Utilization of Time Series Analysis in Forecasting Agricultural Satellite Measurements

Jacob Flynn

Semester: Spring 2024

Course: 5900 Professional Practice

Credit Hours: 4

Faculty Supervisor: Dr. Gopichandh Danala

Company & Sponsor: Data Institute for Societal Challenges (DISC) – Dr. Ebert

Credits: United States Department of Agriculture, Agricultural Research Service (USDA-ARS), El-Reno, Oklahoma, USA

Contents

List of Figures

List of Tables

Glossary:

Enhanced Vegetation Index (EVI): Enhanced Vegetative Index is a quantitative measure of vegetation health and greenness, typically derived from satellite imagery. EVI is used for several assessments, such as determining ecosystem health and monitoring agricultural productivity. It utilizes several spectral bands, including red, blue, and near-infrared wavelengths, to capture vegetative properties and correct for atmospheric conditions/canopy background noise **[1]**.

Land Surface Water Index (LSWI): The Land Surface Water Index is a quantitative measure of the surface water in an environment, typically derived from satellite imagery. LSWI is particularly sensitive to changes in water content, allowing it to detect variations in surface water levels, soil moisture, and vegetation water content. Positive LSWI values typically indicate the presence of water, while negative values correspond to dry or non-vegetated surfaces **[2]**.

Abstract

This project aims to develop and evaluate time series algorithms to accurately forecast two agricultural satellite measurements: Enhanced Vegetative Index (EVI) and Land Surface Water Index (LSWI). These variables provide quantitative measurements for vegetation greenness and the amount of surface water in an environment, both important features for farmers and the USDA alike to maintain soil health and improve crop yields. EVI and LSWI are shown to have a relationship with field measurements found using Eddy Covariance systems. The goal of this project is to assess the feasibility of utilizing machine learning to forecast EVI and LSWI from the Eddy Covariance measurements.

To preprocess the data effectively, several statistical tests were conducted, including the Augmented Dickey-Fuller test for assessing stationarity and the Granger Causality test for feature selection. Vector Auto-Regression (VAR) and Long Short-Term Memory (LSTM) models were developed and compared to identify the optimal forecasting model for the target variables.

The results indicate that the LSTM models outperformed VAR, achieving an R-squared (R^2) value of 0.674 for EVI and 0.569 for LSWI. However, the LSTM models exhibited challenges in capturing noise within the data, suggesting a need for further analysis to address this issue. The models were also trained on a single field in central Oklahoma, possibly making it difficult to generalize to other pastures.

1. Introduction

Farmers rely on various sources of data including weather reports, soil analyses, and satellites to optimize crop yields and soil conditions. Advancements in technology have introduced an era where precise and high-quality data are readily available through soil-measuring instruments and satellites. Using both historical records and real-time data, agricultural experts have begun integrating data science methodologies into existing technologies, aiming to provide more accurate predictions and forecasts for growing seasons and soil conditions. Data science in agriculture can be used for crop monitoring, determining plants' resilience to climate change, or, in the case of this project, predicting future values of water-related indices. Parameters such as temperature, solar radiation, and precipitation play pivotal roles in this endeavor, contributing to the overarching goal of ensuring sustainable food production for future generations.

Currently, soil measurements are gathered using instruments known as Eddy Covariance (EC) systems, or infrared gas analyzers, but their deployment on a national scale entails significant expenses **[3]**. Conversely, water-related indices, such as the Enhanced Vegetative Index (EVI) and Land Surface Water Index (LSWI), are derived from satellite measurements. Previous research has demonstrated the efficacy of EVI and LSWI in predicting various field characteristics **[4]**, indicating their potential for reverse use: employing field measurements from EC systems to predict satellite data. This capability is crucial for filling gaps in satellite data caused by cloud coverage, as well as enabling early forecasting for various applications. The United States Department of Agriculture (USDA) has offered to collaborate with the University of Oklahoma Data Institute of Societal Challenges (DISC) in developing a time series model for this goal. The model will utilize weather and field characteristics as inputs to predict the satellite-derived data. The long-term goal is to use machine learning to mitigate the costs associated with the EC systems, by reducing the number of field sensors and optimizing their placement. The immediate goal, and purpose of this project, is to assess the feasibility of machine learning approaches in estimating satellite-derived measurements.

2. Objectives

The goal of this project is to streamline agricultural property measurements by developing time series analyses for two key vegetative indices, EVI and LSWI, thereby filling gaps in satellite measurements and enabling early forecasts. Additionally, it seeks to equip farmers with the accurate information necessary to maintain optimal soil conditions and achieve high crop yields. By analyzing these indexes a better understanding of pasture conditions can be obtained, influencing agricultural decisions.

To achieve these objectives, a comprehensive understanding of the data and its significance is crucial to identify optimal features for modeling EVI and LSWI. This involves intensive feature selection to discern the causal and associated variables influencing EVI and LSWI. Various time series models will be constructed to determine the most accurate approach to predicting these target variables using historical weather, soil, and satellite measurements. The success of the project will be evaluated through validation metrics such as R^2, Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE).

The completion of this project will be marked by the development of models capable of accurately forecasting EVI and LSWI based on the provided data, thereby helping farmers make informed decisions and better management practices influenced by the future estimates.

3. Data

The USDA provided a dataset comprised of weather, soil, and satellite data from several pastures near El Reno, Oklahoma. The files were subsequently cleaned, imputed, and converted into pickle files, which are now being used in Google Colaboratory. Imputation was needed because the target variables were measured on eight-day intervals, while the weather and soil measurements were measured daily. Linear interpolation was used to estimate the daily averages of EVI and LSWI, thus creating the needed integration of soil/weather and satellite data. Furthermore, a logarithmic transformation to various columns (Cooling Degree Days, Rainfall, Soil Moisture at depths 5 cm and 25 cm, and Rolling Sum of Rain for 7 days, 14 days, 21 days, and 28 days) was necessary, as they were positively skewed.

The dataset for the first pasture (P13) is comprised of 21 continuous features with 7,957 observations collected between March 16, 2000, and December 27, 2021. These features include essential meteorological variables such as air temperatures, relative humidity, and solar radiation, alongside soil parameters. Notably, the dataset contains two key parameters that are to be forecasted: Enhanced Vegetative Index (EVI) and Land Surface Water Index (LSWI). Detailed descriptions of each weather, soil, and satellite variable are provided in **Tables 1** and **2**.

Table 1: Weather and soil data

TMAX, TMIN, TAVG	Max, min, and average air temperature (F)
HAVG	Average relative humidity (%)
VDEF	Average daily vapor deficit (mb)
HDEG	Heating degree days (65 F standard)
CDEG	Cooling degree days (65 F standard)
WSPD	Average wind speed (mph)
ATOT	Solar radiation (Mj/m^2)
RAIN	Daily rainfall (in)
SAVG	Average soil temp. 10 cm under sod (F)
BAVG	Average soil temp. 10 cm under bare soil (F)
TR05, TR25, TR60	Soil moisture calibrated at 5 cm, 25 cm, 60 cm (C)
RAIN_7_Days, 14_Days, 21_Days, 28_Days	Rolling sum of rain for 7 days, 14 days, 21 days, and 28 days (in)

Table 2: Satellite data

EVI	Enhanced Vegetative Index
LSWI	Land Surface Water Index

Below is a box and whisker plot (**Fig. 1**) of each variable in the dataset. There is an observable difference in scales between each variable, which suggests a need to normalize the data and make all variables of equal importance, rather than having certain points dominate in the forecasts compared to others. This can especially be seen with TMAX, which has a range of 16 degrees to 113 degrees Fahrenheit, compared to EVI which has a range of -0.08 to 0.66.

