Missing Data Imputation for GRACE Data Derived from Terrestrial Water Storage (TWS): Statistical and Machine Learning Approaches

Jenna Weldon and Buruk Yimesgen
Advisor: Tamer Elbayoumi
Mentor: Brett Hunter
July 2024

Abstract

This study uses statistical and machine learning techniques to impute missing data from GRACE (Gravity Recovery and Climate Experiment) terrestrial water storage (TWS) datasets. We examined information from 2 of the 62 watersheds across the globe, taking into account variables such as temperature, precipitation, evapotranspiration, NDVI, and ENSO. Creating models to predict TWS and efficiently impute missing values was our aim. We investigated techniques including Extreme Gradient Boosting (XGBoost), Deep Neural Networks (DNN), Generalized Linear Models (GLM), and mean/median imputation. Mean Squared Error (MSE) was used to assess these models' performance to choose the most effective method for precise forecasting and imputation.

Contents

1 Introduction **3**
1.1 Backround Information . . . 3

2 Data Exploration **6**
2.1 About the Factors . . . 7
2.1.1 Terrestrial Water Storage . . . 7
2.1.2 Temperature . . . 8
2.1.3 Evapotranspiration . . . 8
2.1.4 Precipitation . . . 9
2.1.5 Normalized Difference Vegetation Index . . . 9
2.1.6 El Niño/Southern Oscillation . . . 10

3 Methods **11**
3.1 Mean/ Median imputations . . . 12
3.2 Generalized Linear Model (GLM) . . . 13
3.2.1 Brief Description . . . 13
3.2.2 Background Information . . . 13
3.2.3 Structure . . . 13
3.3 Deep Neural Network(DNN) . . . 14
3.3.1 Brief Description . . . 14
3.3.2 Background Information . . . 14
3.3.3 Structure of DNN . . . 15
3.3.4 Elements of DNN . . . 15
3.4 Extreme Gradient Boosting (XGBoost) . . . 16
3.4.1 Brief Description . . . 16
3.4.2 Background Information . . . 17
3.4.3 Structure . . . 17

4 Performance metrics **18**
4.1 Mean Squared Error (MSE) . . . 18

5 Results **21**
5.1 Confidence Intervals . . . 27

6 Conclusion **28**

7 Future Research **29**

1 Introduction

1.1 Backround Information

The Gravity Recovery and Climate Experiment (GRACE), launched on March 17, 2002, consisted of twin satellites that revolutionized our ability to measure Earth's gravity field with unparalleled precision. This mission significantly enhanced our understanding of Earth's dynamic water reservoirs, including those on land, within ice masses, and throughout the oceans. GRACE was the result of a collaborative effort between NASA and the German Aerospace Center (DLR), with critical contributions from the University of Texas Center for Space Research (CSR), the GeoForschungsZentrum (GFZ) Potsdam, and the Jet Propulsion Laboratory (JPL) [6].

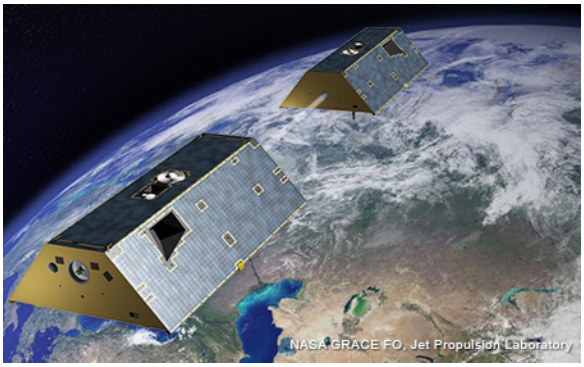

Figure 1: The twin GRACE-FO satellites will follow each other in orbit around the Earth, separated by about 137 miles (220 km). [12]

The primary objective of GRACE was to measure gravity variations by monitoring the distance between the two satellites. These satellites maintained a separation of approximately 220 km [14]. As they orbited the Earth, regions with stronger gravitational pull caused the leading satellite to accelerate slightly, thereby increasing the distance between the two satellites. Conversely, when encountering a region with weaker gravitational pull, the distance between the satellites decreased. These minute changes in distance, which were less than the width of a human hair, could be detected by GRACE's highly sensitive instruments, enabling the detection of changes in relative velocity as small as a micrometer per second [15, 16].

When combined with other data sources and sophisticated models, these detailed gravity measurements provided valuable insights into terrestrial water storage changes, ice-mass variations, ocean bottom pressure changes, and sea-level fluctuations. For example, GRACE data has been used to monitor groundwater depletion in major aquifers, observe ice loss from polar ice sheets and glaciers, and measure the rise in sea levels due to melting ice and thermal expansion of seawater.

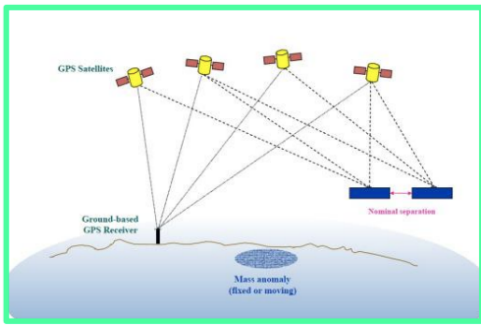

Figure 2: *normal separation [4]*

The follow-on mission, GRACE-FO (Follow-On), launched on May 22, 2018, continues the groundbreaking work of its predecessor. GRACE-FO tracks the movement of Earth's water to monitor changes in underground water storage, the water levels of large lakes and rivers, soil moisture content, the mass of ice sheets and glaciers, and sea level variations. These observations offer a unique and comprehensive view of Earth's climate system, providing crucial data that benefits society by informing water management, agriculture, and climate change adaptation strategies.

