Analyzing Environmental Influences on Corn Yield: A Data-Driven Study in Champaign, Illinois

This manuscript (permalink) was automatically generated from uiceds/project-team-go@3a225c8 on November 18, 2024.

Authors

- **Yung Shun Shih**
  Department of CEE, University of illinois Urbana-Champaign
- **Derek Chen**
  Department of CEE, University of illinois Urbana-Champaign
- **Xinyuan Wang**
  Department of CEE, University of illinois Urbana-Champaign
- **Xiaozhuo Cao**
  Department of CEE, University of illinois Urbana-Champaign

✉ — Correspondence possible via GitHub Issues

Abstract

Project proposal

Crop models are computational tools that assess the effects of environmental variation and cultivation strategies on crop yield (Chapagain et al., 2022; Huang et al., 2019). By incorporating factors such as precipitation, humidity, temperature, fertilization, and soil properties, crop models establish relationships between input parameters and agricultural yield outcomes. From a structural perspective, crop models can be either empirical or mechanistic. Empirical models create statistical relationships based on existing data, while mechanistic models aim to explain relationships by exploring physiological mechanisms and causal connections (Reynolds and Acock, 1985). From a parameter standpoint, crop models generally include weather, soil, and crop-specific parameters to estimate crop biomass. Weather parameters cover solar radiation, precipitation, temperature, and more, while soil parameters focus on humus content, organic matter content, and other soil characteristics. Crop-specific parameters include maximum crop yield, specific nitrogen uptake rate, and related factors. In this project, we aim to replicate Hartmut Bossel's 'Field Crop Cultivation' simulation model as a white-box reference and develop a black-box model using SVD, PCA, and/or Fourier series. The original model is a parsimonious one, primarily focusing on the dynamic effects of precipitation on crop yield across a spectrum of crops in Germany. Initially created in BASIC (Hartmut, 1985) and later in Vensim (Hartmut, 2007) for educational purposes, the model simulates the impact of water and nutrient (nitrogen) availability on plant growth dynamics. Built from first principles, it captures complex interactions between water and nutrient dynamics and can be adapted to different scenarios by applying specific plant and soil parameters.

Data description

The dataset we plan to use is the meteorological records of Champaign, Illinois. We want to predict corn yield by analyzing precipitation and temperature. Data will be obtained from wunderground.com (Savoy, IL Weather History | Weather Underground). And daily temperature and the annual precipitation amount would be needed. The format would be primarily in CSV. The four columns will be temperature (including max, avg and min) and precipitation every day, while the rows will be the date for a whole year.

Exploratory data analysis

1. Background and Research Proposal Crop models are computational tools that assess the effects of environmental variation and cultivation strategies on crop yield (Chapagain et al., 2022; Huang et al., 2019). By incorporating factors such as precipitation, humidity, temperature, fertilization, and soil properties, crop models establish relationships between input parameters and agricultural yield outcomes. From a structural perspective, crop models can be either empirical or mechanistic. Empirical models create statistical relationships based on existing data, while mechanistic models aim to explain relationships by exploring physiological mechanisms and causal connections (Reynolds and Acock, 1985). From a parameter standpoint, crop models generally include weather, soil, and crop-specific parameters to estimate crop biomass. Weather parameters cover solar radiation, precipitation, temperature, and more, while soil parameters focus on humus content, organic matter content, and other soil characteristics. Crop-specific parameters include maximum crop yield, specific nitrogen uptake rate, and related factors.

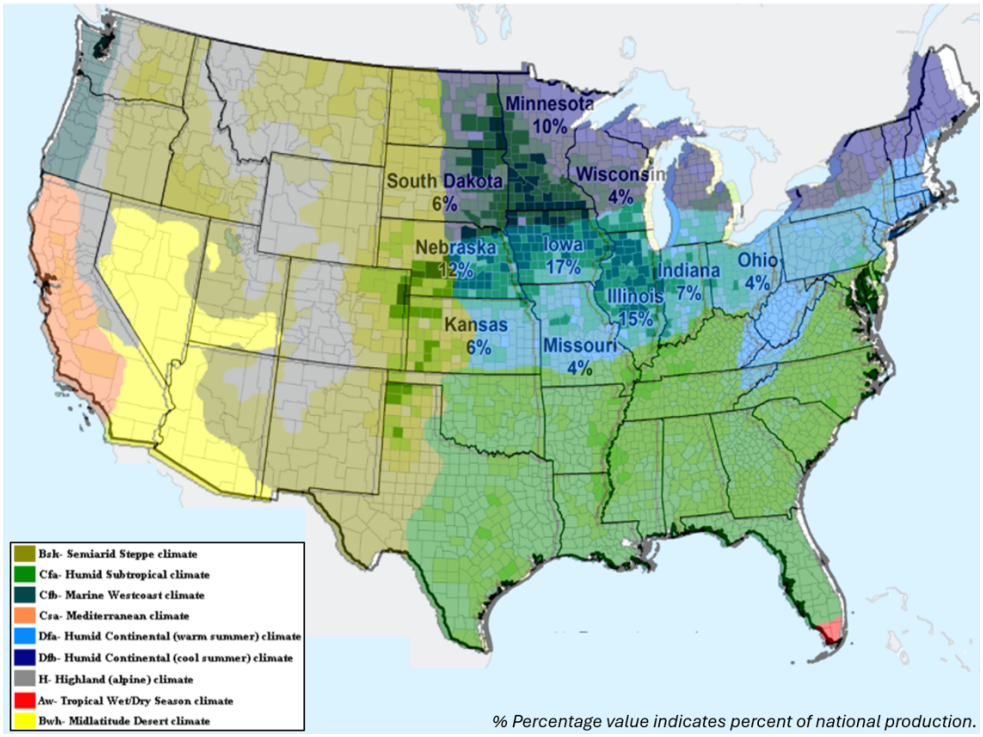
In this project, we aim to replicate Hartmut Bossel's 'Field Crop Cultivation' simulation model as a white-box reference and develop a black-box model using SVD, PCA, and/or Fourier series. The original model is a parsimonious one, primarily focusing on the dynamic effects of precipitation on crop yield across a spectrum of crops in Germany. Initially created in BASIC (Hartmut, 1985) and later in Vensim (Hartmut, 2007) for educational purposes, the model simulates the impact of water and nutrient (nitrogen) availability on plant growth dynamics. Built from first principles, it captures complex interactions between water and nutrient dynamics and can be adapted to different scenarios by applying specific plant and soil parameters.

