Validation, Calibration, and Uncertainty Quantification of the Wofost Crop Simulation Model

by

Sule Kahraman

B.S., Massachusetts Institute of Technology (2020)

Submitted to the Department of Electrical Engineering and Computer Science

in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
Department of Electrical Engineering and Computer Science
May 14, 2021

Certified by .....
Anette "Peko" Hosoi
Associate Dean of the MIT School of Engineering
Neil and Jane Pappalardo Professor of Mechanical Engineering
Professor of Mathematics
Thesis Supervisor

Certified by .....
Munther Dahleh
William A. Coolidge Professor of Electrical Engineering
and Computer Science
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Validation, Calibration, and Uncertainty Quantification of the Wofost Crop Simulation Model

by

Sule Kahraman

Submitted to the Department of Electrical Engineering and Computer Science
on May 14, 2021, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

The *Digital Agriculture in Africa* project [44] aims to unlock the agricultural potential in Africa and help farmers improve their crop productivity by enabling them to make more precise and timely decisions about crop management via better prediction tools. The absence of large, comprehensive, and structured agricultural data, however, limits the performance of these predictions. To address this problem, we propose a Data Platform that aggregates data from a variety of publicly available resources, processes them into reusable data formats, and augments the sparse data via synthetic data generation tools.

In this thesis, we focus on the synthetic generation of agricultural yield data via the WOFOST (WOrld FOod STudies) [20] crop simulation model. Through our validation, calibration, and uncertainty quantification steps, we seek to answer the following question: *How can we reliably generate yield data for different regions of the world using simulation models?*

Due to unavailability of large agricultural data from Africa, we used data from the United States for the validation and calibration steps. Our empirical findings from the validation step demonstrated that the off-the-shelf usage of the WOFOST model, which was originally developed in Europe, may not be suitable for the agricultural studies in the United States, when the input parameters are not precise and accurate. This insight led us to perform the calibration step, where we discovered that the performance of the WOFOST model can be improved by estimating the correct crop parameters using evolutionary algorithms. Through our uncertainty quantification step, we shortlisted a number of input parameters that the model seems most sensitive to and developed a simpler but more tractable and effective model to WOFOST that has an analytical solution. Finally, we provided a quantitative analysis of how the uncertainty from the input parameters propagates through our proposed model to the generated data.

Thesis Supervisor: Anette "Peko" Hosoi
Title: Associate Dean of the MIT School of Engineering
Neil and Jane Pappalardo Professor of Mechanical Engineering
Professor of Mathematics

Thesis Supervisor: Munther Dahleh
Title: William A. Coolidge Professor of Electrical Engineering
and Computer Science

Acknowledgments

I would like to begin by thanking my supervisors, Professors Anette "Peko" Hosoi and Munther A. Dahleh, for their invaluable guidance over the course of the past year, as well as for their warm and friendly attitude, which has made the experience of writing this thesis a smooth, rewarding, and enjoyable process. Peko, thank you for your generous advice, patience, positive energy, and encouragement throughout this work. Munzer, thank you for sharing your insights with me and giving me the opportunity to join your research group. I feel lucky to have been part of this research community and to have worked with such inspiring people at MIT.

I also would like to extend my sincere thanks to Mardavij, Bernardo, Rogerio, Mark, and Flora for helpful discussions on many occasions—it has been a true pleasure and a privilege collaborating with you. Many thanks to Allard de Wit for taking the time to answer all of our questions about WOFOST as well.

Furthermore, I would like to thank all of my friends here at MIT, particularly to those at Number Six, for making this place feel like home and my life here truly memorable. I am especially thankful for my friends Quentin, Christine, Phoebe, Edgar, and Thiago for their emotional and academic support over the years. And I owe an especial debt of gratitude to Driss for carrying me through this work and always being there for me—I could not have made it this far without you.

I want to thank my friends Yasemin, Su, Pınar, Ufuk, and Miraç who have kept me company for the last ten years. I owe an enormous debt of gratitude to Miraç for reading every single line of this thesis, for providing me with valuable feedback, and for supporting me on always doing my best.

Last but not least, I would like to thank my parents and grandparents for all the sacrifices they have made for my education and for their ever-present and ever-growing support and love for me.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Agriculture lies at the heart of Africa's economy. According to the World Economic Form, agriculture constitutes almost a quarter of the gross domestic product (GDP) all across the continent [17]. Even though Africa holds 60% of the world's arable land, it is able to generate only 10% of the global agricultural output [12]. Given the discrepancy between its agricultural potential and its current agricultural production, there is a strong need to develop and implement efficient farming techniques, to invest in capital to support agribusiness enterprises, and to encourage technological innovation in agricultural mechanization.

Digital Agriculture in Africa project, a research collaboration at the MIT Institute for Data, Systems, and Society (IDSS), aims to address this challenge with data. The project aims to unlock the agricultural potential in Africa and help farmers increase their productions by enabling them to make more precise and timely decisions about planting, harvesting, irrigation, and fertilization through better prediction tools. Data-driven precision agriculture has shown to improve yield, reduce cost and ensure sustainability; however, there is not enough granular (at the farm level) data to drive these predictions due to the high cost of manual data collection [37].

One approach to tackle the data scarcity problem in Africa would be to study farming and the impact of agricultural interventions with limited data from technologically advanced farms where there is more data availability. Then this learning can then be translated to predict the value of intervention in under-performing farms. Thus, one of the goals of this research collaboration is to create a platform for sharing data and risk among invested parties, from farmers and lenders to insurers and equipment manufacturers to fertilizer companies [29, 44].

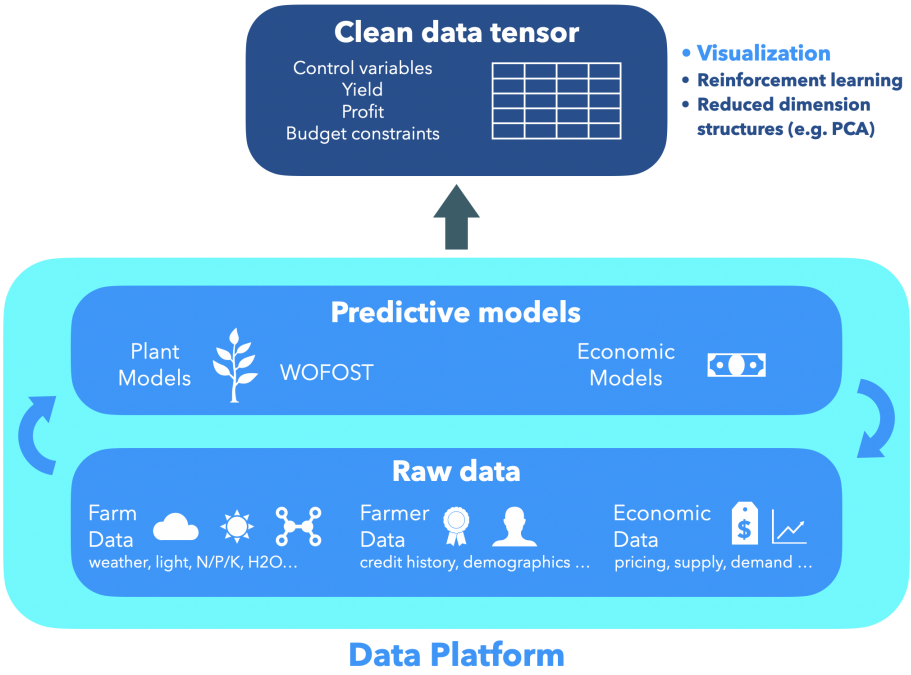


Figure 1-1: Data Platform diagram

The Data Platform, shown in Figure 1-1, consists of two components that are in a continuous feedback loop: "Raw data" and "Simulation and Predictive Models". The first component includes farm data (e.g., weather, fertilization, irrigation practices), farmer data (e.g., credit history, demographics), and economic data (e.g., pricing of commodities, supply and demand). The second component includes plant models such as WOFOST, economic models, and other models that learn efficiently from data. With this positive feedback between the two components of the Data Platform, predictive models perform better as more raw data is available and with better

predictive models we can generate more reliable synthetic data that would help performance of the predictive models. This work pertains to the raw farm data and predictive plant models (left side of Figure 1-1) of the Data Platform.

1.2 Related Work

The Data Platform of the Digital Agriculture in Africa project, consisting of raw data and predictive models, aims to generate crop yields in different regions of the world, especially in sub-Saharan Africa to be understand and quantify the impact of crop management. In this work, we use a plant-based simulation model to generate crop yield data. We focus on using a well-known simulation model in the agriculture community: WOFOST (WOrld FOod STudies) [20], due to its ability to simulate the impact of weather and crop management on growth and development of crops.

Developed and maintained by Wageningen University & Research [20, 34, 48], WOFOST is a mechanistic simulation model for the quantitative analysis of the growth and production of annual field crops. It has been used by many researchers over the world and has been applied for many crops over a large range of climatic and management conditions [20]. It is one of the key components of the European MARS (Monitoring Agricultural Resources) Crop Yield Forecasting System (MCYFS) [33], which is a tool to monitor crop growth development, evaluate short-term effects of anomalous meteorological events, and provide monthly forecasts of crop yield and production [32]. The regional application of the WOFOST model within the MCYFS has been used for climate change impact assessments [13, 41] and yield gap analysis [15, 16]. It is also used to estimate the untapped crop production potential on existing farmland based on current climate and available soil and water resources in the Global Yield Gap Atlas (GYGA) [4].

The WOFOST model can be used to calculate crop production, biomass, and water use for a given location provided knowledge about crop, weather, soil, and manage-

ment of the farm, such as fertilization amounts and dates, irrigation amounts and dates, sowing date and harvest date. In this work, we use WOFOST to calculate the weight of the crop's storage organs at the end of the growing season. We use this weight as a proxy for the crop yield.

Implementation of the WOFOST model is shown in the schematic overview in Figure 1-2 which is obtained from the WOFOST manual [20]. This overview shows the major processes implemented in WOFOST and how they interact with each other.

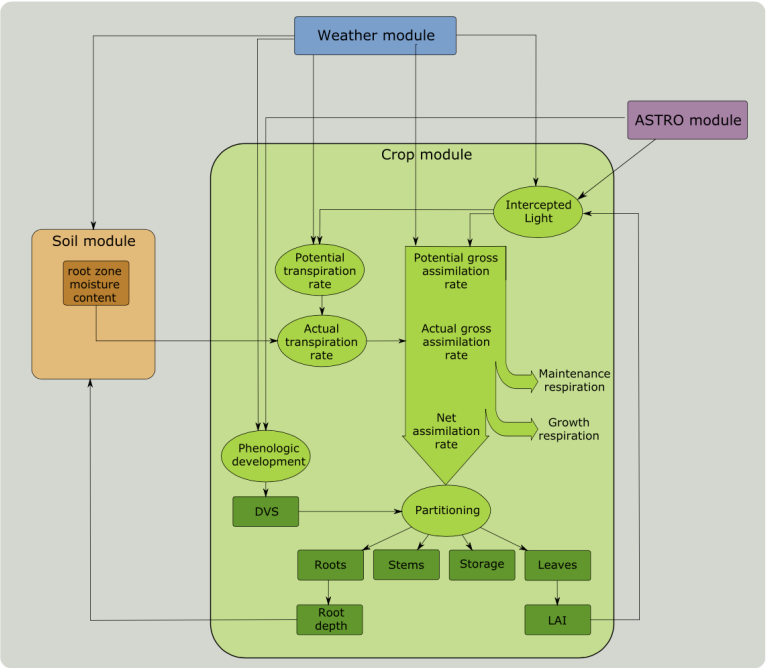


Figure 1-2: Schematic overview of the major processes implemented in WOFOST and their linkages

The WOFOST model calculates yield in two production settings: potential yield and attainable yield. Potential yield is obtained under optimal soil conditions (i.e., efficient water and nutrients for the plant), and attainable yield is obtained under water-limited and nutrient-limited conditions. However, production reducing factors such as pests, weeds, diseases, and pollutants, are not taken into account in neither of these production settings.

Throughout this work, we keep in mind the limitations of the WOFOST model and mitigate them whenever possible. One important limitation of this model is that its outputs may not be reliable in extreme climate conditions, such as drought or flooding. Another one is that the model is deterministic, meaning that it outputs only one value. Because it does not provide confidence intervals or error rates of the results, it is very difficult to assess the quality of its results. Obtaining confidence intervals and error rates for the WOFOST model is a challenging work because there is no database of field experiments with exact sowing date, exact cultivar, exact farm management data to compute error margins.

Another challenge for using the WOFOST model globally and implementing a system similar to the MCYFS [33] in different regions of the world, such as sub-Saharan Africa, is the difficulty of modeling different behaviors of the farmers. One example is that farmers sow a large variety of crops and switch cultivars very often. Different sowing strategies could be implemented in the WOFOST model; however it is another challenge to obtain data about exact cultivars used in farms because farmers might not be willing to share this information with their competitors. Most crop parameter files used today [1] are from years before 2000, hence the WOFOST model does not take it into consideration improvements in technology, such as hybrid seed, improved genetics, microbial soil amendments, etc.

Scarcity of farm level data seriously limits the use of WOFOST as a predictive tool [48]. Many of the environmental data used pertain to average conditions and do not allow the evaluation of the effects of their erratic nature and of extreme conditions. The data constraint is also a problem for the further development of the model because complete and detailed datasets are needed for model validation. Such datasets may be either measured experimental data or data generated by more detailed and validated models; however, only a few datasets appear to be available.

1.3 Our Contributions and Thesis Outline

Through this work, we provide a framework for the Data Platform that facilitates the integration of new farm data to the platform and the generation of reliable data using simulation models. More precisely, we implement the infrastructure for maintaining a database of agricultural data and augmenting this database by generating synthetic data using simulation models. We collect and process raw farm data from various data resource and process into one data matrix to input to the WOFOST simulation model.

As we will witness in the upcoming chapters, assessing the quality of the synthetic data is a crucial part of this work. Thus, we first validate the results of the WOFOST model by comparing the simulated data to the ground truth yield data. Motivated by the results of the validation and we then calibrate the input parameters of the model by using evolutionary algorithms, a gradient-free optimization method.

Even though calibration improves performance of the model, uncertainties remain as major obstacles for the predictive capability and reliability of the simulation models. In order to find stable solutions across a wide range of inputs, and to make better decisions at a known level of confidence, it is necessary to quantify the uncertainty. With uncertainty quantification methods, such as Monte Carlo simulations and perturbation analysis, we identify the input parameters that the WOFOST model is most sensitive to and analyze how the uncertainty propagates through the system of crop growth. For the latter approach, we develop an alternative model to WOFOST that describes the dynamic system of crop growth, which has a simple, analytical form that makes error propagation possible.

The remainder of this thesis is organized as follows. Chapter 2 explains the data generation procedure by detailing the input data needed for the simulation, our data resources and how we process the data into data matrices that can be stored in the Data Platform, as well as how to use the software package to run the WOFOST model.

Chapter 3 talks about the validation of the WOFOST model using all the available data from the United States and presents results analyzing several statistics on the gap between simulated and true yield. Chapter 4 explains our calibration method for the application of WOFOST in different regions of the world and shows the impact of calibration on the generated yield data. Chapter 5 discusses two methods to quantify the uncertainty in the model output: sensitivity analysis, and perturbation analysis. Finally, Chapter 6 summarizes our work and discusses future work.

Chapter 2

Data Collection and Generation

Our work focuses on the agricultural data part of the Data Platform. As a first step to tackle the problem of data sparsity, we propose to generate synthetic crop yield data using simulation models of plant growth based on physical and chemical processes. In this chapter, we explain the data generation methods with the WOFOST crop growth simulation model. We first describe the input parameters to the model. Then we explain how we gather and process data for the input parameters from different resources for each type of input. After explaining the creation of the data matrix and how to run the simulation with the processed data, we end this chapter with a discussion about the data constraints of the WOFOST model.

2.1 Input Parameters of WOFOST

Simulating crop growth with WOFOST requires first defining the farm and the crop as inputs to the model. The farm is defined by its soil and site characteristics, the weather conditions throughout the growing season, and the farm management (e.g. sowing date, harvest date, irrigation and fertilization). The crop is defined by the seed properties used in the farm. If we think of the WOFOST model as a multi-variable function $f(\cdot)$ and its simulated yield output as Y_w, we can then express the entire relation as follows:

$$Y_w = f(c, s, v, x, z) \tag{2.1}$$

where c denotes the crop characteristics, s the soil characteristics, v the site characteristics, x the agromanagement, and z the weather conditions. The inputs to the WOFOST model, namely c, s, x, and z, are all defined by more than 78 parameters in total: 52 crop parameters, 13 soil parameters, 8 weather parameters, and 5+ agromanagement parameters.

Crop parameters describe the initial weight of the organ, phenological properties (the rate of appearance of vegetative and reproductive organs), maintenance respiration rate, death rate of leaves, water use, assimilation rate of the plant, conversion efficiency of assimilates into biomass, among other things. Soil parameters describe the physical soil characteristics such as soil water retention, hydraulic conductivity, soil workability, and soil minerals. The site parameters provide ancillary parameters that are not related to the crop or the soil. Examples include the initial conditions of the water balance, such as the initial soil moisture content (`WAV`) and the initial surface storage (`SSI`), maximum surface storage (`SSMAX`), and the atmospheric CO_2 concentration (`CO2`). Agromanagement parameters are the sowing date, emergence date, maturity date, harvest date, the amounts and dates of irrigation, the amounts and dates of fertilization. Weather parameters are obtained from daily meteorological data during the growing season, and describe the radiation (sunshine), air temperature, rainfall (precipitation), air humidity and wind speed. Further details about the input parameters can be found in Appendix 3 of the WOFOST manual [20].

2.2 Data Collection and Processing

In order to generate realistic crop yield data with the WOFOST simulation model, we need to collect real data for the inputs that describe the the crop characteristics, soil characteristics, weather conditions, and agromanagement from different public databases, such as USDA NASS [8], ISRIC-World Soil Information [14], and NASA Power [6]. We process the data to be in the format WOFOST requires and later with the process data we create a reusable data matrix of the WOFOST inputs.

By inputting the this data matrix, we could then run simulations to generate crop yield data. In this section, we explain our data resources and our methods of data processing.[1]

2.2.1 Yield Data

Crop yield is a measurement of the amount/weight of the crop grown per unit area of land. It is typically measured by kilograms per hectare (kg/ha) or bushels per acre (bsh/ac). The US Department of Agriculture (USDA) takes samples and estimates of crop yields for nearly two-dozen crops in the country and makes its agricultural data available at the Quick Stats Database [8]. This database is a comprehensive tool for accessing agricultural data published by the USDA National Agricultural Statistics Service (NASS) and allows the user to customize a query by commodity, location, or time period. Data can be manipulated and exported as an Excel file. Quick Stats contains official published aggregate estimates related to US agricultural production. The regional scale of the data varies from country level to state level to county level. For selected states, the data includes the total crops and cropping practices for each county, as well as breakouts for irrigated and non-irrigated practices for many crops. The data files contain planted and harvested area, yield per acre, and production. We use the USDA NASS Quick Stats Database to obtain county level crop yield data and state level crop calendar data. Note that the yield data from the US is in unit bushels/acre (bu/ac). We convert it to metric units for yield: tons/hectare or kilograms/hectare (kg/ha) using the conversion rates from [10]. For corn, 1 bu/ac is equal to 62.77 kg/ha, for wheat and soybean 1 bu/ac is equal to 67.25 kg/ha. We extract yield data from the database for two reasons: (1) to aggregate granular yield data for the Data Platform (2) to use it as ground truth yield values for validation (Chapter 3) and calibration (Chapter 4) of the WOFOST model.

[1] GitHub repository `wofost_data` [2] contains the raw and processed data as well as the scripts for data collection and processing.