The satellites in the GRACE-FO mission operate using the same principle as the original GRACE mission. As they orbit Earth, areas with greater mass concentrations (and thus stronger gravity) affect the lead satellite first, causing it to speed up and increase the distance from the trailing satellite. As the trailing satellite passes over the same gravity anomaly, it is pulled toward the lead satellite, decreasing the distance between them. These changes, though imperceptible to the naked eye, are measured with extreme precision using a microwave ranging system.

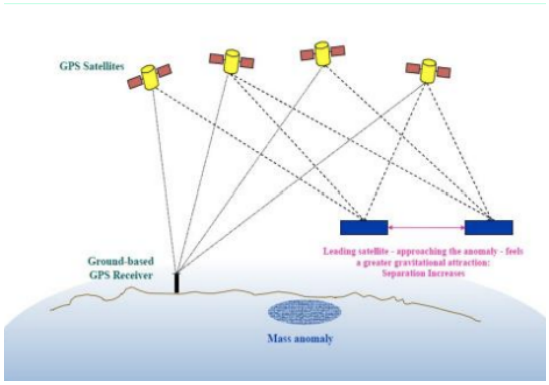

Figure 3: *Separation Increased. As it gets closer to the anomaly, the satellite (A) experiences a stronger gravitational pull.*

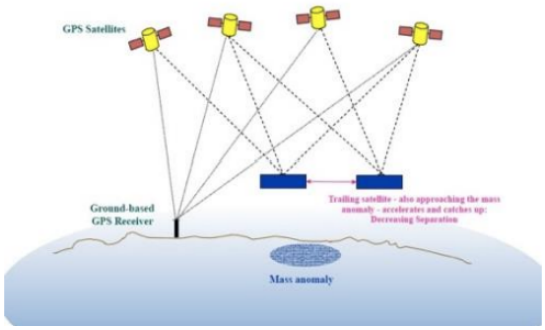

Figure 4: Separation decreased. The satellite (A) not affected by the anomaly while (B) tugged backward by the anomaly.

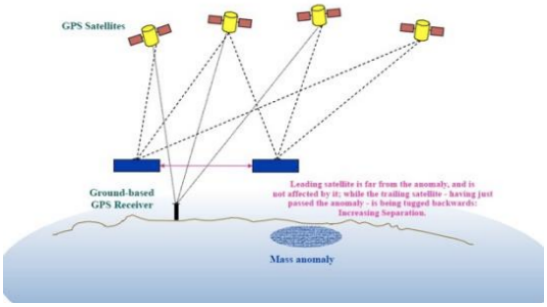

Figure 5: Separation Increased. The satellite (A) not affected by the anomaly while (B) tugged backward by the anomaly.

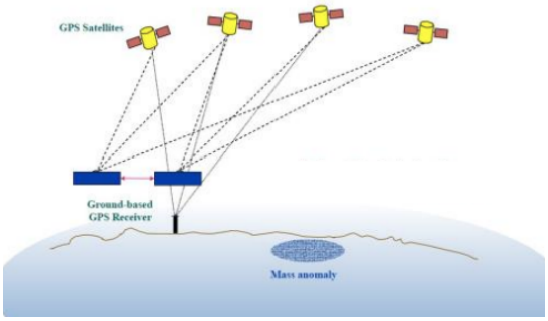

Figure 6: Normal Separation. The satellite (B) leaves the anomaly catches up.

To ensure that only gravitational accelerations are considered, highly accurate accelerometers located

at each satellite's center of mass measure non-gravitational forces such as atmospheric drag and solar radiation pressure. Additionally, GPS receivers on the satellites determine their exact positions over the Earth's surface to within a centimeter, providing a robust framework for interpreting the gravity data.

Overall, the GRACE and GRACE-FO missions have provided invaluable data that enhance our understanding of the Earth's hydrological and climatic processes, offering insights that are essential for managing natural resources and addressing the challenges posed by climate change.
The available GRACE data we have for this project consists of data from 62 watersheds around the world from April 2002 to April 2020. The following map shows the 62 watershed locations around the world. The numbers from 0 to 61 are IDs for these basins, and their names are all listed on a spreadsheet.

