scores. The residuals of the Earth Engine data without pedigree information seem lower than the CIMMYT models except for a single vertical column clustered around a predicted value of 0. Since the data was standardized, this represents fields that were predicted to have average yields but varied considerably. This may be due to the imputation of missing data with average values.

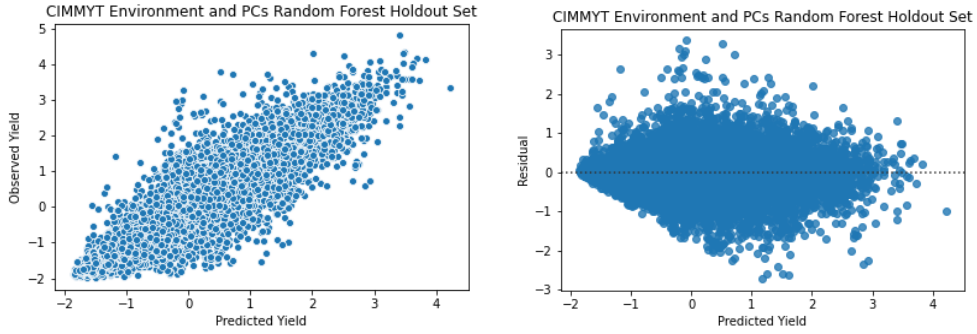
Interestingly, the vertical column of residuals in the GEE data disperses with the introduction of genotypic pedigree data as principal components. There is still a noticeable cluster of residuals slightly below the average yield value, however, indicating that the interaction of environmental and genetic information in the random forest did not help significantly in predicting these values. If these observations could be investigated and resolved, either with more accurate geographic locations or dates, it could significantly improve the accuracy of the models using the remote sensing data.

Figure 18 and 19: Random Forest Model using CIMMYT Environmental Data Scatter Plot (Left) and Residual Plot (Right)



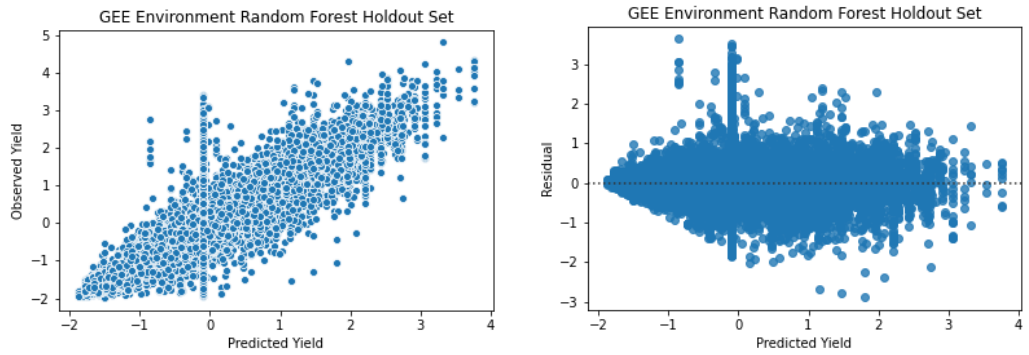
(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Figure 20 and 21: Random Forest Model using CIMMYT Environmental Data and Pedigree Principal Components Scatter Plot (Left) and Residual Plot (Right)



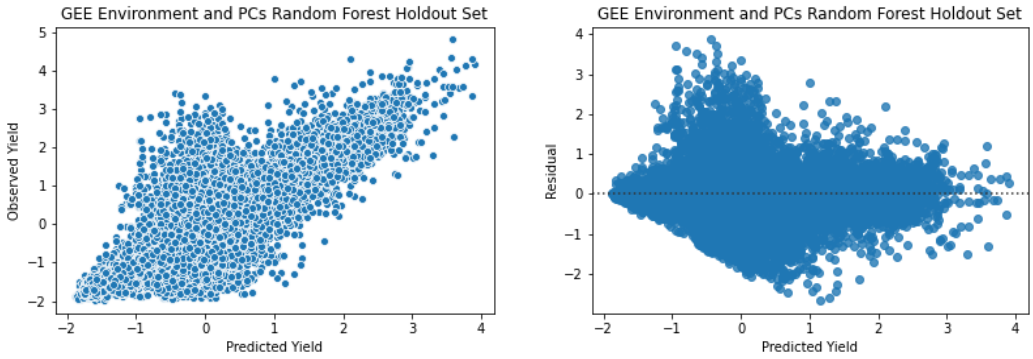
(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Figure 22 and 23: Random Forest Model using GEE Environmental Data Scatter Plot (Left) and Residual Plot (Right)



(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Figure 24 and 25: Random Forest Model using GEE Environmental Data and Pedigree Principal Components Scatter Plot (Left) and Residual Plot (Right)



(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Another interesting finding apparent in the scatterplots is the presence of vertical columns of predicted values in the machine learning models using only environmental data. This seems to indicate that the algorithm received identical input features for a set of observations with various target values and was unable to further distinguish them. Theoretically, this supports the design of the CIMMYT trials: plots in the same trial should have experienced almost identical environmental conditions, so the remaining distribution of residuals should be attributable only to genetic differences between varieties. These vertical columns are also present in the LMM environmental data models, indicating that the linear model was likewise unable to distinguish between varieties in the same location without pedigree data. The vertical columns of points disappear in the models that incorporate pedigree information, supporting the conclusion that the models successfully applied genetic information to help predict yields. It seems reasonable that much of the improvement in accuracy in models that included pedigree data could be attributable to the algorithm resolving the leaf nodes that were not able to split into separate predictions with only environmental data.

Comparison of Random Forest Feature Importance Values

An initial assessment of the features deemed important by the random forest models reveals an interesting disparity between the manually collected CIMMYT and remotely sensed GEE environmental data. CIMMYT models consistently reported altitude and the spatial configuration of rows as most important, followed by precipitation. GEE models, despite having similar altitude and precipitation variables, relied on evapotranspiration, soil moisture, and vapor pressure.

Table 7: Feature Importance Values of Top Ten Features for CIMMYT Environmental Data in the Random Forest Model

Feature Importance for CIMMYT Environmental Data from Random Forest Model	
ALTITUDE	0.1614
SPACE_BTN_ROWS_SOWN	0.0691
LENGTH_OF_ROWS_SOWN	0.0509
LENGTH_OF_ROWS_HARVESTED	0.0496
PPN_11TH_MO_BEFORE_HARVESTED	0.0446
PPN_1ST_MO_BEFORE_HARVESTED	0.0442
PPN_7TH_MO_BEFORE_HARVESTED	0.0431
PPN_4TH_MO_BEFORE_HARVESTED	0.0426
PPN_3RD_MO_BEFORE_HARVESTED	0.0414
PPN_MONTH_OF_HARVESTED	0.0403

Table 8: Feature Importance Values of Top Ten Features for CIMMYT Environmental and Pedigree Data in the Random Forest Model

Feature Importance for CIMMYT Environmental Data with Principal Components from Random Forest Model	
ALTITUDE	0.1427
SPACE_BTN_ROWS_SOWN	0.0616
LENGTH_OF_ROWS_SOWN	0.0446
LENGTH_OF_ROWS_HARVESTED	0.0440
PPN_1ST_MO_BEFORE_HARVESTED	0.0392
PPN_11TH_MO_BEFORE_HARVESTED	0.0383
PPN_7TH_MO_BEFORE_HARVESTED	0.0379

PPN_4TH_MO_BEFORE_HARVESTED	0.0365
PPN_3RD_MO_BEFORE_HARVESTED	0.0360
PPN_MONTH_OF_HARVESTED	0.0350

Table 9: Feature Importance Values of Top Ten Features for GEE Environmental Data in the Random Forest Model

Feature Importance for GEE Environmental Data from Random Forest Model	
(SoilMoi10_40cm_tavg, 8)	0.1133
(vap, 4)	0.0602
(vap, 3)	0.0174
(vap, 7)	0.0169
(SoilMoi00_10cm_tavg, 0)	0.0167
(def, 8)	0.0163
(SoilMoi00_10cm_tavg, 8)	0.0153
(soil, 3)	0.0128
(soil, 5)	0.0098
(soil, 0)	0.0088

Table 10: Feature Importance Values of Top Ten Features for GEE Environmental and Pedigree Data in the Random Forest Model

Feature Importance for GEE Environmental Data with Principal Components from Random Forest Model for Top 40 Features	
(SoilMoi10_40cm_tavg, 8)	0.1133
(vap, 4)	0.0602
(vap, 3)	0.0174
(vap, 7)	0.0169
(SoilMoi00_10cm_tavg, 0)	0.0167
(def, 8)	0.0163
(SoilMoi00_10cm_tavg, 8)	0.0153

(soil, 3)	0.0128
(soil, 5)	0.0098
(soil, 0)	0.0088

For a more complete list of feature importance per model, see Appendix 5.

Reproducibility of Methods

The tertiary line of inquiry revealed several gaps in the reproducibility of crop selection research. The scale of genotype pedigree information makes it impractical to calculate the covariance matrices necessary for a BGLR model on a typical computing system; a specialized virtual machine utilizing a terabyte of RAM was necessary to complete calculations for the entire dataset. Given that genomic data on particular markers--necessary for more advanced quantitative trait loci models-- are orders of magnitude larger than simple genotype relation matrices, it seems that crop selection using standard analytical methods is impractical for non-specialized users. The machine learning methods, on the other hand, were able to calculate predictions for the entire dataset on a freely available server with Google Colab, though the calculations did take several hours to converge in some cases. The issues of data availability with respect to pedigree information and the manual aggregation of CIMMYT trials are also hurdles to reproducibility but ultimately less restrictive than the computational intensity of the Bayesian methods. The complexity of the Earth Engine API for querying environmental data is likewise a hurdle for non-technical users, though once a standard script is developed the process is straightforward and robust to changes in location or timeframe.

Conclusions and Further Research

The results of this project demonstrated that remotely sensed environmental data accessed from Google Earth Engine can provide a low-cost, standardized alternative to manually collected variables for wheat yield trials. The advantages of remote sensing can even improve on the accuracy of limited existing CIMMYT datasets, particularly in their global coverage and temporal completeness, but are not guaranteed to provide better models since they depend on the specificity of the input locations and dates. The increase in dimensionality of the input datasets can be easily compensated for using modern supervised learning techniques, which can improve upon the accuracy of linear mixed models significantly, at the cost of model interpretability. The open-source code workflow offers a feasible alternative to existing inaccessible research methods, except for limited data availability of genotypic pedigree data. The accuracy of yield predictions using this open-source workflow offer promising opportunities to create derivative products such as interactive web applications for determining optimal genotypes based on current climate data that would be more accessible than current offerings for key agricultural stakeholders.