2. Reference model Construction In our project, we have already translated the model into Python as the reference model. We modularized the code into three phases to enhance customization and improve understanding: The first part of the code focuses on preparing input values, which involve defining constants and table functions that are used in the next stage. The second part integrates two sub-models: soil-water and soil-nutrient. The 'Soilwater' model determines the soil water content based on two key factors: water-related parameters and soil parameters. The water-related parameters describe the mass balance of water (precipitation, irrigation, transpiration, evaporation, and percolation). In parallel, the soil parameters define the water-holding capacity of the soil. The other sub-model 'Soilnutrient' contains a mass balance for nitrogen (the interconnection between plant available nitrogen and humus in soil). In summary, soil-water and soil-nutrient are fundamentally important to estimate the yield. In the third part, the output data is presented as diagrams (or in csv files), visualizing the simulation results and enabling analysis of how changing factors influence crop yield.

3. Data Preparation

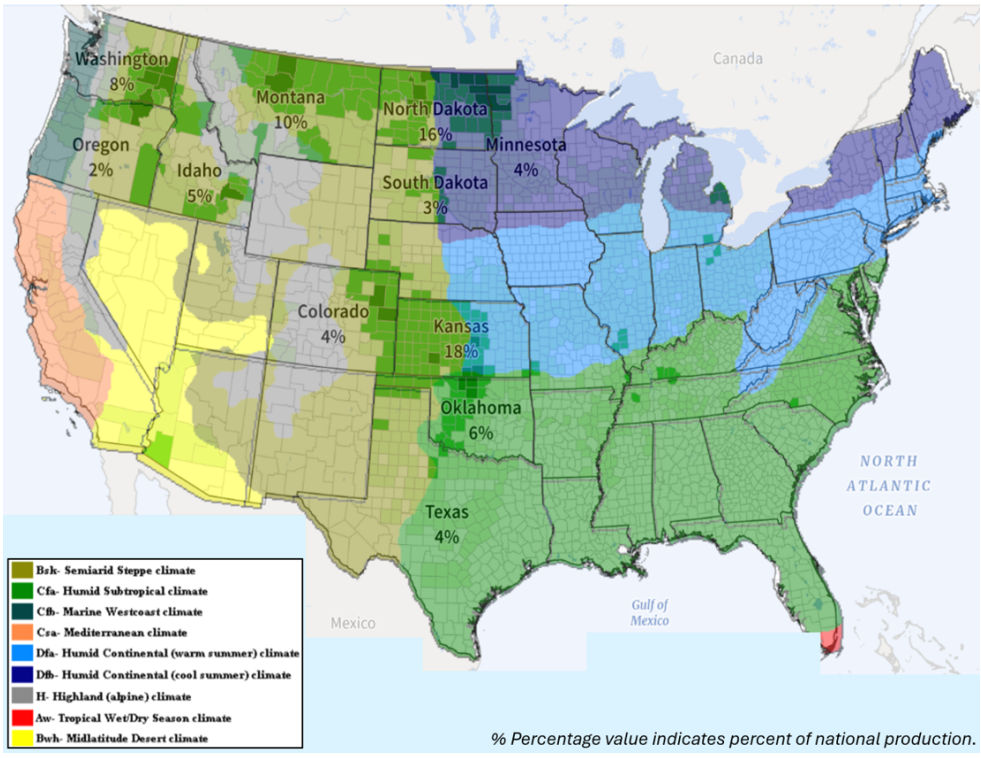
3.1 Selecting Climate Zone and Sampling Points