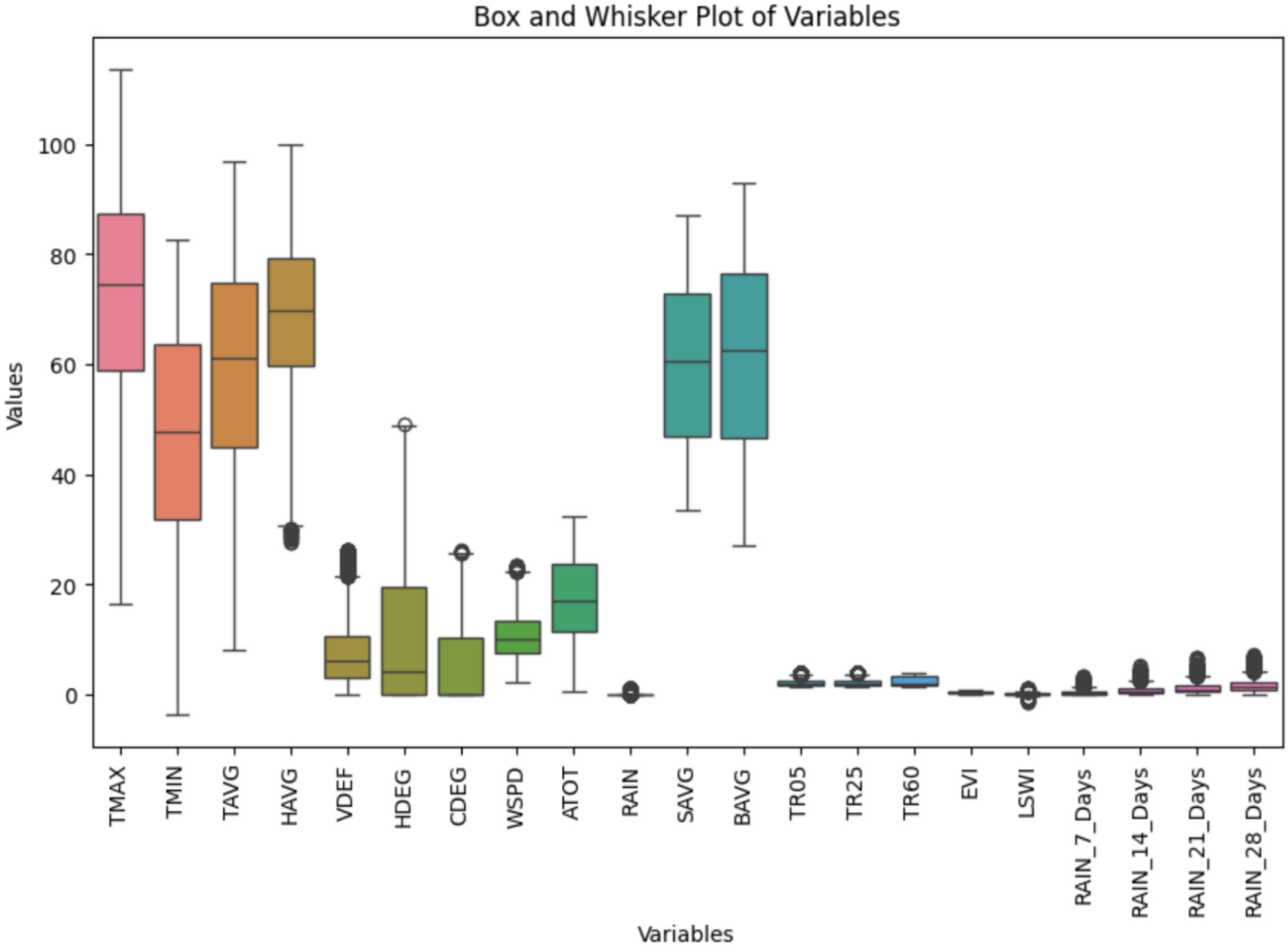

Figure 1: Box and whisker plot of all variables in the dataset

Many of the variables exhibit a wide interquartile range, suggesting high variability within the columns. It can also be seen that a majority of the medians within the boxes are centrally located, suggesting a symmetric distribution. This is important for the models to prevent them from introducing bias, for example towards higher temperatures or lower solar radiation. Upon further inspection, the variables on the right side of the plot contain a significant number of outliers, which is evident in the enhanced violin plot of **Fig. 2**.

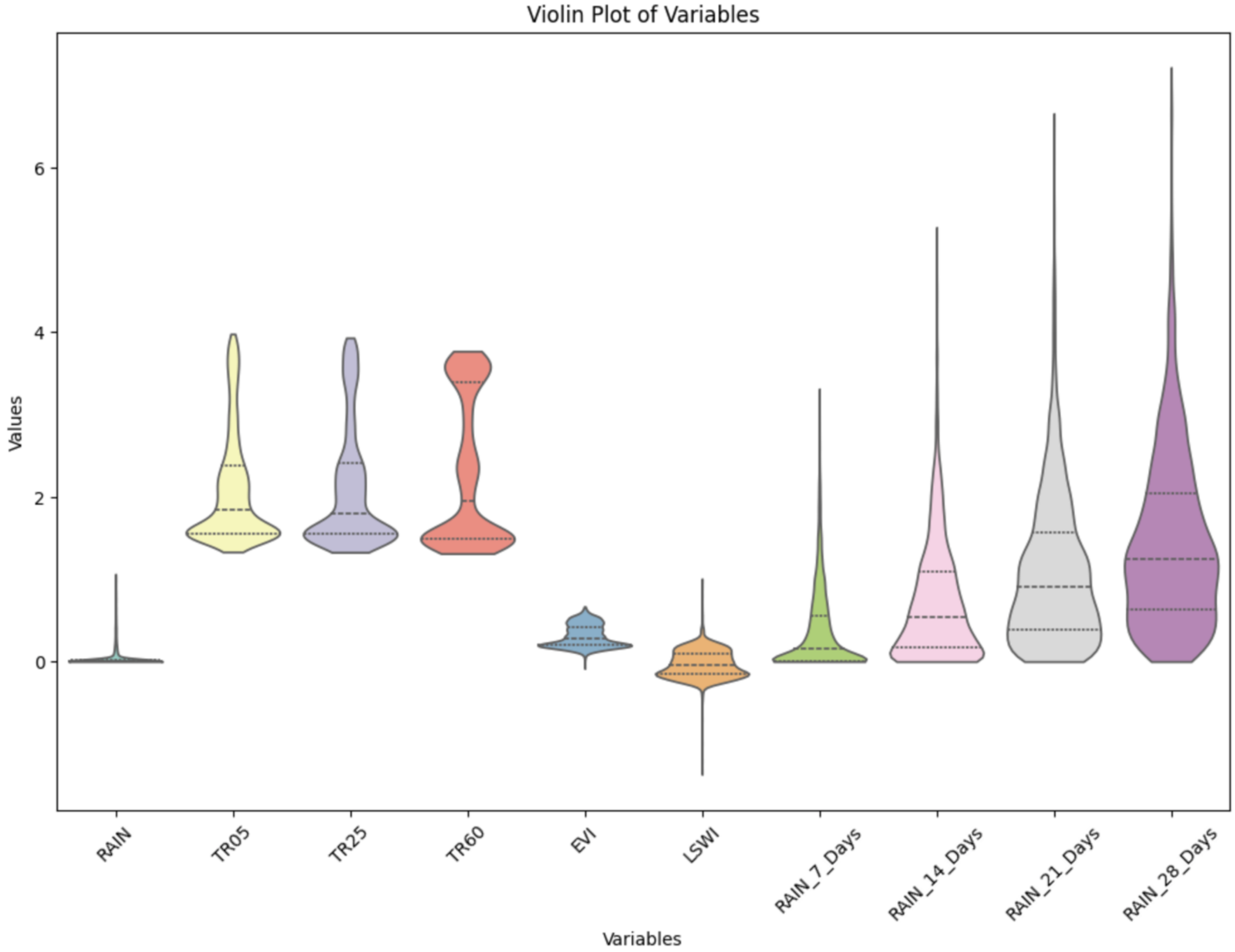

Figure 2: Violin plot of lower range variables

We can see that apart from the target variables, the lower range features are primarily related to moisture. The features are not equally distributed, as seen by the wide densities at the base of the variables, and what appeared as outliers in Fig. 1 are not as they seem. With the exception of the soil moisture variables (TR05, TR25, TR60), most features exhibit a central tendency around zero. This is logical, considering that rainfall can be sporadic, and not every day sees precipitation throughout the year. Consequently, what may have seemed like outliers are, in fact, crucial for the model to accurately gauge the frequency and magnitude of rainfall events, particularly in contexts where the majority of values are expected to be zero. This is particularly true for LSWI, as the main source of water is going to be rainfall for a lot of areas, so periods of more rain or drought will be reflected here.

The next visual, **Fig. 3**, shows line plots of all variables in the dataset.