Additional research opportunities for building upon this work are numerous. New data sources or different crops could be inserted into the data workflow to assess the transferability of the project beyond wheat to maize, sorghum, rice, etc. The utility of remotely sensed data for prediction accuracy in these cases is an open question that could be quickly answered with adequate input data on yields. The pedigree data could also be replaced with more advanced genomic data on quantitative trait loci to improve accuracy further, though this may come at the cost of computational accessibility. Additional remote sensing data could also be included, or alternative spatial or temporal reduction functions could be tested, as well as variations on the spatial buffer distance employed in data querying. The initial CIMMYT data only provided geographic coordinates to the minute level; more detailed geographies could further enhance precision. The inclusion of additional CIMMYT data on newer wheat yield trials could also help improve predictive performance for more recent varieties. There are also many opportunities to test different machine learning models or specifications of hyperparameters; the author does not claim advanced knowledge of machine learning practices and developed the current code as a proof of concept. Finally, climate forecast data could be used to predict yields of certain genotypes in novel locations in the future, enhancing the practical benefit for stakeholders by matching predicted climate scenarios with optimal seed varieties.

An optimized prediction function could also be used to create a classification model that selects the top genotype varieties for a given location based on the predicted yield values simulated from available test sites. For novel locations outside of the provided dataset, environmental input data can be readily queried from remote sensing datasets in Google Earth Engine.

In addition to the conceptual improvements or extensions that can be made to the research, the code itself can be made more efficient and accessible through further revisions. Fortunately, since all the notebooks for importing, processing, and analyzing the data are freely available on GitHub, new users can make suggestions or contributions to the code at any time. The author plans to maintain the code as a living resource for other researchers and stakeholders to build upon and will continue to respond to feedback on the models. Some of the more generally useful notebooks, such as the code to query Earth Engine data for any set of coordinate points and dates as well as the machine learning models, will also be published in a simplified form for further learning and adaptation. The link to the GitHub repository for this project is available in Appendix 1.

Bibliography

Atlin, Gary N., Jill E. Cairns, and Biswanath Das. "Rapid Breeding and Varietal Replacement Are Critical to Adaptation of Cropping Systems in the Developing World to Climate Change." *Global Food Security* 12 (March 1, 2017): 31–37. https://doi.org/10.1016/j.gfs.2017.01.008.

Bernardo, Rex Novero. *Breeding for Quantitative Traits in Plants*. 3rd ed. Woodbury: Stemma Press, 2019.

Brooks, Sally. "Is International Agricultural Research a Global Public Good? The Case of Rice Biofortification." *The Journal of Peasant Studies* 38, no. 1 (January 1, 2011): 67–80. https://doi.org/10.1080/03066150.2010.538581.

Burney, Jennifer A., Steven J. Davis, and David B. Lobell. "Greenhouse Gas Mitigation by Agricultural Intensification." *Proceedings of the National Academy of Sciences* 107, no. 26 (June 29, 2010): 12052–57. https://doi.org/10.1073/pnas.0914216107.

Butler, S. J., J. A. Vickery, and K. Norris. "Farmland Biodiversity and the Footprint of Agriculture." *Science (New York, N.Y.)* 315, no. 5810 (January 19, 2007): 381–84. https://doi.org/10.1126/science.1136607.

Cai, Yaping, Kaiyu Guan, David Lobell, Andries B. Potgieter, Shaowen Wang, Jian Peng, Tianfang Xu, et al. "Integrating Satellite and Climate Data to Predict Wheat Yield in Australia Using Machine Learning Approaches." *Agricultural and Forest Meteorology* 274 (August 15, 2019): 144–59. https://doi.org/10.1016/j.agrformet.2019.03.010.

CGIAR. "Open Access and Open Data." Accessed April 24, 2020. https://www.cgiar.org/how-we-work/accountability/open-access/.

Challinor, A. J., A.-K. Koehler, J. Ramirez-Villegas, S. Whitfield, and B. Das. "Current Warming Will Reduce Yields Unless Maize Breeding and Seed Systems Adapt Immediately." *Nature Climate Change* 6, no. 10 (October 2016): 954–58. https://doi.org/10.1038/nclimate3061.

Chen, Tianqi, and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining - KDD '16, 2016, 785–94. https://doi.org/10.1145/2939672.2939785.

Crespo-Herrera, L. A., J. Crossa, J. Huerta-Espino, M. Vargas, S. Mondal, G. Velu, T. S. Payne, H. Braun, and R. P. Singh. "Genetic Gains for Grain Yield in CIMMYT's Semi-Arid Wheat Yield Trials Grown in Suboptimal Environments." *Crop Science* 58, no. 5 (10/01 2018): 1890–98. https://doi.org/10.2135/cropsci2018.01.0017.

Crespo-Herrera, Leonardo A., Jose Crossa, Julio Huerta-Espino, Enrique Autrique, Suchismita Mondal, Govindan Velu, Mateo Vargas, Hans J. Braun, and Ravi P. Singh. "Genetic Yield Gains In CIMMYT's International Elite Spring Wheat Yield Trials By Modeling The Genotype × Environment Interaction." *Crop Science* 57, no. 2 (2017): 789–801. https://doi.org/10.2135/cropsci2016.06.0553.

Crossa, Jose, Juan Burgueño, Paul L. Cornelius, Graham McLaren, Richard Trethowan, and Anitha Krishnamachari. "Modeling Genotype × Environment Interaction Using Additive Genetic Covariances of Relatives for Predicting Breeding Values of Wheat Genotypes." *Crop Science* 46, no. 4 (July 1, 2006): 1722–33. https://doi.org/10.2135/cropsci2005.11-0427.

Curtis, T, and N G Halford. "Food Security: The Challenge of Increasing Wheat Yield and the Importance of Not Compromising Food Safety." *The Annals of Applied Biology* 164, no. 3 (January 2014): 354–72. https://doi.org/10.1111/aab.12108.

Esposito, Salvatore, Domenico Carputo, Teodoro Cardi, and Pasquale Tripodi. "Applications and Trends of Machine Learning in Genomics and Phenomics for Next-Generation Breeding." *Plants* 9, no. 1 (January 2020): 34. https://doi.org/10.3390/plants9010034.

Gardner, Bruce, and William Lesser. "International Agricultural Research as a Global Public Good." *American Journal of Agricultural Economics* 85, no. 3 (2003): 692–97. https://doi.org/10.1111/1467-8276.00469.

Gebre, Girma Gezimu, Hiroshi Isoda, Dil Bahadur Rahut, Yuichiro Amekawa, and Hisako Nomura. "Gender Differences in the Adoption of Agricultural Technology: The Case of Improved Maize Varieties in Southern Ethiopia." *Women's Studies International Forum* 76 (September 1, 2019): 102264. https://doi.org/10.1016/j.wsif.2019.102264.

Gillberg, Jussi, Pekka Marttinen, Hiroshi Mamitsuka, and Samuel Kaski. "Modelling G×E with Historical Weather Information Improves Genomic Prediction in New Environments." *Bioinformatics* 35, no. 20 (October 15, 2019): 4045–52. https://doi.org/10.1093/bioinformatics/btz197.

Godfray, H. Charles J. "Food for Thought." *Proceedings of the National Academy of Sciences* 108, no. 50 (December 13, 2011): 19845–46. https://doi.org/10.1073/pnas.1118568109.

Gogel, Beverley J., Brian R. Cullis, and Arunus P. Verbyla. "REML Estimation of Multiplicative Effects in Multienvironment Variety Trails." *Biometrics* 51, no. 2 (1995): 744–49. https://doi.org/10.2307/2532960.

González-Camacho, Juan Manuel, Leonardo Ornella, Paulino Pérez-Rodríguez, Daniel Gianola, Susanne Dreisigacker, and José Crossa. "Applications of Machine Learning Methods to Genomic Selection in Breeding Wheat for Rust Resistance." *The Plant Genome* 11, no. 2 (July 2018): 170104. https://doi.org/10.3835/plantgenome2017.11.0104.

Gorelick, Noel, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. "Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone." *Remote Sensing of Environment*, Big Remotely Sensed Data: tools, applications and experiences, 202 (December 1, 2017): 18–27. https://doi.org/10.1016/j.rse.2017.06.031.

Grinberg, Nastasiya F., Oghenejokpeme I. Orhobor, and Ross D. King. "An Evaluation of Machine-Learning for Predicting Phenotype: Studies in Yeast, Rice, and Wheat." *Machine Learning* 109, no. 2 (February 1, 2020): 251–77. https://doi.org/10.1007/s10994-019-05848-5.

Henderson, C. R. "Best Linear Unbiased Estimation and Prediction under a Selection Model." *Biometrics* 31, no. 2 (1975): 423–47. https://doi.org/10.2307/2529430.

Heslot, Nicolas, Hsiao-Pei Yang, Mark E. Sorrells, and Jean-Luc Jannink. "Genomic Selection in Plant Breeding: A Comparison of Models." *Crop Science* 52, no. 1 (2012): 146–60. https://doi.org/10.2135/cropsci2011.06.0297.

Jain, Meha, Singh Balwinder, Preeti Rao, Amit Srivastava, Shishpal Poonia, Jennifer Blesh, George Azzari, Andrew J. McDonald, and David B. Lobell. "The Impact of

Agricultural Interventions Can Be Doubled by Using Satellite Data | Nature Sustainability." Accessed April 9, 2020. https://www.nature.com/articles/s41893-019-0396-x.

Juliana, Philomin, Jesse Poland, Julio Huerta-Espino, Sandesh Shrestha, José Crossa, Leonardo Crespo-Herrera, Fernando Henrique Toledo, et al. "Improving Grain Yield, Stress Resilience and Quality of Bread Wheat Using Large-Scale Genomics." *Nature Genetics* 51, no. 10 (October 2019): 1530–39. https://doi.org/10.1038/s41588-019-0496-6.

Knox, Jerry, Tim Hess, Andre Daccache, and Tim Wheeler. "Climate Change Impacts on Crop Productivity in Africa and South Asia." *Environmental Research Letters* 7, no. 3 (September 2012): 034032. https://doi.org/10.1088/1748-9326/7/3/034032.

Lammerts van Bueren, Edith T., Paul C. Struik, Nick van Eekeren, and Edwin Nuijten. "Towards Resilience through Systems-Based Plant Breeding. A Review." *Agronomy for Sustainable Development* 38, no. 5 (August 22, 2018): 42. https://doi.org/10.1007/s13593-018-0522-6.

Lobell, David B. "Errors in Climate Datasets and Their Effects on Statistical Crop Models." *Agricultural and Forest Meteorology*, Agricultural prediction using climate model ensembles, 170 (March 15, 2013): 58–66. https://doi.org/10.1016/j.agrformet.2012.05.013.

Makate, Clifton, and Marshall Makate. "Interceding Role of Institutional Extension Services on the Livelihood Impacts of Drought Tolerant Maize Technology Adoption in Zimbabwe." *Technology in Society* 56 (February 1, 2019): 126–33. https://doi.org/10.1016/j.techsoc.2018.09.011.