United States: Corn Production

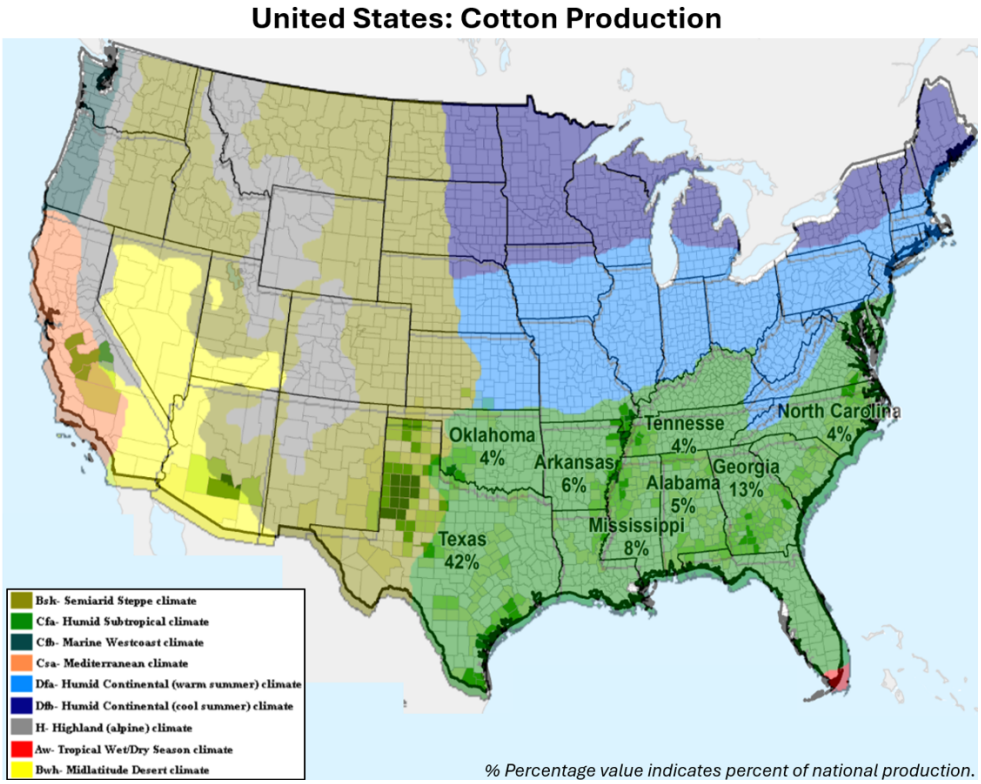


% Percentage value indicates percent of national production.

United States: Wheat Production



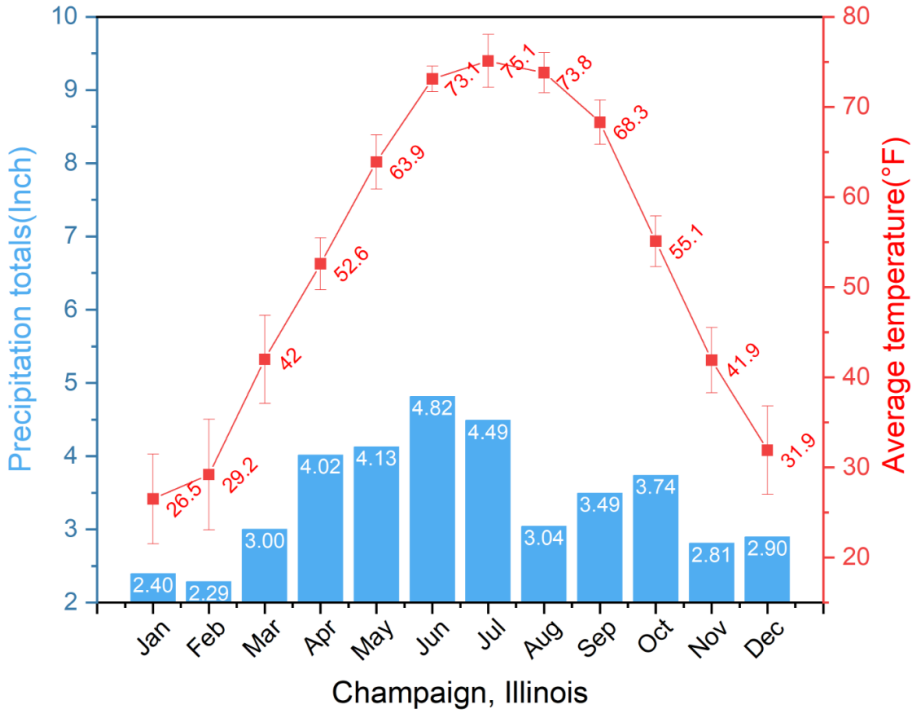
% Percentage value indicates percent of national production.