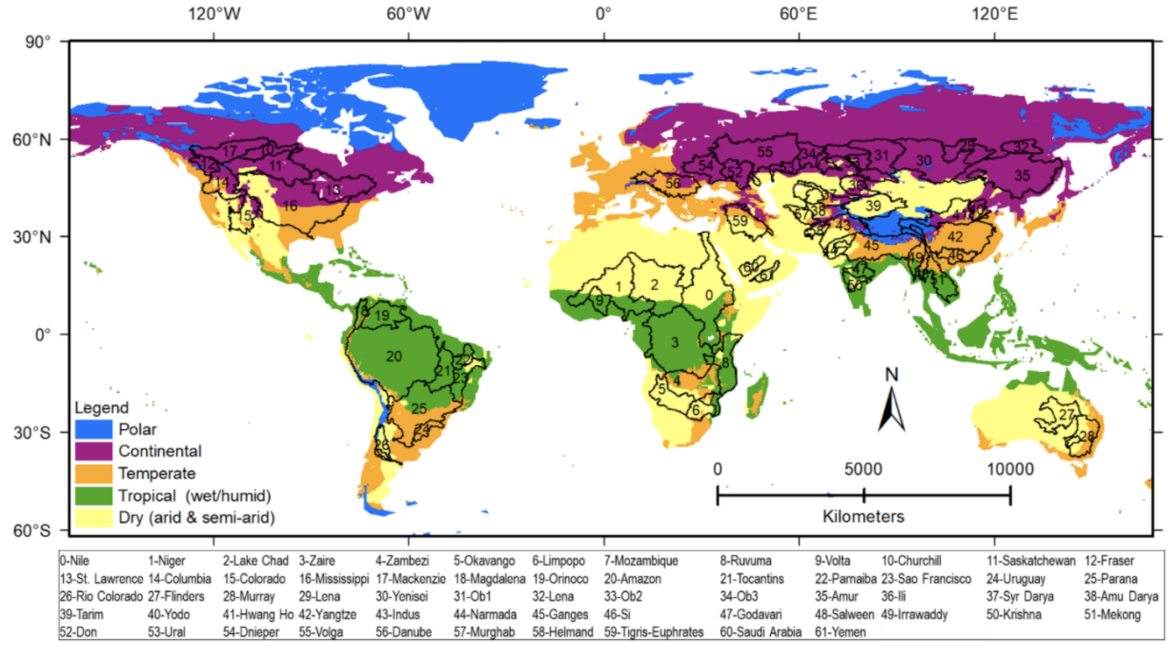

Figure 7: Spatial distribution of the 62 global watersheds (names are shown at the bottom) examined. [8]

GRACE data has some missing values in some months because sometimes the batteries of the satellites die or they turn them off to conserve battery. Our goal for this project was to generate relationships or models that can describe the GRACE-TWS data with the other factors. We each chose data from one of the 62 basins and analyzed the relationships between GRACE data and the other factors seeking to find a statistical/mathematical model that can explain these relations and use it in forecasting GRACE data in the future. From there, we wanted to fill in the missing values in GRACE data, analyze the performance of the models, and find the best model that can impute the missing values in GRACE data. We want to use a statistical model to test the data we already have to get the best model, and with some confidence, we can impute and predict the missing values and reduce bias from the missing values. To find the best model we used a tool called the Mean Squared error (MSE) and the goal of this was to

get the MSE as small as we could. MSE is the square of the averages of all differences and we need to minimize the MSE. We need to see how the MSE is changing from these factors, impute missing values for each model, and use that to find MSE. Our goal was to find one with the lowest MSE by removing missing values, imputing the mean into the missing values, and imputing the median into the missing values. We needed to train each model in these three different ways, predict the missing data points, then graph the predicted model and the actual and see how accurate we could get the model to be. After we tested these models, we imputed only the missing values into the original GRACE and created a data frame with each of the models imputed into the missing values. We then tested those columns of data with the original models we were using to see if we could get more accurate results.

2 Data Exploration

To explore the data, we each picked one of the basins to work with and used those data sets. We chose Sao Francisco, which is on the eastern coast of South America and has a tropical climate. We can see wet and dry periods throughout the years on the time series graph representation of GRACE data. We also chose Saudi Arabia, which is in Asia and has a dry climate. The time series graph shows a decline in GRACE throughout the years, because of the lack of water. We chose these because they are two very different datasets, as Sao Francisco's values are much higher than Saudi Arabia's. Also, they have different climates, as Sao Francisco is tropical and Saudi Arabia is dry. Also, Saudi Arabia has an interesting downhill slope throughout the years, and Sao Francisco has more periods of wet and dry, which we found interesting. We wanted to test two basins that were very different from each other, and we thought that these were two good basins to use for that. We started exploring these two basins by first looking at the data and seeing a basic summary of our datasets.

DATE	ID	Basin	GRACE	TWS	TMP	ET	NDVI	PRECIP	ENSO	CLIMATE
2003-01-01	61	Saudi Arabia	3.846688198	6.29678470	289	0.0284258970	0.944322860	1.607325880	0.8	Dry
2004-01-01	61	Saudi Arabia	2.555154679	4.80348921	291	0.0827683179	0.994255860	1.195745925	0.2	Dry
2005-01-01	61	Saudi Arabia	2.112645230	7.2588409	288	0.143883760	0.949868025	1.373506602	0.1	Dry
2006-01-01	61	Saudi Arabia	2.222751863	8.0633419	289	0.005115271	0.919209201	1.024232987	-0.7	Dry
2007-01-01	61	Saudi Arabia	2.459965223	8.2610403	287	0.186118240	0.928914607	1.079696011	0.6	Dry
2008-01-01	61	Saudi Arabia	2.021570665	8.2426926	288	0.099846000	0.938264580	1.026703719	-1.1	Dry
2009-01-01	61	Saudi Arabia	1.565431013	7.260947510	288	0.119646048	0.933368489	1.476662631	-1.0	Dry
2010-01-01	61	Saudi Arabia	2.247593788	5.87855714	289	0.049902916	0.945792507	1.195526474	0.9	Dry