Malosetti, Marcos, Jean-Marcel Ribaut, and Fred A. van Eeuwijk. "The Statistical Analysis of Multi-Environment Data: Modeling Genotype-by-Environment Interaction and Its Genetic Basis." *Frontiers in Physiology* 4 (March 12, 2013). https://doi.org/10.3389/fphys.2013.00044.

Meyer, Karin. "Factor-Analytic Models for Genotype × Environment Type Problems and Structured Covariance Matrices." *Genetics Selection Evolution* 41, no. 1 (January 30, 2009): 21. https://doi.org/10.1186/1297-9686-41-21.

Ogundari, Kolawole, and Olufemi D. Bolarinwa. “Impact of Agricultural Innovation Adoption: A Meta-Analysis.” *Australian Journal of Agricultural and Resource Economics* 62, no. 2 (2018): 217–36. https://doi.org/10.1111/1467-8489.12247.

Ornella, Leonardo, Gideon Kruseman, and Jose Crossa. “Satellite Data and Supervised Learning to Prevent Impact of Drought on Crop Production: Meteorological Drought.” *Drought - Detection and Solutions*, June 6, 2019. https://doi.org/10.5772/intechopen.85471.

Ornella, Leonardo & Cervigni, Gerardo & Tapia, Elizabeth. “Applications of Machine Learning for Maize Breeding.” In *Crop Stress and Its Management: Perspectives and Strategies*. Springer, 2013. https://www.researchgate.net/publication/235952888_Applications_of_Machine_Learning_for_Maize_Breeding.

Pan, Yao, Stephen C. Smith, and Munshi Sulaiman. “Agricultural Extension and Technology Adoption for Food Security: Evidence from Uganda.” *American Journal of Agricultural Economics* 100, no. 4 (July 1, 2018): 1012–31. https://doi.org/10.1093/ajae/aay012.

Pérez-Rodríguez, Paulino, José Crossa, Jessica Rutkoski, Jesse Poland, Ravi Singh, Andrés Legarra, Enrique Autrique, Gustavo de los Campos, Juan Burgueño, and Susanne Dreisigacker. “Single-Step Genomic and Pedigree Genotype × Environment Interaction Models for Predicting Wheat Lines in International Environments.” *The Plant Genome* 10, no. 2 (2017): plantgenome2016.09.0089. https://doi.org/10.3835/plantgenome2016.09.0089.

Piepho, H.-P. “Analyzing Genotype-Environment Data by Mixed Models with Multiplicative Terms.” *Biometrics* 53, no. 2 (1997): 761–66. https://doi.org/10.2307/2533976.

Poku, Adu-Gyamfi, Regina Birner, and Saurabh Gupta. “Why Do Maize Farmers in Ghana Have a Limited Choice of Improved Seed Varieties? An Assessment of the Governance Challenges in Seed Supply.” *Food Security* 10, no. 1 (February 1, 2018): 27–46. https://doi.org/10.1007/s12571-017-0749-0.

Ramirez-Villegas, Julian, Anabel Molero Milan, Nickolai Alexandrov, Senthold Asseng, Andrew J. Challinor, Jose Crossa, Fred van Eeuwijk, et al. “CGIAR Modeling

Approaches for Resource-Constrained Scenarios: I. Accelerating Crop Breeding for a Changing Climate." *Crop Science* n/a, no. n/a. Accessed April 12, 2020. https://doi.org/10.1002/csc2.20048.

Ramstein, Guillaume P., Sarah E. Jensen, and Edward S. Buckler. "Breaking the Curse of Dimensionality to Identify Causal Variants in Breeding 4." *Theoretical and Applied Genetics* 132, no. 3 (March 1, 2019): 559–67. https://doi.org/10.1007/s00122-018-3267-3.

Raschka, Sebastian, and Vahid Mirjalili. *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2, 3rd Edition*. Packt Publishing Ltd, 2019.

Reynolds, Matthew, John Foulkes, Robert Furbank, Simon Griffiths, Julie King, Erik Murchie, Martin Parry, and Gustavo Slafer. "Achieving Yield Gains in Wheat." *Plant, Cell & Environment* 35, no. 10 (October 2012): 1799–1823. https://doi.org/10.1111/j.1365-3040.2012.02588.x.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-Agnostic Interpretability of Machine Learning." *ArXiv:1606.05386 [Cs, Stat]*, June 16, 2016. http://arxiv.org/abs/1606.05386.

Ritchie, Hannah, and Max Roser. "Crop Yields." *Our World in Data*, October 17, 2013. https://ourworldindata.org/crop-yields.

Rudin, Cynthia. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1, no. 5 (May 2019): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

Scherr, Sara J, and Jeffrey A McNeely. "Biodiversity Conservation and Agricultural Sustainability: Towards a New Paradigm of 'Ecoagriculture' Landscapes." *Philosophical Transactions of the Royal Society B: Biological Sciences* 363, no. 1491 (February 12, 2008): 477–94. https://doi.org/10.1098/rstb.2007.2165.

Shekoofa, Avat, Yahya Emam, Navid Shekoufa, Mansour Ebrahimi, and Esmaeil Ebrahimie. "Determining the Most Important Physiological and Agronomic Traits Contributing to Maize Grain Yield through Machine Learning Algorithms: A New Avenue in Intelligent Agriculture." *PLoS ONE* 9, no. 5 (May 15, 2014). https://doi.org/10.1371/journal.pone.0097288.

Shelestov, Andrii, Mykola Lavreniuk, Nataliia Kussul, Alexei Novikov, and Sergii Skakun. “Exploring Google Earth Engine Platform for Big Data Processing: Classification of Multi-Temporal Satellite Imagery for Crop Mapping.” *Frontiers in Earth Science* 5 (2017). https://doi.org/10.3389/feart.2017.00017.

Smith, A. B., B. R. Cullis, and R. Thompson. “The Analysis of Crop Cultivar Breeding and Evaluation Trials: An Overview of Current Mixed Model Approaches.” *The Journal of Agricultural Science* 143, no. 6 (December 2005): 449–62. https://doi.org/10.1017/S0021859605005587.

Sukumaran, Sivakumar, Jose Crossa, Diego Jarquín, and Matthew Reynolds. “Pedigree-Based Prediction Models with Genotype × Environment Interaction in Multienvironment Trials of CIMMYT Wheat.” *Crop Science* 57, no. 4 (2017): 1865–80. https://doi.org/10.2135/cropsci2016.06.0558.

Tilman, David, Christian Balzer, Jason Hill, and Belinda L. Befort. “Global Food Demand and the Sustainable Intensification of Agriculture.” *Proceedings of the National Academy of Sciences* 108, no. 50 (December 13, 2011): 20260–64. https://doi.org/10.1073/pnas.1116437108.

Trnka, Miroslav, Song Feng, Mikhail A. Semenov, Jørgen E. Olesen, Kurt Christian Kersebaum, Reimund P. Rötter, Daniela Semerádová, et al. “Mitigation Efforts Will Not Fully Alleviate the Increase in Water Scarcity Occurrence Probability in Wheat-Producing Areas.” *Science Advances* 5, no. 9 (September 1, 2019): eaau2406. https://doi.org/10.1126/sciadv.aau2406.

Tscharntke, Teja, Yann Clough, Thomas C. Wanger, Louise Jackson, Iris Motzke, Ivette Perfecto, John Vandermeer, and Anthony Whitbread. “Global Food Security, Biodiversity Conservation and the Future of Agricultural Intensification.” *Biological Conservation*, ADVANCING ENVIRONMENTAL CONSERVATION: ESSAYS IN HONOR OF NAVJOT SODHI, 151, no. 1 (July 1, 2012): 53–59. https://doi.org/10.1016/j.biocon.2012.01.068.

U.S. Agency for International Development. “USAID’s Legacy in Agriculture Development: 50 Years of Progress,” 2016, 182.

U.S. Wheat Associates. “First Look at 2019/20 by USDA Sees Another Record World Wheat Crop.” Accessed May 10, 2020.

https://www.uswheat.org/wheatletter/first-look-at-2019-20-by-usda-sees-another-record-world-wheat-crop/.

Wu, Xiao-Lin, Timothy M. Beissinger, Stewart Bauck, Brent Woodwart, Guilherme J. M. Rosa, Natalia De Leon Gatti, Kent A. Weigel, and Daniel Gianola. "A Primer on High-Throughput Computing for Genomic Selection." *Frontiers in Genetics* 2 (2011). https://doi.org/10.3389/fgene.2011.00004.

Yao, Yiqi, and Alejandro Ochoa. "Testing the Effectiveness of Principal Components in Adjusting for Relatedness in Genetic Association Studies." *BioRxiv*, November 29, 2019, 858399. https://doi.org/10.1101/858399.

Data Citations

CIMMYT Wheat Yield Trials

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2017, "1st to 10th Semi-Arid Wheat Yield Trial", hdl:11529/10876, CIMMYT Research Data & Software Repository Network, V5, UNF:6:rRrYxa+IhE8+L8MKihpgYA==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "11th Semi-Arid Wheat Yield Trial", hdl:11529/10548300, CIMMYT Research Data & Software Repository Network, V1, UNF:6:2AsD1Ezx3/JR+Trtk2O6tg==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "12th Semi-Arid Wheat Yield Trial", hdl:11529/10548301, CIMMYT Research Data & Software Repository Network, V1, UNF:6:XXkUjAuLkD6SBWAaB05Lzw==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2015, "13th Semi-Arid Wheat Yield Trial", hdl:11529/10314, CIMMYT Research Data & Software Repository Network, V5, UNF:6:lVms0QkojFCuDG4GjBkpng==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2015, "14th Semi-Arid Wheat Yield Trial", hdl:11529/10297, CIMMYT Research Data & Software Repository Network, V5, UNF:6:nEWfU0mDxAP4M8u6qecCkw==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2015, "15th Semi-Arid Wheat Yield Trial", hdl:11529/10293, CIMMYT Research Data & Software Repository Network, V5, UNF:6:GpJjMITzRqWx0dDjy7XQlQ==