2.2.2 Crop Data

Crop characteristics depend on the specific seed used in the given farm. Because the seed varies widely by region, farm, and year, it is difficult to find real data for the crop parameters of WOFOST. Hence, we use the default crop parameters provided by [1] which contains the parameter sets for 22 crops, including, but not limited to, barley, chickpea, cassava, maize, wheat, sugarbeet, rice. The parameters are stored in a data serialization format `YAML`, and and within each `YAML` file, different crop ecotypes and crop varieties can be defined. To accommodate the definition of different crop varieties in one file, the parameter files have been organized in a clear structure. An example `YAML` file for the crop barley is shown in Figure 2-1.

```yaml
Version: 1.0.0
CropParameters:
    GenericC3: &GenericC3

        ## All parameters for C3 crops go here

    GenericC4: &GenericC4

        ## All parameters for C4 crops go here

    EcoTypes:
        springbarley: &springbarley
            <<: *GenericC3          # Ecotype springbarley inherits from GenericC3

            ## All parameters specific for springbarley go here

    Varieties:
        Spring_barley_301:
            <<: *springbarley       # Variety Spring_barley_301 inherits from ecotype springbarley
            TSUM1:
            -  800
            - temperature sum from emergence to anthesis
            - ['C.d']
            TSUM2:
            -  750
            - temperature sum from anthesis to maturity
            - ['C.d']
```

Figure 2-1: Example `YAML` file containing the crop parameters for barley.

Since we use the crop parameters from [1], the crop data does not require any processing. Let us note that these default crop parameters were developed by experts in Europe and might not be applicable to seeds in different regions in sub-Saharan Africa or the United States. We discuss mitigation methods for calibrating crops for different regions without expert knowledge in Section 4.

2.2.3 Soil Data

To find realistic soil properties, we used the `ISRIC-Wise30sec` dataset from ISRIC-World Soil Information [14], which is the host of the World Data Center for Soils (WDC-Soils). This dataset includes about 21,000 soil profiles described in 19 soil variables. Due to the difficulty of finding real values for all the WOFOST soil parameters, we use default values for the parameters `SMTAB, CRAIRC, CONTAB, RDMSOL` and replace the values for the parameters `SMW, SMFCF, K0, SOPE, KSUB` with the values we obtained by processing soil data from the `ISRIC-Wise30sec` dataset. To calculate these five parameters we used different sources: [26] for `SMW` , [40] for `SMFCF`, [35] for `SOPE` and [27] for `K0`.

```
** SOIL DATA FILE for use with WOFOST Version 5.0, June 1990
**
** EC-4 fine

SOLNAM='EC4-fine'

** physical soil characteristics

** soil water retention
SMTAB    =  -1.000,   0.570,    ! vol. soil moisture content
             1.000,   0.533,    ! as function of pF [log (cm); cm3 cm-3]
             1.300,   0.524,
             1.491,   0.515,
             2.000,   0.486,
             2.400,   0.451,
             2.700,   0.420,
             3.400,   0.350,
             4.204,   0.300,
             6.000,   0.270
SMW      =   0.300             !  soil moisture content at wilting point [cm3/cm3]
SMFCF    =   0.460             !  soil moisture content at field capacity [cm3/cm3]
SM0      =   0.570             !  soil moisture content at saturation [cm3/cm3]
CRAIRC   =   0.050             ! critical soil air content for aeration [cm3/cm3]

** hydraulic conductivity
CONTAB   =   0.000,   1.033,    ! 10-log hydraulic conductivity
             1.000,  -0.824,    ! as function of pF [log (cm); log (cm/day)]
             1.300,  -1.155,
             1.491,  -1.398,
             1.700,  -1.523,
             2.000,  -1.959,
             2.400,  -2.495,
             2.700,  -2.886,
             3.000,  -3.276,
             3.400,  -3.770,
             3.700,  -4.131,
             4.000,  -4.481,
             4.204,  -4.745

RDMSOL   =   120.              ! soil maximum rootable depth
K0       =   10.789            ! hydraulic conductivity of saturated soil [cm day-1]
SOPE     =   0.55              ! maximum percolation rate root zone[cm day-1]
KSUB     =   0.37              ! maximum percolation rate subsoil [cm day-1]

** soil workability parameters
SPADS    =   0.050             !  1st topsoil seepage parameter deep seedbed
SPODS    =   0.025             !  2nd topsoil seepage parameter deep seedbed
SPASS    =   0.100             !  1st topsoil seepage parameter shallow seedbed
SPOSS    =   0.040             !  2nd topsoil seepage parameter shallow seedbed
DEFLIM   =  -0.300             !  required moisture deficit deep seedbed
```

Figure 2-2: Example soil parameters for medium fine soil

An example set of soil parameters are shown in Figure 2-2. Parameters shown in grey are from the default soil file and the parameters shown in blue are updated using the `ISRIC-Wise30sec` dataset.

2.2.4 Site Data

Due to the difficulty of finding the exact measurements for the site parameters at each farm, we use default values for the site parameters. Required site parameters are the atmospheric CO_2 concentration (`CO2`) and the initial soil moisture content (`WAV`) to run the simulation and we use default values of `CO2` = 360 and `WAV`= 100 in all of all the simulation runs.

2.2.5 Weather Data

For the weather data, we use the gridded data provided by National Aeronautics and Space Administration - Prediction of World Wide Energy Resources (NASA-POWER) [6]. NASA POWER database presents a global coverage of complete weather data at horizontal resolution of 1 degree latitude–longitude (about 100 km). Assessment of the NASA POWER database [39] also found that it is a useful database to investigate the impacts of climatic variability on crop yield throughout the years with reasonable confidence, once its complete long-term database is freely available. We chose this database to obtain weather data because of its resolution and simplicity of querying the weather for a given pair of coordinates.

We note that daily weather data can be incomplete because some weather stations may not provide continuous 24-hour reporting. We use a simple imputation method: Linear interpolation between the available data. For example, if the weather data is missing for the date `4/15/2020` to `4/18/2020`, we can use data from `4/14/2020` and `4/19/2020` to fill in the missing values using linear interpolation. It would, however, be dangerous to linearly interpolate for longer periods of time, since the weather can

change erratically. Hence, we use this linear interpolation method for missing values only up to seven days. For simplicity, we ignore data from a growing season if it still contains missing values after the linear interpolation method described. Because precipitation can be too erratic, it should not interpolated but rather set to zero when missing.

It is also possible to derive the missing values in a different way. For instance, data from a weather station in the vicinity or the long term daily averages from the same station can be used. Note that this imputation method however is not implemented in our work. GYGA Protocol for Weather Data [11] also outlines a step-by-step procedure for imputation of weather data.

2.2.6 Agromanagement Data

To find the agromanagement parameters sowing date, emergence date, and harvest date, we utilize the crop calendar data from USDA NASS [8]. This data is provided at the state-level and on an annual basis. Due to the difficulty of finding irrigation and fertilization data, we assume no irrigation and no fertilization for our simulations. Depending on the type of yield data (irrigated vs. non-irrigated), we use different production settings (potential vs. water-limited).

This entails that when we are simulating for regions where we have irrigated yield data, we assume optimal irrigation; so, the model is in the potential yield setting. When we are simulating for regions with non-irrigated yield data, we assume zero irrigation; so, we do not include any irrigation action in the agromanagement parameters and run the model in the water-limited setting. When using data from the United States, we also assume optimal fertilization; therefore, we do not run simulations in the nutrient-limited setting. When there is fertilization data available, the dates and amounts of fertilization should be included in the agromanagement parameters and the model should be run the nutrient-limited setting.

```yaml
AgroManagement:
- 2019-01-01:
    CropCalendar:
        crop_name: wheat
        variety_name: winter-wheat
        crop_start_date: 2019-01-01
        crop_start_type: emergence
        crop_end_date: 2019-04-11
        crop_end_type: maturity
        max_duration: 100
    TimedEvents:
    -   event_signal: irrigate
        name: Irrigation application table
        comment: All irrigation amounts in cm
        events_table:
        - 2019-01-15: {amount: 10, efficiency: 0.7}
        - 2019-02-15: {amount:  5, efficiency: 0.7}
    -   event_signal: apply_npk
        name: Timed N/P/K application table
        comment: All fertilizer amounts in kg/ha
        events_table:
        - 2019-01-01: {N_amount: 15, P_amount: 15, K_amount: 15, N_recovery: 0.7, P_recovery: 0.7, K_recovery: 0.7}
        - 2019-02-01: {N_amount: 20, P_amount: 10, K_amount: 10, N_recovery: 0.7, P_recovery: 0.7, K_recovery: 0.7}
    StateEvents:
    -   event_signal: apply_npk
        event_state: DVS
        zero_condition: rising
        name: DVS-based N/P/K application table
        comment: all fertilizer amounts in kg/ha
        events_table:
            - 0.3: {N_amount : 1, P_amount: 3, K_amount: 4}
            - 0.6: {N_amount: 11, P_amount: 13, K_amount: 14}
            - 1.12: {N_amount: 21, P_amount: 23, K_amount: 24}
    -   event_signal: irrigate
        event_state: SM
        zero_condition: falling
        name: Soil moisture driven irrigation scheduling
        comment: all irrigation amounts in cm of water
        events_table:
            - 0.15: {irrigation_amount: 20}
```

Figure 2-3: Example `YAML` file for agromanagement parameters

Figure 2-3 shows an example `YAML` file for the agromanagement parameters. `CropCalendar` field defines the crop name, variety, and important dates for the growing season. `TimedEvents` field defines the irrigation and fertilization dates and amounts. `StateEvents` field can be used to define irrigation and fertilization patterns using state variables such as development stage `DVS`, or soil moisture `SM` compared to using predetermined dates as in `TimedEvents`.

2.3 Data Matrix

We bring together all the aforementioned data in one matrix for ease of use and storage. Figure 2-4 demonstrates the data matrix we created from the crop, soil, site, weather and agromanagement, as well as the true yield data from the data resources explained in Section 2.2. Each row in this multidimensional matrix corresponds to

a location, denoted as Farm 1 ... Farm N. Different colors illustrate the types of data: crop, soil, site, weather, agromanagement, and true yield data. Each column corresponds to an input parameter and each cell contains the value of the parameter for the given farm.

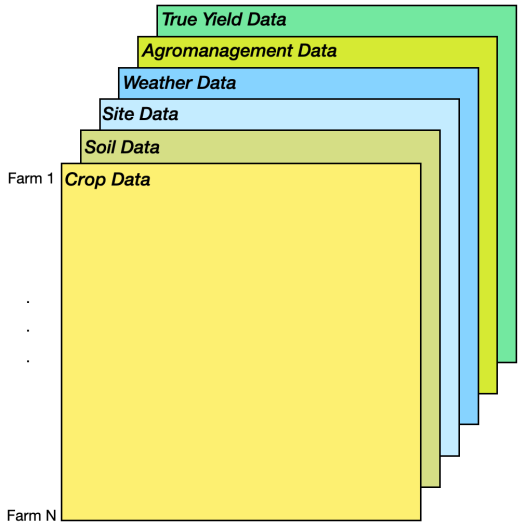


Figure 2-4: Representation of the data matrix for the WOFOST input parameters and true yield data

In order to combine different types of data, further data processing is necessary. While the soil data includes a set of coordinates where a certain soil profile is present, the yield data only contains the county and state names, no coordinates. We find the coordinates for the county centers and use it as the county coordinates. Then we find the closest (minimum Euclidean distance) soil profile to the county coordinates using a k-d tree consisting of soil profile coordinates for quick nearest-neighbor lookup [38]. Once we find the soil profiles for each county, we use query the weather data for the county coordinates. Next, we explain how we run the simulation using the prepared data matrix.

2.4 Running the Simulation Model

PCSE (Python Crop Simulation Environment) [21] is a Python package for building crop simulation models developed in Wageningen (Netherlands). PCSE provides the environment to implement crop simulation models, the tools for reading ancillary data (weather, soil, agromanagement) and the components for simulating biophysical processes such as phenology, respiration and evapotranspiration. PCSE includes implementations of the WOFOST and Lintul3 crop simulation models, and we use its WOFOST implementation in this work.

We run the simulation by first creating a parameter object with all input parameters and then starting the simulation engine. Running a year long simulation for a given location takes about a few seconds. When the simulation run is complete, with the `get_summary_output()` method on the WOFOST object we can output some final results such as the emergence (`DOE`), anthesis (`DOA`), maturity (`DOM`), harvest (`DOH`) dates, total biomass (`TAGP`), weight of roots (`TWRT`), leaves (`TWLV`), stem (`TWST`), and storage organs (`TWSO`), maximum LAI (`LAIMAX`), etc. We can also retrieve the time series of daily simulation output using the `get_output()` method such as the time series for weights, leaf area index, etc. as shown in Figure 2-5. For yield simulation, retrieving weight of the storage organs (`TWSO`) is all that is necessary.

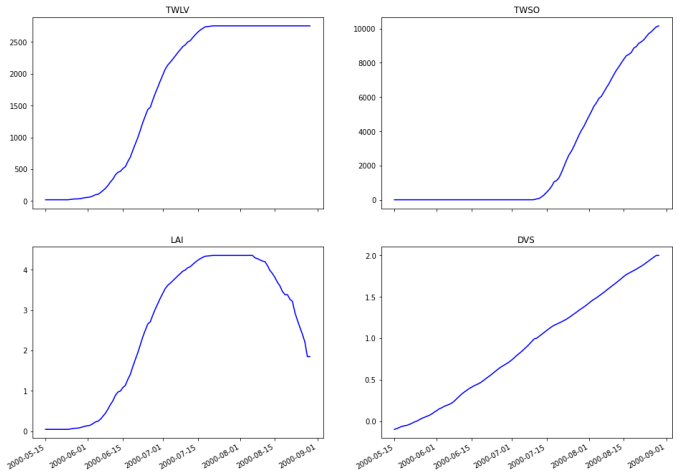


Figure 2-5: Examples of WOFOST outputs as time-series

2.5 Discussion

Even though we chose a well-studied crop, corn, and a region with publicly available data, the US, we could not find real-world data for all the input parameters of the model. Thus, the large number of input parameters of WOFOST make it difficult to apply it directly in different regions of the world where exact measurements from the fields are not easy to obtain, such as in sub-Saharan Africa. This challenge of finding real-world data for the input parameters reemphasizes the need for the Data Platform where raw data from different regions and from different stakeholders can be aggregated. We also make a note that the soil parameters we derive from the ISRIC soil database needs to be verified with some ground truth data because the derivation formulas were gathered from various sources and may not be accurate.

Chapter 3

Validation

Understanding the quality of the generated data is crucial especially when the data is used in precision agriculture and decision-making algorithms to assist the public and the government officials with agricultural interventions and policy-making. To that end, we assess the performance of WOFOST by juxtaposing its simulated results with the true yield data. We compare the simulated yield data in the United States to the crop yield data obtained from USDA NASS [8]. For this comparison to be meaningful and reliable, we assume that the crops with yield data from USDA NASS contains no limitations to crop growth by nutrients and no yield reductions due to weeds, pests, or diseases because the WOFOST simulation model does not include these external factors. We think these assumptions hold true in many regions in the US for some crops, such as corn and wheat, hence we focus on these crops.

3.1 Methodology

As shown in Figure 3-1, validation workflow starts with aggregation and processing of raw data (as described in Chapter 2). We run the simulation for each data point separately for one growing season and generate simulated yield data. By comparing the simulated yield to the true yield data, we achieve validation results as we present in Section 3.2

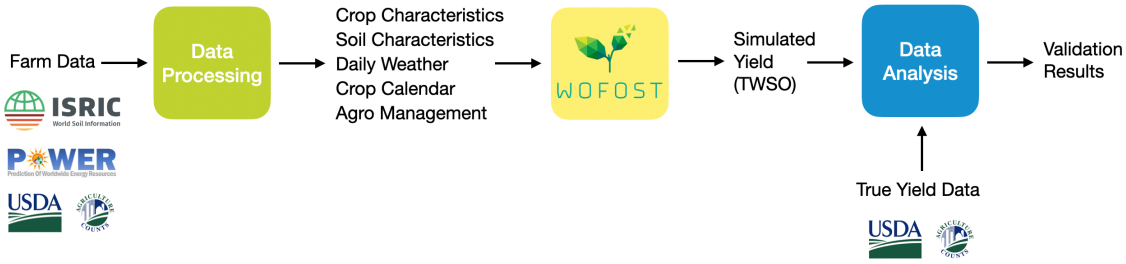


Figure 3-1: Validation overflow

For validation, we generate the yield data by running WOFOST in two different settings: (1) potential yield setting for irrigated crops and (2) water-limited yield setting for rain-fed (non-irrigated) crops.

Potential Production

Potential production represents the production ceiling for a given crop when grown in a given area under specific weather conditions. It is determined by the crop's photosynthesis response to CO_2 and the temperature, and solar radiation regimes during the growing season [20]. In practice, this ceiling can only be reached with application of high amounts of fertilizers and irrigation, and thorough pest and weed control. In addition, there should be no losses caused by traffic or grazing, and there should be no damage to the crop by wind, hail, and frosts. Because potential yield is also determined by crop properties, yield potential varies over crop varieties and can be increased by breeding. In the WOFOST model, potential yield depends on the choice of crop variety, sowing date and weather data.

For counties that grow irrigated crops, we choose the potential production setting due to unavailability of data for irrigation and fertilization patterns. Given the high production rates of corn in the US, we assume optimal irrigation and fertilization by farmers. Therefore, by running the simulation in the potential production setting we only see the effect of crop variety, crop calendar and weather conditions on the yield.

Water-Limited Production

By taking into account the reductions by water and nutrients, WOFOST includes two attainable production levels: water-limited and nutrient-limited (or water and nutrient limited). At the attainable production level, the yield of the crop is limited by the availability of water or nutrients during the growing season. The water-limited yield represents the maximum yield that can be obtained under rain-fed conditions but with optimal nutrient supply [20].

The nutrient limited production corresponds to a situation where water is not limited but the nutrients in the soil are insufficient to cover the crop's demands. Examples to nutrient insufficiency include when the soil does not contain enough Nitrogen, and when part of the fertilizer applied is lost through leaching, volatilization or denitrification.

Out of water-limited, nutrient-limited, and water-nutrient-limited production settings, we choose the water-limited setting for the non-irrigated crops, mainly because of the observation that there is no irrigation or fertilization data available that can be used as an input to the simulation as agromanagement data. For the yield data that is reported as non-irrigated in the US, we assume that farmers performed no irrigation but optimal fertilization; therefore, we do not include the limitation of the nutrients.

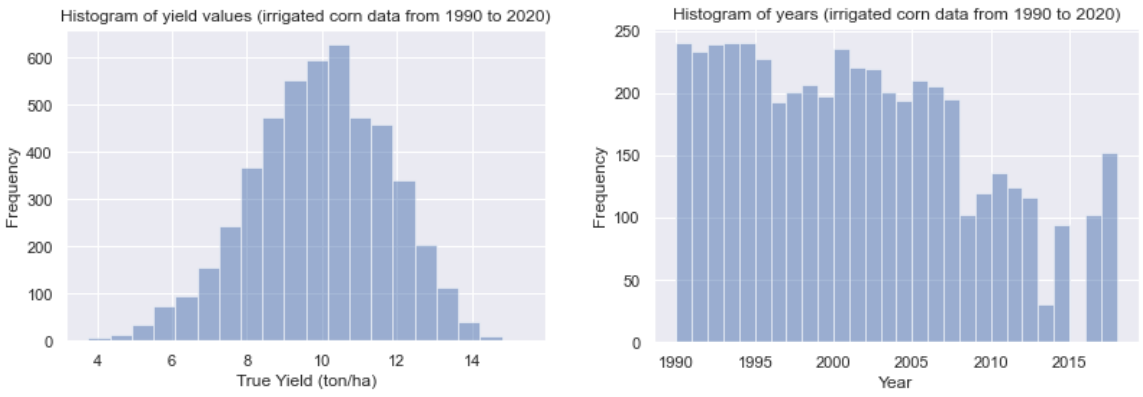
After running the simulations in the given production settings, we validate the results of the WOFOST model by comparing the simulation output `TWSO` (total weight of storage organs), denoted as Y_w, to the true yield data, denoted as Y_t, for irrigated and non-irrigated crops from USDA NASS database. We analyze certain statistics of the difference between the true yield and the simulated yield, $Y_t - Y_w$, which is referred to as the yield gap, Y_{gap}, hereinafter.