Table 1: Saudi Arabia Data

DATE	ID	Basin	GRACE	TWS	TMP	ET	NDVI	PRECIP	ENSO	CLIMATE
4/1/2002	24	Sao Francisco	0.487643375	25.8119163	298	0.3638507	0.617807150	0.04625925	-0.4	Tropical
5/1/2002	24	Sao Francisco	2.717365464	26.843415	297	0.1714162	0.553474540	0.04029183	-0.1	Tropical
6/1/2002	24	Sao Francisco	NA	47.2835561	296	0.0675631	0.500041200	0.01876504	0.4	Tropical
7/1/2002	24	Sao Francisco	NA	29.0389784	296	0.4116829	0.444848960	0.017469506	0.4	Tropical
8/1/2002	24	Sao Francisco	14.44029374	13.5652343	297	1.9894260	0.399988149	0.01056410	1.0	Tropical
9/1/2002	24	Sao Francisco	16.18749451	12.631923	298	2.5444534	0.37628994	0.01709035	0.8	Tropical
10/1/2002	24	Sao Francisco	17.58745024	16.9636364	300	0.1858459	0.38621027	0.03664697	0.8	Tropical
11/1/2002	24	Sao Francisco	16.08335874	57.302009	299	0.8545271	0.470595440	0.14537186	0.8	Tropical

Table 2: Sao Francisco Data

Above and below there is a small sample of the data structure used in our project. This sample represents the first 8 rows of the datasets from both the Sao Francisco and Saudi Arabia Basins, showcasing key variables and their respective values.

```
     DATE                ID          Basin              GRACE          
 Length:218         Min.   :61   Length:218         Min.   :-7.7828  
 Class :character   1st Qu.:61   Class :character   1st Qu.:-4.9035  
 Mode  :character   Median :61   Mode  :character   Median :-2.7264  
                    Mean   :61                      Mean   :-2.2995  
                    3rd Qu.:61                      3rd Qu.: 0.2932  
                    Max.   :61                      Max.   : 2.9247  
                                                    NA's   :33       
      TWS              TMP               ET                NDVI        
 Min.   :354.1   Min.   :287.0   Min.   : 0.004164   Min.   :0.09111  
 1st Qu.:365.2   1st Qu.:293.0   1st Qu.: 0.043336   1st Qu.:0.09436  
 Median :376.9   Median :299.5   Median : 0.332676   Median :0.09579  
 Mean   :377.6   Mean   :299.1   Mean   : 1.761101   Mean   :0.09590  
 3rd Qu.:387.8   3rd Qu.:305.0   3rd Qu.: 2.750572   3rd Qu.:0.09716  
 Max.   :411.1   Max.   :308.0   Max.   :14.165358   Max.   :0.10556  

     PRECIP              ENSO          CLIMATE         
 Min.   :0.0000793   Min.   :-2.4000   Length:218        
 1st Qu.:0.0012451   1st Qu.:-0.7000   Class :character  
 Median :0.0041649   Median :-0.1000   Mode  :character  
 Mean   :0.0092072   Mean   :-0.1445                     
 3rd Qu.:0.0106896   3rd Qu.: 0.4000                     
 Max.   :0.1237223   Max.   : 2.2000                     
```

Figure 8: Saudi Arabia 5 Number Summaries

```
     DATE                ID          Basin              GRACE          
 Length:218         Min.   :24   Length:218         Min.   :-44.5252 
 Class :character   1st Qu.:24   Class :character   1st Qu.:-19.5722 
 Mode  :character   Median :24   Mode  :character   Median : -8.5002 
                    Mean   :24                      Mean   : -9.9903 
                    3rd Qu.:24                      3rd Qu.:  0.3123 
                    Max.   :24                      Max.   : 18.9597 
                                                    NA's   :33       
      TWS              TMP               ET                NDVI        
 Min.   :381.0   Min.   :294.0   Min.   : 40.69   Min.   :0.3354   
 1st Qu.:426.8   1st Qu.:297.0   1st Qu.: 60.60   1st Qu.:0.4392   
 Median :475.7   Median :298.0   Median : 78.38   Median :0.5422   
 Mean   :483.0   Mean   :297.6   Mean   : 76.78   Mean   :0.5263   
 3rd Qu.:535.2   3rd Qu.:299.0   3rd Qu.: 92.92   3rd Qu.:0.6145   
 Max.   :669.7   Max.   :301.0   Max.   :115.10   Max.   :0.6999   

     PRECIP              ENSO          CLIMATE         
 Min.   :0.00362   Min.   :-2.4000   Length:218        
 1st Qu.:0.02215   1st Qu.:-0.7000   Class :character  
 Median :0.06795   Median :-0.1000   Mode  :character  
 Mean   :0.11109   Mean   :-0.1445                     
 3rd Qu.:0.17891   3rd Qu.: 0.4000                     
 Max.   :0.54577   Max.   : 2.2000                     
```

Figure 9: Sao Francisco 5 Number Summaries

We then began creating different graphs to see which models we thought would be best. We first made graphs of time series of all of the factors to see how these changed over time. The six factors that we also looked at were the different features that we used when testing different models, and this was the X variable or the characteristics used to help train our data. The Y variable is the target or dependent variable, which is GRACE, and the Y variable is the target of what we are trying to predict and impute.