Global Wheat Program ; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2015, "16th Semi-Arid Wheat Yield trial", hdl:11529/10271, CIMMYT Research Data & Software Repository Network, V4, UNF:6:2lbCGiLYUrZRETqNgyC2/A==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2015, "17th Semi-Arid Wheat Yield Trial", hdl:11529/10265, CIMMYT Research Data & Software Repository Network, V7, UNF:6:0Gb6ZXprLH8zWcvceTb08A==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2015, "18th Semi-Arid Wheat Yield Trial", hdl:11529/10170, CIMMYT Research Data & Software Repository Network, V5, UNF:6:LP57zSyliPbQl0AMxkm+Kg==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2015, "19th Semi-Arid Wheat Yield Trial", hdl:11529/10337, CIMMYT Research Data & Software Repository Network, V6, UNF:6:7UpjEr5Ti85KlcalnMX7xw==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2015, "20th Semi-Arid Wheat Yield Trial", hdl:11529/10338, CIMMYT Research Data & Software Repository Network, V4, UNF:6:LtJpovMQ4KSQoc0zrHJtmw==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2015, "21st Semi-Arid Wheat Yield Trial", hdl:11529/10340, CIMMYT Research Data & Software Repository Network, V3, UNF:6:+A1t+N10NGp7MWYMiEkVZQ==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2017, "22nd Semi-Arid Wheat Yield Trial", hdl:11529/10985, CIMMYT Research Data & Software Repository Network, V2, UNF:6:7C5NKTR3fJpBaxBEABmJ4Q==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2017, "23rd Semi-Arid Wheat Yield Trial", hdl:11529/10987, CIMMYT Research Data & Software Repository Network, V3, UNF:6:IS3hQGvnKqnNfzPJVQ9AnA==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2018, "24th Semi-Arid Wheat Yield Trial", hdl:11529/10548042, CIMMYT Research Data & Software Repository Network, V2, UNF:6:29VJH9ankKMpV/+vs202aA==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2018, "25th Semi-Arid Wheat Yield Trial", hdl:11529/10548135, CIMMYT Research Data & Software Repository Network, V2

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "26th Semi-Arid Wheat Yield Trial", hdl:11529/10548302, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2017, "1st to 23rd Elite Selection Wheat Yield Trial", hdl:11529/10893, CIMMYT Research Data & Software Repository Network, V4, UNF:6:kFCZwf78DKq1+K2DvEaIHA==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "24th Elite Selection Wheat Yield Trial", hdl:11529/10548346, CIMMYT Research Data & Software Repository Network, V1, UNF:6:pdyKMSqyX8CgDbmb3WZLmw==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "25th Elite Selection Wheat Yield Trial", hdl:11529/10548347, CIMMYT Research Data & Software Repository Network, V1, UNF:6:cgprVTYeqOlys+SjUXnzLg==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "26th Elite Selection Wheat Yield Trial", hdl:11529/10548349, CIMMYT Research Data & Software Repository Network, V1, UNF:6:kh85pa2m0pQqJZxbz9XpwA==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "27th Elite Selection Wheat Yield Trial", hdl:11529/10548350, CIMMYT Research Data & Software Repository Network, V1, UNF:6:sZEikO35KXcDcRoDJo5Kdg==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "28th Elite Selection Wheat Yield Trial", hdl:11529/10548351, CIMMYT Research Data & Software Repository Network, V1, UNF:6:Lmphi0Y6/8imiVpB1RNBTQ==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2017, "29th Elite Selection Wheat Yield Trial", hdl:11529/10397, CIMMYT Research Data & Software Repository Network, V3, UNF:6:/xW09fAuTsurYXhEjFYoqQ==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2017, "30th Elite Selection Wheat Yield Trial", hdl:11529/10404, CIMMYT Research Data & Software Repository Network, V3, UNF:6:eqPxHXkap6tCjpMDNP1jqQ==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2017, "31st Elite Selection Wheat Yield Trial", hdl:11529/10400, CIMMYT Research Data & Software Repository Network, V3, UNF:6:8HifOUvdKaYIV54jt3hwoQ==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2017, "32nd Elite Selection Wheat Yield Trial", hdl:11529/10401, CIMMYT Research Data & Software Repository Network, V5, UNF:6:eJDiQfs/zzd+0KFjESHdOA==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2017, "33rd Elite Selection Wheat Yield Trial", hdl:11529/10402, CIMMYT Research Data & Software Repository Network, V3, UNF:6:f2xZFwPp4UZugokcov1w6Q==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2017, "34th Elite Selection Wheat Yield Trial", hdl:11529/10403, CIMMYT Research Data & Software Repository Network, V3, UNF:6:AfZvLHe7tead5GJ77pm+NQ==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2017, "35th Elite Selection Wheat Yield Trial", hdl:11529/10988, CIMMYT Research Data & Software Repository Network, V4, UNF:6:2IKKDwQUC+x31JkKRTnqjA==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2018, "36th Elite Selection Wheat Yield Trial", hdl:11529/10989, CIMMYT Research Data & Software Repository Network, V3, UNF:6:kLxFsNDa8qNxvzNvutGQiQ==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2018, "37th Elite Selection Wheat Yield Trial", hdl:11529/10548041, CIMMYT Research Data & Software Repository Network, V5, UNF:6:N03p5u1IyxfClCfBOWK+Vw==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "38th Elite Selection Wheat Yield Trial", hdl:11529/10548343, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "39th Elite Selection Wheat Yield Trial", hdl:11529/10548345, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "1st to 10th High Rainfall Wheat Yield Trial", hdl:11529/10548195, CIMMYT Research Data & Software Repository Network, V4, UNF:6:1bUNVa1OpwFsogRFMSF4vQ==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "11th High Rainfall Wheat Yield Trial", hdl:11529/10548202, CIMMYT Research Data & Software Repository Network, V2, UNF:6:8B2lZ7f/GAnAmOz9N/I9GA==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "12th High Rainfall Wheat Yield Trial", hdl:11529/10548204, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "13th High Rainfall Wheat Yield Trial", hdl:11529/10548205, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "14th High Rainfall Wheat Yield Trial", hdl:11529/10548206, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "15th High Rainfall Wheat Yield Trial", hdl:11529/10548207, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "16th High Rainfall Wheat Yield Trial", hdl:11529/10548208, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "17th High Rainfall Wheat Yield Trial", hdl:11529/10548209, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "18th High Rainfall Wheat Yield Trial", hdl:11529/10548210, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "19th High Rainfall Wheat Yield Trial", hdl:11529/10548211, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "20th High Rainfall Wheat Yield Trial", hdl:11529/10548212, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "21st High Rainfall Wheat Yield Trial", hdl:11529/10548213, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "22nd High Rainfall Wheat Yield Trial", hdl:11529/10548223, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "23rd High Rainfall Wheat Yield Trial", hdl:11529/10548224, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "24th High Rainfall Wheat Yield Trial", hdl:11529/10548225, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "25th High Rainfall Wheat Yield Trial", hdl:11529/10548226, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "26th High Rainfall Wheat Yield Trial", hdl:11529/10548227, CIMMYT Research Data & Software Repository Network, V1

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "1st to 10th High Temperature Wheat Yield Trial", hdl:11529/11089, CIMMYT Research Data & Software Repository Network, V3, UNF:6:LSthho4eES7Tm960eaiEwA==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "11th High Temperature Wheat Yield Trial", hdl:11529/10548246, CIMMYT Research Data & Software Repository Network, V2, UNF:6:oJ2DuGSy9ABipG56/5AKRQ==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "12th High Temperature Wheat Yield Trial", hdl:11529/10548247, CIMMYT Research Data & Software Repository Network, V2, UNF:6:qOQyhzQDRLvWPpWkPYia1A==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "13th High Temperature Wheat Yield Trial", hdl:11529/10548193, CIMMYT Research Data & Software Repository Network, V2, UNF:6:Vxdy7e0lWhMPXFSREyGA5A==

Global Wheat Program; IWIN Collaborators; Singh, Ravi ; Payne, Thomas, 2019, "14th High Temperature Wheat Yield Trial", hdl:11529/10548192, CIMMYT Research Data & Software Repository Network, V2, UNF:6:nUT7R+5rJbBpoezCfASFFg==

Global Wheat Program; IWIN Collaborators ; Singh, Ravi; Payne, Thomas, 2019, "15th High Temperature Wheat Yield Trial", hdl:11529/10548063, CIMMYT Research Data & Software Repository Network, V2, UNF:6:8YdQ7RxCYkOZnfwm/eVM2Q==

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "16th High Temperature Wheat Yield Trial", hdl:11529/10548194, CIMMYT Research Data & Software Repository Network, V2

Global Wheat Program; IWIN Collaborators; Singh, Ravi; Payne, Thomas, 2019, "17th High Temperature Wheat Yield Trial", hdl:11529/10548314, CIMMYT Research Data & Software Repository Network, V2

Google Earth Engine Data

Abatzoglou, John, Solomon Dobrowski, Sean Parks, and Katherine Hegewisch. "Monthly Climate and Climatic Water Balance for Global Terrestrial Surfaces from 1958-2015," 2017. https://doi.org/10.7923/G43J3B0R.

Ebita, Ayataka, Shinya Kobayashi, Yukinari Ota, Masami Moriya, Ryoji Kumabe, Kazutoshi Onogi, Yayoi Harada, et al. "The Japanese 55-Year Reanalysis 'JRA-55': An Interim Report." *SOLA* 7 (2011): 149–52. https://doi.org/10.2151/sola.2011-038.

Farr, Tom G., Paul A. Rosen, Edward Caro, Robert Crippen, Riley Duren, Scott Hensley, Michael Kobrick, et al. "The Shuttle Radar Topography Mission." *Reviews of Geophysics* 45, no. 2 (May 19, 2007): RG2004. https://doi.org/10.1029/2005RG000183.

Google Developers. "ERA5 Daily Aggregates - Latest Climate Reanalysis Produced by ECMWF / Copernicus Climate Change Service." Accessed April 12, 2020. https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_DAILY.

Google Developers. "FLDAS: Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System." Accessed April 12, 2020.

https://developers.google.com/earth-engine/datasets/catalog/NASA_FLDAS_NOAH01_C_GL_M_V001.

Google Developers. "GLDAS-2.1: Global Land Data Assimilation System." Accessed April 12, 2020. https://developers.google.com/earth-engine/datasets/catalog/NASA_GLDAS_V021_NOAH_G025_T3H.

Google Developers. "SRTM Digital Elevation Data 30m | Earth Engine Data Catalog." Accessed May 9, 2020. https://developers.google.com/earth-engine/datasets/catalog/USGS_SRTMGL1_003.

Google Developers. "TerraClimate: Monthly Climate and Climatic Water Balance for Global Terrestrial Surfaces, University of Idaho." Accessed April 12, 2020. https://developers.google.com/earth-engine/datasets/catalog/IDAHO_EPSCOR_TERRACLIMATE.

Harris, Ian, Timothy J. Osborn, Phil Jones, and David Lister. "Version 4 of the CRU TS Monthly High-Resolution Gridded Multivariate Climate Dataset." *Scientific Data* 7, no. 1 (April 3, 2020): 1–18. https://doi.org/10.1038/s41597-020-0453-3.

McNally, Amy, Kristi Arsenault, Sujay Kumar, Shraddhanand Shukla, Pete Peterson, Shugong Wang, Chris Funk, Christa D. Peters-Lidard, and James P. Verdin. "A Land Data Assimilation System for Sub-Saharan Africa Food and Water Security Applications." *Scientific Data* 4, no. 1 (December 2017): 170012. https://doi.org/10.1038/sdata.2017.12.