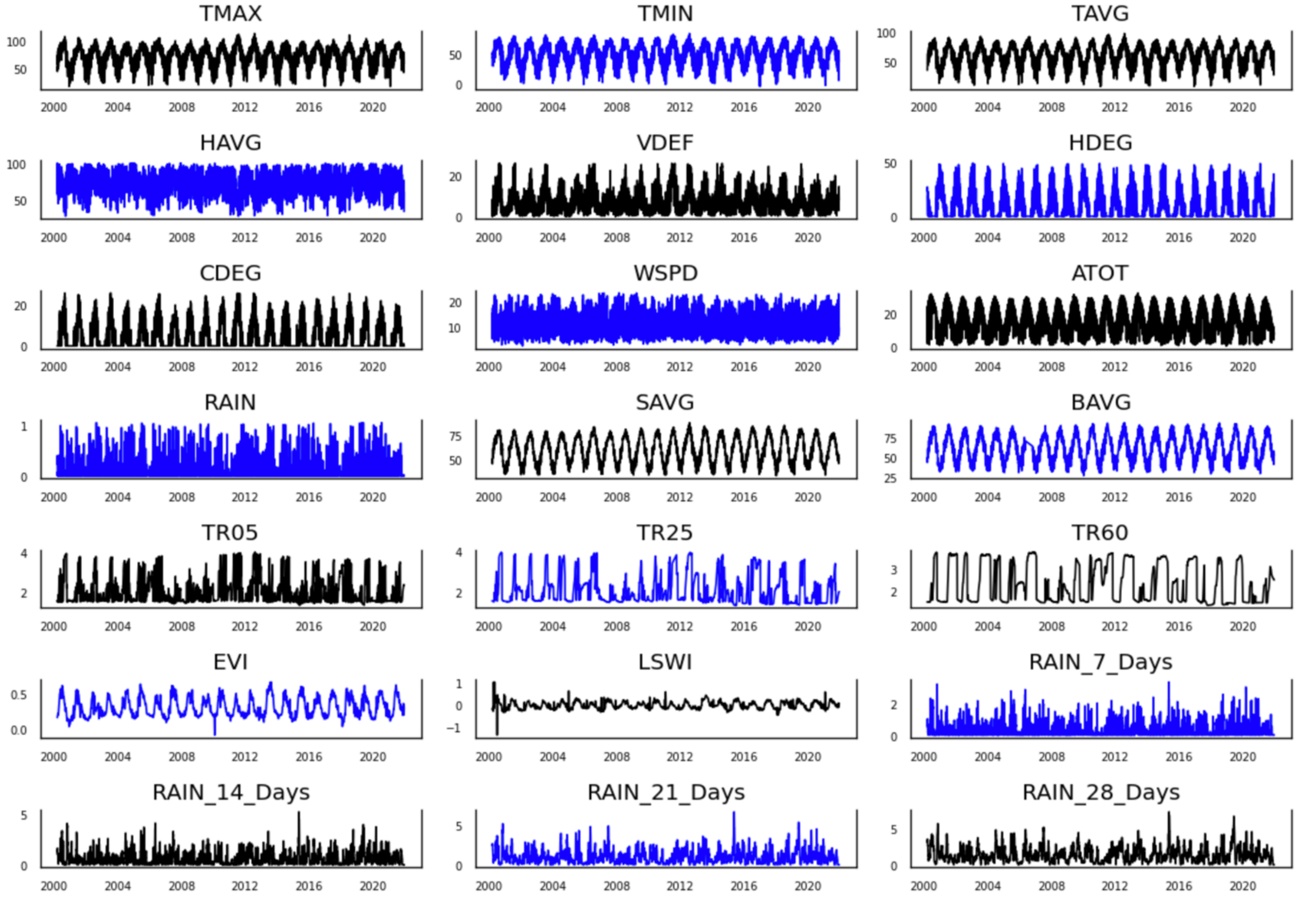

Figure 3: Line plots of all variables in the dataset between 2000 and the end of 2021

Many of the variables exhibit a sinusoidal pattern, which indicates an annual seasonality within the data. This aligns in the context of weather data, and how weather exposure affects the soil. Seasonal patterns can help improve the forecasts' accuracy by accounting for the annual variations and long-term trends in these conditions **[5]**. This is especially true due to the long time periods used to train the models. Plots of variables such as TMAX, HDEG, and ATOT, seen above, clearly show this annual pattern with four distinct peaks between each date on the X-axis. Other variables, such as rain, WSPD, and HAVG, do not show this pattern as clearly, but may still influence the target variables which warrants further investigation.

Lastly we have **Fig. 4**, which is the line plot of the two target variables, EVI and LSWI.

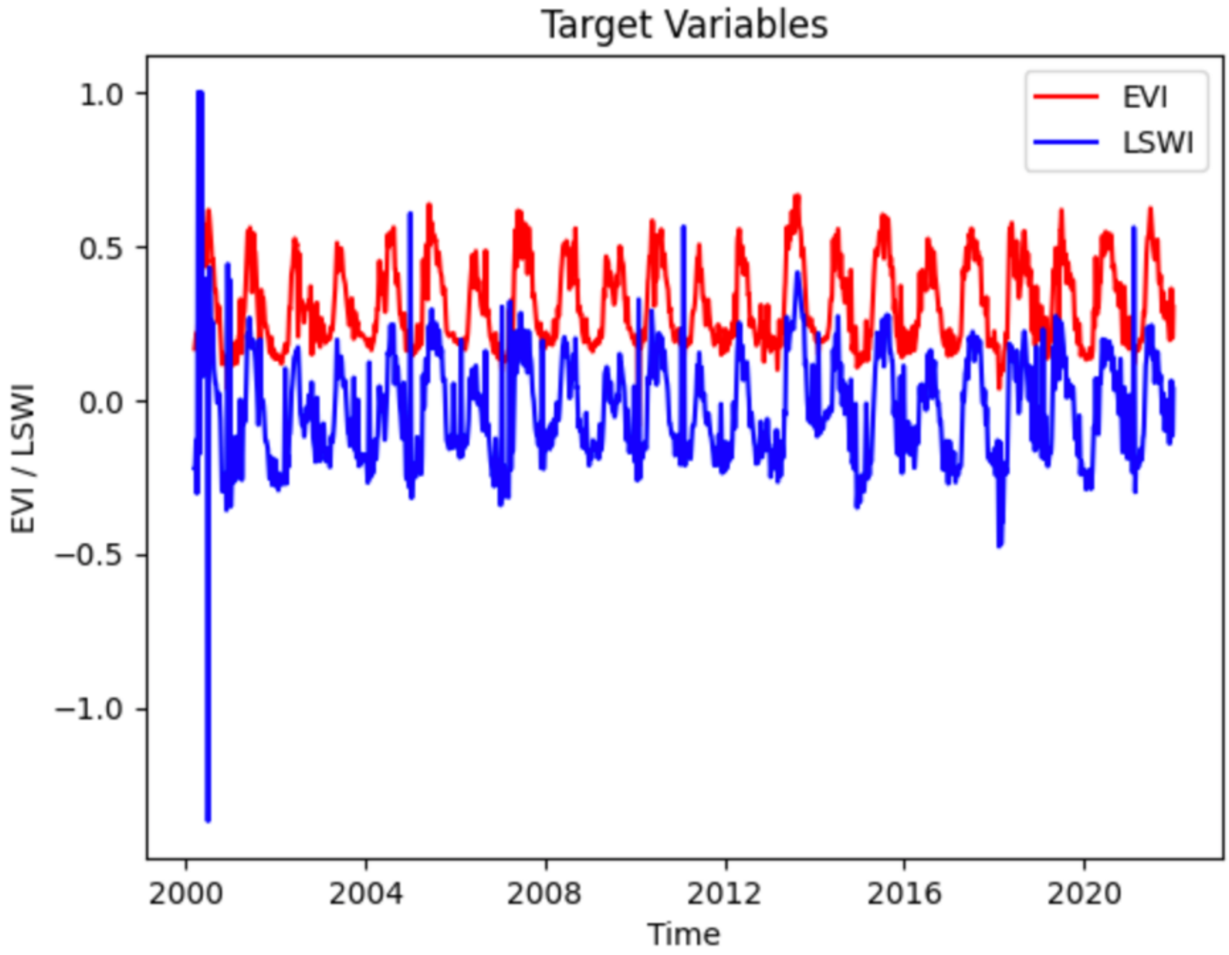

Figure 4: Line plot of target variables EVI and LSWI

Both EVI and LSWI show consistent sinusoidal patterns, aligning with the growing seasons typically observed in the middle of the year. This visual similarity emphasizes the strong relationship between weather, soil, and vegetation dynamics, supporting the potential of weather and soil data for forecasting these variables. Despite the relatively low variability within the data, as depicted in Fig. 1 and Fig. 4, normalizing the data can further enhance the comparability of the variables and make for a more accurate forecast. Thus, the MinMaxScaler from the sklearn library was used to transform each feature to have a range of [0, 1]. As previously mentioned, this step was necessary because the features originated from different scales, including Fahrenheit, inches, and Mj/m^2. Standardizing the data to a uniform scale ensures balance across all features, which is important for models sensitive to variations in feature scales.