3.1.1 Data

Currently, the US is the largest corn producer in the world, with approximately 96 million acres of land reserved for corn production [43]. Because of the high production of corn and data collection efforts of the USDA, there is a plethora of corn data coming from many states in the US; hence, we chose to use corn yield data from the US for validation of WOFOST results.

Irrigated Yield Data

Irrigated corn yield data size has 15265 (county, year) pairs from year 1944 to 2020 for counties in states Delaware, Colorado, Kansas, Nebraska, Texas, South Dakota, New Mexico, North Dakota, Oklahoma, Wyoming, Montana, and Idaho. Because there is no weather data available before 1990 [6], we only consider the yield data after 1990. We further remove data points that have yield zero because zero, in this content, indicates that the county did not plant any crops or that the data was missing. After the removal of zero yield values, the data size becomes 5924. We further remove data points for where the weather data contains too many missing values (viz., for more than 7 consecutive days). With all the removals, the final data size becomes 4875. Figures 3-2 and 3-3 summarize the distribution of the raw yield data for irrigated corn.



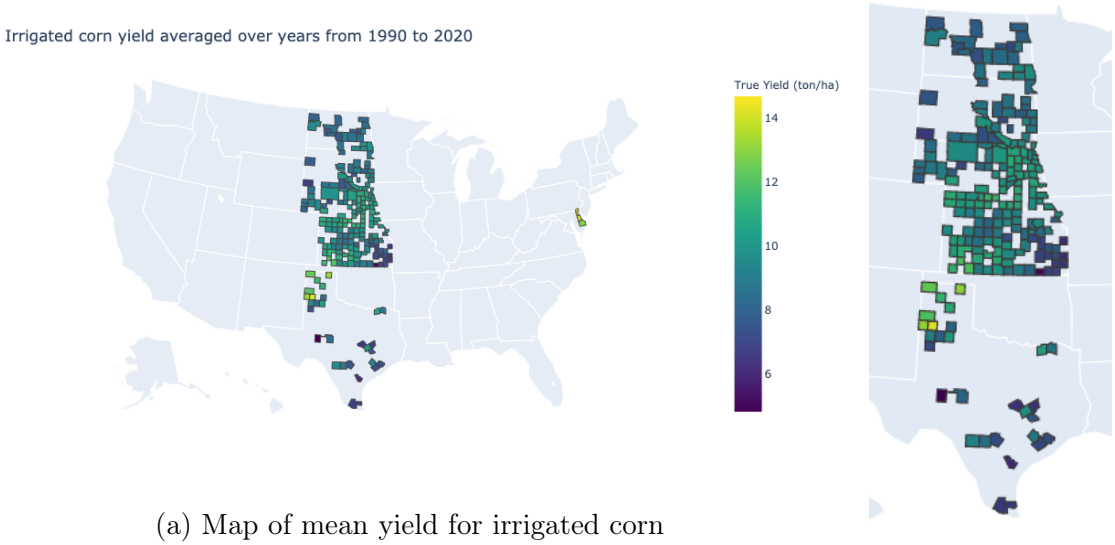
(a) Histogram of yield values

(b) Histogram of year values

Figure 3-2: Distribution of irrigated yield data

Figure 3-2a shows the distribution of the yield values in the irrigated corn yield data. We observe that the yield values are mostly around 10 ton/ha. Figure 3-2b shows the distribution of the year values in the irrigated corn yield data. We observe that there are less data points available after 2008 in the dataset. This does not mean that the country has been getting less corn yield, it just shows the availability of reported data.

Figure 3-3a shows the geographical distribution of the available data on a choropleth map at the county level. We observe that there is some data available from the states Delaware, Colorado, Kansas, Nebraska, Texas, South Dakota, and North Dakota. Figure 3-3b is a zoomed-in version of Figure 3-3a for easier inspection. Note that the zoomed in image does not include the state of Delaware.



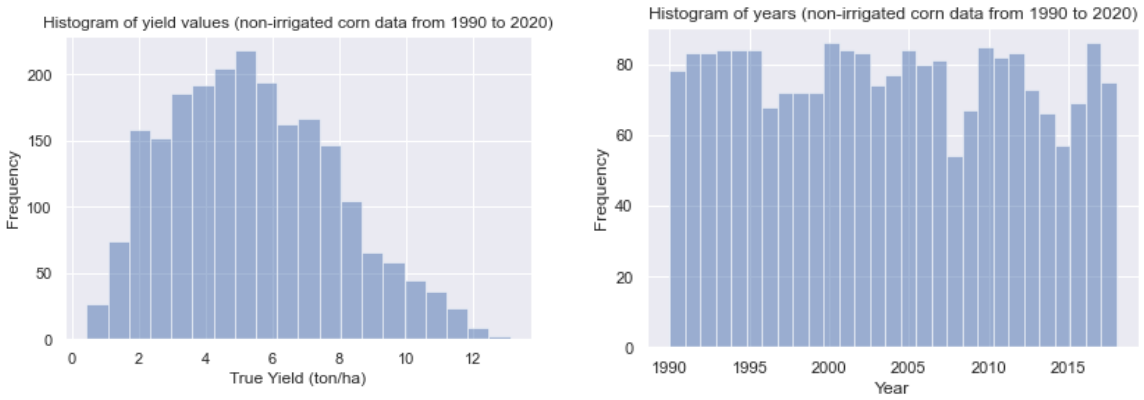
(a) Map of mean yield for irrigated corn

(b) Zoomed-in map

Figure 3-3: Map of yield data averaged over years for irrigated corn

Non-irrigated Yield Data

Non-irrigated corn yield data size has 13609 (county, year) pairs from 1944 to 2020 for the states of Delaware, Colorado, Kansas, Nebraska, Texas, South Dakota, North Dakota, Oklahoma, Wyoming, New Mexico, and Montana. Since there is no weather data available before 1990 [6], we only consider the yield data after 1990. Then, the non-irrigated yield data size becomes 5848 and we lose the data points from Montana. We also remove the data points that have too many missing values in the weather data and end up with a data size of 2226 with data from states Colorado, Kansas, Nebraska, and Texas. Figure 3-4 summarizes the distribution of the raw yield data. From Figure 3-4, we observe that the non-irrigated yield data volume is much smaller



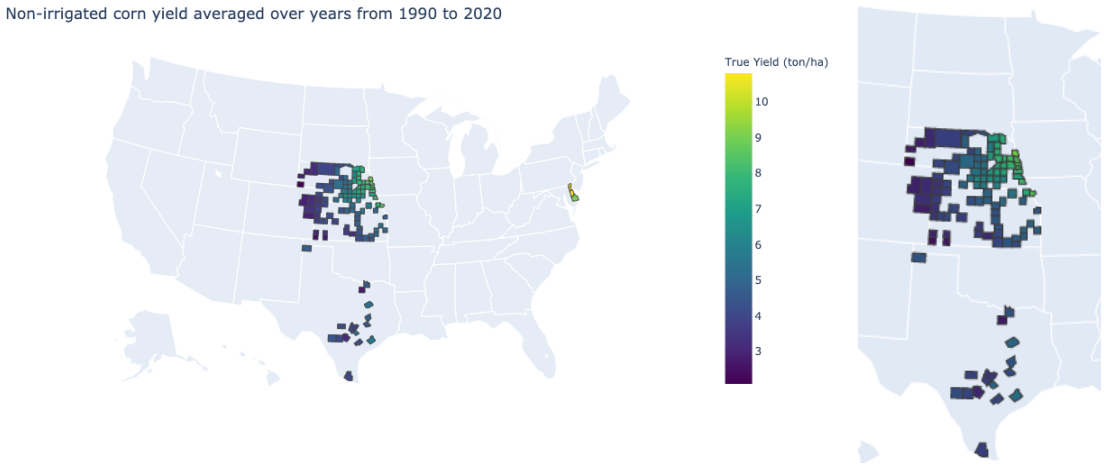
(a) Histogram of yield values

(b) Histogram of year values

Figure 3-4: Distribution of non-irrigated yield data

than the irrigated yield data. We also observe that the true yield for non-irrigated corn is around 5 ton/ha, whereas the irrigated corn yield distribution was centered around 10 ton/ha. This observation makes sense because irrigation is an intervention that is known to help the growth of the crop and thus explains the increase in yield.

From Figure 3-5 we observe that the non-irrigated yield data is from only 117 counties, most of which are in the states Kansas and Nebraska.



(a) Map of mean yield for non-irrigated corn (b) Zoomed-in map

Figure 3-5: Map of yield data averaged over years for non-irrigated corn

3.2 Results

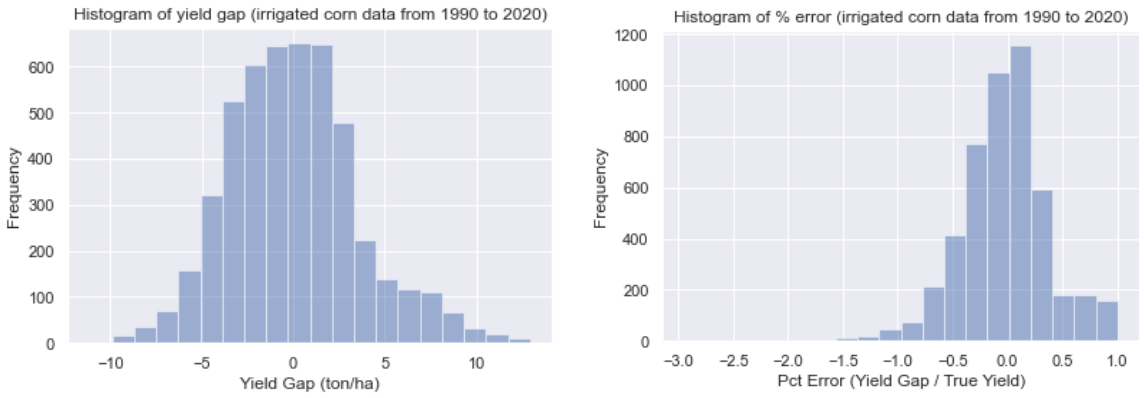
After generating yield data using the WOFOST simulation at all data points available in the US as explained in Section 3.1.1, we compare the simulated data to the true yield data based on the following metrics:

- Distributions of yield gap $Y_{\text{gap}} = Y_t - Y_w$ and percentage error $Y_{\text{pct}} = \frac{Y_{\text{gap}}}{Y_t}$, measuring how far the simulated yield Y_w is from the true yield Y_t,
- Time series of yield gap and percentage error averaged over counties for each year,
- Maps of yield gap and percentage error averaged over years for each county,
- R^2 (coefficient of determination), and p-value for the correlation between
  - simulated yield and true yield, and
  - mean yield gap (averaged over counties) and year.

3.2.1 Irrigated Yield Results

Let's first take a look at the distributions of the yield gap and percentage error as shown in Figure 3-6. In the left plot, we observe that the yield gap distribution is centered near zero with mean 0.01 ton/ha and standard deviation 3.57 ton/ha. Close

to zero mean of yield gap indicates that the WOFOST model simulated the yield close to the true yield on average. However the standard deviation of the yield gap is quite large, indicating that the WOFOST model is not very consistent. Percentage error distribution is also centered near zero with mean -0.03 and standard deviation 0.41. Again standard deviation of 41% indicates quite a large error for the simulation model.



(a) Histogram of yield gap

(b) Histogram of percentage error

Figure 3-6: Histogram of yield gap and percentage error between the true and simulated yield for irrigated corn

The bar plot in Figure 3-7 shows the time series of simulated yield and true yield averaged over counties. Height of each bar indicates the average of all data points in a given year. Blue bar indicates the true yield and orange line indicates the simulated yield values in ton/ha. Note that this plot does not show the yield trend in the US over the years since the number and the location of available data points are different for each year. We observe in this plot that on average the simulated yield is in the same order of magnitude as the true yield; however they do not have the same trend in time.

Figure 3-8 shows the yield gap and percentage error averaged over counties for each year. We observe that the yield gap and percentage error are mostly negative before year 1996 and mostly positive after year 2003. This indicates that the WOFOST

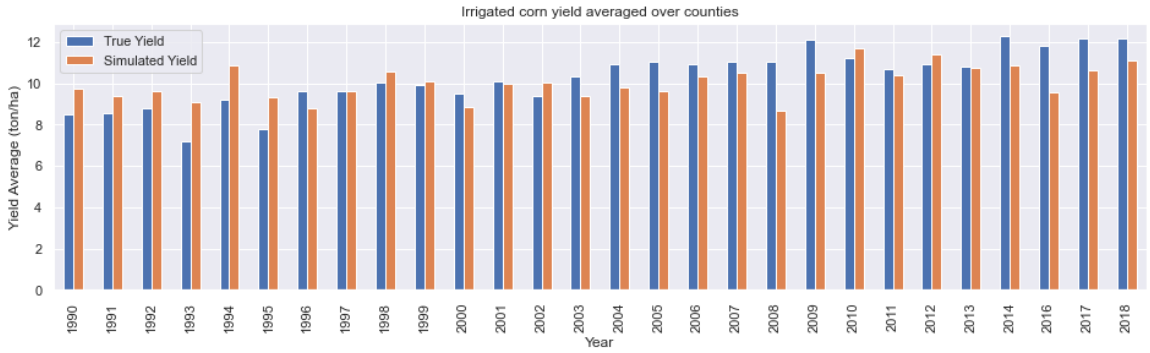
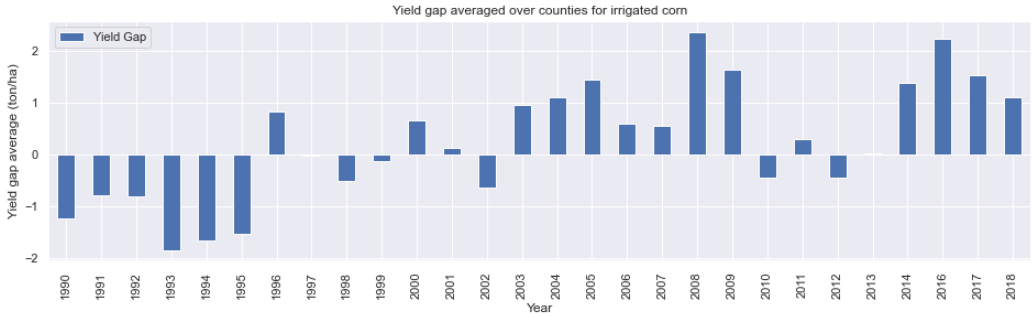
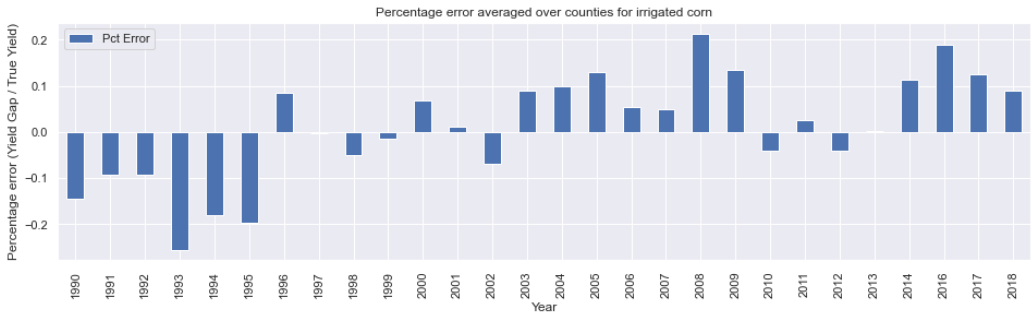


Figure 3-7: Simulated yield compared to true yield (averaged over counties) for irrigated corn

model underestimates the yield in earlier years while overestimating it in more recent years. Years 1997 and 2013 have close to zero yield gap and percentage error. Years 1999 and 2001 also have pretty low yield gap and percentage error. This means that the simulation model comes very close to the reality in predicting the yield in some years, which gives hope that the WOFOST results are not completely random.



(a) Yield gap



(b) Percentage error

Figure 3-8: Yield gap and percentage error as time series averaged over years for irrigated corn data

Next, we look at the trend of mean yield gap over the years to try to understand our observations from Figure 3-8. Figure 3-15 shows the scatter plot of mean yield gap (averaged over counties) vs. year, meaning that each point in the plot represents the average of all data points in a given year. As the regression line indicates, we observe a positive trend of the yield gap over the years with $R^2 = 0.473$ and p-value $= 0$. R^2 is not too low (>0.3), thus there is some effect of year in the yield gap; however it is also not very high (<0.5), thus the effect of year in the yield gap is only weak. This significant but weak relationship between the mean yield gap and year makes us also wonder about whether yield gap has a relationship with geographical location.