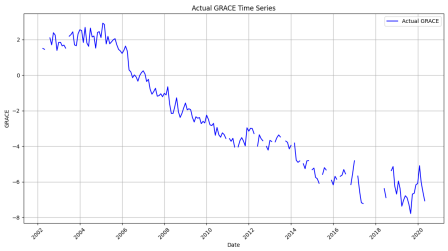

Figure 10: Saudi Arabia GRACE

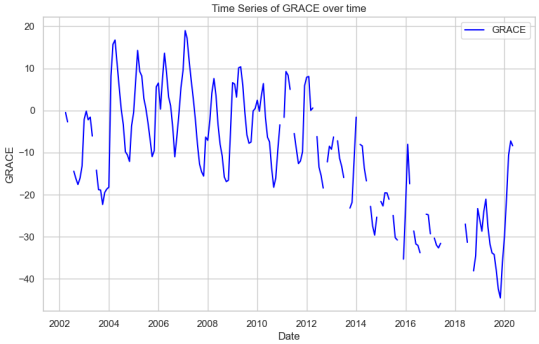

Figure 11: Sao Francisco GRACE

2.1 About the Factors

We also looked at exogenous factors and explanations on how to read the data. We researched more about these factors to learn how they affect and correlate with GRACE. We also looked at each of these factors over time to see how the factor changed over time, and to see if it impacted how GRACE changed over time.

2.1.1 Terrestrial Water Storage

Terrestrial water storage (TWS) is the sum of all water on the land surface and in the subsurface. It is measured to understand the water balance and its changes over time due to natural processes and

human activities. It is used for satellite remote sensing and these changes have been observed by GRACE since 2002. Since GRACE uses gravity to infer changes in TWS and measures TWS, it makes it a very crucial factor.

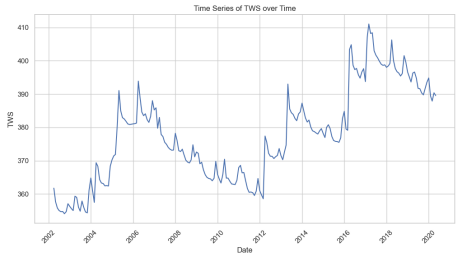

Figure 12: Saudi Arabia TWS

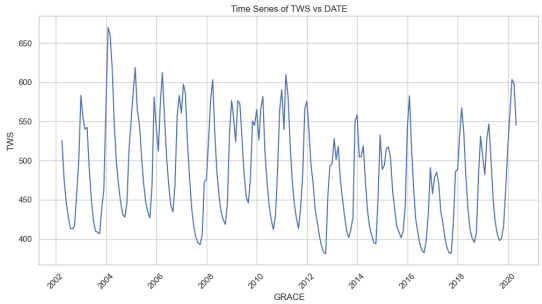

Figure 13: Sao Francisco TWS

2.1.2 Temperature

Temperature (TMP) is another factor of GRACE because it is an important part of the Earth's climate system. Temperature can influence GRACE because of seasonal changes, ocean temperatures and currents, snow and ice melt, and many other changes in temperature that can impact the distribution of water on Earth.

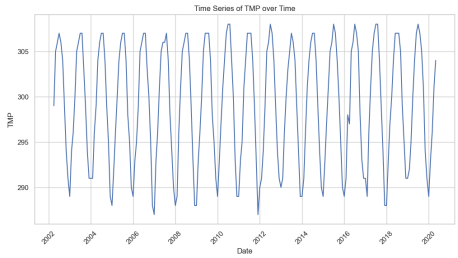

Figure 14: Saudi Arabia Temperature

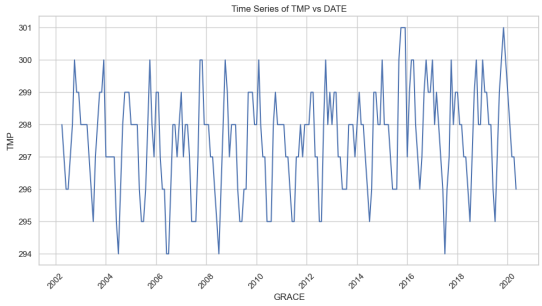

Figure 15: Sao Francisco Temperature

2.1.3 Evapotranspiration

Evapotranspiration (ET) is the process of water being transferred from the land to the atmosphere by evaporation from the soil and other surfaces and by transpiration from plants. GRACE does not directly measure evapotranspiration, but evapotranspiration can provide us with critical data on TWS that allow for the estimation and analysis of evapotranspiration.

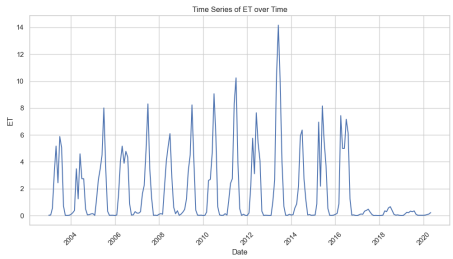

Figure 16: Saudi Arabia ET

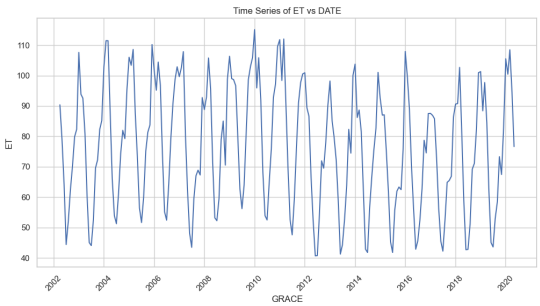

Figure 17: Sao Francisco ET

2.1.4 Precipitation

Precipitation is another crucial factor of GRACE because it impacts the distribution and movement of water on and below the surface of the Earth. Precipitation can increase the surface water. Rain fall and snow melts add to surface water and causes changes in GRACE. Also, other factors can influence GRACE like groundwater levels, soil moisture, and river flow. These are all variations in the gravitational field, which can affect GRACE data in many ways.