4. Methodology

4.1 Tests

After the data had been scaled, a stationarity test was conducted using the Augmented Dickey-Fuller (ADF) test, which was learned under the guidance of Dr. Danala as part of Practicum. This test verifies whether a variable has constant variance within the data, a prerequisite for an accurately forecasted model. The ADF test was chosen because it is one of the most widely used stationarity tests in time series analysis; it evaluates the null hypothesis that there is a unit root in an auto-regressive model, indicating that the data is nonstationary **[6]**. Nonstationary data can lead to unreliable predictions, as models may not anticipate the patterns inherent in the data. Given that one of the selected models for this objective is Vector Auto-Regression (VAR), which assumes stationarity in the data, it was imperative to confirm this assumption beforehand. Fortunately, the test revealed that all the data was already stationary, making further transformations unnecessary.

Once it was confirmed that the data was stationary, the Granger Causality test was employed for feature selection. The Granger Causality test was chosen because it determines whether one variable can meaningfully describe another through a linear relationship **[7]**, as learned under the supervision of Dr. Danala. The test also includes a lag parameter that specifies how many past observations are needed to capture any dependencies between variables. For this test, a lag of 10 was chosen in order to capture short-term dependencies between EVI and LSWI, meaning that fluctuations in temperature or rain could affect the target variables within this delay. In this context, variables with a $p < 0.05$ were considered significant for feature selection. The interpretation for hypothesis testing was learned in Dr. Barker's Fundamentals of Engineering Statistical Analysis course.

The Granger Causality test is a method used to determine whether past values of one variable (X) can aid in predicting future values of another variable (Y) beyond what can be predicted using past values of Y alone. To conduct the test, a Vector Auto-Regressive (VAR) model with lagged terms for both variables is used to specify the number of terms included in the model. In the autoregressive equations for Y and X, intercept terms (α_0, β_0), lagged coefficients ($\alpha_1, \beta_1, \dots$), lagged values ($Y_{t-1}, X_{t-1}, \dots$), and error terms (ε_t, η_t) are included in the linear equations shown below, with a lag order equal to three **[8]**.

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \alpha_3 Y_{t-3} + \varepsilon_t$$
$$X_t = \beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + \eta_t$$

These equations allow the assessment of the predictive power in past values of X on future values of Y by comparing the regression models. An F-test is used to determine whether the inclusion of lagged values of X improves the predictive performance of Y by having a p-value less than 0.05. It's also important to note that the Granger Causality test relies on several assumptions, including stationarity of the data, linearity between variables, and discrete time periods. However, it's worth noting that the data used for this test may challenge the assumption of discrete time, given that the field and weather measurements were recorded daily, while the satellite observations were taken every 8 days and subsequently interpolated using linear interpolation.

A correlation matrix was also used to aid in feature selection, as shown by the heatmap in **Fig. 5** on the next page.

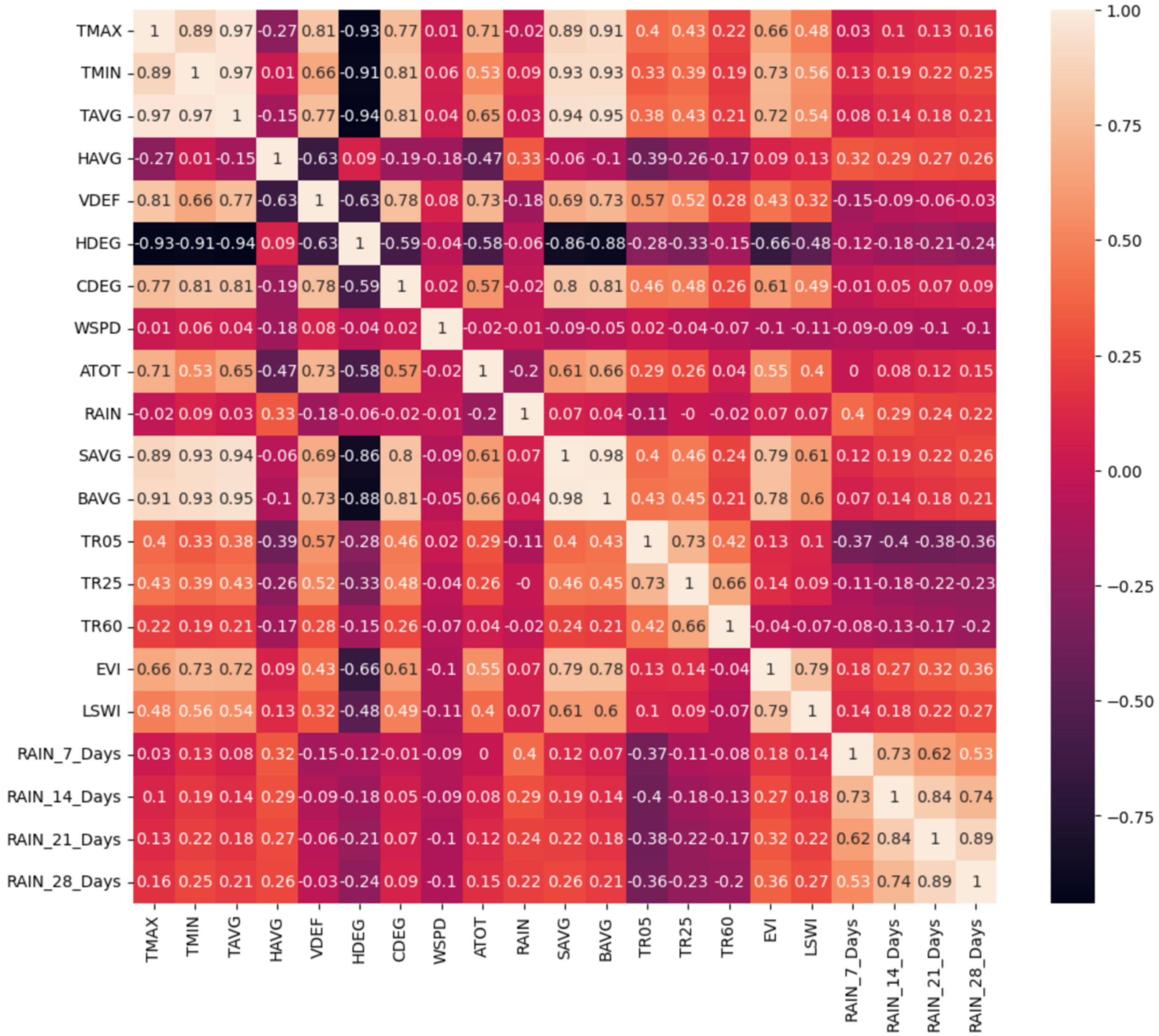

Figure 5: Heatmap of all features

While the Granger Causality test shows causal relationships between variables, the correlation matrix provides linear associations between them, thus adding additional insights into how the variables are related and how strong the relationships are. The correlation matrix uses Pearson's correlation to calculate the strength of the linear relationship between -1 and 1. Pearson's was chosen because it measures linear relationships and is robust to outliers, meaning extreme values do not influence the coefficients. Both of these qualifications made Pearson's optimal, since the Granger Causality test also assumes a linear relationship, and outliers seen in extreme weather events would not hold a strong influence over the target variables. The heatmap, as well as Pearson's correlation, were learned in Dr. Nicholson's Intelligent Data Analytics course.

After completing the feature identification process, six variables were identified to aid in forecasting EVI as shown in **Table 3**, while eight variables were selected to forecast LSWI in **Table 4**.

Table 3: Variables chosen to aid in forecasting EVI

TAVG	Average air temperature
HDEG	Heating degree days
WSPD	Windspeed
ATOT	Solar radiation
RAIN	Daily rainfall
TR60	Soil moisture calibrated at 60 cm

Table 4: Variables chosen to aid in forecasting LSWI

TMIN	Min air temperature
TAVG	Average air temperature
HDEG	Heating degree days
ATOT	Solar radiation
SAVG	Average soil temp. 10 cm under sod
BAVG	Average soil temp. 10 cm under bare soil
TR60	Soil moisture calibrated at 60 cm
RAIN_28_Days	Rolling sum of rain for 28 days

Using these variables, two types of models were determined to be fit for the project objective: Vector Auto-Regression (VAR) and Long Short-Term Memory (LSTM). Both methods are supervised learning approaches, as the models are trained on historical data with known EVI and LSWI values to make predictions.