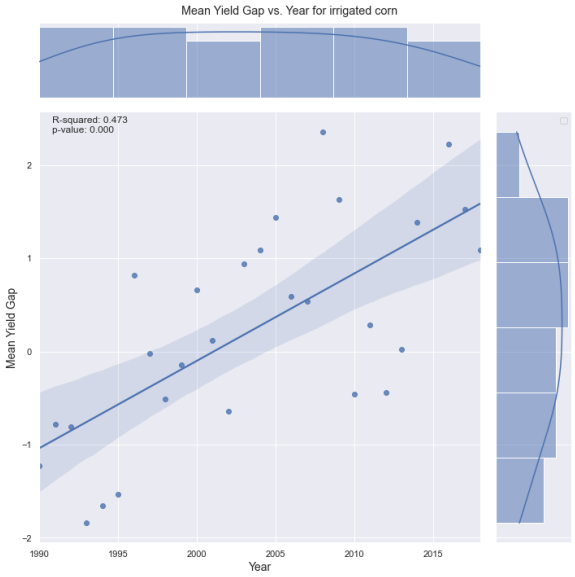
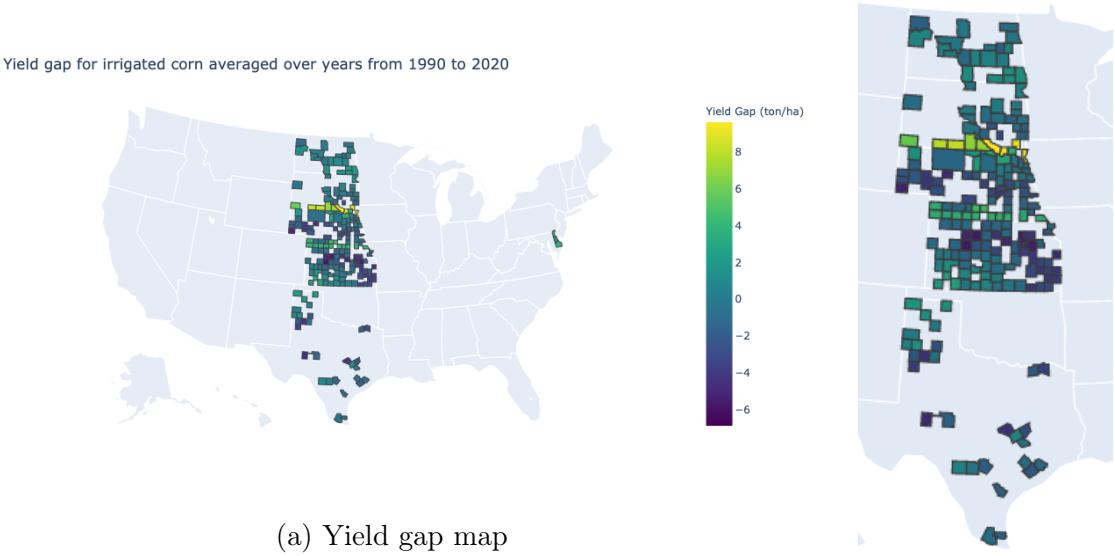


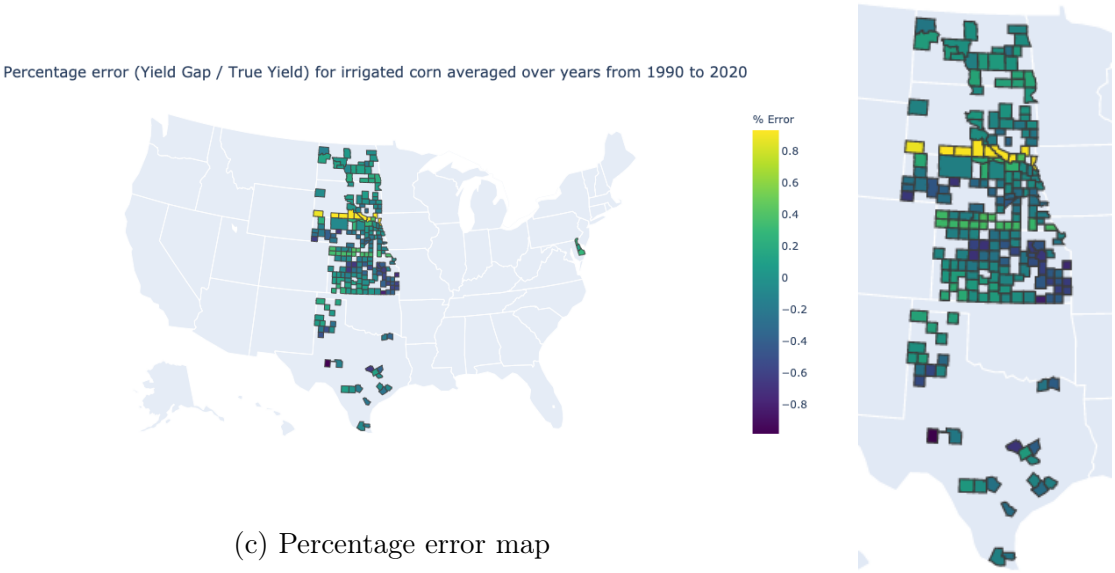
Figure 3-9: Scatter plot of mean yield gap (averaged over counties) vs. year

Figure 3-10 shows the geographical distributions of the yield gap and percentage error results on a county-level choropleth map of the United States. The colors indicate the yield gap and percentage error averaged over years for each county. Zoomed-in images of the maps are also included on the right to allow for closer inspection of the results. In the map, dark blue color indicates counties where the yield gap is on average negative, meaning that the WOFOST model overestimates the yield. Bright green and yellow color indicate the counties where the yield gap is on average



(a) Yield gap map

(b) Zoomed-in map



(c) Percentage error map

(d) Zoomed-in map

Figure 3-10: Maps summarizing the yield gap and percentage error of the simulated yield compared to true yield for irrigated corn data

positive, meaning that the WOFOST model underestimates the yield. Blueish green and greenish blue colors indicate that the yield gap is close to zero, meaning that

the WOFOST model on average closely estimates the yield. We further observe that counties in the southern border of South Dakota with Nebraska are colored in yellow; hence, the the WOFOST model underestimates the most in this region (more than 8 ton/ha yield gap and more than 80% error). Other parts of South Dakota and most of North Dakota, the WOFOST model seems to get close results to the true yield while underestimating a little (less than 2 ton/ha yield gap and less than 20% error). The dark blue colored counties are mostly in the southeast region of Kansas, and in a few counties in Texas and in Nebraska, where the WOFOST model overestimates by a lot (more than 6 ton/ha yield gap and more than 60% error).

Next, we investigate whether the simulated yield has a linear relationship with the true yield. Figure 3-11 shows the scatter plot of simulated yield vs. true yield using all the data points from the irrigated corn dataset. $R^2 = 0.014$ and the p-value $= 0$ indicating that there is no strong correlation between the simulated and the true yield. Dashed blue line with slope 1 shows the regression line that we would expect if the WOFOST model accurately simulated the yield. In order to understand the lack of strong relation between the simulated and true yield, we color the scatter plot by the state the data points belong. We observe that most of the data points that belong to the South Dakota have simulated yield close to zero, hinting at a possible issue with the input parameter values for counties in South Dakota.

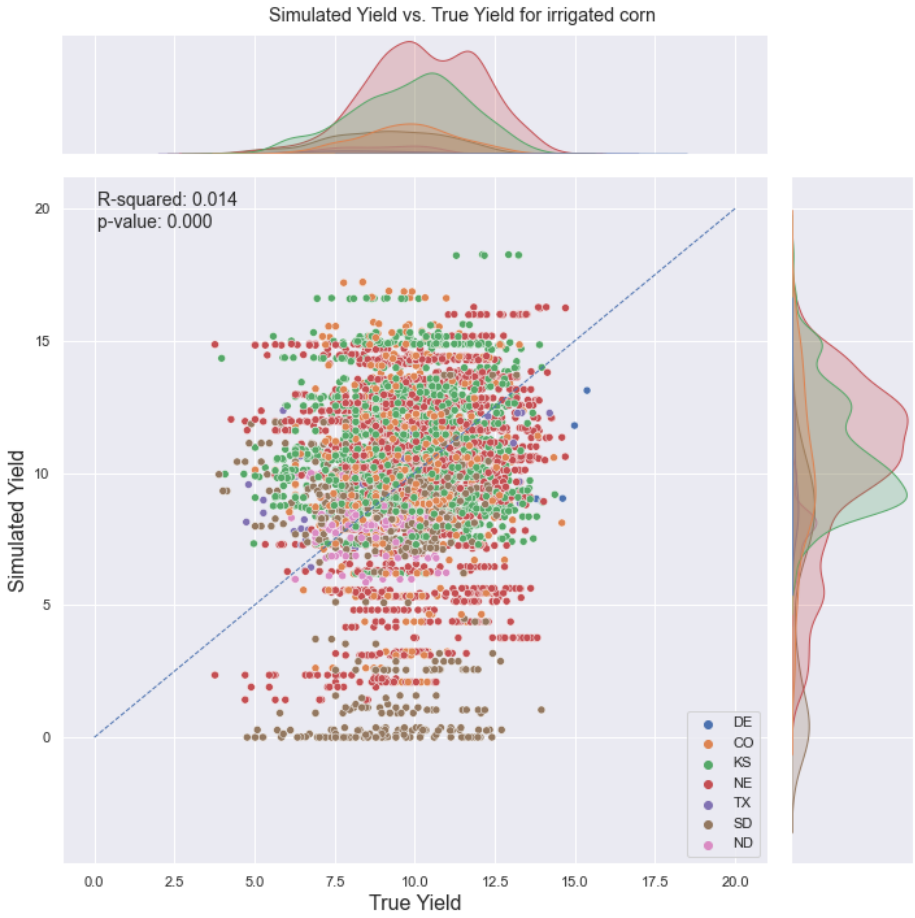
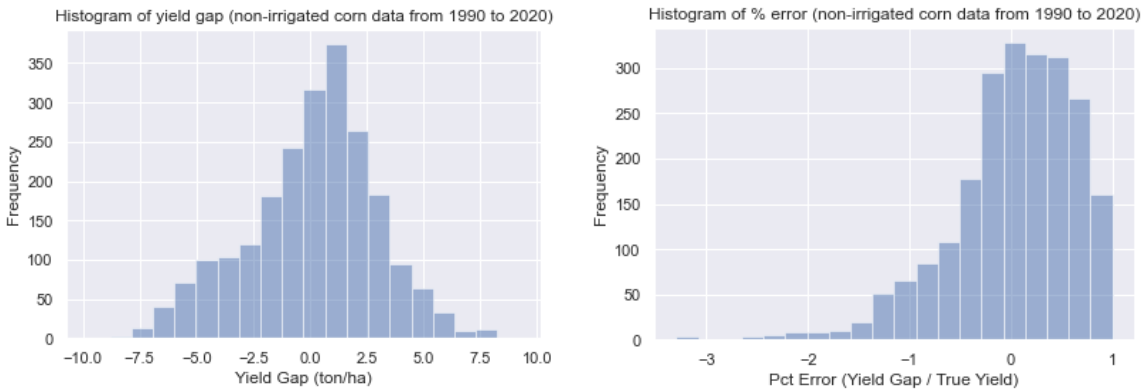


Figure 3-11: Scatter plot of simulated vs. true yield for all irrigated corn data (colored by state)

3.2.2 Non-irrigated Yield Results

We validate the results of the WOFOST simulation model for the non-irrigated corn data the same way we did for the irrigated corn data in Section 3.2.1. Here, the results are generated by running the simulation in the potential production setting compared to the water-limited setting used for the irrigated corn data.

Figure 3-12 shows the distributions of the yield gap and percentage error. We observe that the yield gap distribution is centered to the right of zero with mean 0.99 ton/ha and standard deviation 2.87 ton/ha. The positive mean indicates that the WOFOST model in the potential production setting underestimates the yield on average. The standard deviation of the yield gap is less than that of the standard deviation for the irrigated yield data; however it is still quite large considering the yield values are less for the non-irrigated data than those for the irrigated data. Percentage error distribution is centered to the right of zero, with mean 0.02 and standard deviation 0.62. Standard deviation of the percentage error is larger than that of the irrigated yield (41%) and this large standard deviation indicates that the simulated yield results vary a lot.



(a) Histogram of yield gap

(b) Histogram of percentage error

Figure 3-12: Histogram of yield gap and percentage error between the true and simulated yield for non-irrigated corn

Time series of simulated yield and true yield averaged over counties is shown in Figure 3-13. Again, note that this plot does not show the yield trend in the US over the years since the number and the location of available data points are different for each year. Blue bar indicates the true yield and orange line indicates the simulated yield values in ton/ha. This bar plot shows that true and simulated yield do not follow the same trend in time. We also observe that the difference between the simulated and true yield is much larger in certain years such as 1992, 1993, 1994, 2004, 2016 and 2017.

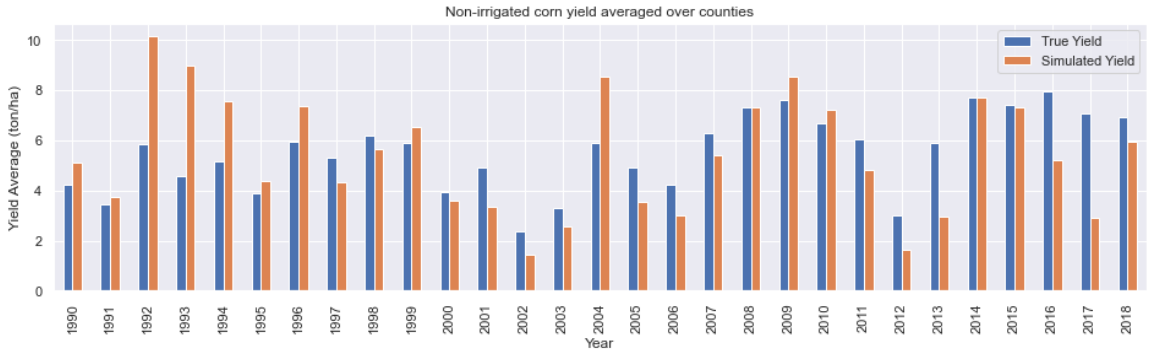
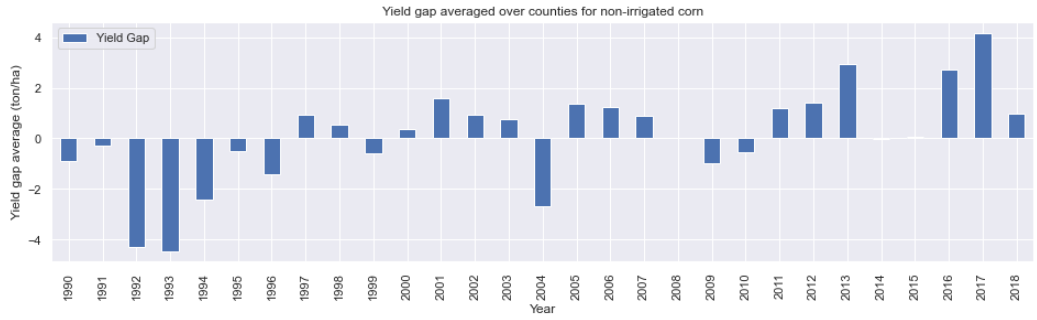


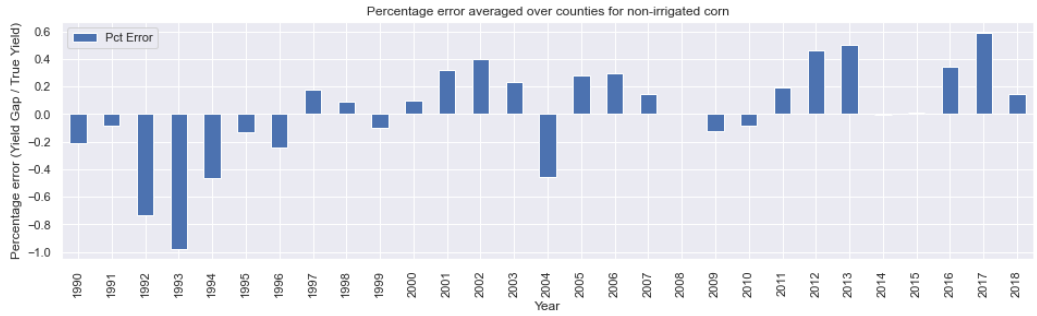
Figure 3-13: Simulated yield compared to true yield (averaged over counties) for non-irrigated corn

Figure 3-14 shows the yield gap and percentage error averaged over counties for each year. Similar to the observation for the irrigated yield data, the difference between true yield and simulated yield changes with year. The WOFOST model again overestimates in earlier years and underestimates in more recent years. We observe that the yield gap and percentage error are mostly negative before year 1997 (compared to 1996 for irrigated yield) and mostly positive after year 2011 (compared to 2003 for irrigated yield). Years 2008, 2013 and 2014 have close to zero yield gap and percentage errors. Years 1991, 1995, 1998 and 2000 also have pretty low yield gap and percentage error. This means that the simulation model in the potential yield setting also comes very close to the reality in predicting the yield in some years, which again gives hope that the WOFOST results are not completely random. We also observe

very high percentage errors in certain years, such as 1992 (close to 70%) and 1993 (close to 100%). These high errors also indicate that the simulated yield data in the potential setting can be very unreliable in certain years.



(a) Yield gap



(b) Percentage error

Figure 3-14: Yield gap and percentage error as time series averaged over years for non-irrigated corn data

We, again, find the trend of mean yield gap over the years to try to understand differences in yield gap over years. Figure 3-15 shows the scatter plot of mean yield gap (averaged over counties) vs. year. The regression line indicates a positive trend of the yield gap over the years with $R^2 = 0.393$ and p-value $= 0$. Similar to the irrigated yield results, we see a significant but weak relationship between the mean yield gap and year.

We also inspect the geographical distributions of the yield gap and percentage error results on a county-level choropleth map, shown in Figure 3-16. The colors indicate the yield gap and percentage error averaged over years for each county. Zoomed-in images of the maps are again included on the right to allow for closer inspection of the

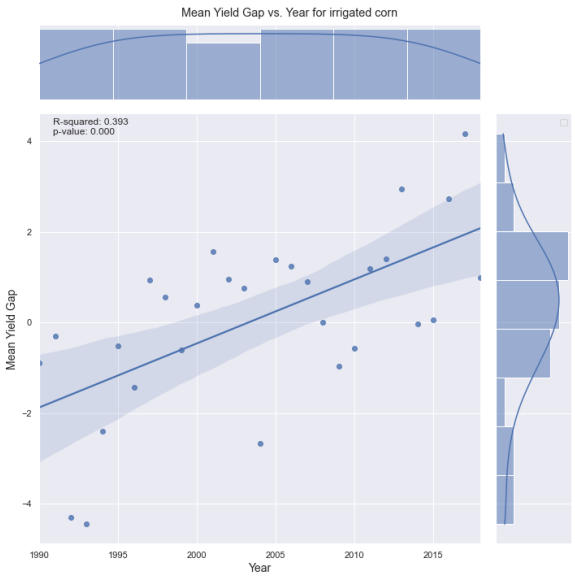
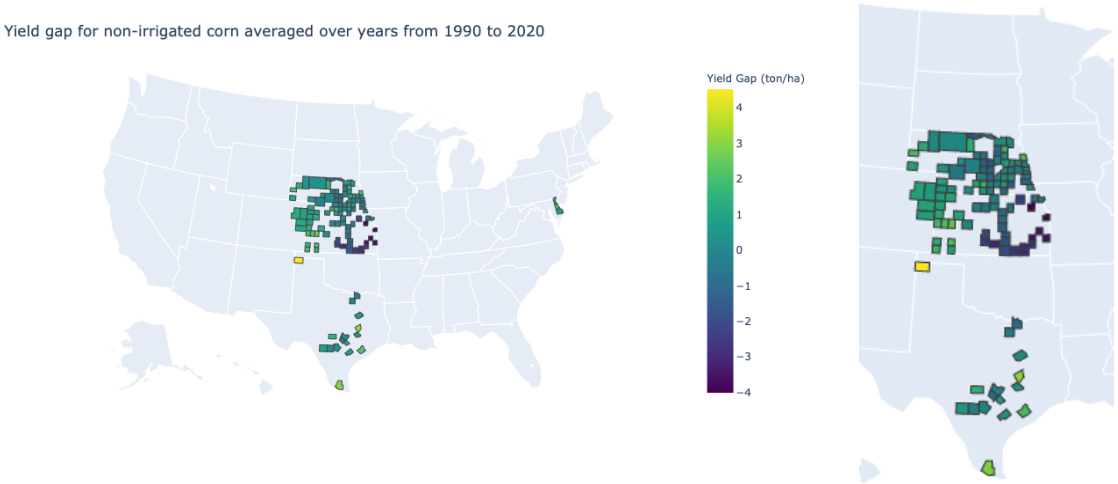


Figure 3-15: Scatter plot of mean yield gap (averaged over counties) vs. year

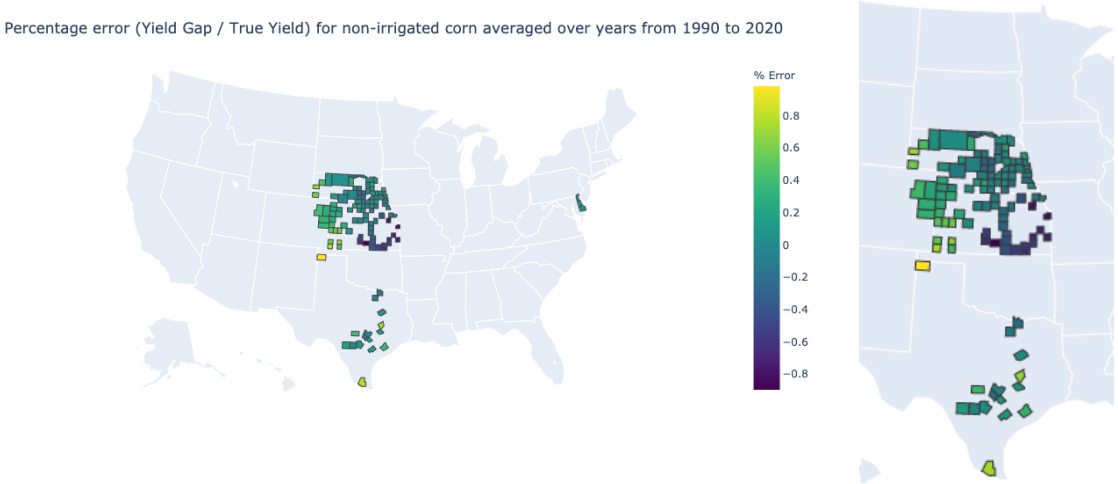
results. We observe that the WOFOST model underestimates the most in one county on the north-west of Texas (more than 4 ton/ha of yield gap and more than 80% of error) while most of the underestimated counties are located in the south-east region Nebraska (more than 3 ton/ha of yield gap and more than 60% of error). These maps show that the difference between simulated and true yield also has a geographical dependence.