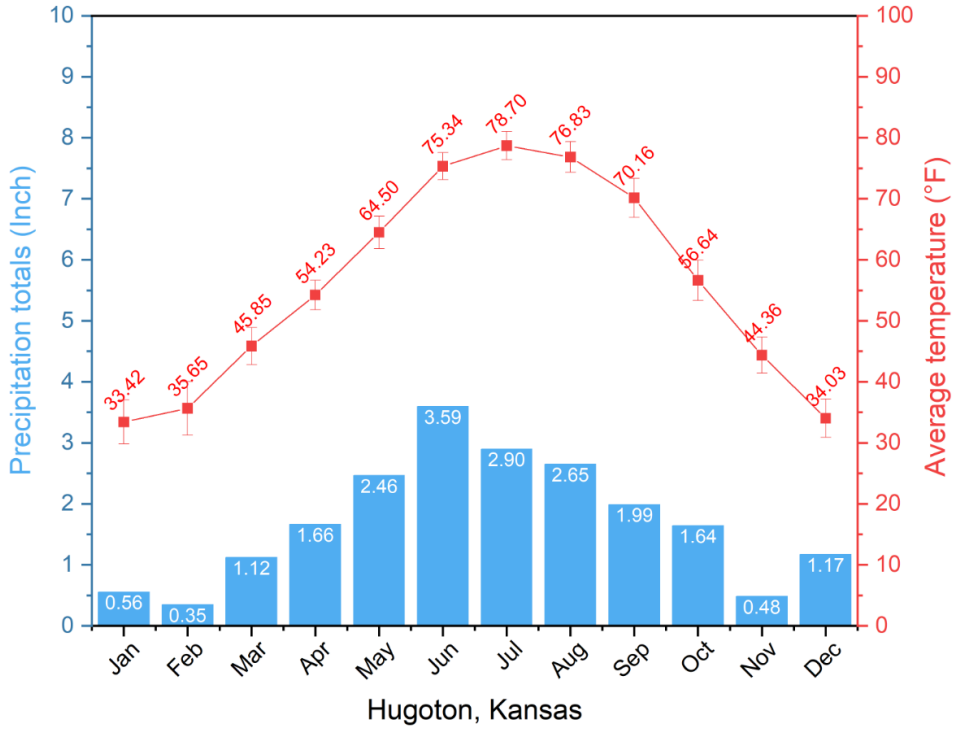


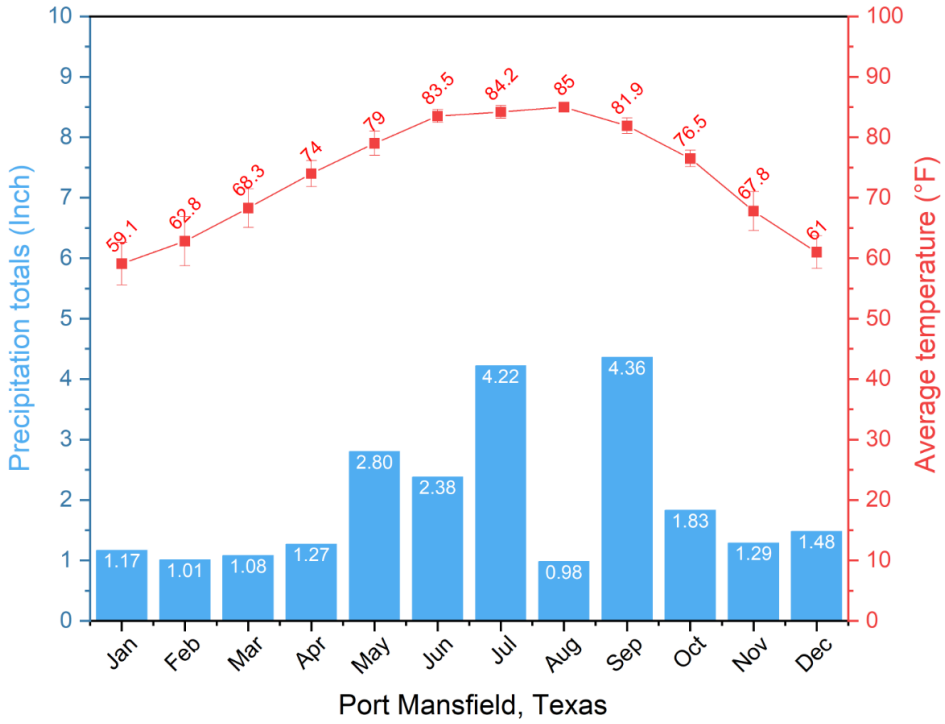
The figures above show an overlay analysis of the distribution percentages of corn, wheat, and cotton yields in the United States with climate zones. Our team determined the range of temperature and precipitation data needed by examining these overlays.

We can clearly see the locations where each crop's high-yield regions intersect with various climate zones, enabling us to understand how climate factors influence each crop's growth conditions.