4.2.1 Models – VAR

Vector Auto-Regression (VAR) was chosen because it operates on the principle that each variable in the system is a linear function of its own past lags, as well as the lags of other variables **[9]**. VAR is used when variables are interrelated and influence each other over time. This implies that the model leverages historical data of multiple parameters to predict future values. Thus, it was thought that variables such as rain, heating degree days, and the soil moisture were interrelated for EVI and LSWI based on the water and plant cycles, making VAR a suitable choice. It was further thought that VAR was suitable on the basis of variable dependence. Each variable in the system can function as an independent or dependent variable, making the model flexible in determining forecasts other than the target feature if necessary.

In employing Vector Auto-Regression, several key assumptions were made to provide reliable estimates between the time series variables. Firstly, VAR models assume that the relationship between variables can be captured using linear equations. As seen in the heat map in Fig. 5, and the features chosen in Tables 3 and 4, there does exist some linearity between the target variables and those chosen in feature selection. While they are not all perfect, other variables were chosen based on the Granger Causality test, which account for the causation of the target variables. Secondly, the assumption of stationarity, as verified by the Augmented Dickey-Fuller (ADF) test, asserts that the mean and variance of the variables remain constant over time—a condition met by each feature in our dataset. Lastly, the assumption of interrelatedness says that the variables influence each other over time. This was confirmed through the Granger Causality test, which demonstrated the causal influence of each selected variable on EVI and LSWI. Together, these assessments validate the underlying assumptions of the VAR modeling approach and provide a foundation for the analysis.

Vector Auto-Regression also has several limitations to consider. One downside of VAR is its sensitivity to the lag order. The lag determines how many past observations of each variable are used to predict the next value. In this project's case, a lag of 365 was chosen to be appropriate, as it allows the model to utilize values from the same month in the previous year for predictions, rather than those from a week or quarter prior. This captures the yearly, seasonal patterns that result from Earth's orbit around the sun. Another limitation in VAR is that it is primarily used to forecast short-term data, rather than long-term. This is primarily due to the assumption of stationarity in the data, as longer-term data can become non-stationary over time due to higher complexity or abrupt shifts. Furthermore, VAR models offer limited flexibility in parameter tuning. Apart from the lag order, the model's hyperparameters are sparse, with only the data frequency—such as daily measurements in our case—available for adjustment. This lack of additional tuning parameters may constrain the model's ability to capture more nuanced dynamics beyond the lag value.

The architecture for VAR is consistent across both the EVI and LSWI models, differing primarily in input variables, with EVI having six variables and LSWI having eight. Both models adopt a lag order of 365, representing one year of daily observations. Unlike LSTM models, which will be discussed in **Section 4.2.2**, VAR architecture is rooted in a system of linear equations, where each variable is regressed on its lagged values and those of other variables in the system **[10]**. VAR models lack hidden states or memory cells. They rely on linear regression to model the relationships among variables and do not use layers to capture complex dynamics or long-term dependencies in time series data. Key components of VAR architecture include selecting the lag order and fitting the model using ordinary least squares.

4.2.2 Models – LSTM

While VAR models are preferable for capturing linear patterns in time series data, Long Short-Term Memory (LSTM) models are capable of expressing nonlinear patterns and long-term dependencies, thus complementing the modeling approach. Unlike traditional statistical models, LSTMs leverage recurrent neural networks (RNNs) to effectively capture dependencies between successive time steps of data **[11]**. One key strength, and what made it a favorable modeling approach, is its flexibility in capturing diverse patterns and the dependencies within the data. LSTM layers are designed to learn and retain information over long sequences of data, effectively modeling irregular intervals and highly variable data. This made LSTM feasible in training 20-years' worth of data in order to predict the last two years.

LSTM models are ideally suited for capturing underlying patterns within sequential data, making them particularly effective for analyzing long-term trends and recurring patterns. While LSTMs are capable of handling nonstationary data, stationarity aids the model in discerning the fundamental patterns within the dataset. Fortunately, the dataset exhibits stationarity across all variables, as confirmed by the Augmented Dickey-Fuller (ADF) test. Additionally, LSTMs operate under the assumption that the features provided to the model are relevant and contribute meaningfully to the extraction of patterns in the target variable. This assumption was corroborated by the Granger Causality test, which identified causal relationships between variables, and by identifying features from the correlation matrix. Lastly, the effectiveness of LSTM models relies on the availability of sufficient training data to learn the complex relationships inherent in the data. In this study, the LSTM models were trained and validated using 20 years of data, with the last 2 years reserved for testing. Further details on the train/test split will be discussed in **Section 4.3**.

The following paragraph summarizes each of the hyperparameters used, as described by Ralf Staudemeyer and Eric Morris in "Understanding LSTM a tutorial into Long Short-Term Memory Recurrent Neural Networks." LSTMs utilize several tuning parameters, including the learning rate, number of neurons, number of epochs, activation function, and regularization. Additionally, LSTMs can incorporate multiple layers, allowing for the integration of Dropout and Dense layers. The learning rate determines the magnitude of weight updates during training, impacting the convergence speed and stability of the model. Neurons refer to the memory capacity of the LSTM, influencing its ability to capture patterns in the data. Epochs denote the number of iterations over the training data, with each iteration adjusting model parameters to minimize the loss function. The activation function introduces nonlinearity to the model, facilitating the capture of complex relationships in the data. Regularization techniques such as L1 or L2 regularization penalize large weights to prevent overfitting. Each LSTM layer serves as a fundamental component of the recurrent neural network, applying the specified hyperparameters to retain information across sequences. Dropout regularization mitigates overfitting by randomly deactivating neurons, while Dense layers connect preceding neuron layers to produce the final output prediction **[12]**.

LSTM models are a type of RNN designed to handle sequential data, exceling at capturing patterns and long-term dependencies. This is done through the use of memory cells which contain an input gate, a forget gate, and an output gate, all of which control the flow of information **[11]**. This flow is managed by the architecture of the LSTM units, which include the input variables, the number of neurons, and various other hyperparameters.

For the LSTM architecture used in this project, separate models were developed for EVI and LSWI data analysis. The EVI model incorporated six selected features per time step, chosen based on the results of the Granger Causality test. These features were inputted into the initial LSTM layer comprising 30 neurons, determined through a grid search that optimized the validation loss. Following this layer, a dropout layer with a 30% dropout rate was applied to mitigate overfitting; this was determined through iterative model runs focusing on minimizing RMSE and maximizing R^2 metrics. Subsequently, two additional LSTM layers were introduced with five and two neurons respectively. These were aimed at capturing nonlinearities within the data and extracting higher-level representations. The decrease in number of neurons additionally acts as a dimension reduction, which helps prevent overfitting and is easier to process; they were found through experimentation of the model while also taking computational efficiency under consideration.

Finally, a dense layer with 1 neuron was added to generate predictions. The "tanh" activation function demonstrated superior performance compared to alternatives, such as “relu” or “sigmoid.” The model was optimized using the Adam optimizer with a learning rate of 0.001, trained over 20 epochs, with a batch size of 32.

The LSTM model for LSWI underwent a comparable process to the LSTM model for EVI. It incorporated eight features per time step, selected based on the results of the Granger Causality test. In the initial LSTM layer, the model employed 50 neurons, followed by a dropout layer with a 40% dropout rate to address potential overfitting. After these layers, two additional LSTM layers were added, consisting of five and two neurons, respectively. These layers aimed to capture nuanced patterns within the data and extract higher-level representations. The model concluded with a dense layer containing a single neuron, which underwent 1% L2 regularization to further mitigate overfitting. Finally, the model was compiled using the Adam optimizer with a learning rate set at 0.01.

4.3 Data Splitting and Hyperparameter Tuning

Due to the differences in their operational mechanisms, two train-test splits were employed during model development. For the VAR model, the training data spanned the initial 20 years from 2000 to the conclusion of 2019, while the subsequent two years from 2020 to the end of 2021 were designated as the test set. This split was chosen to ensure enough information was provided to the model that the seasonal pattern could be recognized and forecasted.

In contrast, the LSTM model utilized a validation set to facilitate the measurement of validation loss, while continuing to follow the same rationale as above. Consequently, the training data encompassed the period from 2000 to the end of 2017, the validation set extended from 2018 to the conclusion of 2019, and the testing set covered the years 2020 to 2021.

The decision to forgo cross-validation stemmed from the recognition that the inherent seasonality within the data could be compromised. By employing a hold-out validation approach, the continuity between past and future data points was preserved, thereby enhancing the model’s ability to capture the dynamic nature of the dataset.