"ERA5 Hourly Data on Single Levels from 1979 to Present." Accessed May 9, 2020. https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview.

"WorldClim - Global Climate Data (WORLDCLIM) | Land & Water | Food and Agriculture Organization of the United Nations | Land & Water | Food and Agriculture Organization of the United Nations." Accessed May 9, 2020. http://www.fao.org/land-water/land/land-governance/land-resources-planning-toolbox/category/details/en/c/1043064/.

ICIS Coefficient of Parentage Matrix

"ICISWiki." Accessed May 10, 2020. https://cropforge.github.io/iciswiki/index.html.

Software Citations

Emerson, John W, and Michael J Kane. "The R Package Bigmemory: Supporting Efficient Computation and Concurrent Programming with Large Data Sets." *Journal of Statistical Software*, n.d., 16.

Gorelick, Noel, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. "Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone." *Remote Sensing of Environment*, Big Remotely Sensed Data: tools, applications and experiences, 202 (December 1, 2017): 18–27. https://doi.org/10.1016/j.rse.2017.06.031.

Jeff Reback, Wes McKinney, jbrockmendel, Joris Van den Bossche, Tom Augspurger, Phillip Cloud, Mortada Mehyar. (2020, March 18). pandas-dev/pandas: Pandas 1.0.3 (Version v1.0.3). Zenodo. http://doi.org/10.5281/zenodo.3715232

Kingma, Diederik P., and Jimmy Ba. "Adam: A Method for Stochastic Optimization." *ArXiv:1412.6980 [Cs]*, January 29, 2017. http://arxiv.org/abs/1412.6980.

Pérez-Rodríguez, Paulino. "BGLR: A Statistical Package for Whole Genome Regression and Prediction," n.d., 30.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Wang, Qian, Xianyi Zhang, Yunquan Zhang, and Qing Yi. "AUGEM: Automatically Generate High Performance Dense Linear Algebra Kernels on X86 CPUs." In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis on - SC '13*, 1–12. Denver, Colorado: ACM Press, 2013. https://doi.org/10.1145/2503210.2503219.

Xianyi, Zhang, Wang Qian, and Zhang Yunquan. "Model-Driven Level 3 BLAS Performance Optimization on Loongson 3A Processor." In *2012 IEEE 18th*

International Conference on Parallel and Distributed Systems, 684–91, 2012. https://doi.org/10.1109/ICPADS.2012.97.

"AI Platform | Google Cloud." Accessed April 23, 2020. https://cloud.google.com/ai-platform.

"Compute Engine: Virtual Machines (VMs) | Google Cloud." Accessed April 23, 2020. https://cloud.google.com/compute.

Appendices

Appendix 1 - GitHub Repository of Notebook Files

A complete GitHub repository of the notebooks and source data necessary to replicate this project are available publicly at the following URL: https://github.com/AaronScherf/wheat_yield_prediction_gee

The repository will be updated to include the interim data, figures, and model results, as well as a copy of this report.

Appendix 2 - Links to Data Sources

All CIMMYT wheat yield trial data is available on their public dataverse, hosted at the time of writing at the following URL: https://data.cimmyt.org/dataverse/cimmytdatadvn

The Earth Engine data is linked in Table 3, repeated below. In order to query the specific data for the locations and dates in this project, use the following notebook: https://github.com/AaronScherf/wheat_yield_prediction_gee/blob/master/notebooks/4_query_GEE_server.ipynb

Table 3: Google Earth Engine Raster Datasets and Variables

Spatial Resolution	Dataset Name	Date range	Example Variables
30 meters (0.016 arc minutes)	NASA SRTM Digital Elevation Data	2000	• Altitude
2.5 arc minutes (4.63km)	TerraClimate: Monthly Climate and Climatic Water Balance for Global Terrestrial Surfaces, University of Idaho	1958 to present	• Actual evapotranspiration • Climate water deficit • Palmer Drought Severity Index • Precipitation accumulation • Soil moisture • Vapor pressure
0.1 arc degrees (6 arc minutes, 11.1km)	FLDAS: Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System	1982 - present	• Evapotranspiration • Surface pressure • Total precipitation rate • Snowfall rate • Baseflow-groundwater runoff • Near-surface air temperature • Soil heat flux • Soil temperature
0.25 arc degrees (15 arc	ERA5 Daily aggregates - Latest climate reanalysis produced by	1979 - present	• Maximum air temperature at 2m • Average air temperature at 2m • Minimum air temperature at 2m

minutes, 27.78km)	ECMWF / Copernicus Climate Change Service		• Dewpoint temperature at 2m • Total precipitation • Surface pressure

Finally, the ICIS coefficient of parentage data was not available on their website and instead had to be requested from a scientist at CIMMYT, Dr. Jose Crossa. The author has contacted ICIS to inquire about a publicly accessible version of the dataset. Until then, the data is available in the GitHub repo in the "A_matrix.csv" file.

Appendix 3 - Research Dissemination Plan

The ultimate objective of this project is to inform crop selection modelling and its application for the distribution of improved seed varieties. As such, it targets a range of stakeholders, each of which must be addressed through different channels of communication. The utility of the information ultimately delivered to each user must also be considered, so that each stakeholder receives the component of the project most relevant to their needs without excluding them from accessing and understanding the entire process. In considering this challenge, the project's author has identified four main sets of stakeholders: crop breeding researchers, organizations funding those researchers, public agricultural regulators, and agricultural extension officers.

While other users may find the work useful or interesting, these three groups seemed most relevant to the impact objective of increasing intensive crop yields for farmers in historically less productive regions. While individual farmers or farming organizations may benefit from this research, they were not considered as a direct user, but rather as an indirect stakeholder affected through their relationships with the three groups above.

In most cases, crop breeding research is funded by governments, multilateral organizations, or private companies. The researchers in turn work with public regulators, extension officers, and branches of agricultural supply firms to disseminate their improved varieties to farmers, for further testing and deployment. The entire system is designed to maximize benefits for farmers and improve their production, ultimately generating greater agricultural products for consumers around the world. As van Bueren et al (2018) point out, however, this concentration of research and development activities in a few multinational corporations and multilateral organizations introduces barriers to dissemination, either from intentional withholding of information through intellectual property protections or incidental inefficiencies of communication.

This project seeks to improve the accessibility of predictive modelling for researchers working for public or non-profit entities, such as CIMMYT, who share their results openly. While these resources may be available, it does not necessarily mean they are accessible. The sheer number of potential seed varieties, the complex nature of their interactions with constantly shifting environments, and the overlapping sources of information create a large and complex decision space for regulators and extension officers. Since one of the major gaps identified in delivering improved varieties to

farmers is the deployment of publicly available genotypes to regions with less funding available for internal crop research, this dissemination strategy will focus on open-access platforms, ideally made available at little or no cost to the users. The goal is to enable less resourced stakeholders at the national and local level to make more informed decisions on which seed varieties to approve and distribute to farmers.

Research Funding Organizations

Developing and testing new varieties of crops requires extensive resources and expertise as well as large-scale international cooperation. The economies of scale of research activities--whereby the best researchers and most funding accrue to organizations with consistent results and a strong public mandate--tend naturally towards monopolization. Private crop breeding companies such as Bayer, Syngenta and Corteva have concentrated the market for seed varieties through patented genetic information and sterile hybrid seeds, which prevent farmers from reproducing their own seed stock. Publicly funded breeding organizations, on the other hand, have a multilateral or government mandate to distribute their seeds openly and place fewer restrictions on their reproduction. Some of the major public organizations funding breeding research include CGIAR, the Food and Agriculture Organization (FAO) of the United Nations, the International Fund for Agricultural Development (IFAD), and the U.S. Agency for International Development (USAID) Bureau for Resilience and Food Security.

Communicating the results of this project to research funding organizations requires an individualized, high-touch outreach strategy. The nodes of information sharing and decision making in these organizations are highly centralized, meaning that if a few individuals distribute the results then the information will be widely available. As such, the author will pursue direct communication via email or in-person meetings with representatives from these organizations. A brief document summarizing the results in an accessible and visual format can be distributed with links to the full paper, code and data resources, and any derivative visualization tools. There may also be opportunities to present to key sub-committees within these organizations, to explain in greater detail the potential of open-source analysis, remotely sensed data, and machine learning approaches for crop research. Many of these key individuals are also likely to attend academic or trade conferences and read their journals, so presentations and submission for publication offer a more diffuse communication opportunity.

The intended impacts of the research dissemination on breeding funding organizations include: the adoption or enforcement of open-source data and reproducible methods requirements for funded projects, the adoption of remote sensing methods for data collection in crop trials and impact evaluations, and the use of predictive algorithms for seed variety matching to locations based on environmental conditions.

Crop Breeding Researchers

The development and testing of new seed varieties clearly takes highly specialized expertise, cooperation between disparate research and test sites, and continuously evolving knowledge on which varieties and traits are contributing to crop productivity

and resilience. It is thus unsurprising that research activities can be distributed among several groups, organized by crop type, regional focus, or funding source. Researchers working for private seed development and supply companies are organizationally more proximal to their stakeholders; publicly funded researchers, on the other hand, must navigate a large network of research groups and distribution channels. Many of the most well-known global organizations are extensions of CGIAR, including CIMMYT, the International Rice Research Institute (IRRI), the International Center for Agricultural Research in the Dry Areas (ICARDA), the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), the International Food Policy Research Institute (IFPRI), and Africa Rice. Other notable organizations include the European Association for Research on Plant Breeding, the Indian Council of Agricultural Research (ICAR), the Max Planck Institute for Plant Breeding Research, and many national level research centers, such as the John Ines Center in the UK, the Bangladesh Agricultural Research Institute (BARI), and the Crop Research Institute in Ghana.

Crop research and breeding organizations are more diffuse than their funders, making an individualized outreach strategy more difficult. The scientists working for these organizations are the primary stakeholder group of interest, however, and they can be reached through internal newsletters, leading journals, and conferences. Individual emails may also be effective, though they will likely hold more weight if distributed through funding institutions or other notable researchers.

The intended impacts of the dissemination of this project on crop breeding researchers include: the adoption of open-source data sharing and reproducible analysis code, the adoption of remote sensing data for improved predictive modelling and lower cost field trial data collection, and the consideration of machine learning modelling approaches as a replacement for complicated linear mixed models. The aggregated data used for this project may also be of interest to researchers focused on wheat, particularly at CIMMYT. The method for querying remote sensing data from Google Earth Engine may also be useful for practitioners with less experience in satellite data or JavaScript coding. The coding notebooks and reproduction workflow for the project, therefore, will be shared with these stakeholders, along with the paper and an abstract with key results.