3.2 Data Preparation


Precipitation totals(Inch)
Average temperature(°F)
Champaign, Illinois
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2.40
2.29
3.00
4.02
4.13
4.82
4.49
3.04
3.49
3.74
2.81
2.90
26.5
29.2
42
52.6
63.9
73.1
75.1
73.8
68.3
55.1
41.9
31.9


Precipitation totals (Inch)
Average temperature (°F)
Hugoton, Kansas
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
0.56
0.35
1.12
1.66
2.46
3.59
2.90
2.65
1.99
1.64
0.48
1.17
33.42
35.65
45.85
54.23
64.50
75.34
78.70
76.83
70.16
56.64
44.36
34.03



In this study, three typical U.S. crops, corn, cotton, and wheat, were selected as examples in our analysis. These crops are grown in different climate zones and play a vital role in U.S. agriculture. To scientifically analyze the impact of climate on crop yields, the study first chose climate data sampling points based on the United States crop production maps (USDA United States - Crop Production Maps), which clearly shows the key production areas for different crops. Climate data, including monthly average temperature and monthly precipitation, were collected from three different weather monitoring stations within each key production area. Climate data from 2004 to 2024, within a 20-year period of time, were collected and used in this study for model analysis.

Here, we chose to represent the climatic characteristics of a region using the monthly average temperature and monthly average precipitation for each year. For example, the State of Illinois, of which mostly is humid continental climate, is a major corn production area in the US. The characteristics of this climate type is presented by temperature and precipitation data from Champaign. Similarly, we chose Nobel County in Minnesota as a typical wheat-producing area to illustrate its climate characteristics, and Port Mansfield in Texas to represent the climate characteristics of a typical cotton-growing area.

4. Reference Model Results Analysis and Questions

4.1 Exploratory Data Analysis on Reference Model Results, Humid Subtropical Climate

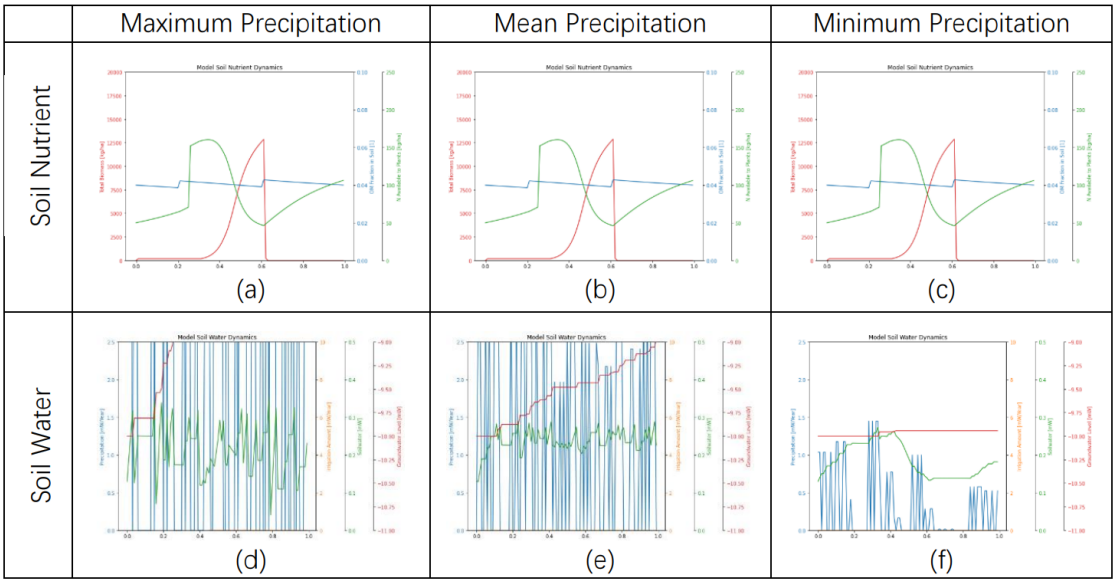


Figure 1. Model Results Under Varying Precipitation in Humid Subtropical Climate (Stoneville, MS)

Figure 1. shows the mechanistic model results for soil water and nutrient dynamics in a humid subtropical climate (Stoneville, Mississippi) under varying precipitation scenarios (maximum, mean, and minimum). The results are shown in two sets of plots. The x-axis represents time within one year, ranging from 0 to 1. Plots in the first row (a, b, c) display the changes in soil nutrients over time, specifically total biomass (red), nitrogen available to plants (green), and organic matter fraction in soil (blue). Plots in the second row (d, e, f) illustrate the in precipitation with randomized weather event (blue line), groundwater levels (red line), and soil water content (blue line) across three precipitation scenarios.

From the soil nutrient data, we can observe the seasonal dynamics of biomass levels as well as plant-available nitrogen in the soil. At the beginning of cultivation, with the application of fertilizer, nitrogen levels reach their peak and then decrease as the crop continues to grow. From the soil water data, we can see that a water surplus exists in both the max and mean rainfall scenarios, leading to a significant rise in groundwater levels (d, e). Additionally, since the reference model did not account for surface runoff, there could be a significant overestimation of precipitation's contribution to groundwater levels.