In developing the LSTM models, a tuning grid was used to determine the optimal hyperparameters. The hyperparameters included the learning rate, the number of neurons, the number of epochs, the activation function, and the regularization term. These hyperparameters were tested against each other, with the combination yielding the lowest validation loss selected as the optimal configuration.

The hyperparameters used in making the LSTM models can be found in **Table 5** for each target variable.

Table 5: LSTM hyperparameters for each target variable

LSTM Hyperparameters					
	Learning Rate	**Neurons**	**Epochs**	**Activation Function**	**Regularization**
EVI	0.001	30	20	tanh	None
LSWI	0.01	50	20	tanh	L2(0.01)

5 Results and Analysis

In summary, the goal of this project was to develop a model capable of predicting EVI and LSWI from field measurements using time series analysis techniques. This approach seeks to leverage historical EC system measurements in forecasting traditionally satellite-based observations, thereby filling gaps in the data and enabling early projections. The data underwent preprocessing steps including scaling, ensuring stationarity, feature selection, and tuning for VAR and LSTM models, rendering them ready for analysis.

Both the VAR and LSTM models exhibit similar trends in predicting EVI and LSWI. While both models capture the seasonal patterns observed in the data, they struggle to accurately capture individual peaks and troughs, particularly in predicting LSWI. The LSTM models appear to capture more noise compared to the VAR models, rendering them problematic in determining individual values.

Fig. 6 on the next page illustrates the performance of the VAR model using a lag of 365 to predict EVI.

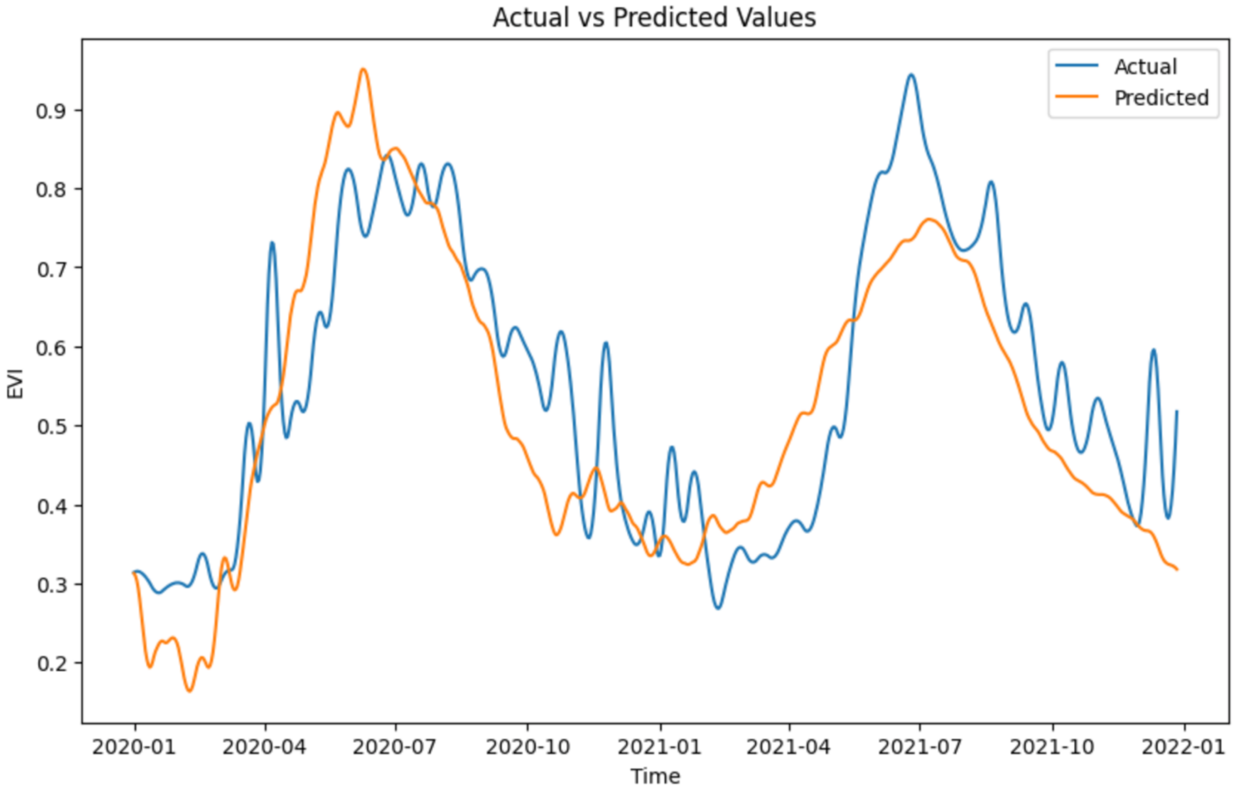

Figure 6: Performance of the VAR model using a lag of 365 to predict EVI.

The plot displays the actual values represented by blue lines and the predicted values by orange lines. While the general trend is captured, there are fluctuations between overestimation and underestimation of peaks and troughs, particularly noticeable around July 2020 and August 2021. These fluctuations may reflect the complexity and irregularity of weather patterns. Overall, the model achieved an R-squared (R^2) of 0.671 and a Root Mean Squared Error (RMSE) of 0.105. This indicates that the model explains over half of the variability in the data but struggles to accurately predict the values of EVI. These results can be further supplemented with the Mean Absolute Error (MAE) of 0.088, Mean Square Error (MSE) of 0.011, and the Mean Absolute Percentage Error (MAPE) of 17.315%. MAE measures the average magnitude of errors between predicted and actual values, thus an MAE of 0.088 indicates that the model on average has an absolute error of 0.088 EVI units. Similarly to RMSE, MSE is the average squared difference between predicted and actual values but is more sensitive to outliers, as it squares the errors. As such, an MSE of 0.011 suggests that the model is close to the ideal values, with 0 being perfect. Finally, MAPE measures the percent difference between the predicted and actual values, meaning the model deviates about 17.315% from the actual values.

The observed errors could cause problems with farmers over/underestimating the health or saturation of their field. This is particularly an issue when the model overestimates the EVI value, leading to farmers thinking that their pasture is healthier than it truly is, when it would actually benefit from supplementation of nutrients or water. Underestimations could also cause problems: incorrectly interpreting that pastures could benefit from additional water could lead to oversaturating the field and causing crop damage.

Following the VAR forecast for EVI, LSTM was utilized to predict EVI as seen in **Fig. 7** below.

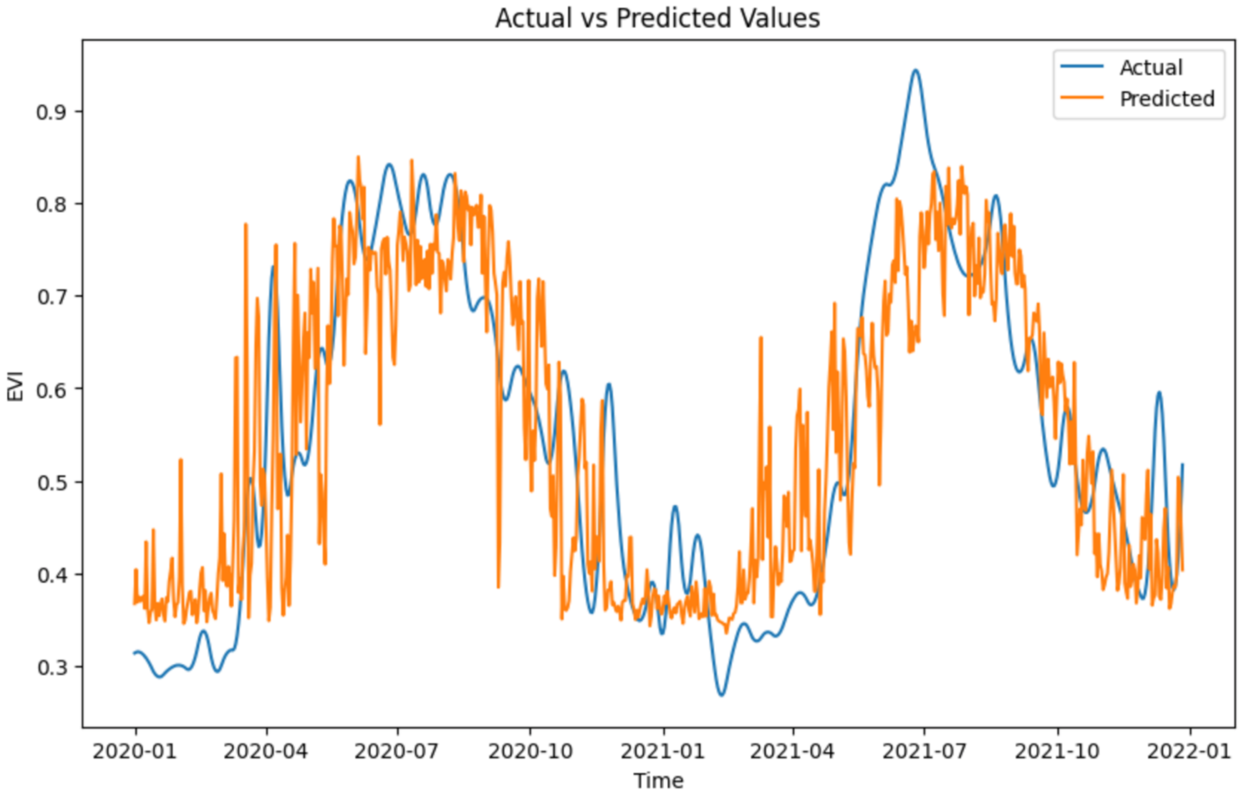

Figure 7: Time series forecasting for EVI using LSTM during the period 2020-2022.