To see whether the simulated yield has a linear relationship with the true yield, we plot the true yield and the simulated yield using all the data points from the non-irrigated corn dataset as shown in Figure 3-17. $R^2 = 0.375$ and the p-value $= 0$ indicate that there is a week correlation between the simulated and the true yield (larger than that of the irrigated corn with $R^2 = 0.014$). Dashed blue line with slope 1 shows the regression line that we would expect if the WOFOST model accurately simulated the yield. By coloring the scatter plot by the state the data points belong to, we also observe that all states have a similar trajectory. Nebraska seems to have the largest variation in simulated yield.



(a) Yield gap map

(b) Zoomed in map



(c) Percentage error map

(d) Zoomed in map

Figure 3-16: Maps summarizing the yield gap and percentage error of the simulated yield compared to true yield for non-irrigated corn data

3.3 Discussion

One important takeaway from the validation results for both irrigated and non-irrigated corn yield is the weak relationship between the mean yield gap and year for both irrigated and non-irrigated yield data. Another takeaway is that the difference between true yield and simulated yield also varies with geographic location. These differences could be due to different crop varieties used by farmers, accuracy of weather data, and/or accuracy of the soil data. The trend in mean yield gap could also be due to technological advancements in agriculture throughout years, such as

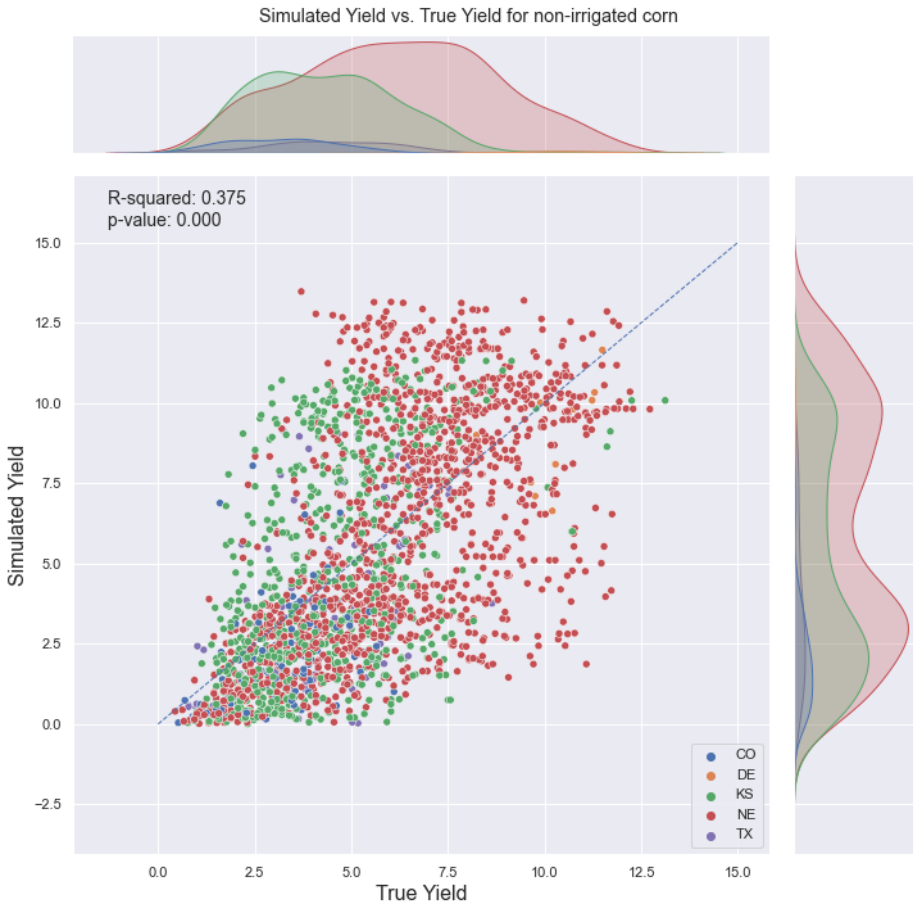


Figure 3-17: Scatter plot of simulated vs. true yield for all non-irrigated corn data (colored by state)

seed technology. Because we do not have any available data for the specific crop varieties used in a given county at a given year, we use constant crop parameters. However, crop cultivars show great variance from farm to farm and from year to year. In the validation simulations, our model does not capture the genetic differences between cultivars used in different locations because we use only one crop variety for all the counties and years. Observed trend in yield gap throughout the years and the geographical distribution of the error could be explained on the grounds of genetic differences of cultivars used in different counties and in different years. Thus, the for more meaningful results, the crop parameters of the WOFOST model should be calibrated for a region and a year.

Chapter 4

Calibration

The previous chapter has illustrated that an off-the-shelf usage of the WOFOST model to simulate yield in a new region does not generate reliable results. Because the regional application of WOFOST strongly depends on the values of crop parameters [18], using default crop parameters that developed for the seeds used in Europe 40 years ago is not a reliable method.

Cultivars can vary in their characteristics, such as their requirements for reaching maturity or flowering, their response to heat, cold, drought, and their ability to resist specific diseases and pests. Cultivars change region to region, climate to climate, country to country, due to environmental factors and local seed technologies. Hence, crop parameters are typically calibrated based on location-specific observations and optimized for areas with relatively homogeneous conditions [49]. Previous work [9, 19, 22] require experimental data for at least two years to calibrate crop parameters. However, the cost of data collection and lack of experimental data from many regions, especially regions in sub-Saharan Africa, make the calibration with observational data rather difficult.

To evaluate the crop simulation results of WOFOST at potential and water-limited production settings in the absence of observed data [22] developed a procedure, similar to that of the Global Yield Gap Atlas [9], checking the plausibility and consistency

of the WOFOST simulation results. This procedure requires an expert in agriculture evaluating visual maps of WOFOST outputs, such as mid-season and end-season leaf area index (LAI), yield, harvest index, and flowering and maturity dates. Since this calibration procedure requires domain knowledge and is tedious for the purpose of yield prediction, we propose a simpler approach.

In this work, for reliable estimation of yields at regional and farm scale, we develop a methodology to calibrate crop parameters with available yield data. In Section 4.1 we explain our formulation for the calibration objective and our optimization method. In Section 4.1.4 we detail our implementation of the optimization method and in Section 4.2 we present the results of the calibration.

4.1 Methodology

With calibration, our objective is to find the model parameters that generate outputs that are in well-alignment with the real-world data. Since we have no direct observation of the reality, that is, the specific crop seed used on the field, we use yield as a proxy.

Let us define the parameters that we want to calibrate as the vector θ. We formulate the objective function $J_{yield}(\theta)$ as the mean squared error (MSE) between the true yield Y_A and the simulated yield Y_W:

$$J_{yield}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(Y_A(i) - Y_W(i;\theta))^2, \tag{4.1}$$

where θ is the vector with parameters to calibrate, Y_A is the true yield, Y_W is the simulated yield, $i \in \mathcal{D}$ is a data point defined by a (location, year) pair in dataset $\mathcal{D}$, and $N = |\mathcal{D}|$. This objective function describes the distance between the simulated and true yield according to the dataset $\mathcal{D}$. Therefore, we minimize the objection function 4.1 to find the optimal parameters that are most descriptive of the data.

Following subsections further define the calibration objective function, by first explaining the choice of parameters to calibrate, i.e. the vector θ in Section 4.1.1. Then we discuss the criteria for choosing a dataset D for the calibration task in Section 4.1.2, and in Section 4.1.3 explain in detail our method of solving the optimization problem.

4.1.1 Parameter Selection for Calibration

Because WOFOST is a complex model controlled by 52 crop parameters, we first investigate which of these parameters are the most important for calibration. Previous work [22, 19] developed calibration procedures for the phenological development of the model. Phenological development describes successive stages that a crop passes through. In the WOFOST model these stages are expressed in degree-days and defined by two parameters: `TSUM1` and `TSUM2`. The `TSUM1` parameter defines the number of degree-days for the emergence-anthesis period. The `TSUM2` parameter defines the number of degree-days for the anthesis-maturity period [20]. These two parameters describe the crop seed variety. Hence for the purpose of finding the parameter values that best describe the crop variety used in a region, we choose these two parameters for the calibration objective. We define $\theta = [\theta_0, \theta_1]$, where $\theta_0 =$ `TSUM1` and $\theta_1 =$ `TSUM2`.

4.1.2 Data

With calibration, we are trying to find the parameters that best explain the available yield data in a region, therefore choosing the dataset $\mathcal{D}$ to calibrate on is an important task. Data points $i \in \mathcal{D}$ in the objective function 4.1 are defined by a (location, time) pair, therefore choosing the dataset $\mathcal{D}$ correspond to making assumptions about the regional scale and the time scale of seed (crop variety) changes.

Our first assumption is on the regional scale of the dataset $\mathcal{D}$. Realistically seeds (crop variety) used in each farm can be different from each other. Crop testing data

from [5] demonstrates that there can be more than 50 different seed types (crop varieties) used in a given county in Iowa. This is because each farmer has access to, and prefers, different types of seeds for their fields. Therefore, the most realistic regional scale for calibration would be at the farm-level, or a region that is known to use same crop variety. Since farm-level data is difficult to obtain, and since the most granular yield data we have is at the county level, we assume that farmers in a given county use similar seeds (so that we calibrate the crop parameters at a county-level). With the assumption that the counties in a given state from the US Corn Belt states are similar to each other, we use data points from all the counties of a given state. This assumption is necessary since the only available data for crop calendars are at the state-level.

Our second assumption is on the time scale of the dataset $\mathcal{D}$. Data from USDA NASS [8] demonstrates that the US corn grain yields have steadily increased since the late 1930s [36]—and more than half of the yield gains are attributed to genetic improvements achieved by plant breeders [24]. We need to take into account the crop variety changes amongst years for a more realistic and reliable simulation. Since crop varieties are switched out after 4-5 years in the Corn Belt, we choose a sliding window size of 4 years for the calibration dataset.

To summarize our assumptions, the regional scale for the calibration dataset is statewide county yield data and the time-scale is 4 years. This means that to calibrate crop parameters for a county for a given year, we use yield data of the previous 4 years from all counties in the given state. In Section 4.2, we show the results of the calibration on our dataset.

4.1.3 Evolutionary Algorithms

Because the objective function does not have an explicit derivative and it is not easy to compute the gradient numerically, gradient-based optimization algorithms cannot be deployed to find the optimal parameters that minimize the objective function. As

a gradient-free (stochastic) optimization method, we decided to use the evolutionary algorithms, which are inspired by the principle of Darwinian natural selection.

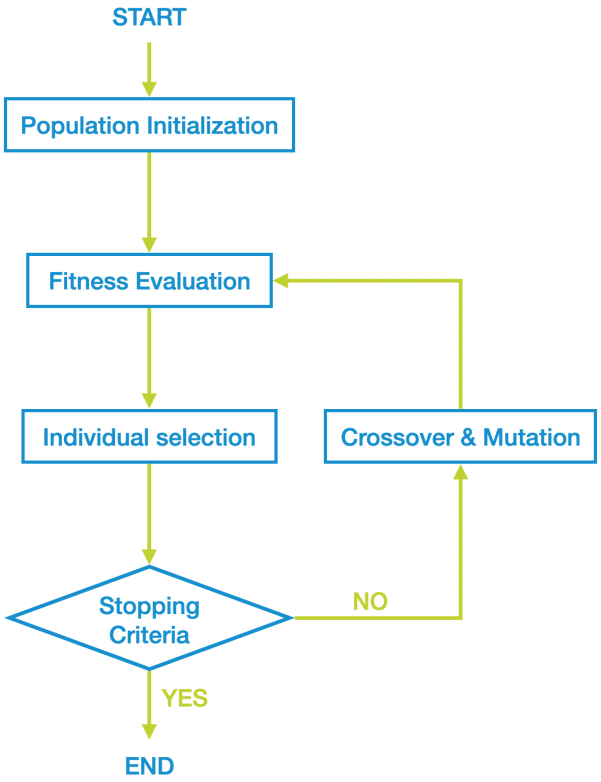


Figure 4-1: Evolutionary algorithm diagram

Evolutionary algorithms generate many possible solutions for a problem and then evaluate how well the solutions solve the problem using a given fitness function. The *fittest* solutions procreate with some randomness. Figure 4-1 provides a high-level summary of the evolutionary algorithms.

A typical evolutionary algorithm requires a genetic representation of the solution domain, and a fitness function to evaluate the solution. First, an initial population consisting of candidate solutions—that is, individuals with varying genes—is created. The fitness function is then used to evaluate each individual solution. The individu-

als with lowest fitness scores are then removed from the population. If the stopping criteria is met, the algorithm terminates and outputs the fittest individuals so far. If not, however, the selected individuals (also called, survivors) from the remaining population crossover. During crossover, mutations can happen, albeit with a low probability. After crossover and mutation, the new population consists of the older generation with the best fitness and their offspring. The algorithm continues this fitness evaluation and generation of new populations until the stopping criteria is finally met. Stopping criteria could be that a solution that minimizes the objective function is found, or that a fixed number of generations has been reached, or that the highest ranking solution's fitness has reached a plateau such that successive iterations no longer produce necessarily better results, or a combination thereof.

In the setting of crop parameter calibration, our goal is to find the θ that best explains the data; hence, we use the θ vector as our gene representation for the evolutionary algorithm, as illustrated in Figure 4-2.

Figure 4-2: Gene representation of crop parameters for the evolutionary algorithm

4.1.4 Implementation

We implement the evolutionary algorithm using the `DEAP` software package, which is an evolutionary computation framework that works with multiprocessing mechanisms such as `multiprocessing` and `SCOOP` [25]. We formulate the genetic representation for our problem as $\theta = [\texttt{TSUM1}, \texttt{TSUM2}]$ and the fitness function as the objection function 4.1. Another advantage of using an evolutionary algorithm is that it does not contain many hyperparameters to tune. The only hyperparameters are the population size and the number of generations. We chose the stopping criteria to be the number of generations reaching the set threshold of 15, because we observed that running the algorithm for 15 generations was sufficient enough to have a convergence. We choose

a initial population size of 10, five of which are pre-determined "good" candidates and the other five are randomly generated. "Good" candidates are the five varieties of corn defined in [1] which correspond to pairs of values for (`TSUM1, TSUM2`).

4.2 Results

Considering the regional and time scale explained in Section 4.1.2, we tested our calibration method on the corn yield data from counties in one single state. According to the USDA NASS data, four states that had produced over one billion bushels of corn in 2020 are Iowa, Illinois, Nebraska, and Minnesota [8]. Iowa was the state with the most corn production [42]. Because it is one of the states in the US Corn Belt and dominates corn production in the US, we chose Iowa to test our calibration method. Figure 4-3 shows a map of the true corn yield average by county in Iowa from years 1990 to 2019. On the map, darker green indicates more yield and lighter green indicates less yield while white indicates no data.

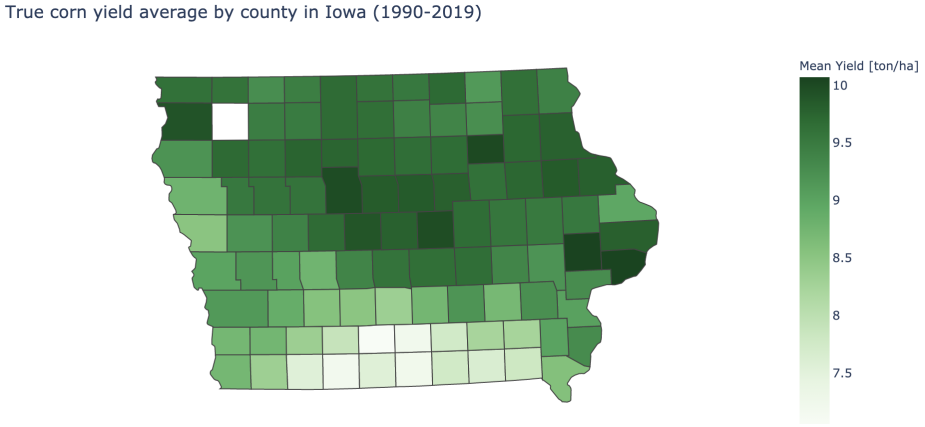


Figure 4-3: True corn yield average by county in Iowa (1990-2019)

Figure 4-4 shows the calibration results where columns `TSUM1` and `TSUM2` show the optimal crop parameters determined by the evolutionary algorithm using yield data from all the counties of years indicated by the *Training Years* column. Calibration is done for the year given in the *Test Year* column using data from the *Training Years* column. Column *Train MSE* is the final mean squared error on the training

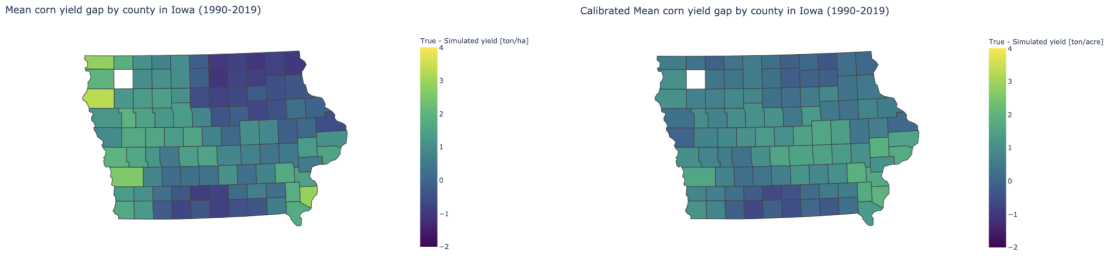
set achieved by the values given in TSUM1 and TSUM2 columns. Notice that in this table, some rows contain less than four training years. This is because there was no available data for some years in the previous four years of the test year.