4.2 Further Questions on Reference Model Results, Humid Subtropical Climate 1. Model Glitch: Notice that, despite the differences in precipitation levels, plots a, b, and c are largely identical. It appears that the sub-model 'Soilwater' is not successfully linked with the other sub-model 'Soilwater' in the simulation. Further debugging is required to calibrate the reference model. 2. Optimal Precipitation for Corn Growth: Even under minimum precipitation, there is no significant water deficit; thus, all crops grow under optimal water conditions. However, how does each scenario impact actual corn yield in Stoneville? Can the model accurately reflect the optimal precipitation range for maximum crop growth? 3. Nutrient Leaching: Will excessive rainfall affect nutrient level through leaching end erosion? The original model was designed to represent moderate conditions at or below optimal precipitation, but could high precipitation levels cause additional nutrient loss in other pathways? 4. Model Validation: How well do these model outputs align with real crop yield data from similar climate zones? (This may exceed the scope of our project.) 5. Long-Term Soil Health Under Crop Rotation: Over multiple growth cycles and by applying sustainable practice such as crop rotation, what would be the cumulative effect in the long-term on soil nutrient content and water content? (Exceeding the scope.)

4.3 Exploratory Data Analysis on Reference Model Results, Humid Continental (warm summer)

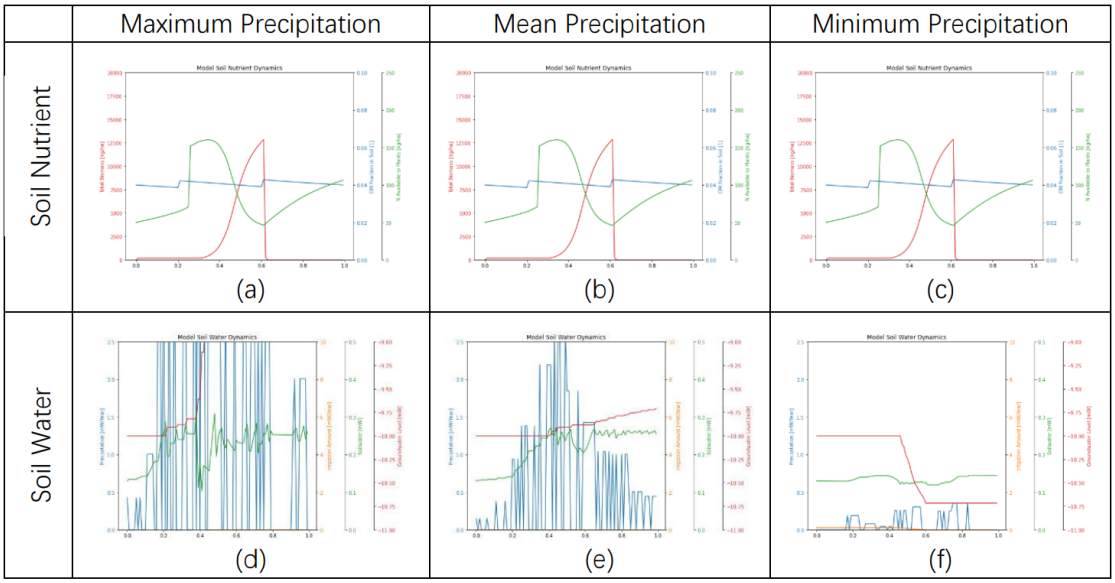


Figure 2. Model Results Under Varying Precipitation in Humid Continental (warm summer) (Arnold, IA)

Figure 2. shows the mechanistic model results for soil water and nutrient dynamics in a humid subtropical climate (Arnold, Iowa) under varying precipitation scenarios (maximum, mean, and minimum). The similarity of soil nutrient results is probably cause by model error. In soil water, we can see three precipitation cases covers highwater excess, minor water deficit and large water deficit, indication that the locational conditions can be a good setting for us to use this reference model to explain the situation. Also, notice that the precipitation pattern (blue) is different from Stoneville, and hence causing different dynamics in soil water content (green), for example, not significant seasonal variation .

4.4 Further Questions on Reference Model Results, Humid Subtropical Climate 1. Model Glitch: Given that precipitation in this scenario is below the optimum level, it can be confirmed that there is a model glitch. In the original model, when water availability is below optimal, both crop growth and microbial activities (which affect the transformation of organic matter into plant-available nitrogen) in the soil should be reduced. Therefore, the overall biomass curve should shrink vertically (red), the available nitrogen level should remain low instead of increasing (green) after harvesting, and the organic matter level should remain high (blue) after harvesting. 2. Dataset Scope: In the future, we may limit our climate zones to Humid Continental or drier regions to avoid the structural limitations of the mechanistic model in explaining crop yield under heavy rainfall conditions and accounting for surface runoff in the region's water balance. 3. Evapotranspiration (ET) Rate Correction: Currently, the reference model characterizes the ET rate as a constant. Since we are using this model in different locations with varying humidity, temperature, and wind speed, we could enhance the model's accuracy by incorporating location-specific ET rate.