The LSTM model exhibits a higher level of noise in the data compared to the VAR model, as depicted in the figure. However, it closely follows the general trend of the data, accurately capturing the peaks and troughs of EVI. The increased noise forecasted by the LSTM model may be attributed to the absence of regularization, which warrants further investigation. Notably, the tuning process did not necessitate regularization for achieving optimal validation loss. Despite its noise, the LSTM model achieved an R^2 value of 0.674 and an RMSE of 0.104. These metrics indicate slightly higher accuracy in explaining the variance compared to VAR and marginally better performance in predicting values. However, visually the LSTM model appears to capture excessive noise. The model did receive better error and accuracy scores overall with an MAE of 0.082, an MSE of 0.011, and an MAPE of 16.228%.

This model suffers from a different problem than VAR; instead of over- and underestimating the values on an individual peak and trough basis, it continuously over- and underestimates individual values themselves. On a day-by-day basis, this would lead to confusion and warrant continuous investigation of the health of pastures, thus making it infeasible to use in this regard. It does follow the overall trend of the data closely, making it of possible use for long term forecasts. It could be used to follow the direction of EVI, for instance if the average EVI values tend to rise or fall in different seasons or years forward.

Now, moving on to LSWI, the first VAR model can be seen below in **Fig. 8**.

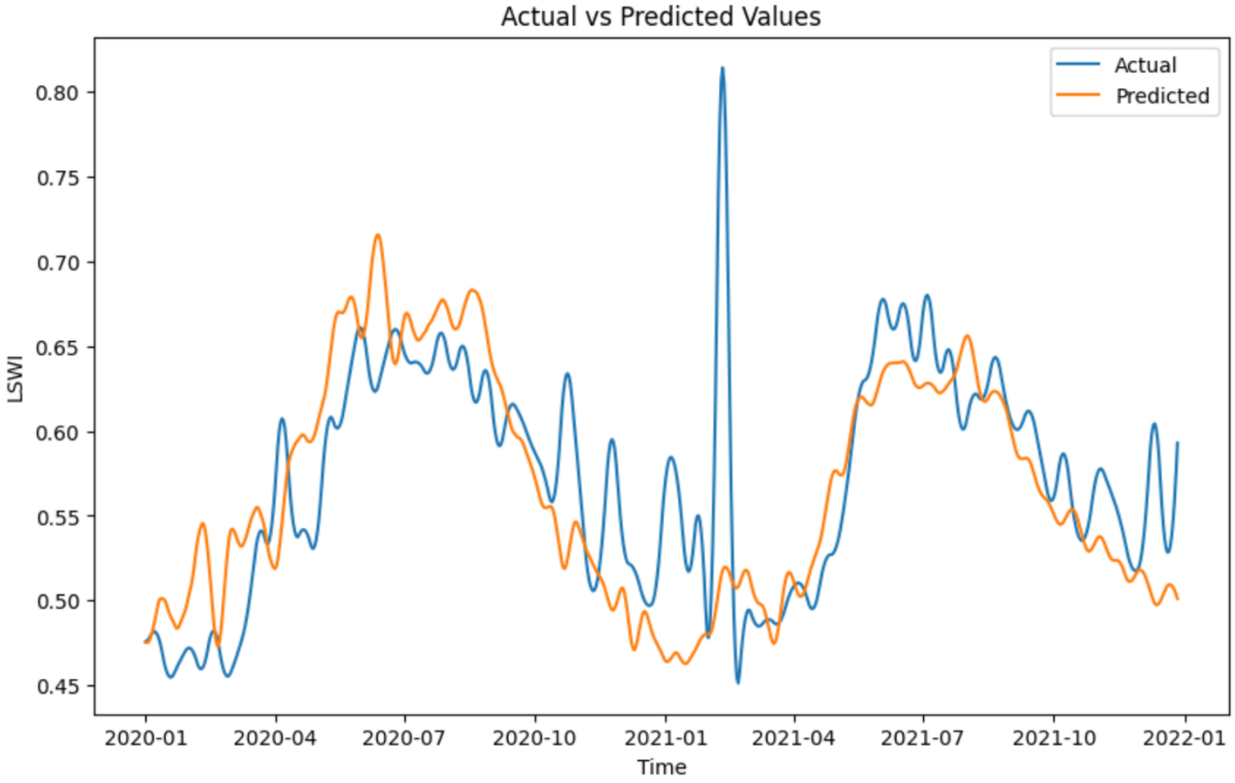

Figure 8: Performance of the VAR model using a lag of 365 to predict LSWI.

The VAR model exhibits a similar pattern in forecasting both EVI and LSWI. Initially, it overestimates the peak values between April 2020 and October 2020, followed by a rapid decline below the actual values for the remainder of the forecast period. Although there is a modest attempt to capture the substantial peak observed around February 2021, the model largely misses it, resulting in a discrepancy in LSWI of approximately 0.33 units. This model yielded the poorest performance metrics among the four models, with an R^2 value of 0.415 and an RMSE of 0.049. These metrics indicate limited explanatory power for the variance but relatively better accuracy in value prediction compared to both EVI models. Additional statistics show that the model achieved an MAE of 0.035, an MSE of 0.002, and a MAPE of 6.127%.

Visually, the actual values for LSWI exhibit greater variability on a month-by-month basis, which may account for the challenges encountered when utilizing VAR to model a larger time span. Similarly to the application of VAR for EVI, the observed over- and underestimations of LSWI could have significant implications for crop yield. For instance, the presence of substantial peaks in the plot suggests higher surface water levels in the fields than initially assumed. This discrepancy could prompt unnecessary irrigation practices, potentially exacerbating the health of the crops rather than improving it.

Fig. 9 shows the final model, using LSTM to model LSWI.