Public Agricultural Regulators

Regional and national government regulators of agricultural products and seed varieties play an essential role in protecting public safety and food sovereignty. The decision on which seed varieties to allow or prioritize for distribution is often made by a few key individuals, based on available scientific evidence. These stakeholders may not always have access to updated information on which varieties are most likely to optimize agricultural outcomes given current or expected environmental conditions. Providing trustworthy information on environmental forecasts and the response of different seeds tailored to their constituent geography is thus vital for their decision making.

Regulatory authority is often controlled by the ministry of agriculture at the national level or in regional trade blocs such as the European Union or Southern African Development Community (SADC). The decision space for these regulators often involves complex trade-offs between potential benefits for farmers and consumers weighed against the potential risks of introducing new species. Political economic influences,

such as international relations or concerns over seed and food sovereignty, may also factor into the decision of which varieties to allow. Many countries are justifiably sceptical of private seed companies seeking to distribute new varieties, since there are few methods to hold these actors accountable beyond access to markets. Multilateral entities such as CIMMYT or national development agencies such as USAID may also be viewed with suspicion, depending on the position of the regulator's government in geopolitical systems.

There is no universal approach to distributing tools or information to such a diverse array of regulators and public servants. Emails to regulatory offices are not likely to be very effective unless accompanied by the reputational authority of a major research organization or funder. The most effective channel of communication for this project, therefore, is to rely on the research groups and their funders to introduce open-source analytical approaches that improve transparency with regulators. If reproducible and accessible research becomes more of a global standard, regulators may begin demanding complete access to the data and methods behind arguments for certain seed varieties. A standard methodological framework for assessing different seed varieties based on past field trials could thus represent a key resource for agricultural regulators to verify and adapt models to their needs.

Reduced informational asymmetries between researchers and regulators is also expected to contribute to the biodiversity and food sovereignty of less-developed countries, who currently may have limited power or influence over research methods. Researchers within these countries, who are limited by resource constraints, may also represent a key stakeholder group who can help advocate for more open data and methods, since they would be the most likely to benefit by offering interpretations or adaptations of publicly available models. The distribution of the data and code notebooks used by this project to researchers in less-resourced regions is therefore a key part of this dissemination strategy.

Agricultural Extension Officers

Agricultural extension offices connect national agriculture ministries and the actual farmers they serve. By distributing information on new technologies, practices, and seed varieties, extension officers form a key stakeholder group in improving the genetic stock of any region's crop production. The efficacy of extension officers for increased productivity and food security has been well studied (Pan et al, 2018; Makate and Makate, 2019) and has even been shown to affect the gender dynamics of an agricultural system (Gebre et al, 2019). Poku et al (2018) found that farmers in Ghana were not adopting new seed varieties even after they had been approved due to resource constraints in the agricultural regulator's office and an underdeveloped public extension system. Ensuring that there are enough extension officers reaching as many farmers as possible is a key component of any seed distribution strategy.

Extension officers are typically organized by the national ministry of agriculture or local branches thereof, so communicating with them as stakeholders would most easily be accomplished through the public ministries. Given the anticipated difficulty of communication, however, it seems more effective to equip the development organizations and non-profit groups which often assist extension officers with improved

information resources. Extension officers are not expected to have advanced technical abilities in carrying out crop breeding research; as such, it is critical to develop accessible tools that convey the benefits of different seed varieties without obfuscating them behind layers of complex programming.

A simplified digital map, describing the optimal seed varieties for each location in a region that are approved for use, could offer a streamlined decision-assistance tool for extension officers that is targeted to their geographic scope. If the genotype variety predictions made by this tool could be exported to a static file, like a PDF, it would be easier to share among stakeholders that do not have the time or interest in interactive digital tools. If the agricultural regulators or development organizations likewise lack the capacity to produce a geographically specific map from an interactive application, there could instead be a request tool to message a researcher who can assist them.

If agricultural extension officers can be integrated into a more targeted and efficient seed variety distribution system, the ultimate transfer of the varieties to farmers will be far more likely. By producing accurate, geographically specific, and constantly updated tools for researchers, funders, regulators, and extension officers to make decisions on which varieties to distribute where, this project hopes to improve the rate and coverage of the dissemination of publicly available seed genotypes.

Additional Potential Users

Beyond the stakeholders immediately involved in the seed distribution system, this project aims to make the entire field of crop breeding research more accessible to an array of potential users. Two audiences that the open-source and reproducible code may be of interest to are students and educators. Genotype-by-environment crop trials are an important aspect of crop breeding that is difficult to understand without readily available and realistic data. By providing an aggregated set of wheat yield trials and corresponding pedigree information, this project hopes to make it easier for future crop researchers to explore the process of predicting phenotypic traits.

In addition to its submission to the UC Berkeley Library, this project will be published online as an open-source GitHub repository and series of Google Colab notebooks. While some data, such as the coefficient of parentage (COP) matrix, is not available in a readily available format online, the author will request permission to host the data publicly or at least provide detailed instructions for potential users to acquire it. Once the data is included, the entire project should be replicable. For the sections which are more complicated or require longer runtimes, such as the scraping of GID values from the CIMMYT germplasm bank website or imputation of missing environmental data, the intermediate data files will also be provided to speed up the process of replication. The two notebooks which offer the most value for future adaptation or use--the query of Earth Engine data using a table of points and prediction of yield values using machine learning--will be adapted into a simplified format with clear documentation of the code. It is hoped that these tutorial notebooks can be used for similar applications or instructional purposes.

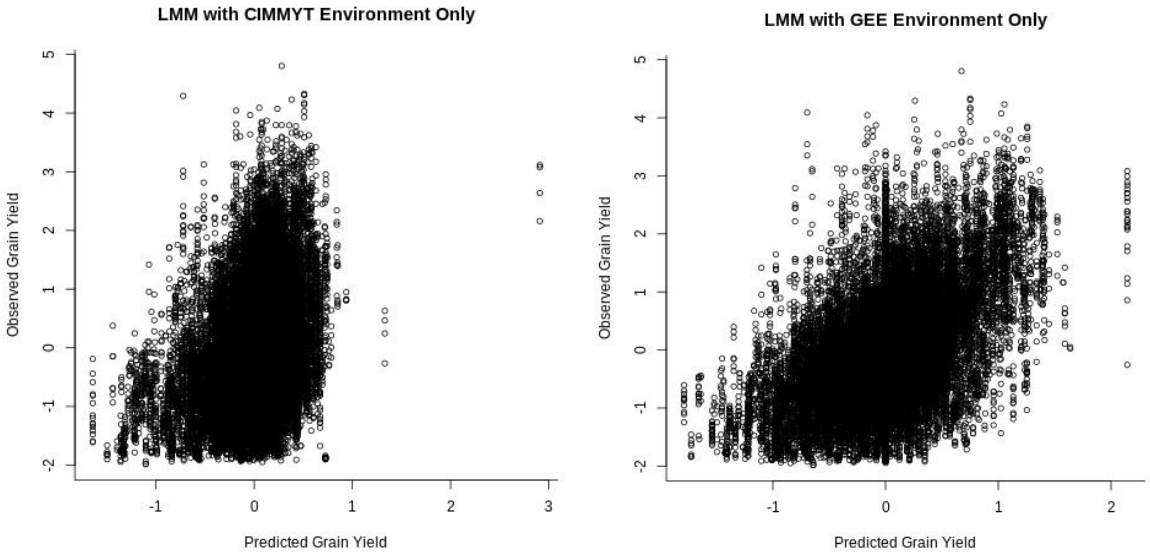
Appendix 4 - Prediction Model Results and Plots

Table 6: RMSE and R-Squared for LMM and ML Models for All Environment and Genotype Dataset Combinations

Holdout RMSE (R-Squared)		BGLR LMM GxE Interaction	Random Forest	XGBoost	Multi-Layer Perceptron Network
CIMMYT Env Data	Environment Data Only	0.941 (0.113)	0.414 (0.829)	0.423 (0.822)	0.419 (0.824)
	Env & Pedigree	0.928 (0.139)	**0.401 (0.839)**	0.423 (0.821)	0.456 (0.792)
GEE Env Data	Environment Data Only	0.869 (0.243)	0.504 (0.746)	0.507 (0.743)	0.514 (0.736)
	Env & Pedigree	0.859 (0.261)	0.509 (0.741)	0.499 (0.751)	0.509 (0.741)

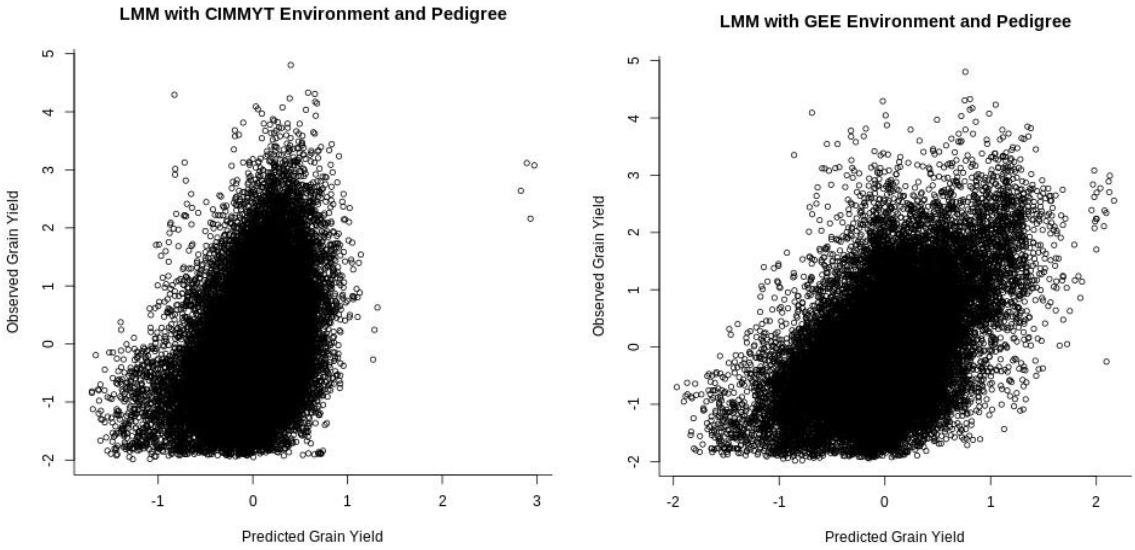
(Source: Aggregated CIMMYT Data and Google Earth Engine Data; Calculations by Author)

Figure 14 and 15: Scatterplots of Actual vs Predicted Test Values using a BGLR LMM Model with Environmental Variables using CIMMYT (Left) and Earth Engine (Right)