	TSUM1	TSUM2	Training Years	Train MSE	Test Year	Test MSE
0	1025	1535	[1990]	0.83	1991	1.16
1	1016	1329	[1990, 1991]	0.77	1992	3.03
2	392	1301	[1990, 1991, 1992]	1.18	1993	4.91
3	365	1525	[1990, 1991, 1992, 1993]	1.88	1994	0.92
4	409	625	[1991, 1992, 1993, 1994]	2.19	1995	1.45
5	372	1113	[1992, 1993, 1994, 1995]	1.95	1996	1.43
6	1025	1281	[1993, 1994, 1995, 1996]	2.05	1997	1.91
7	995	1220	[1994, 1995, 1996, 1997]	1.26	1998	1.4
8	995	1170	[1995, 1996, 1997, 1998]	1.44	1999	1.24
9	1018	778	[1996, 1997, 1998, 1999]	1.14	2000	0.98
10	996	755	[1997, 1998, 1999, 2000]	0.85	2001	1.62
11	1005	620	[1998, 1999, 2000, 2001]	0.93	2002	2.4
12	990	610	[1999, 2000, 2001, 2002]	1.25	2003	3.42
13	855	715	[2000, 2001, 2002, 2003]	1.36	2004	2.94
14	915	615	[2001, 2002, 2003, 2004]	1.63	2005	6.23
15	850	625	[2002, 2003, 2004, 2005]	2.52	2006	7.01
16	670	610	[2003, 2004, 2005, 2006]	3.5	2007	0.62
17	675	625	[2005, 2006, 2007]	2.11	2009	11.07
18	550	625	[2006, 2007, 2009]	4.28	2010	2.4
19	531	610	[2007, 2009, 2010]	2.52	2011	6.03
20	995	766	[2009, 2010, 2011]	2.7	2012	9.65
21	600	700	[2010, 2011, 2012]	2.81	2014	2.29
22	596	632	[2012, 2014]	2.21	2016	6.7
23	683	840	[2014, 2016]	4.36	2017	3.9

Figure 4-4: Calibration results for maize in Iowa

Figure 4-5 shows the maps of yield gap before and after calibration. Color scale of the two plots are the same. In these maps, color yellow indicates a high and positive yield gap whereas dark blue indicates a high and negative yield gap. Colors in between green and blue indicate yield gaps close to zero. We used the same color scale for both maps to show the effect of calibration. On the left map, which shows the yield gap before calibration, we observe that the difference between true yield and simulated yield varies over regions widely. However, with calibration, the yield gap became more uniform around the state as shown on the right map. Inspecting the color differences between the pre-calibration and post-calibration maps, we also observe that the yield gap and the percentage error values were halved after the cal-

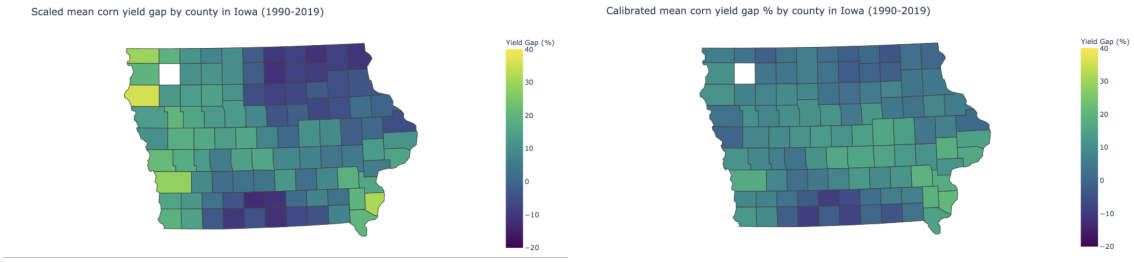
ibration. Thus, we can say that the calibration of the crop parameters improved the quality of the simulated yield data.



(a) pre-calibration (b) post-calibration

Figure 4-5: Maps of yield gap in Iowa by county for corn

Figure 4-6 also demonstrates the similar calibration effect on the percentage error (which is calculated as the difference between true yield and simulated yield divided by the true yield).



(a) pre-calibration (b) post-calibration

Figure 4-6: Maps of percentage error in Iowa by county corn

4.3 Discussion

With calibration, our objective was to find the model parameters that generate yield data that are in well-alignment with the real-world yield data. We formulated this calibration task as minimizing the mean squared error between simulated yield and the true yield. Using the gradient-free optimization method evolutionary algorithms, we calibrated the parameters of a given year by minimizing the mean squared error of the yield gap for the previous four years in all counties in the state. Using a sliding

window of size four, we found the optimal crop parameters for each year that are most likely to produce the true yield data from Iowa. We observed that the calibration significantly reduced the gap between the simulated and true yield in all the counties, but not to zero. We chose the time scale of the data set to use for optimization to be four years, however it would valuable to run the calibration with a data set of time scale $\in 2, 3, 4, 5$ years and compare the MSEs and the optimal parameters for each time scale.

Another choice we made with the data set was the regional scale. We chose the entire state as the regional scale of the data, however crop varieties differ from county to county. If we were to use larger time scale for the data, we could use smaller regional scale such as county. It would also make sense to use ecoregions as the regional scale. Ecoregions denote areas of general similarity in ecosystems and in the type, quality, and quantity of environmental resources. An ecoregion is identified through patterns and composition of both biological and physical characteristics, including geology, physiography, vegetation, climate, soils, land use, wildlife, and hydrology.

Further evaluation of the calibration method could be possible by comparing the ground truth values for the crop parameters in certain counties or states to the optimal values found by the evolutionary algorithm. However, we could not find any available data for corn parameters `TSUM1` and `TSUM2`.

Chapter 5

Uncertainty Quantification

It is essential that the consumers of the present data, including, but not limited to, farmers, stakeholders, and governments, understand and appreciate the significance and the limitations of the data proposed in the Data Platform. Let us consider a farm, for instance. When we input its weather for the entire crop calendar, the crop the farmer is using, the characteristics of the soil, farmer's irrigation and fertilization pattern to WOFOST, the simulation model outputs a single number as the simulated yield at the end of the crop season. How much can the farmer trust this number? What if the farmer was mistaken about the exact type of seed he was using? What if they could not to not irrigate on the exact date they planned to?

The Data Platform should provide information about how much the actual yield can deviate from the simulated yield due to variations in the inputs of the model. Hence, the main goal of uncertainty quantification work is to identify the sources of yield uncertainty. Our approach is to statistically vary the input parameters of WOFOST through sensitivity analysis, and to analytically propagate uncertainty from inputs to outputs in the dynamical system of plant growth, to shortlist a handful of parameters that essentially control the crop yield prediction.

This chapter presents these two methods for obtaining confidence intervals for the output of the simulation model. Section 5.1 discusses the sensitivity analysis approach,

and Section 5.2 presents our simplified model for plant growth (an approximate model to WOFOST) and results of error propagation on our model.

5.1 Sensitivity Analysis

For sensitivity analysis of the input parameters, we use the Sobol method [45], which is intended to determine how much of the variability in model output is dependent upon each of the input parameters, either upon a single parameter or upon an interaction between different parameters Sobol method computes sensitivity indices by Monte Carlo (or quasi-Monte Carlo) methods, which are used for estimating the influence of individual variables or groups of variables on the model output. Sobol sensitivity analysis is not intended to identify the cause of the input variability. It just indicates what impact and to what extent it will have on model output. The Sobol sensitivity index is defined as follows:

$$S_i = \frac{\mathrm{Var}(\mathbb{E}[Y|Q_i])}{\mathrm{Var}(Y)} \tag{5.1}$$

where, $\mathbb{E}[Y|Q_i]$ denotes the expected value of the output Y when parameterQ_i is fixed. The first order sensitivity index tells us the expected reduction in the variance of the model when we fix parameter Q_i. The sum of the first order Sobol sensitivity indices can not exceed one [28]. The total Sobol sensitivity index S_{Ti} includes the sensitivity of both first order effects as well as the sensitivity due to interactions (covariance) between a given parameter Q_i and all other parameters [31]. It is defined as:

$$S_{Ti} = 1 - \frac{\mathrm{Var}(\mathbb{E}[Y|Q_{-i}])}{\mathrm{Var}(Y)} \tag{5.2}$$

where Q_{-i} denotes all parameters except Q_i. The sum of the total Sobol sensitivity indices is equal to or greater than one [28]. If no higher order interactions are present, the sum of both the first and total order Sobol indices are equal to one.

To compute the Sobol sensitivity indices, we use `SALib` (Sensitivity Analysis Library in Python) [7]. We first generate 1000 equally distanced samples for each parameter

Q_i. We chose the number of samples to be 1000 because higher values provide better estimates of sensitivity even though they increase computation time. Then, by using the Saltelli sampler [3] to we generate a number of combinations of parameter values. One limitation of this method is that it is only applicable for the scalar parameters in WOFOST, which also has a number of tabular parameters defined as a function of development stage or temperature. The sensitivity of these tabular parameters cannot be properly analyzed with this approach. We omit the sensitivity analysis for weather parameters because because weather parameters are all in the form of vectors containing time-series weather data for the entire growing season. We also omit the sensitivity analysis of soil parameters because the simulation throws errors when a combination of soil parameters do not work well together.

5.1.1 Sensitivity Indices of Crop Parameters

Figure 5-1 shows the sensitivity indices for the scalar crop parameters of WOFOST. We used 1000 equally distanced samples in the range specified in the WOFOST manual [20] for each scalar parameter. We didn't include tabular parameters to not complicate the analysis. In the plot, blue bars indicate the first order sensitivity indices and the orange bars indicate the total order sensitivity indices. We can interpret the sensitivity indices in Figure 5-1 as how much of the variance in the model output each parameter is responsible for. Parameters such as `Q10`, `CFET`, `RML`, `RMR`, `PERDL`, have zero first order and total order sensitivity indices, therefore variations of these parameters result in comparatively small variations in the final model output. On the other hand, parameters `SPAN`, `TSUM1`, `CVO` and `TBASE` have high sensitivity indices compared to all the other parameters, therefore we can say that a change in these parameters lead to a more dramatic change in the model output.

Next we calculate the second order sensitivities to see the effect of varying two parameters Q_i and Q_j simultaneously, additional to the effect of their individual variations. Figure 5-2 shows that the second order sensitivities are close to zero for all parameter pairs. This indicates that the fractional contribution of parameter interactions to the

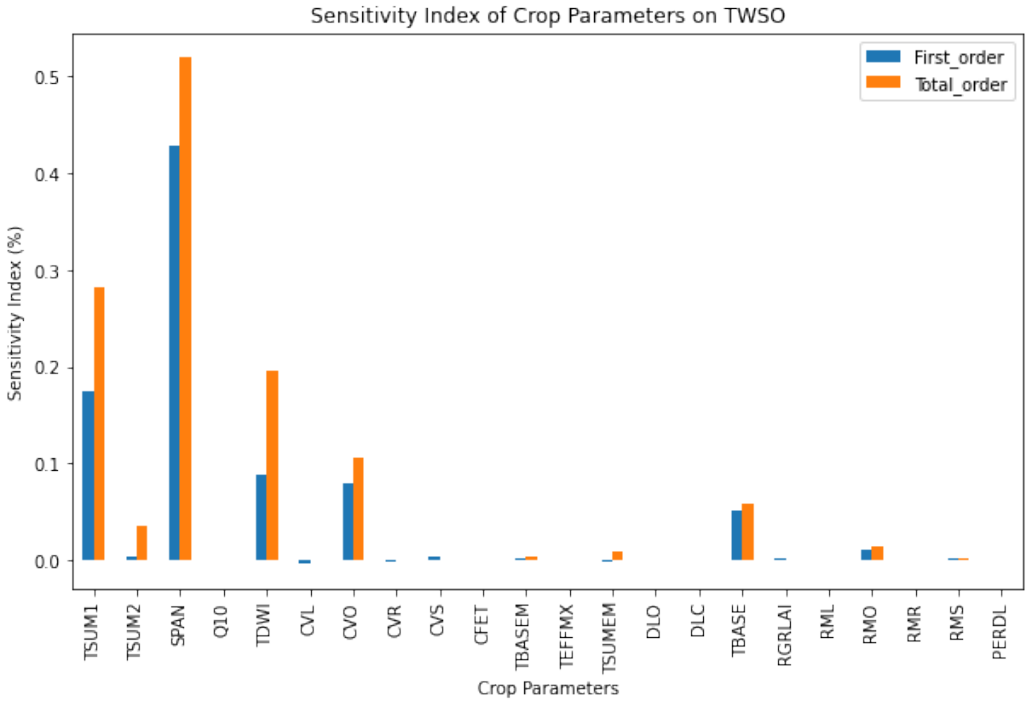


Figure 5-1: Sensitivity indices for scalar crop parameters

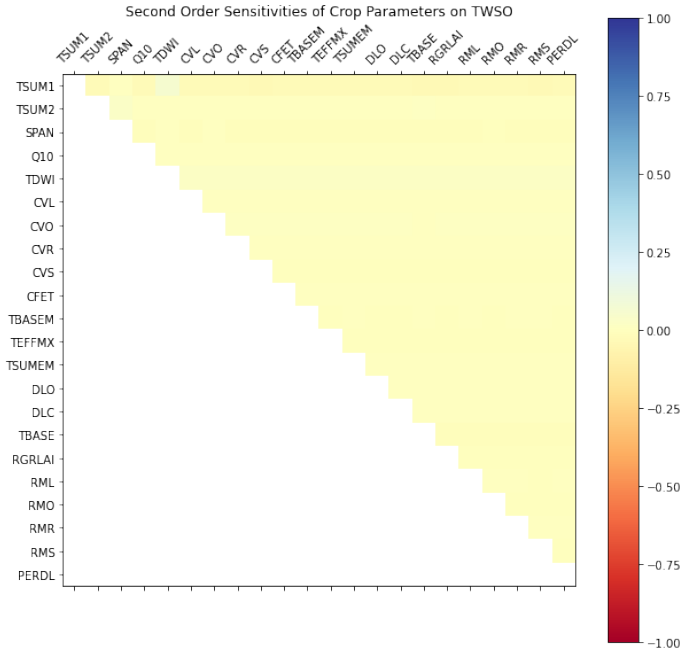


Figure 5-2: Second order sensitivity indices for scalar crop parameters

output variance are not significant.

5.2 Perturbation analysis

To quantify the uncertainty of the output of the simulation model, our second approach is to analytically propagate error through the dynamical system of the plant growth that the simulation model describes. One challenge is that the dynamics of plant growth described by the WOFOST model is very complex, therefore perturbation to the system is not possible to solve analytically. Because WOFOST is setup as a simulation model, there is no explicit analytical solution that can be written from the rate equations. In this section, we first explain the exact system WOFOST model describes in Section 5.2.2 and then we propose a simplification to the model in Section 5.2.3 that we can write an analytical solution to. Finally, we present the methodology and results for the perturbation on the simplified model in Section 5.2.7.

5.2.1 Notation

Notation we use for the rest of the chapter can be summarized as follows.

- WOFOST assumes the plant has four organs and we use $i \in [1, 2, 3, 4]$ to indicate the plant organ, $i = 1$ corresponding to the root, $i = 2$ to the leaves, $i = 3$ to the stem, and $i = 4$ to the storage organs.

- $t \in [0, ..., T]$ indicates time and is measured in days. $t = 0$ corresponds to the day the simulation is started at.

- We use capital letters to indicate state variables, which are the total plant weight W, total plant assimilation rate $A(t, W)$, total plant respiration rate $R(t, W)$, the unit assimilation rate $U(t, W)$, etc. This means that the capital letters indicate dependency to the plant weight W.

- We use lowercase letters to indicate known functions of time, e.g. $c(t)$, $z(t)$.

- We use Greek letters to indicate input parameters to the model, e.g. the partitioning factor $\pi_i(t)$, ξ_i.

5.2.2 WOFOST model for plant growth

Disclaimer: This section summarizes our understanding of the WOFOST model. Please refer to the WOFOST manual [20] by de Wit et al. for the original equations.

The WOFOST model is defined by the total plant growth rate equation:

$$\frac{\partial W}{\partial t} = c(t)(g(t)\, A(t, W) - R(t, W)), \tag{5.3}$$

where $\frac{\partial W}{\partial t}$ is the growth rate of the entire plant, i.e. change in total weight of the plant, $c(t)$ is the conversion efficiency factor of assimilates, $g(t)$ is the reduction factor due to plant transpiration (i.e. a function of soil moisture), $A(t, W)$ is the total plant assimilation rate and $R(t, W)$ is the total plant respiration rate.

The sum of the weight of the organs are equal to the total weight of the plant and the rate of change of weight of each organ $\frac{\partial W_i}{\partial t}$ is directly proportional to the rate of change of total weight of the plant $\frac{\partial W}{\partial t}$:

$$W(t) = \sum_{i=1}^{4} W_i(t) \tag{5.4}$$

$$\frac{\partial W_i}{\partial t} = \pi_i(t) \frac{\partial W}{\partial t}, \tag{5.5}$$

Respiration rate $R(t, W)$ is a measure of the loss to the plant breathing and can be expressed as follows:

$$R(t, W) = b(t) \sum_{i=1}^{4} \zeta_i W_i(t) \tag{5.6}$$

$$b(t) = \phi^{\frac{\tau_t - \tau_r}{10}} \tag{5.7}$$

where τ_t is the (average) temperature on day t, $\tau_r = 25°C$ is the reference temperature, and ϕ is a scalar crop parameter, ζ_i is the maintenance coefficient of organ i.

The conversion efficiency factor of assimilates $c(t)$ is a known function of time and crop parameters:

$$c(t) = \left(\frac{\pi_1(t)}{\xi_1} + (1 - \pi_1(t)) \sum_{i=2}^{4} \frac{\pi_i(t)}{\xi_i} \right)^{-1}, \tag{5.8}$$

where ξ_i are scalar crop parameters, and $\pi_i(t)$ are tabular crop parameters. π_i depends on the development stage of the crop $d(t, \tau)$, which describes whether the plant is emerging, flowering, reached maturity, etc. Development stage $d(t, \tau)$ is a known function of temperature from day 0 to t (indicated as $\tau = [\tau_0, ..., \tau_t]$), and omitted in this section for brevity.

Assimilation rate is a measure of the photosynthetic efficiency of the plant and it describes the rate of glucose ($C_6H_{12}O_6$) produced in the following chemical equation.

$$6\,CO_2 + 6\,H_2O + (\text{energy}) \longrightarrow C_6H_{12}O_6 + 6\,O_2 \tag{5.9}$$

Assimilation rate of the plant is an average over the day D, and along the canopy C. Averages are calculated using the 3-point Gaussian integration method, i.e. three points (hours) during the day ($h \in [-1, 0, 1]$) and three points (levels) on the canopy ($l \in [-1, 0, 1]$).

$$A(t, W) = \delta(t) \int_D L(t, W) \int_C U(t, W, l, h)\, dl\, dh \tag{5.10}$$

where $\delta(t)$ is the daylength obtained from daily weather data and $L(t, W)$ is the leaf area index.

Leaf area index (LAI) is a measure for the total area of leaves of the plant per unit ground area and it's unitless. WOFOST model describes $L(t, W)$ in two different forms, depending on the growth stage of the plant. $L_{exp}(t)$ describes the exponential growth stage and L_{sc} describes the source-limited growth stage.

$$\frac{\partial L_{exp}}{\partial t} = \lambda(t)\, \tau_e\, L_{exp}(t), \tag{5.11}$$

where $\lambda(t)$ is the maximum relative increase of LAI, τ_{eff} is the daily effective temperature (i.e. the number of degrees above the base temperature τ_{base} which is a scalar crop parameter).

$$\frac{\partial L_{sc}}{\partial t} = \mu(t)\,\frac{\partial W_2}{\partial t}, \tag{5.12}$$

where $\mu(t)$ is the specific leaf area, the ratio of leaf area to leaf dry mass, and W_2 is the weight of leaves.

$U(t, W)$ is the assimilation rate per unit leaf area, referred to as the net assimilation rate (NAR) in the literature [23], [30], [46]. WOFOST model calculates the NAR $U(t, W)$ as a weighted average of the assimilation by sunlit leaves $U_{sl}(t)$ and the assimilation by shaded leaves $U_{sh}(t, W)$ [47]. The weights are the fraction of sunlit leaves $z(t)$ and shaded leaves $(1 - z(t))$.

$$U(t, W, l, h) = z(t, l)\,U_{sl}(t, W, l, h) + (1 - z(t, l))\,U_{sh}(t, W, l, h) \tag{5.13}$$

$$z(t, l) = e^{-\kappa_{dr,bl} L_l(t,W)} \tag{5.14}$$

$$L_l(t, W) = (0.5 + l\sqrt{0.15})\,L(t) \quad \forall\, l \in [-1, 0, 1]\,, \tag{5.15}$$

where $L_l(t, W)$ is the leaf area index on the level l of the canopy.

Unit assimilation rates $U_{sl}(t, W)$ and $U_{sh}(t, W)$ are calculated using the light absorbed by sunlit leaves $I_{sl}(t, l, h)$ and by shaded leaves $I_{sh}(t, l, h)$.

$$U_{sh}(t, W, l, h) = \alpha(t)\,(1 - e^{-\frac{\epsilon(t)\,I_{sh}(t,l,h)}{\alpha(t)}}) \tag{5.16}$$

$$U_{sl}(t, W, l, h) = \alpha(t)\,\left(1 - (\alpha(t) - U_{sh}(t, W))\,\frac{1 - e^{-\frac{\epsilon(t)\,I_{sl}(t,l,h)}{\alpha(t)}}}{\epsilon(t)\,I_{sl}(t, l, h)}\right), \tag{5.17}$$

where $\alpha(t)$ is the maximum assimilation rate, $\epsilon(t)$ is the initial light use efficiency, both of which are tabular crop parameters.