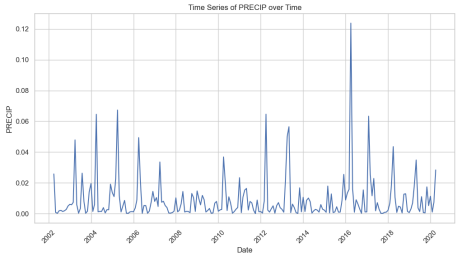

Figure 18: Saudi Arabia Precipitation

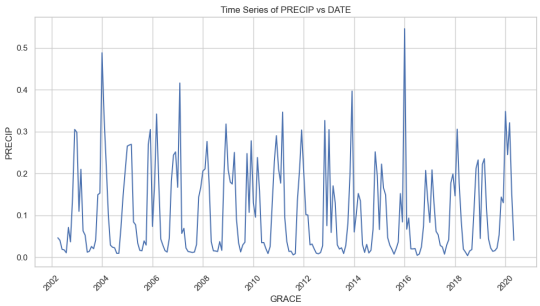

Figure 19: Sao Francisco Precipitation

2.1.5 Normalized Difference Vegetation Index

Normalized difference vegetation index (NDVI) is the measurement of surface vegetation coverage activity. NDVI and GRACE data provide information that when put together can better our understanding of the interactions between terrestrial water storage and vegetation dynamics.

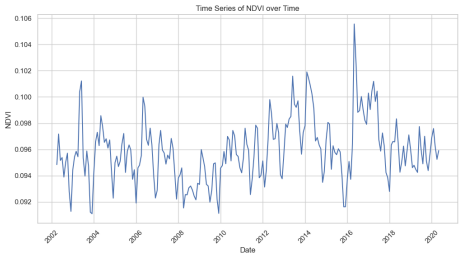

Figure 20: Saudi Arabia NDVI

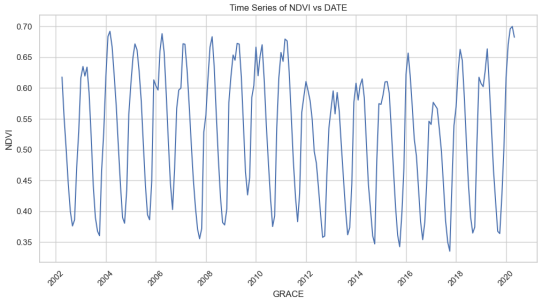

Figure 21: Sao Francisco NDVI

2.1.6 El Niño/Southern Oscillation

El Niño/Southern Oscillation (ENSO) is a warming of the ocean surface, or above-average sea surface temperatures (SST), in the central and eastern tropical Pacific Ocean. It is a climate phenomenon characterized by periodic fluctuations in sea surface temperatures and atmospheric pressure in the equatorial Pacific Ocean, and it causes some regions to face more rainfall every so often. It is a critical factor influencing global climate patterns, which affects the distribution and movement of water and ice on Earth, and GRACE's ability to measure changes in Earth's gravitational field allows scientists to observe and study these effects, which can help us learn about the connection between climate phenomena and the water on Earth and its processes.

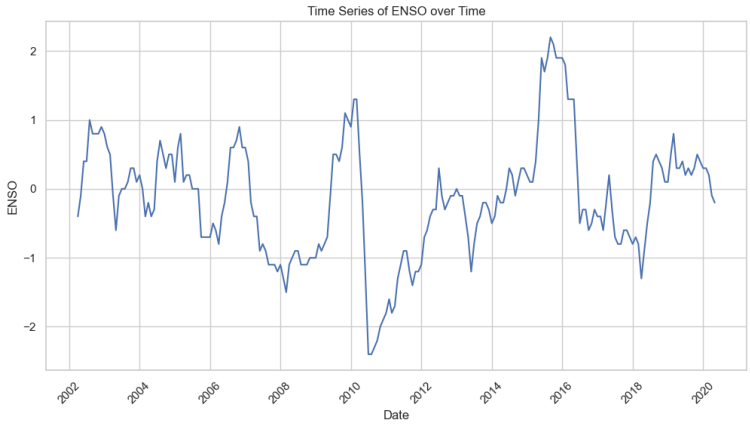

Figure 22: ENSO time series (ENSO is always the same across the globe meaning there is no need for two graphs)

We also made time series graphs for a distribution of where all of the missing values are. the red is where the data is missing notice that both basins have the same missing dates as do all the basins.