5 Predictive Modeling Plan The aim of this project is to create a statistical model to produce similar estimates as the mechanistic model. The benefit of this simplification is to reduce computational costs. Another potential outcome of this approach is that, by using regression analysis, we can test correlations and rank the inputs that have the most significant effect on the output, thereby helping to determine the dominant factors influencing crop growth and decision-making in crop management. To do this, we will first ensure that the reference mechanistic model functions correctly, making it capable of generating predictive yield based on the precipitation data. Then, to create sufficient data, we can randomly generate 1,000 (or more) precipitation curves for each scenario using the mean value and standard deviation obtained from the data in section 3. Third, the generated precipitation curves will be stored as CSV files, ready for model input. Fourth, we will translate the current Python model into Julia and automatically run simulations to obtain the corresponding yield for each precipitation scenario. Finally, we will compile a new DataFrame that includes both the precipitation

and yield data, allowing us to apply SVD, PCA, and/or Fourier series to identify the dominant eigenvectors and underlying patterns in the data.

Preliminary Predictive Modeling

1 Data Description of One Scenario

1.1 Explanation of Columns

The dataset includes several key columns, each playing an essential role in the analysis. The *sim_index* column represents the simulation timeline, allowing the data to be tracked sequentially. While not directly used in the model, it helps visualize time-dependent trends. The *MULTIPLIER_FOR_RAINFALL* column is a scaling factor applied to raw rainfall data, reflecting environmental adjustments or experimental conditions. Using this multiplier, the rain_amount column is calculated as the cumulative rainfall over time, serving as the independent variable in the regression analysis to explore its impact on biomass production. The *precipitation* column indicates the level of rainfall at each time point, providing additional environmental context. Similarly, the *soil_moisture* column captures the moisture levels in the soil, influenced by rainfall and precipitation. This column is not directly used in the predictive model now. Finally, the *total_biomass* column represents the dependent variable, measuring the biomass produced under given conditions. This serves as the target variable in the regression model, with predictions based on the *rain_amount* variable. Together, these columns create a comprehensive dataset for analyzing the interplay between rainfall and biomass in varying environmental conditions.

1.2 Relationships and Usage in Code

The dataset's variables are used in specific ways to build and analyze the predictive model. The *rain_amount column*, derived by multiplying *MULTIPLIER_FOR_RAINFALL* and *raw rainfall* data, serves as the core predictor to model its relationship with total_biomass. This dependent variable acts as the target for the regression analysis, allowing the model to evaluate its predictive accuracy. Supporting variables such as *precipitation* and *soil_moisture* provide additional environmental context, which could be leveraged for feature engineering in more advanced models. The *sim_index* ensures that the data can be tracked sequentially for exploratory analysis and visualization. These relationships enable the construction of a decision tree regression model, which uses *rain_amount* to predict *total_biomass*, and its performance is validated through visualizations and comparisons with the observed data.

2 Model Function Description

The core of this project is to implement a simple decision tree regression model from scratch without relying on external machine learning libraries. The basic idea of decision tree regression is to recursively split the dataset into homogeneous subsets and estimate the mean of each subset to predict the target variable. Specifically, the model consists of the following modules:

2.1 Decision Tree Construction Function

The goal of this function is to construct a regression tree model based on the feature data (P) and target data (B).

Stopping Criteria: The recursion stops when the sample size is less than or equal to the minimum split sample size (min_samples_split), or when the maximum depth (depth) is reached. In this case, the mean of the target variable is used as the prediction value.

Finding the Best Split Point: The model attempts to iterate over all unique values of the feature to find the split point that minimizes the error (sum of squared losses) for the left and right subsets. The smaller the squared loss, the higher the homogeneity of the dataset.

Recursive Splitting: Once the best split point is found, the model splits the data into left and right subtrees and recursively constructs the subtrees until the stopping criteria are met.

2.2 Model Workflow

Training Phase: The decision_tree_regression function is used to recursively split the training dataset and construct the decision tree model. At each step of the split, the possible split points are iterated over, and the squared loss is calculated to select the optimal split point, dividing the dataset into two homogeneous subsets.

Prediction Phase: The predict_tree function is used to predict new data. Each new feature value is directed through the tree's split rules to find the corresponding leaf node, and the mean value of that node is output as the final prediction.

2.3 Experimental Results and Analysis

By testing the decision tree regression model on the rainfall data and the biomass data, it was observed that the model effectively performed segmented predictions based on the given data, which continuously split the feature space to minimize the variance of the target variable as much as possible. The goodness of fit is used to estimate the prediction outcome, which is calculated as follows:

$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{tot}}}{\mathrm{SS}_{\mathrm{res}}}$, in which $\mathrm{SS}_{\mathrm{tot}}$ is Total Sum of Squares, $\mathrm{SS}_{\mathrm{res}}$ is Residual Sum of Squares.

Although this implementation is relatively simplified, it effectively demonstrates the core ideas and basic construction process of decision tree regression.

image

Figure 1 is the prediction result of model with depth 100, which has a goodness of fit 97.14%.

image

Figure 2 is the result of model with depth 3 which has a goodness of fit 96.91%. It can be observed that as the number of layers in the decision tree increases, its fitting performance in the early stages improves. In fact, the final goodness of fit is also higher.