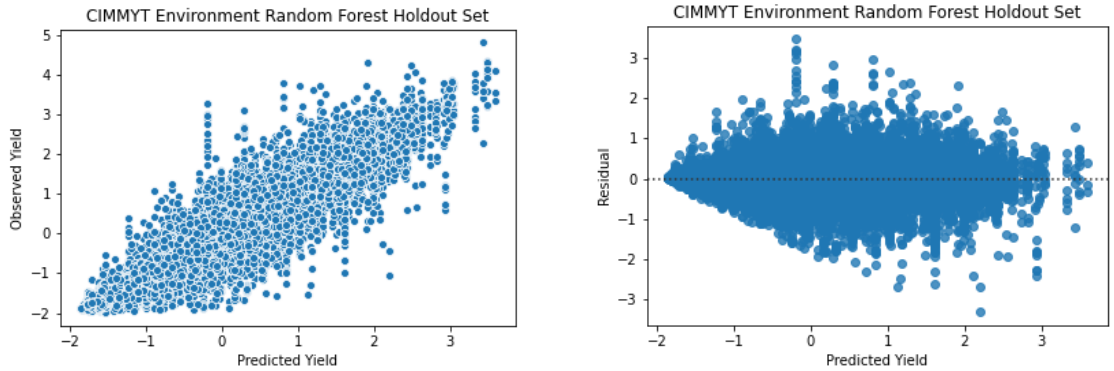
(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Figure 16 and 17: Scatterplots of Actual vs Predicted Test Values for Grain Yield using a BGLR LMM Model with Pedigree by Environment Interactions using CIMMYT Data (Left) and Earth Engine Data (Right)



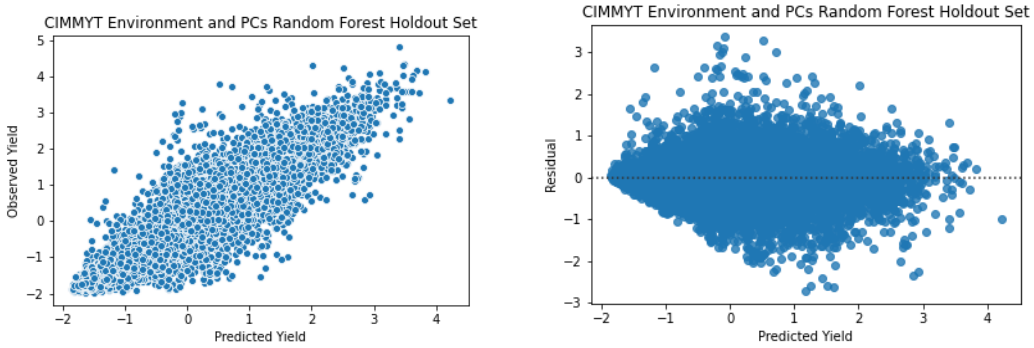
(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Figure 18 and 19: Predicted vs Actual Values of Wheat Yield for Random Forest Model using CIMMYT Environmental Data Scatter Plot (Left) and Residual Plot (Right)



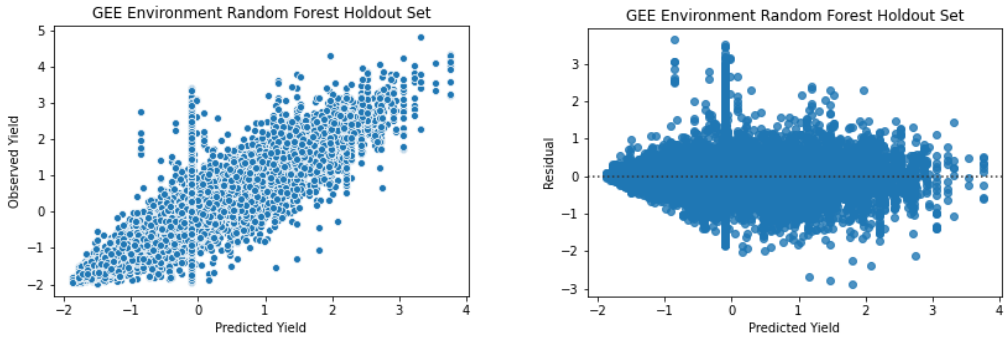
(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Figure 20 and 21: Predicted vs Actual Values of Wheat Yield for Random Forest Model using CIMMYT Environmental Data and Pedigree Principal Components Scatter Plot (Left) and Residual Plot (Right)



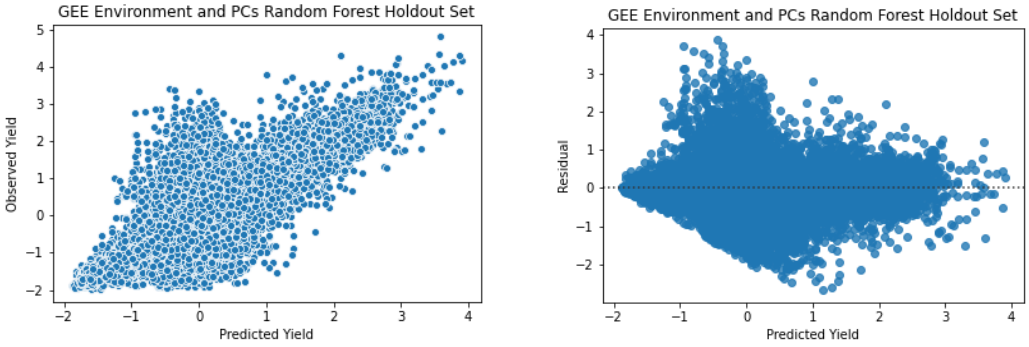
(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Figure 22 and 23: Predicted vs Actual Values of Wheat Yield for Random Forest Model using GEE Environmental Data Scatter Plot (Left) and Residual Plot (Right)



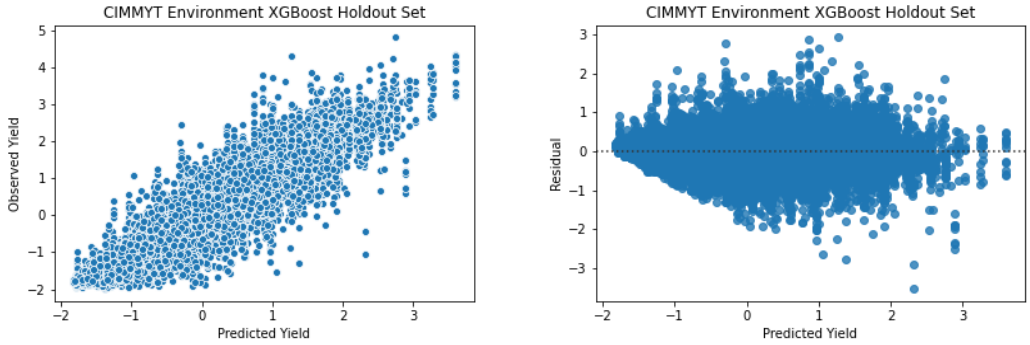
(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Figure 24 and 25: Predicted vs Actual Values of Wheat Yield for Random Forest Model using GEE Environmental Data and Pedigree Principal Components Scatter Plot (Left) and Residual Plot (Right)



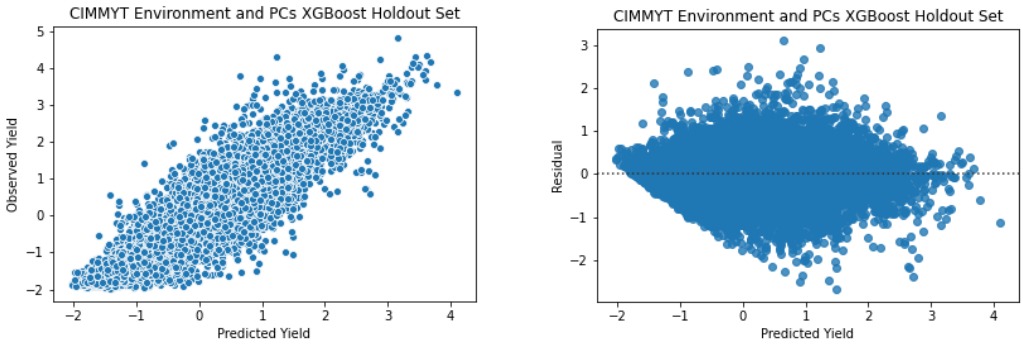
(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Figure 26 and 27: Predicted vs Actual Values of Wheat Yield for XGBoost Model using CIMMYT Environmental Data Scatter Plot (Left) and Residual Plot (Right)



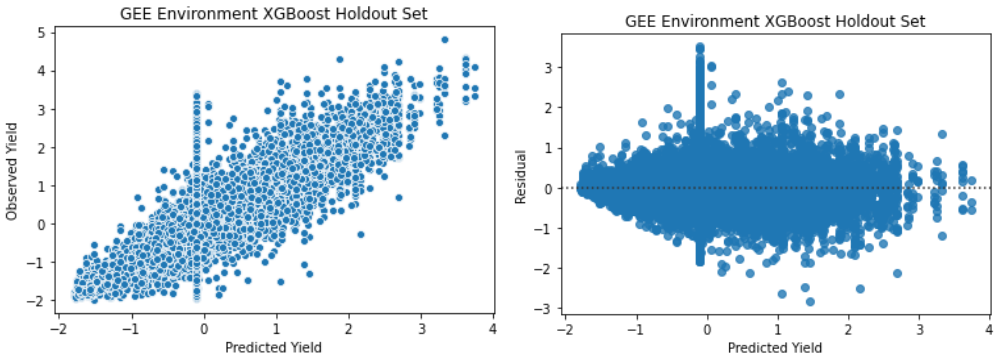
(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Figure 28 and 29: Predicted vs Actual Values of Wheat Yield for XGBoost Model using CIMMYT Environmental Data and Pedigree Principal Components Scatter Plot (Left) and Residual Plot (Right)



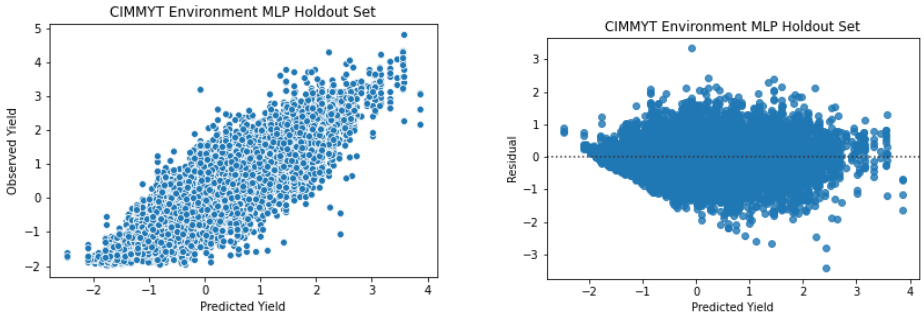
(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Figure 30 and 31: Predicted vs Actual Values of Wheat Yield for XGBoost Model using GEE Environmental Data Scatter Plot (Left) and Residual Plot (Right)