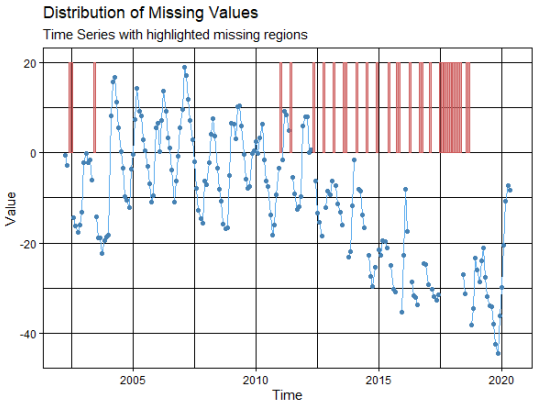

Figure 23: Sao Francisco missing data

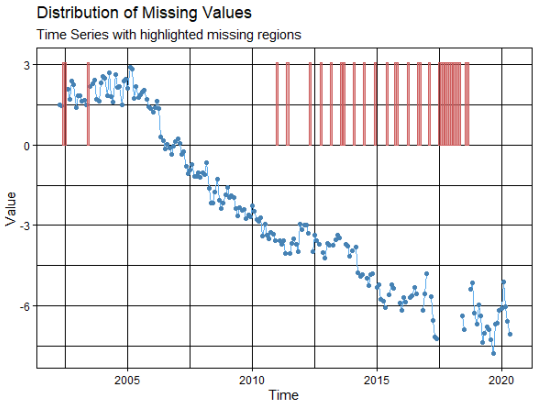

Figure 24: Saudi Arabia missing data

Every model used for data imputation, except mean/median imputation, followed the structure of the following equation. The GRACE value at a specific time (t) is represented as a combination of all the exogenous factors and an error term (ϵ) all at t.

$$GRACE_t = TWS_t + TMP_t + ET_t + ENSO_t + NDVI_t + PRECIP_t + \epsilon_t \tag{1}$$

Next, we tested to see if the graphs had any linear correlation to any of the other GRACE factors, and to do this we created correlation matrices and scatterplots for each factor compared with GRACE, and we did this for each of our basins. If the correlation is close to 1 in the matrix, we could assume that the two factors are correlated.

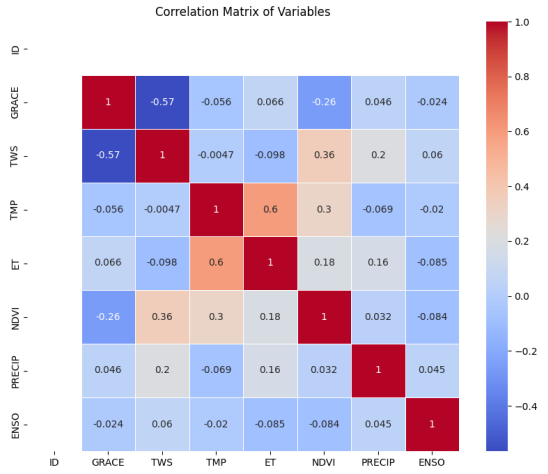

Figure 25: Saudi Arabia Correlation Matrix

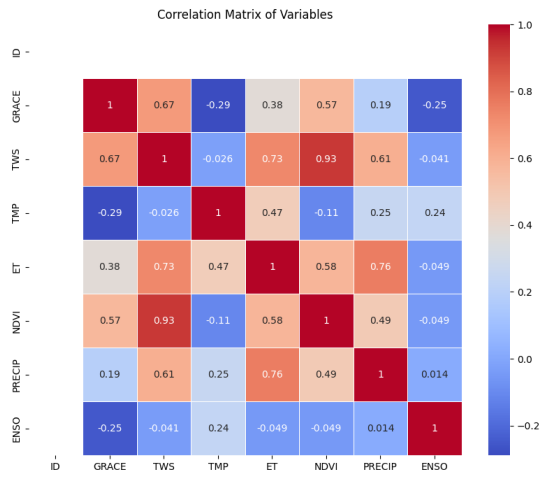

Figure 26: Sao Francisco Correlation Matrix

From these graphs, we concluded that none of the factors were linearly correlated to GRACE data and that every factor is independent. Also, since we know that it is not linear, we can also conclude that GRACE is not normally distributed. Multi-Linear Regression (MLR) uses the assumption that each of the factors have no correlation between the factors. There are a few factors with a high correlation; thus the use of MLR as one of our models is not ideal. To use MLR the distribution of the dependent variable, GRACE, would have to be normal, and we discovered that it is not for either of our basins, so we know that we cannot use MLR. From here we determined which methods and models would be the best approach for us.

3 Methods

In the process of addressing missing data in time series analysis, selecting the appropriate imputation methods is crucial for maintaining the integrity and predictive power of the dataset. We have chosen to explore mean/median imputation, Generalized Linear Models (GLM), Deep Neural Networks (DNN), and XGBoost due to their distinct advantages and varying complexity. Mean and median imputation

serve as straightforward, yet effective baseline methods, providing a simple way to handle missing values without introducing significant bias. GLM offers a statistical approach that leverages relationships within the data to predict missing values, making it suitable for linear dependencies. We chose GLM because it is one of the most basic machine learning models, and before we tested this we thought GLM would not work as well in Saudi Arabia as it would in Sao Francisco, because in the Saudi Arabia graph, it is not symmetrical we do not see set periods and cycles. Since the graph goes downhill, we thought there would not be a good enough pattern to obtain a good GLM with it. In Sao Francisco, since the graph somewhat travels up and down in cycles more, we thought we could obtain better results with that data set. DNNs, with their ability to capture complex, non-linear patterns through multiple layers of abstraction, are well-suited for handling intricate temporal dependencies in the data. XGBoost, a powerful ensemble learning method, combines the strengths of multiple decision trees to enhance prediction accuracy and robustness. We thought that both DNN and XGBoost would work well because we thought that their complexity and hyperparameters would allow the imputations to be close to the actual GRACE values. By comparing these diverse techniques, we aim to assess their effectiveness and identify the most reliable method for imputing missing data in time series, ultimately improving the overall quality and utility of the dataset.