3 Data Description for All Scenario

We aim to recreate a simplified surrogate model to reduce the computation time of the mechanistic model. In the mechanistic model, 17 different variables are calculated for every iteration. Each variable represents time-series data consisting of 100 data points over a 1-year range. Among these variables, 'precipitation' and 'multiplier for precipitation' serve as inputs, and their combination constitutes a new scenario. 'Total biomass' refers to the yield of corn, which is the final output. The remaining 14 variables are intermediate variables used in the calculations. In summary, for every scenario, the outputs include 17 time-series variables, each with 100 data points, accumulating to a total of 1,700 data points per scenario.

To generate a spectrum of scenarios for better estimation of the corn yield in the US, we overlapped the corn production map with the climate zone map and select 9 sampling locations across three different major climate zones. For each location, the precipitation data is gathered in time range of 20 to 21 years. To generate more scenarios, we use the corrected precipitation, which is the product of 'precipitation' from meteological data and 'multiplier for precipitation' from 0.1 to 0.9 (less than optimum). In summary, the total scenario generated is the product of location number (9), precipitation in different years (20-21) and multiplier for precipication, in total 1692 scenarios. Figure 1 shows that the simulation data can be retrieved even after this data transformation process.

To generate a spectrum of scenarios for better estimation of corn yield across the U.S., we overlap the major corn production map with the climate zone map and select 9 sampling locations representing three major climate zones. For each location, precipitation data is gathered over a 20- to 21-year period. Additional scenarios are generated by using corrected precipitation, calculated as the product of meteorological 'precipitation' data and a 'multiplier for precipitation' ranging from 0.1 to 0.9 (representing suboptimal conditions). This approach results in a total of 1,692 scenarios, which is the product of the number of locations (9), years of precipitation data (20-21), and multipliers for precipitation (9). Gif 1-3 demonstrates that the simulation data remains intact even after this data transformation process. In summary, the dataset consists of 1692 scenarios and 1700 datapoints for each scenario. Using the mechanistic model to generate this dataset and stored as a csv file, consisting inputs, output and intermediate variables.


image

image


image


image

4 Dimension Reduction by SVD

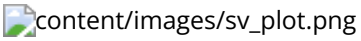
The original mechanistic model consists of 87 equations and above 100 variables for each iteration step, 15 of which are integral equations updated at each iteration. This high dimensionality and computational complexity increase computational time and make the model harder to interpret. To address these challenges, we apply Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) to reduce the dimensionality of the dataset, evaluate the contribution of the most important principal components, and recreate the dataset using the compressed eigenvectors.

To prepare the dataset for SVD, we first reorganized the data by stretching all the data points in one scenario into a single column in the DataFrame. Each variable has 100 time-series elements and the number of columns equals the total number of scenarios. Second, we calculated the average scenario by horizontally taking the mean value across scenarios, then subtracted this average scenario from

the dataset itself to obtain the centered data (X). Third, we performed SVD on the centered data to obtain the three singular components (U, S, and V').

Figure 2 illustrates the singular values ($F.S$) plotted on a logarithmic scale. It shows that the dataset's variance starts relatively small and decreases rapidly at the initial stage. Figure 3 identifies the variance explained by the first five principal components (PCA modes), indicating that almost all variance can be captured by the first three PCA modes. Additionally, Figure 4 visualizes the first 20 eigen-scenarios (columns of $F.U$). Finally, the dataset was reconstructed using the compressed data gained from the SVD process, and Figure 5 displays the reconstructed scenario.

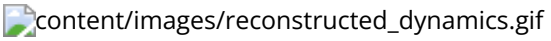
Figure 1 illustrates the singular values ($F.S$) plotted on a logarithmic scale, highlighting that the dataset's variance starts relatively small and decreases sharply at the initial stage. Figure 2 shows the variance explained by the first five principal components (PCA modes), indicating that almost all variance can be captured by the first three PCA modes. Additionally, Figure 3 visualizes the first 20 eigen-scenarios (columns of $F.U$), providing insights into the dataset's principal structures. Finally, the dataset was reconstructed using the compressed data from the SVD process, and Figure 4 displays the reconstructed scenario.



content/images/sv_plot.png

content/images/variance 1-5 SVD indices.png

content/images/eigensce_plot_10sce.png



content/images/reconstructed_dynamics.gif

References

Chapagain, R., Remenyi, T. A., Harris, R. M., Mohammed, C. L., Huth, N., Wallach, D., ... & Ojeda, J. J. (2022). Decomposing crop model uncertainty: A systematic review. Field Crops Research, 279, 108448.

Huang, J., Gómez-Dans, J. L., Huang, H., Ma, H., Wu, Q., Lewis, P. E., ... & Xie, X. (2019). Assimilation of remote sensing into crop growth models: Current status and perspectives. Agricultural and forest meteorology, 276, 107609.

Reynolds, J. F., & Acock, B. (1985). Predicting the response of plants to increasing carbon dioxide: a critique of plant growth models. Ecological Modelling, 29(1-4), 107-129.