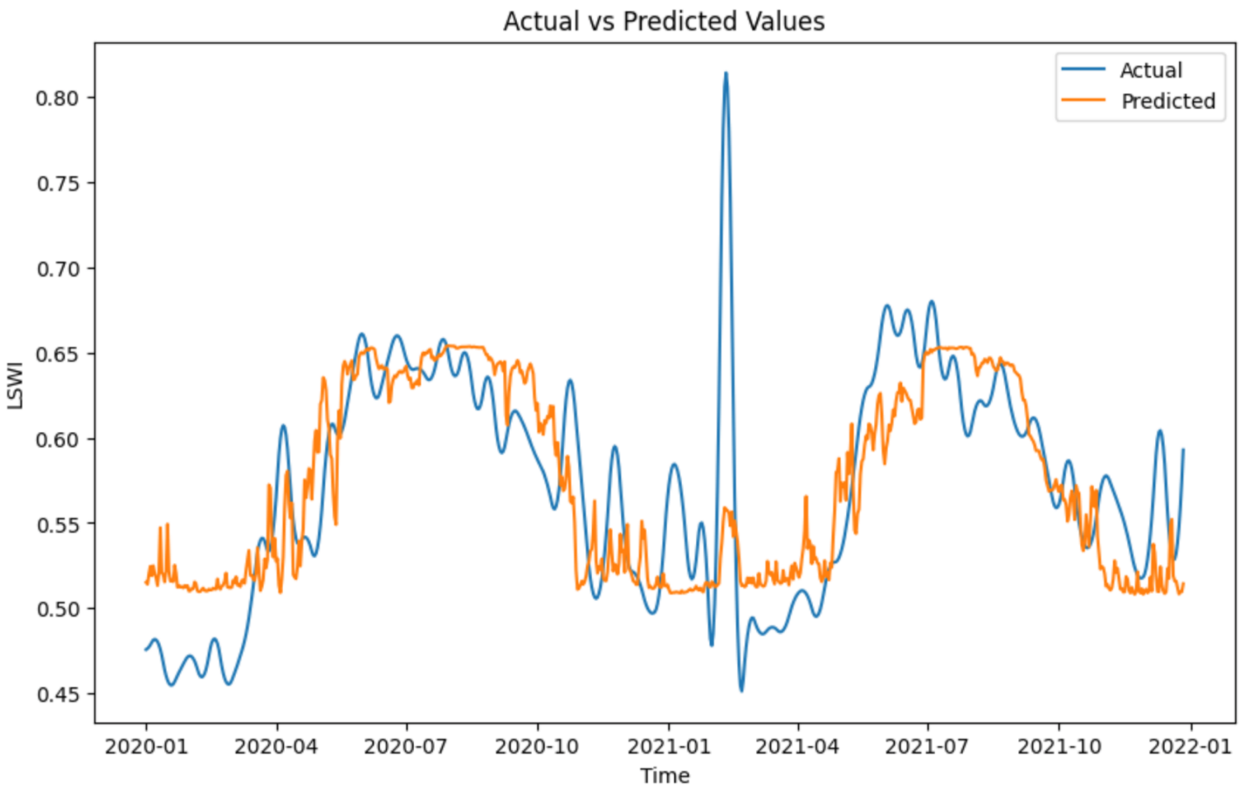

Figure 9: Time series forecasting for LSWI using LSTM during the period 2020-2022.

The LSTM model for LSWI closely tracks the data trend, exhibiting less noise compared to the LSTM model for EVI, albeit still retaining some. Notably, there is an improved attempt to capture the peak in February, though still falling short by approximately 0.28 units. The LSTM model for LSWI demonstrates higher variance capture compared to VAR, achieving an R^2 of 0.569, and slightly more accurate modeling with an RMSE of 0.042. It was also more successful in the other metrics with an MAE of 0.031, an MSE of 0.001, and a MAPE of 5.603%.

Though Fig. 9 received a poor R^2 score, it visually fits the test data rather well compared to VAR for LSWI. It still has some trouble over- and underestimating LSWI but to a much smaller degree. With the exception of February 2021, the typical predicted LSWI values are less than 0.50 units off from the actual, and the model continues with the same relative accuracy for the full two years of testing data. The R^2 and RMSE metrics are not great for this model, but with further investigation and tuning, it could yield better results.

A summary of each model statistic can be seen below in **Tables 6** and **7**.

Table 6: Summary statistics for EVI using VAR and LSTM

EVI					
	MAE	**MSE**	**RMSE**	**MAPE**	**R^2**
VAR	0.088	0.011	0.105	17.315	0.671
LSTM	0.082	0.011	0.104	16.228	0.674

Table 7: Summary statistics for LSWI using VAR and LSTM

LSWI					
	MAE	**MSE**	**RMSE**	**MAPE**	**R^2**
VAR	0.035	0.002	0.049	6.127	0.415
LSTM	0.031	0.001	0.042	5.603	0.569

In both EVI and LSWI, the LSTM model demonstrates slightly lower error metrics (MAE, MSE, RMSE) and higher accuracy metrics (MAPE, R^2) compared to the VAR model, suggesting better overall performance. However, the LSTM model exhibits challenges in capturing noise in the data, particularly noticeable in visual inspection.

The models show similar relative accuracy for EVI, as indicated by the marginal difference in MAPE and delta R^2 scores. In contrast, for LSWI the difference in MAPE is even smaller, suggesting closer forecasting accuracy between the models. However, the lower R^2 values indicate that both models struggle to capture the variance in LSWI, highlighting potential limitations in the feature selection process.

6 Deliverables

The outcome of this project was developing a robust time series model capable of accurately forecasting Enhanced Vegetation Index (EVI) and Land Surface Water Index (LSWI) based on weather and soil variables. To achieve this objective, two sophisticated models for each target variable were meticulously constructed, using Vector Auto-Regression (VAR) and Long Short-Term Memory (LSTM). The USDA hopes to incorporate these models in their objective of mitigating the use of the Eddy Covariance systems, which are of considerable cost both financially and environmentally.

These models offer significant implications for both research and business objectives. For the USDA and farmers alike, the predictive power of time series analysis is invaluable. It enables the forecasting of pasture conditions, empowering stakeholders to make well-informed decisions regarding pasture management, resource allocation, and crop yield optimization based on real-time field conditions.

Moreover, the utilization of VAR in this context extends the scope beyond EVI and LSWI predictions. It enables the forecasting of all predictor variables incorporated in the model. For example, by leveraging LSWI data, forecasts for solar radiation can also be derived. This expanded utility enhances the versatility and applicability of the model, further amplifying its impact on research and business objectives within the agricultural domain.

7 References

[1] USGS, “Landsat Enhanced Vegetation Index.” 2024. https://www.usgs.gov/landsat-missions/landsat-enhanced-vegetation-index

[2] Christian, J., Basara, J., Lowman, L., et al. “Flash Drought Identification from Satellite-Based Land Surface Water Index.” 2024. Elsevier. https://doi.org/10.1016/j.rsase.2022.100770

[3] Wagle, P., Gowda, P., et al. “Response of Tallgrass Prairie to Management in the U.S. Southern Great Plains: Site Descriptions, Management Practices, and Eddy Covariance Instrumentation for a Long-Term Experiment.” USDA, Agricultural Research Service. 2019. doi: https://doi.org/10.3390/rs11171988

[4] Son, N., Chen, C. et al. “Classification of Multitemporal Sentinel-2 Data for Field-Level Monitoring of Rice Cropping Practices in Taiwan.” Elsevier. 2019. doi: https://doi.org/10.1016/j.asr.2020.01.028

[5] Ghysels, E., Osborn, D., Rodrigues, P. “Chapter 13 Forecasting Seasonal Time Series.” Elsevier. 2006. doi: https://doi.org/10.1016/S1574-0706(05)01013-X

[6] Jalil, A. Rao, N., “How to Write about Economics and Public Policy – Dickey Fuller Test.” Elsevier. 2018.

[7] Stern, D. “Encyclopedia of Energy – Economic Growth and Energy.” Elsevier. 2004. doi: https://doi.org/10.1016/B0-12-176480-X/00147-9

[8] Andrikopoulos, P., et al. “Handbook of Frontier Markets: The European and African Evidence.” 2016. https://doi.org/10.1016/C2015-0-00473-7

[9] Penn State. “11.2 Vector Autoregressive models VAR(p) models.” Penn State Eberly College of Science. 2023. https://online.stat.psu.edu/stat510/lesson/11/11.2

[10] Stock, J., and Watson, M. “Vector Autoregressions.” Journal of Economic Perspectives. Vol. 15, 4. 2001. https://www.princeton.edu/~mwatson/papers/Stock_Watson_JEP_2001.pdf

[11] MathWorks. “Long Short-Term Memory Neural Networks.” 2024. https://www.mathworks.com/help/deeplearning/ug/long-short-term-memory-networks.html

[12] Staudemeyer, R., Morris, E. “Understanding LSTM a tutorial into Long Short-Term Memory Recurrent Neural Networks.”Arxiv. 2019. https://arxiv.org/pdf/1909.09586

8 Self-Assessment

My primary learning objectives for this project were to gain proficiency in time series analysis and apply my academic knowledge to real-world problems using the CRISP-DM methodology. While the final results were not as precise as anticipated, I successfully deepened my understanding of time series analysis and data manipulation. Throughout the project, I acquired valuable insights into modeling techniques such as VAR and LSTM, which I applied while collaborating with the Data Institute for Societal Challenges (DISC) and the USDA.

I expanded my knowledge of statistical tests like the ADF test and the Granger Causality test, which were essential in analyzing agricultural data and understanding the factors influencing growing seasons and vegetation patterns. Undertaking this project for 4 credit hours, it was an unpaid research opportunity supervised by Dr. Gopichandh Danala at DISC, who can be contacted at danala@ou.edu.

In terms of data science and analytics skills, the project demanded proficiency in time series analysis, data visualization, and statistical modeling. Additionally, I had to independently learn about specific techniques and datasets relevant to agricultural analysis. Overall, this project provided a valuable opportunity to apply my academic knowledge to practical scenarios while further developing my skills in data analysis and interpretation.