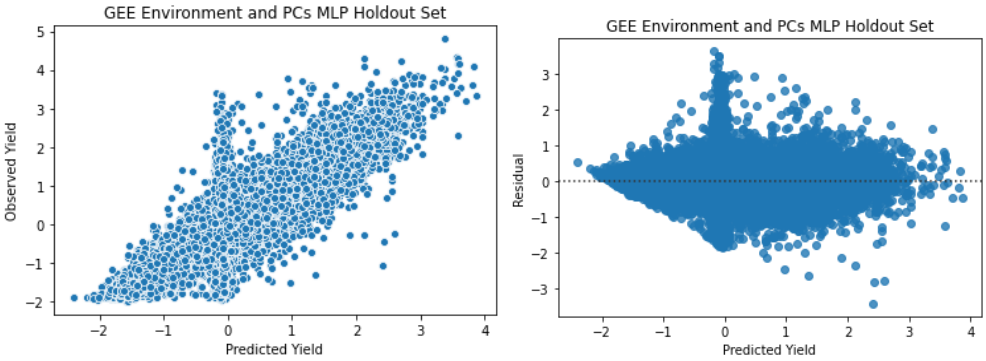
(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Figure 32 and 33: Predicted vs Actual Values of Wheat Yield for Multi-Layer Perceptron Model using CIMMYT Environmental Data Scatter Plot (Left) and Residual Plot (Right)



(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Figure 34 and 35: Predicted vs Actual Values of Wheat Yield for Multi-Layer Perceptron Model using GEE Environmental Data and Pedigree Principal Components Scatter Plot (Left) and Residual Plot (Right)



(Source: CIMMYT Wheat Trial Data and Google Earth Engine Data; Calculations by Author)

Appendix 5 - Feature Importance Values from Random Forest Model

Table 11: Feature Importance Values of Top Forty Features for CIMMYT Environmental Data in the Random Forest Model

Feature Importance for CIMMYT Environmental Data from Random Forest Model for Top 40 Features	
ALTITUDE	0.1614
SPACE_BTN_ROWS_SOWN	0.0691
LENGTH_OF_ROWS_SOWN	0.0509
LENGTH_OF_ROWS_HARVESTED	0.0496
PPN_11TH_MO_BEFORE_HARVESTED	0.0446
PPN_1ST_MO_BEFORE_HARVESTED	0.0442

PPN_7TH_MO_BEFORE_HARVESTED	0.0431
PPN_4TH_MO_BEFORE_HARVESTED	0.0426
PPN_3RD_MO_BEFORE_HARVESTED	0.0414
PPN_MONTH_OF_HARVESTED	0.0403
PPN_10TH_MO_BEFORE_HARVESTED	0.0395
PPN_9TH_MO_BEFORE_HARVESTED	0.0392
PPN_6TH_MO_BEFORE_HARVESTED	0.0372
PPN_8TH_MO_BEFORE_HARVESTED	0.0369
PRECIPITATION_FROM_SOWING_TO_MATURITY	0.0362
PPN_2ND_MO_BEFORE_HARVESTED	0.0339
PPN_5TH_MO_BEFORE_HARVESTED	0.0336
NO_OF_ROWS_HARVESTED	0.0312
TOTAL_PRECIPIT_IN_12_MONTHS	0.0216
NO_OF_ROWS_SOWN	0.0208
IRRIGATED_YES	0.0186
FOLIAR_DISEASE_DEVELOPMENT_TRACES	0.0055
LODGING_SLIGHT	0.0054
INSECT_DAMAGE_TRACES	0.0035
WEED_PROBLEM_TRACES	0.0035
WEED_PROBLEM_SLIGHT	0.0033
FOLIAR_DISEASE_DEVELOPMENT_SLIGHT	0.0033
LODGING_TRACES	0.0032
FERTILIZER_APPLIED_YES	0.0030
INSECT_DAMAGE_SLIGHT	0.0030
FOLIAR_DISEASE_DEVELOPMENT_MODERATE	0.0030
BIRD_DAMAGE_SLIGHT	0.0029
WEED_PROBLEM_MODERATE	0.0027
BIRD_DAMAGE_TRACES	0.0027
CROP_STAND_OR_DENSITY_DENSE	0.0023

ROOT_DISEASE_DEVELOPMENT_TRACES	0.0022
LODGING_MODERATE	0.0018
FOLIAR_DISEASE_DEVELOPMENT_SEVERE	0.0016
INSECT_DAMAGE_MODERATE	0.0012
ROOT_DISEASE_DEVELOPMENT_SLIGHT	0.0012

Table 12: Feature Importance Values of Top Forty Features for CIMMYT Environmental and Pedigree Data in the Random Forest Model

Feature Importance for CIMMYT Environmental Data with Principal Components from Random Forest Model for Top 40 Features	
ALTITUDE	0.1427
SPACE_BTN_ROWS_SOWN	0.0616
LENGTH_OF_ROWS_SOWN	0.0446
LENGTH_OF_ROWS_HARVESTED	0.0440
PPN_1ST_MO_BEFORE_HARVESTED	0.0392
PPN_11TH_MO_BEFORE_HARVESTED	0.0383
PPN_7TH_MO_BEFORE_HARVESTED	0.0379
PPN_4TH_MO_BEFORE_HARVESTED	0.0365
PPN_3RD_MO_BEFORE_HARVESTED	0.0360
PPN_MONTH_OF_HARVESTED	0.0350
PPN_10TH_MO_BEFORE_HARVESTED	0.0347
PPN_9TH_MO_BEFORE_HARVESTED	0.0340
PPN_6TH_MO_BEFORE_HARVESTED	0.0330
PPN_8TH_MO_BEFORE_HARVESTED	0.0324
PRECIPITATION_FROM_SOWING_TO_MATURITY	0.0322
PPN_2ND_MO_BEFORE_HARVESTED	0.0298
PPN_5TH_MO_BEFORE_HARVESTED	0.0293
NO_OF_ROWS_HARVESTED	0.0281
TOTAL_PRECIPIT_IN_12_MONTHS	0.0187

NO_OF_ROWS_SOWN	0.0182
IRRIGATED_YES	0.0163
FOLIAR_DISEASE_DEVELOPMENT_TRACES	0.0050
LODGING_SLIGHT	0.0047
INSECT_DAMAGE_TRACES	0.0031
PC_0	0.0031
WEED_PROBLEM_SLIGHT	0.0030
WEED_PROBLEM_TRACES	0.0030
FOLIAR_DISEASE_DEVELOPMENT_SLIGHT	0.0030
LODGING_TRACES	0.0028
INSECT_DAMAGE_SLIGHT	0.0027
FERTILIZER_APPLIED_YES	0.0026
PC_2	0.0026
FOLIAR_DISEASE_DEVELOPMENT_MODERATE	0.0025
BIRD_DAMAGE_SLIGHT	0.0024
BIRD_DAMAGE_TRACES	0.0023
WEED_PROBLEM_MODERATE	0.0022
PC_5	0.0022
CROP_STAND_OR_DENSITY_DENSE	0.0021
PC_3	0.0020
ROOT_DISEASE_DEVELOPMENT_TRACES	0.0019

Table 13: Feature Importance Values of Top Forty Features for GEE Environmental Data in the Random Forest Model

Feature Importance for GEE Environmental Data from Random Forest Model for Top 40 Features	
(SoilMoi10_40cm_tavg, 8)	0.1133
(vap, 4)	0.0602
(vap, 3)	0.0174

(vap, 7)	0.0169
(SoilMoi00_10cm_tavg, 0)	0.0167
(def, 8)	0.0163
(SoilMoi00_10cm_tavg, 8)	0.0153
(soil, 3)	0.0128
(soil, 5)	0.0098
(soil, 0)	0.0088
(Qg_tavg, 3)	0.0083
(soil, 6)	0.0079
(aet, 8)	0.0078
(Psurf_f_tavg, 10)	0.0067
(Qair_f_tavg, 1)	0.0067
(Qair_f_tavg, 12)	0.0067
(maximum_2m_air_temperature, 4)	0.0063
(Qg_tavg, 2)	0.0062
(total_precipitation, 1)	0.0062
(Evap_tavg, 6)	0.0057
(total_precipitation, 6)	0.0055
(Rainf_f_tavg, 5)	0.0055
(soil, 1)	0.0054
(aet, 11)	0.0052
(def, 2)	0.0052
(minimum_2m_air_temperature, 10)	0.0051
(def, 12)	0.0051
(total_precipitation, 0)	0.0050
(aet, 2)	0.0050
(Qg_tavg, 4)	0.0048
(total_precipitation, 7)	0.0047

(total_precipitation, 4)	0.0047
(vap, 8)	0.0046
(Rainf_f_tavg, 2)	0.0046
(total_precipitation, 8)	0.0046
(Qair_f_tavg, 2)	0.0044
(aet, 10)	0.0044
(total_precipitation, 3)	0.0043
(tmmx, 7)	0.0043
(Evap_tavg, 10)	0.0043

Table 13: Feature Importance Values of Top Forty Features for GEE Environmental and Pedigree Data in the Random Forest Model

Feature Importance for GEE Environmental Data with Principal Components from Random Forest Model for Top 40 Features	
(SoilMoi00_10cm_tavg, 8)	0.1012
(vap, 4)	0.0486
(Qair_f_tavg, 8)	0.0287
(SoilMoi00_10cm_tavg, 0)	0.0214
(def, 8)	0.0187
(vap, 3)	0.0161
(soil, 3)	0.0127
(soil, 0)	0.0111
(soil, 5)	0.0082
(aet, 2)	0.0078
(Psurf_f_tavg, 10)	0.0072
(def, 11)	0.0065
(Qg_tavg, 2)	0.0064
(tmmx, 7)	0.0063
(Qg_tavg, 4)	0.0061

(Qair_f_tavg, 1)	0.0061
(Rainf_f_tavg, 2)	0.0061
(Rainf_f_tavg, 5)	0.0057
(Qair_f_tavg, 12)	0.0055
(aet, 8)	0.0054
(aet, 11)	0.0053
(soil, 1)	0.0049
(Qg_tavg, 3)	0.0049
(soil, 11)	0.0048
(Evap_tavg, 10)	0.0048
(Psurf_f_tavg, 1)	0.0048
(Rainf_f_tavg, 6)	0.0048
(def, 2)	0.0047
(maximum_2m_air_temperature, 4)	0.0047
(vap, 0)	0.0047
(minimum_2m_air_temperature, 9)	0.0042
(soil, 6)	0.0042
(tmmn, 9)	0.0042
(Psurf_f_tavg, 12)	0.0041
(Qair_f_tavg, 7)	0.0041
(Evap_tavg, 6)	0.0041
(Qg_tavg, 8)	0.0041
(aet, 6)	0.0040
(Tair_f_tavg, 11)	0.0039
(Evap_tavg, 5)	0.0039