The light absorbed by sunlit and shaded leaves are calculated as follows:

$$I_{sl}(t,l,h) = I_{sl}(t,h) = \frac{(1-\sigma)\, I_{0,dr}(t,h)}{\sin\beta} \tag{5.18}$$

$$I_{sh}(t,l,h) = I_{a,df}(t,l,h) + I_{a,dr,t}(t,l,h) - I_{a,dr,dr}(t,l,h), \tag{5.19}$$

where $I_{0,dr}$ is the direct part of the photosynthetically active radiation (PAR) flux at the top of the canopy at point h during the day, *sigma* is the scattering coefficient, and β is the solar elevation. $I_{a,\cdot}(t,l,h)$ indicates the absorbed PAR at point l on the canopy and at hour h during the day. $I_{a,df}$ for the diffuse flux, $I_{a,dr,t}$ is for the total direct flux and $I_{a,dr,bl}$ is for the direct flux on non-reflective ('black') leaves.

Intensity of the aforementioned types of radiation is calculated as follows:

$$I_{a,df}(t,l,h) = \kappa_{df}(1-\rho) I_{0,df}(t,h)\, e^{-\kappa_{df} L_l(t,W)} \tag{5.20}$$

$$I_{a,dr,t}(t,l,h) = \kappa_{dr,t}(1-\rho) I_{0,dr}(t,h)\, e^{-\kappa_{dr,t} L_l(t,W)} \tag{5.21}$$

$$I_{a,dr,tbl}(t,l,h) = \kappa_{dr,bl}(1-\sigma) I_{0,dr}(t,h)\, e^{-\kappa_{dr,bl} L_l(t,W)} \tag{5.22}$$

where $I_{0,df}(t,h)$ is the diffuse part of PAR flux, κ is the extinction coefficient for the specified PAR flux, and ρ is the reflection coefficient of the canopy. $I_{0,dr}$ and $I_{0,df}$ are derived quantities by the *astro* module of WOFOST from the daily weather data. We omit other information about the WOFOST model because our work focuses on the mentioned equations only. Further detail can be found in the Wofost system description [20].

Bringing all these equations together we can write the system of equations for the

WOFOST model as follows:

$$\frac{\partial W}{\partial t} = c(t)(g(t)\,A(t,W) - R(t,W)) \tag{5.23}$$

$$R(t,W) = b(t)\sum_{i=1}^{4}\zeta_i W_i(t) \tag{5.24}$$

$$A(t,W) = \delta(t)L(t,W)\int_D\int_C U(t,W,l,h)\,dl\,dh \tag{5.25}$$

$$U(t,W,l,h) = z(t,l)\,U_{sl}(t,W,l,h) + (1-z(t,l))\,U_{sh}(t,W,l,h) \tag{5.26}$$

$$U_{sh}(t,W,l,h) = \alpha(t)\,(1-e^{-\frac{\epsilon(t)\,I_{sh}(t,l,h)}{\alpha(t)}}) \tag{5.27}$$

$$U_{sl}(t,W,l,h) = \alpha(t)\left(1-(\alpha(t)-U_{sh}(t,W))\,\frac{1-e^{-\frac{\epsilon(t)\,I_{sl}(t,l,h)}{\alpha(t)}}}{\epsilon(t)\,I_{sl}(t,l,h)}\right) \tag{5.28}$$

$$z(t,l) = e^{-\kappa_{dr,bl}L_l(t,W)} \tag{5.29}$$

$$L_l(t,W) = (0.5 + l\sqrt{0.15})\,L(t) \quad \forall\, l \in [-1,0,1] \tag{5.30}$$

5.2.3 Simplified model for plant growth

As shown in the previous section, the WOFOST crop growth model describes a complex dynamical system that we cannot analyze the perturbations on. With a few assumptions and linear approximations to the many exponential functions in the system, the system can be simplified. In this section we describe the assumptions and approximations we make and show that the results obtained with the simplified model is close enough to the results of the WOFOST simulation model.

Assumptions

- Instead of the averaging over the day and the canopy with the 3-point Gaussian integration method, we take one point in the day and one level on the canopy is a proxy for the plant growth. Hence we can rewrite the following expressions

without the l and h variables and the integrals over day and canopy.

$$I_{0,dr}(t,h) = I_{0,dr}(t,h=0) = I_{0,dr}(t) \tag{5.31}$$

$$I_{sl}(t,h) = I_{sl}(t,h=0) = I_{sl}(t) \tag{5.32}$$

$$I_{a,df}(t,l,h) = I_{a,df}(t,l=0,h=0) = I_{a,df}(t) \tag{5.33}$$

$$I_{a,dr,t}(t,l,h) = I_{a,dr,t}(t,l=0,h=0) = I_{a,dr,t}(t) \tag{5.34}$$

$$I_{a,dr,dr}(t,l,h) = I_{a,dr,dr}(t,l=0,h=0) = I_{a,dr,dr}(t) \tag{5.35}$$

$$I_{sh}(t,l,h) = I_{sh}(t,l=0,h=0) = I_{sh}(t) \tag{5.36}$$

$$L_l(t,W) = L_{l=0}(t) = \frac{L(t,W)}{2} \tag{5.37}$$

$$U(t,W,l,h) = U(t,W,l=0,h=0) = U(t,W) \tag{5.38}$$

$$A(t,W) = \delta(t)\, L(t,W)\, U(t,W) \tag{5.39}$$

- We assume in the middle of the plant $l = 0$, fraction of sunlit and shaded leaves are equal to each other, hence the fraction of sunlit leaves is constant and Equation 5.14 becomes:

$$z(t,l=0) = z = \frac{1}{2} \tag{5.40}$$

- We only use the source-limited growth expression Equation 5.12 for the leaf area index rate and rewrite it in the following form:

$$L(t,W) = L_{sc}(t,W) \tag{5.41}$$

$$\frac{\partial L_{sc}}{\partial t} = \mu(t)\,\frac{\partial W_2}{\partial t} \tag{5.42}$$

$$L_{sc}(t) = \mu(t)\, W_2(t) - \int_0^t \frac{\partial \mu}{\partial t'}\, W_2(t')dt' \tag{5.43}$$

$$\approx \mu(t)\, W_2(t) \tag{5.44}$$

We can make the approximation on the last line because the second term is negligible due to $\frac{\partial \mu}{\partial t'}$ being close to zero.

- By integrating the differential equation 5.5, we can express $W_i(t)$ as follows:

$$\frac{\partial W_i}{\partial t} = \pi_i(t)\,\frac{\partial W}{\partial t} \tag{5.45}$$

$$W_i(t) = \pi_i(t)\,W(t) - \int_0^t \frac{\partial \pi_i}{\partial t'}\,W(t')dt' \tag{5.46}$$

Now we assume that the integral term can be expressed as $\eta_i(t)$ which is a numerical correction to $\pi_i(t)$.

$$\pi_i^*(t) = \pi_i(t) - \eta_i(t) \tag{5.47}$$

$$W_i(t) \approx \pi_i^*(t)\,W(t) \tag{5.48}$$

This assumption in Equation 5.48 is necessary to be able to write an analytical solution to Equation 5.3. Hence we estimate $\eta_i(t)$ numerically from a random run of WOFOST simulation.

- We approximate the NAR for shaded leaves by using the linear approximation for the exponential function near $I_{sh}(t, W) = 0$. Equation 5.16 becomes:

$$U_{sh}(t, W) = \alpha(t)\,(1 - e^{-\frac{\epsilon(t)\,I_{sh}(t)}{\alpha(t)}}) \tag{5.49}$$

$$\approx \epsilon(t)\,I_{sh}(t, W) \tag{5.50}$$

We then assume that the radiation absorbed by shaded leaves $I_{sh}(t, W)$ is equal to the radiation absorbed from diffuse radiation $I_{a,df}$, meaning that we ignore the direct radiation component. This assumption makes sense because we are already taking into account the direct part of the PAR for the sunlit leaves. With the assumption $l = 0$, we wrote $L_l(t, W) = \frac{L(t)}{2}$, then Equation 5.19 becomes:

$$I_{sh}(t, W) \approx I_{a,df}(t) = \kappa_{df}\,(1 - \rho) I_{0,df}(t)\,e^{-\frac{\kappa_{df}\,L(t,W)}{2}} \tag{5.51}$$

Let $a = -\frac{\kappa_{df}}{2}$, $x = L(t, W)$ and $f(x) = e^{ax}$. Using the first two terms of the

Taylor series let's approximate $f(x)$.

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x - x_0) \\ &= f(x_0) - f'(x_0)\, x_0 + f'(x_0)\, x \\ &= c_0 + c_1 \end{aligned}$$

where

$$\begin{aligned} f(x_0) &= e^{a\, x_0} \quad \text{and} \quad f'(x_0) = a\, e^{a\, x_0} \\ c_0 &= f(x_0) - x_0 f'(x_0) = e^{a\, x_0} - x_0\, a\, e^{a\, x_0} \\ &= e^{-\frac{\kappa_{df}}{2}\, x_0} + \frac{\kappa_{df}}{2} x_0\, e^{-\frac{\kappa_{df}}{2} x_0} \\ c_1 &= f'(x_0) = a\, e^{a\, x_0} \\ &= -\frac{\kappa_{df}}{2}\, e^{-\frac{\kappa_{df}}{2} x_0} \end{aligned}$$

Then, we can rewrite 5.51 as follows:

$$I_{sh}(t, W) \approx \kappa_{df}\, (1 - \rho) I_{0,df}(t)\, (c_0 + c_1 L(t, W)) \tag{5.52}$$

$$= k(t)(c_0 + c_1 L(t, W)) \tag{5.53}$$

$$k(t) = \kappa_{df}\, (1 - \rho) I_{0,df}(t) \tag{5.54}$$

- We ignore the $U_{sh}(t, W)$ term in Equation 5.17 and $U_{sl}(t, W)$ loses its dependency on W, therefore we switch to the lowercase notation $u_{sl}(t)$ and rewrite it as follows:

$$u_{sl}(t) = \alpha(t) \left(1 - (\alpha(t))\, \frac{1 - e^{-\frac{\epsilon(t)\, I_{sl}(t)}{\alpha(t)}}}{\epsilon(t)\, I_{sl}(t)} \right) \tag{5.55}$$

$$I_{sl}(t) = \frac{(1 - \sigma)\, I_{0,dr}(t)}{\sin \beta} \tag{5.56}$$

$U_{sl}(t)$ is now just a known function of crop parameters and weather input $I_{0,dr}(t)$.

- By plugging the approximation for $W_i(t)$ from Equation 5.48 into Equation 5.6, we can rewrite the respiration rate $R(t, W)$ as follows:

$$R(t, W) = b(t) \sum_{i=1}^{4} \zeta_i \, \pi_i^*(t) \, W(t) \tag{5.57}$$

$$= r(t) \, W(t) \tag{5.58}$$

$$r(t) = b(t) \sum_{i=1}^{4} \zeta_i \, \pi_i^*(t) \tag{5.59}$$

Given the list of assumptions and approximations we make to simplify the WOFOST model, we can now write the new system of equations for the Simple Model.

5.2.4 System of Equations for the Simple Model

Bringing together all the assumptions and Equations 5.31 - 5.57, we can rewrite the system of equations that describe the dynamics of plant growth as follows:

$$\frac{\partial W}{\partial t} = c(t)(g(t) \, A(t, W) - R(t, W)) \tag{5.60}$$

$$R(t, W) \approx r(t) \, W(t) \tag{5.61}$$

$$A(t, W) \approx \delta(t) \, L(t, W) \, U(t, W) \tag{5.62}$$

$$U(t, W) \approx z \, u_{sl}(t) + (1 - z) \, U_{sh}(t, W) \tag{5.63}$$

$$U_{sh}(t, W) \approx \epsilon(t) \, I_{sh}(t, W) \tag{5.64}$$

$$I_{sh}(t, W) \approx k(t)(c_0 + c_1 L(t, W)) \tag{5.65}$$

$$L(t, W) \approx \mu(t) \, W_2(t) \tag{5.66}$$

$$W_2(t) \approx \pi_2^*(t) \, W(t) \tag{5.67}$$

In one line, we can write:

$$\frac{\partial W}{\partial t} = c(t) \, (g(t) \, (\delta(t)\mu(t)\pi_2^*(t) \, W(t) \, (z \, u_{sl}(t) + (1 - z) \, \epsilon(t)k(t)(c_0 + c_1\mu(t)\pi_2^*(t) \, W(t)))) - r(t) \, W(t))$$

Rearranging we get:

$$\frac{\partial W}{\partial t} = p(t)W(t) + q(t)W^2(t) \tag{5.68}$$

where

$$p(t) = c(t)g(t)\delta(t)\mu(t)\pi_2^*(t)\left(zu_{sl}(t) + (1-z)\epsilon(t)k(t)c_0\right) - c(t)r(t) \tag{5.69}$$

$$q(t) = c(t)g(t)\delta(t)\mu^2(t)(\pi_2^*)^2(t)(1-z)\epsilon(t)k(t)c_1 \tag{5.70}$$

Simple model described by Equation 5.68 is an approximation to the WOFOST model described in Equations 5.23 - 5.30. Again, we made all the assumptions and approximations explained in this section with the goal of finding an analytical solution to the rate equation for plant growth. Next we calculate this solution.

5.2.5 Analytical Solution

By taking the integral with respect to time on each side of Equation 5.68 we get:

$$\int_0^t \frac{\partial W}{\partial t} dt = \int_0^t (p(t)W(t) + q(t)W^2(t))dt \tag{5.71}$$

$$W(t) = \frac{e^{\int_0^t p(t')dt'}}{C_0 - \int_0^t e^{\int_0^z p(z')dz'} q(z)dz} \tag{5.72}$$

where $C_0 = \frac{1}{W_0}$ and W_0 is a crop parameter describing the initial weight of the plant. Next, we compare results we get using the solution to the simple model against the results of the WOFOST simulation model.

5.2.6 Results

To compare Simple Model to the WOFOST model, we ran the WOFOST simulation in the potential yield setting for the crop maize and a random county in Iowa for the year 2000 as it was one of the years with complete weather data. We computed the coefficients $p(t)$ and $q(t)$ in Equation 5.72 with the time-series data of crop parameters and weather obtained from the WOFOST simulation. We plotted the progression of the total dry matter weight $W(t)$ from the emergence of the plant (day 0) to the day

it was harvested.

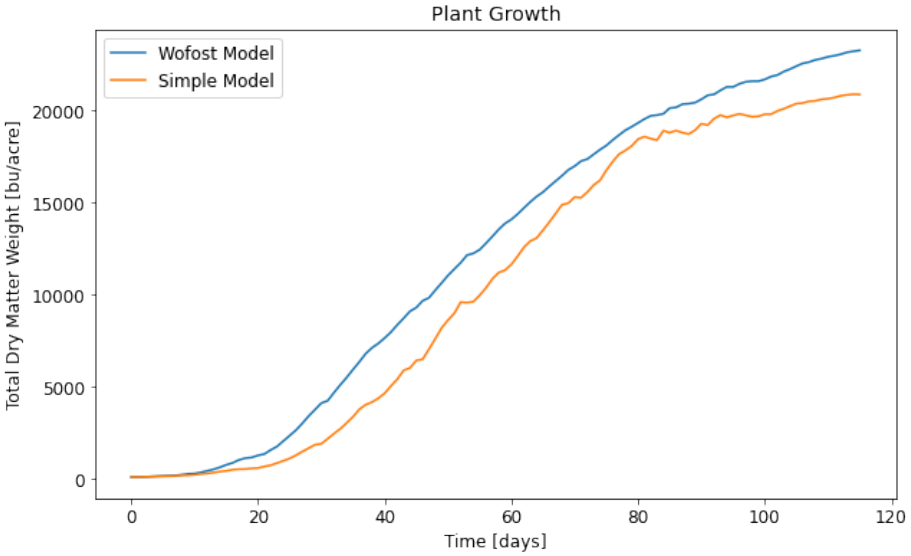


Figure 5-3: Comparison of total dry matter weight

Figure 5-3 shows the simple model's solution for the total plant weight (orange) compared to the total weight evolution obtained from the WOFOST simulation model (blue). We observe that the Simple Model's solution is close to the WOFOST's simulation when the plant it very young and it starts underestimating compared to the WOFOST model starting around day 10. This gap is mostly due to the underestimation of the leaf area index (LAI) which is computed from the total weight of leaves $W_2(t)$. Note that the discrepancy between the models is not constant because the errors caused by different assumptions are not constant in time.

Figure 5-4 shows the partitioning of the dry matter weight to the plant organs. Dashed lines indicate the results achieved by the simple model and solid lines indicate the WOFOST model's results. Blue lines indicate the weight of the roots, green the leaves, magenta the stem and the red the storage organs. We see that the Simple model consistently underestimates the weight of roots and leaves compared to the WOFOST model. Even though it starts by underestimating the weight of the stem too, it gets very close the WOFOST model by the maturity stage of the plant.

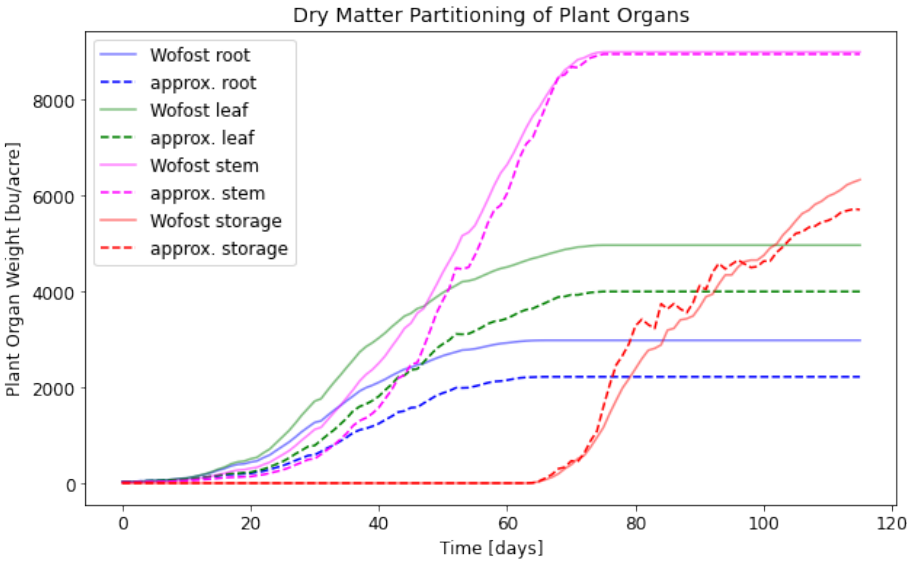


Figure 5-4: Comparison of plant organ weights

The weight of storage organs is the quantity we're interested in the most since it corresponds to the crop yield. We see that both the Simple model and the WOFOST model have zero weight for the storage organs the plant gets more mature. Later the Simple model overestimates and then underestimates compared to the WOFOST model. At the end of the growing season, the weight of storage organs, i.e. the yield, is estimated to be slightly lower than what the WOFOST model simulates.

To reiterate, the goal of developing the Simple model is to have a system close to the WOFOST model but easier to analyze than the exact WOFOST model. With the Simple model, now we can quantify the uncertainty by finding how much the system can deviate from the optimal trajectory due to inputs of the model, such as weather, crop characteristics, soil characteristics, etc.

5.2.7 Uncertainty Propagation

We propagate uncertainty from input parameters through the system described by the Simple Model in Equation 5.68. We start by propagating uncertainties in $q(t)$ and $p(t)$ because they are derived from many input parameters as expressed in Equations 5.69 and 5.70. This means that our goal is to calculate the uncertainty in weight of

the total plant $W(t)$ given the uncertainties in $q(t)$ and $p(t)$ using Equation 5.72. We denote uncertainty in a given quantity Q as δQ. Uncertainty in $W(t)$ is denoted as δW, uncertainty in $p(t)$ denoted as δp and uncertainty in $q(t)$ denoted as δq.

Suppose we know $p(t)$ with uncertainty δp and $q(t)$ is known with high confidence, i.e. $\delta q = 0$. We can calculate the uncertainty δW as follows:
Let $I = \int_0^z p(z')dz' = \sum_{z'=0}^{z} p(z')$. This equality also holds because z' is a discrete variable as it represents days. Then,

$$\delta I = \delta p\sqrt{z} \tag{5.73}$$

Let $F = e^I$, then

$$\delta F = |e^{\delta I}|\,|\delta I| = |e^{\delta p\sqrt{z}}|\,|\delta p\sqrt{z}| \tag{5.74}$$

Let $M = Fq(z)$, then

$$\delta M = |q(z)|\delta F = |q(z)|\,|e^{\delta p\sqrt{z}}|\,|\delta p\sqrt{z}| \tag{5.75}$$

Let $S = \int_0^t M dz = \sum_{z=0}^{t}$, then

$$(\delta S)^2 = \sum_{z=0}^{t}(\delta M)^2 = \sum_{z=0}^{t}\left(|q(z)|\,|e^{\delta p\sqrt{z}}|\,|\delta p\sqrt{z}|\right)^2 \tag{5.76}$$

$$\delta S = \sqrt{\sum_{z=0}^{t}\left(|q(z)|\,|e^{\delta p\sqrt{z}}|\,|\delta p\sqrt{z}|\right)^2} \tag{5.77}$$

Using the above definitions, we can express Equation 5.72 as:

$$W(t) = \frac{F(t)}{C_0 - S(t)} \tag{5.78}$$

Then the fractional uncertainty in $W(t)$ can be written as:

$$\frac{\delta W}{|W|} = \sqrt{\left(\frac{\delta F}{F}\right)^2 + \left(\frac{\delta S}{S}\right)^2 - 2\sigma_{S,F}}, \tag{5.79}$$

where $\sigma_{S,F}$ is the covariance of S and F, which can be calculated using $p(t)$ and $q(t)$ from Equations 5.69 and 5.70.

Chapter 6

Conclusion and Future Work

6.1 Summary of Our Empirical Findings

In this thesis, we set out to answer the question: *How can we reliably generate yield data for different regions of the world using simulation models?* To answer this question, we used the WOFOST crop simulation model to generate yield data and assessed the reliability of the generated data through validation, calibration and uncertainty quantification of the model.

Motivated by applications of the WOFOST model in the European Monitoring Agricultural Resources Crop Yield Forecasting System (MCYFS) [33], we first investigated its applicability in different regions of the world, such as the US. Due to the ample availability of yield data for the corn crop in the US, we used county-level annual corn yield data for the validation of the WOFOST model. In order to run the simulation in each county, we aggregated real-world data from various data sources to be given as the inputs parameters of the model. These data sources describing the crop characteristics, soil characteristics, weather conditions, sowing dates, and harvest dates varied in their formats and completeness. We processed these different types of data from different sources for each county and formatted as one data matrix that can be given as an input to the WOFOST simulation.

After constructing the data matrix for counties and years that we had corn yield data for, we generated yield data by running the WOFOST simulation model for each county in the US for the growing season each year. In order to validate the results of the model, we compared the simulated yield to the true yield data. While we observed the WOFOST model simulated yields close to the true yield in some years and counties, there were still some years and counties with large yield gaps. Investigating the inconsistency of WOFOST, we observed a weak but significant positive trend in the mean yield gap and the percentage error throughout the years. Inspecting the yield and percentage error on county level maps, we also observed some geographical correlations. We hypothesized that these correlations could be due to the inaccuracy of some of the input parameters. Because we did not have any available data for the specific crop varieties used in a given county at a given year, we used constant crop parameters. The fact is that crop cultivars show great variance from farm to farm and from year to year. Thus, we needed to calibrate the crop parameters of the model match the reality more closely.

With calibration, our objective was to find the model parameters that generate yield data that are in well-alignment with the real-world yield data. We formulated this calibration task as minimizing the mean squared error between simulated yield and the true yield. Using the gradient-free optimization method evolutionary algorithms, we found the crop parameters that are most likely to produce the true yield data from Iowa. We observed that the calibration significantly reduced the gap between the simulated and true yield throughout the state and reduced the variations of yield gaps amongst counties.

In this work, we further captured the problem of assessing the quality of the simulated data due to the deterministic nature of the WOFOST model. We provided two approaches to quantify the uncertainty of the WOFOST outputs: Monte-Carlo based sensitivity analysis and error propagation through an alternative model to WOFOST. We developed this alternative model to achieve a simple analytical equation for the

plant growth dynamics, because it was intractable to analyze the uncertainty using the complex system the WOFOST model describes.

6.2 Future Work

6.2.1 Data Collection and Generation

Finding soil, crop, weather, and yield data from different regions of the world—especially from the sub-Saharan regions of Africa—and aggregating them in the Data Platform is an important next step in the *Digital Agriculture in Africa* project. Finding data from field experiments in farms instead of counties would be another valuable contribution to the Data Platform, because a county is a much larger region than farm and data from different farms in a county can show great variance. Due to the limited scope of this thesis, we could focus only on the corn data; however, there are plenty of data for crops, such as wheat, barley, and soybean in the USDA NASS database, and generating data matrices for different crops and running simulations for these crops would be another important contribution to the field.

6.2.2 Validation

Validation was our first step towards understanding understanding the quality of WOFOST simulated yield data. Our empirical findings about the yield gap and percentage error of corn in the US were surprising and led us to the calibration work. It would be interesting to repeat the validation process for a crop other than corn, such as wheat and soybean, to see whether our observations with the other crops would be similar to those with the corn data. To have maximize the number of data points, we evaluated the validation metrics on multiple states. Future work could investigate the metrics on a state-by-state basis to obtain new insights about the variance of the yield gap and percentage error.

6.2.3 Calibration

For calibration, we chose the gene representation $\theta = [\texttt{TSUM1}, \texttt{TSUM2}]$, because these are the two parameters that are different between cultivars and that are used by previous work for calibration. Sensitivity indices in Figure 5-1 can be used to determine other candidates to include in the gene representation, θ, for the evolutionary algorithm used in calibration. Parameters with higher sensitivity indices, such as `SPAN`, `TDWI`, `CVO` and `TBASE`, can be included as well as the `TSUM1` and `TSUM2` parameters. Another useful future work might be to measure how well the calibrated parameters for the state of Iowa transfers to other states in the Corn Belt, such as Illinois.

6.2.4 Uncertainty Quantification

While developing the alternative model to WOFOST (the Simple Model), we only compared the progression of the weights in the WOFOST model to the Simple Model at a given location and year. Future work could generate yield data using the Simple Model for all the counties and years, like we did in Chapter 3 for validation. Similar to Figures 3-11 and 3-17, plotting the yield generated by the Simple Model against the true yield could show whether the error of the Simple Model is similar to that of the WOFOST model. Another important future work would be to learn the parameters $p(t)$ and $q(t)$ with enough data, or some other parameters that are easier to measure than the large number of WOFOST parameters to make it easier to accurately measure inputs to the model and to reduce input uncertainty.

Bibliography

[1] Github repository. URL: https://github.com/ajwdewit/WOFOST_crop_parameters.

[2] Github repository. URL: https://github.com/sulekahraman/wofost_data.

[3] *Global Sensitivity Analysis. The Primer*. URL: http://www.andreasaltelli.eu/file/repository/A_Saltelli_Marco_Ratto_Terry_Andres_Francesca_Campolongo_Jessica_Cariboni_Debora_Gatelli_Michaela_Saisana_Stefano_Tarantola_Global_Sensitivity_Analysis_The_Primer_Wiley_Interscience_2008_.pdf.

[4] Global yield gap atlas. URL: http://www.yieldgap.org.

[5] Iowa crop performance tests (icpt). URL: http://www.croptesting.iastate.edu/Corn/CornSingleLocation.aspx.

[6] Nasa prediction of worldwide energy resources - the power project. URL: https://power.larc.nasa.gov/.

[7] Salib - sensitivity analysis library in python. URL: https://salib.readthedocs.io/en/latest/.

[8] Usda nass quickstats. united states department of agriculture national agricultural statistics service, washington, d.c. URL: https://quickstats.nass.usda.gov/.

[9] How good is good enough? data requirements for reliable crop yield simulations and yield-gap analysis. *Field Crops Research*, 177:49–63, 2015. URL: https://www.sciencedirect.com/science/article/pii/S0378429015000866, doi:https://doi.org/10.1016/j.fcr.2015.03.004.

[10] Iowa State University Ag Decision Maker. Metric conversion, 2013. URL: https://www.extension.iastate.edu/agdm/wholefarm/pdf/c6-80.pdf.

[11] Global Yield Gap Atlas. Gyga protocol for weather data. URL: https://www.yieldgap.org/methods-weather-data.

[12] World Bank. Fact sheet: The world bank and agriculture in africa, 2016. URL: http://web.worldbank.org/WBSITE/EXTERNAL/COUNTRIES/AFRICAEXT/0,,contentMDK:21935583~pagePK:146736~piPK:146830~theSitePK:258644,00.html.

[13] B. Barnabás, K. Jäger, and A. Fehér. The effect of drought and heat stress on reproductive processes in cereals. *Plant, Cell and Environment*, 31(1):11–38, 2008. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-36749081620&doi=10.1111%2fj.1365-3040.2007.01727.x&partnerID=40&md5=562ae8ae7e80bfd51c4c01e153e54efd, doi:10.1111/j.1365-3040.2007.01727.x.

[14] Niels H. Batjes. World soil property estimates for broad-scale modelling (wise30sec, ver. 1.0), 2015. (available via: http://www.isric.org/data/data-download). URL: https://epic.awi.de/id/eprint/40205/1/ReadMe1st__ISRIC-WISE30sec-soildataset.pdf.

[15] M. Blanco, F. Ramos, B. Van Doorslaer, P. Martínez, D. Fumagalli, A. Ceglar, and F.J. Fernández. Climate change impacts on eu agriculture: A regionalized perspective taking into account market-driven adjustments. *Agricultural Systems*, 156:52–66, 2017. cited By 20. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85020287785&doi=10.1016%2fj.agsy.2017.05.013&partnerID=40&md5=8615dd552b456d80dbd38b6eeb9e31fc, doi:10.1016/j.agsy.2017.05.013.

[16] H. Boogaard, J. Wolf, I. Supit, S. Niemeyer, and M. van Ittersum. A regional implementation of wofost for calculating yield gaps of autumn-sown wheat across the european union. *Field Crops Research*, 143:130–142, 2013. cited By 82. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84875551287&doi=10.1016%2fj.fcr.2012.11.005&partnerID=40&md5=6a8c625e50de28241c61d5ea67acc9d0, doi:10.1016/j.fcr.2012.11.005.

[17] Science Technology Calestous Juma Director, Belfer Center for Science Globalization, and Harvard Kennedy School of Government International Affairs. What is africa's agriculture potential? URL: https://www.weforum.org/agenda/2015/09/what-is-africas-agriculture-potential/#:~:text=The%20World%20Bank%20projects%20that,of%20GDP%20across%20the%20continent.

[18] A. Ceglar and L. Kajfež-Bogataj. Simulation of maize yield in current and changed climatic conditions: Addressing modelling uncertainties and the importance of bias correction in climate model simulations. *European Journal of Agronomy*, 37(1):83–95, 2012. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-82155176178&doi=10.1016%2fj.eja.2011.11.005&partnerID=40&md5=9c7f1e3e08ddc5d39a5a49a2689f833b, doi:10.1016/j.eja.2011.11.005.

[19] A. Ceglar, R. van der Wijngaart, A. de Wit, R. Lecerf, H. Boogaard, L. Seguini, M. van den Berg, A. Toreti, M. Zampieri, D. Fumagalli, and B. Baruth. Improving wofost model to simulate winter wheat phenology in europe: Evaluation and effects on yield. *Agricultural Systems*, 168:168–180, 2019. URL: https://www.sciencedirect.com/science/article/pii/S0308521X17309897, doi:https://doi.org/10.1016/j.agsy.2018.05.002.

[20] Allard de Wit. Wofost - world food studies: System description of the wofost 7.2 cropping systems model. URL: https://www.wur.nl/en/Research-Results/Research-Institutes/Environmental-Research/Facilities-Tools/Software-models-and-databases/WOFOST.htm.

[21] Allard de Wit. Pcse: The python crop simulation environment. pcse documentation, release 5.4, May 2020. https://pcse.readthedocs.io/_/downloads/en/stable/pdf/. URL: https://pcse.readthedocs.io/en/stable.

[22] Allard de Wit, Hendrik Boogaard, Davide Fumagalli, Sander Janssen, Rob Knapen, Daniel van Kraalingen, Iwan Supit, Raymond van der Wijngaart, and Kees van Diepen. 25 years of the wofost cropping systems model. *Agricultural Systems*, 168:154–167, 2019. URL: https://www.sciencedirect.com/science/article/pii/S0308521X17310107, doi:https://doi.org/10.1016/j.agsy.2018.06.018.

[23] Allison J. A Ernon, A. Method of calculating net assimilation rate. 1963. URL: https://doi.org/10.1038/200814a0.

[24] Jorge Fernandez-Cornejo. The seed industry in u.s. agriculture: An exploration of data and information on crop seed markets, regulation, industry structure, and research and development. *Agricultural Information Bulletin No. (AIB-786) 81 pp*, February 2004. URL: https://www.ers.usda.gov/publications/pub-details/?pubid=42531.

[25] Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. Deap: Evolutionary algorithms made easy. *Journal of Machine Learning Research*, 13:2171–2175. URL: https://github.com/deap/deap.

[26] R. Gamie and F. De Smedt. Experimental and statistical study of saturated hydraulic conductivity and relations with other soil properties of a desert soil. *European Journal of Soil Science*, 69(2):256–264. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ejss.12519, doi:https://doi.org/10.1111/ejss.12519.

[27] R. Gamie and F. De Smedt. Experimental and statistical study of saturated hydraulic conductivity and relations with other soil properties of a desert soil. *European Journal of Soil Science*, 69(2):256–264. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ejss.12519, arXiv:https:

//onlinelibrary.wiley.com/doi/pdf/10.1111/ejss.12519, doi:https://doi.org/10.1111/ejss.12519.

[28] Graham Glen and Kristin Isaacs. Estimating sobol sensitivity indices using correlations. *Environmental Modelling Software*, 37:157–166, 2012. URL: https://www.sciencedirect.com/science/article/pii/S1364815212001065, doi:https://doi.org/10.1016/j.envsoft.2012.03.014.

[29] OCP Group. Sociotechnical systems research center. URL: https://ssrc.mit.edu/programs/ocp-group/.

[30] A. Yokota H. Ashida. *Comprehensive Biotechnology (Second Edition)*. 2011.

[31] Toshimitsu Homma and Andrea Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering System Safety*, 52(1):1–17, 1996. URL: https://www.sciencedirect.com/science/article/pii/0951832096000026, doi:https://doi.org/10.1016/0951-8320(96)00002-6.

[32] EU Science Hub. Crop yield forecasting. URL: https://ec.europa.eu/jrc/en/research-topic/crop-yield-forecasting.

[33] EU Science Hub. Monitoring agricultural resources (mars). URL: https://ec.europa.eu/jrc/en/mars.

[34] C.A. Van Diepen (Eds.) I. Supit, A.A. Hooijer. System description of the wofost 6.0 crop simulation model implemented in cgms, european communities (eur15956en). 1994. URL: https://op.europa.eu/en/publication-detail/-/publication/a99325a7-c776-11e6-a6db-01aa75ed71a1.

[35] KATHRYN KERBY. The average percolation rate for various soil types. URL: https://www.hunker.com/13406958/the-average-percolation-rate-for-various-soil-types.

[36] Christopher J. Kucharik and Navin Ramankutty. Trends and variability in u.s. corn yields over the twentieth century. *Earth Interactions*, 9(1):1 – 29, 2005. URL: https://journals.ametsoc.org/view/journals/eint/9/1/ei098.1.xml, doi:10.1175/EI098.1.

[37] J. Lowenberg-DeBoer. The precision agriculture revolution: Making the modern farmer. June 2015. URL: https://www.foreignaffairs.com/articles/united-states/2015-04-20/precision-agriculture-revolution.

[38] Songrit Maneewongvatana and David M. Mount. Analysis of approximate nearest neighbor searching with clustered point sets, 1999.

[39] Leonardo A. Monteiro, Paulo C. Sentelhas, and George U. Pedra. Assessment of nasa/power satellite-based weather system for brazilian conditions and its impact on sugarcane yield simulation. *International Journal of Climatology*, 38(3):1571–1581. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/

10.1002/joc.5282, arXiv:https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.5282, doi:https://doi.org/10.1002/joc.5282.

[40] Ohio Environmental Protection Agency (OhioEPA) Division of Drinking and Ground Waters. Technical guidance for ground water investigations. chapter 3 characterization of site hydrogeology. October 2006. URL: https://www.epa.state.oh.us/portals/28/documents/TGM-3.pdf.

[41] J.R. Porter and M.A. Semenov. Crop responses to climatic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1463):2021–2035, 2005. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-33144474167&doi=10.1098%2frstb.2005.1752&partnerID=40&md5=c21686f0b38d45d5029d45d770e0d5e7, doi:10.1098/rstb.2005.1752.

[42] USDA NASS QuickStats. 2020 state agriculture overview iowa. URL: https://www.nass.usda.gov/Quick_Stats/Ag_Overview/stateOverview.php?state=IOWA.

[43] World Population Review. Corn production by state 2021, 2021. URL: https://worldpopulationreview.com/state-rankings/corn-production-by-state.

[44] Systems Scott Murray, Institute for Data and Society. Empowering african farmers with data, May 2019. URL: https://news.mit.edu/2019/empowering-african-farmers-with-data-0530.

[45] I.M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 55(1):271–280, 2001. The Second IMACS Seminar on Monte Carlo Methods. URL: https://www.sciencedirect.com/science/article/pii/S0378475400002706, doi:https://doi.org/10.1016/S0378-4754(00)00270-6.

[46] J-E. Hallgreen S.P. Long. *Techniques in Bioproductivity and Photosynthesis (Second Edition)*. 1985.

[47] C.J.T. Spitters. Separating the diffuse and direct component of global radiation and its implications for modeling canopy photosynthesis part ii. calculation of canopy photosynthesis. *Agricultural and Forest Meteorology*, 38(1):231–242, 1986. URL: https://www.sciencedirect.com/science/article/pii/0168192386900614, doi:https://doi.org/10.1016/0168-1923(86)90061-4.

[48] C.A. van Diepen, J. Wolf, H. van Keulen, and C. Rappoldt. Wofost: a simulation model of crop production. *Soil Use and Management*, 5(1):16–24, 1989. cited By 392. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0024483115&doi=10.1111%2fj.1475-2743.1989.tb00755.x&partnerID=40&md5=2bcecd07189ca59b3253135db1840931, doi:10.1111/j.1475-2743.1989.tb00755.x.

[49] Xiong Wei, Conway Declan, Lin Erda, Xu Yinlong, Ju Hui, Jiang Jinhe, Holman Ian, and Li Yan. Future cereal production in china: The interaction of climate change, water availability and socio-economic scenarios. *Global Environmental Change*, 19(1):34–44, 2009. URL: https://www.sciencedirect.com/science/article/pii/S0959378008000988, doi:https://doi.org/10.1016/j.gloenvcha.2008.10.006.