3.1 Mean/ Median imputations

When given a data set with missing data mean and median imputations are two simple straightforward methods that are used a lot in different situations. Mean imputation replaces missing values with the mean of the non-missing data, while median imputation is the same but with the median of the non-missing data. The biggest advantage of this method is the simplicity and ease of implementation. With both, the imputation does not affect the mean/median of the data which can be very helpful with some data sets. These imputations are also computationally efficient, making them suitable for extensive data sets where more complex methods might be out of the picture.

However, mean/median imputation has several disadvantages. One significant problem is that because the imputed values are less variable than the original data points, they lessen the dataset's variance. This decrease in variance may cause inferences to be made that are potentially misleading and an underestimation of variability. Furthermore, both approaches disregard the correlations across variables, which may lead to skewed approximations, particularly in cases where the data is not entirely missing completely at random (MCAR). Since they do not consider other factors, it is usually not accurate enough to use as a model. Another issue is that they may skew the distribution in datasets containing notable outliers. While median imputation might not be able to represent the subtleties of multimodal distributions accurately, mean imputation tends to do so. These distortions can skew findings and judgments by compromising the validity of any further study.

3.2 Generalized Linear Model (GLM)

3.2.1 Brief Description

Generalized Linear Models (GLMs) are an extension of traditional linear regression models that allow for modeling response variables with various distributions. Unlike ordinary linear regression, which assumes that the response variable is normally distributed and the relationship between the response and predictors is linear, GLMs can handle different distributions such as binomial, Poisson, and gamma distributions.

3.2.2 Background Information

The concept of GLM evolved by many researchers for much time, but the term was officially introduced by John Nelder and Robert Wedderburn (1972) and they were able to apply statistical knowledge in a way that shaped how statisticians think about generalizations of linear models. They were able to group different distributions as members of the exponential family, apply the least squares algorithm to the family, introduce the terminology "generalized linear models", and suggest that this unification would be a huge improvement and help simplify this topic. Later, the first available software to implement GLMs was the Fortran based GLIM system which was developed by the Royal Statistical Society's Working Party on Statistical Computing, released in 1974, and developed through 1993. [17]

3.2.3 Structure

The flexibility of being able to use this model with distributions that are not normal is achieved through three components: a linear predictor, a link function, and an error structure. [18]

The linear predictor is a combination of the input variables and their coefficients. The linear predictor is a key component that puts together the explanatory variables and their corresponding coefficients (parameters) into a single value, plus an intercept term β_0. This value is then used in the link function to predict the response variable. It is mathematically expressed as:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

The link function relates the mean of the response variable to the linear predictor, enabling the model to handle non-linear relationships. The variance function states how the variance of the response variable depends on its mean. By including these elements, GLMs give a framework for modeling a wide range of data types and relationships, making them a crucial tool in statistical modeling. Also, every error structure has a link function associated with it, and in glm(), the family argument specifies the error distribution and link function.

The error structure tells us how the errors are distributed. There are different types of error structures, for example, count data, shown as integers with a minimum value of zero, have Poisson-distributed errors. Proportions, which range from zero to one, have binomial-distributed errors. Our data has Gaussian-distributed errors since it is the most straightforward assumption for continuous data.

3.3 Deep Neural Network(DNN)

3.3.1 Brief Description

DNN structure is a more complex type of Artificial Neural Network (ANN), and it automatically extracts deeper and more complicated relationships between model inputs and outputs. [8] A DNN is a neural network with a significant depth, meaning it has more than one hidden layer between the input and output. These hidden layers allow the network to learn complex patterns in data by gradually extracting higher-level features from raw input. The DNN is composed of interconnected organized layered neurons. The first layer of neurons is called the input node. It receives information from input data and combines and passes the data to the next layer through transformation. The last layer in the network is called the output layer and produces the output of the model. The layers between input and output are called hidden layers, and more hidden layers lead to greater depth, allowing the network to find more complicated patterns.

3.3.2 Background Information

The concept of neural networks dates back to the 1940s when Warren S. McCulloch and Walter Pitts created the first computational model of a neuron inspired by the human brain. [5] Then, the perceptron, introduced by Frank Rosenblatt in 1958, is one of the earliest types of neural networks. The backpropagation algorithm was originally introduced in the 1970s, but its importance wasn't fully seen until a 1986 paper by David Rumelhart, Geoffrey Hinton, and Ronald Williams explained several neural networks where backpropagation works more efficiently than earlier approaches to learning. [13] This made it possible to use neural networks to solve problems that could not be solved previously. Back-propagation allowed for the training of multi-layer perceptrons (MLPs), which could handle nonlinear problems. Later, the term "deep learning" became more known in the 2000s and 2010s as researchers like Geoffrey Hinton, Yoshua Bengio, and Yann LeCun demonstrated the capability of deep neural networks with multiple hidden layers.

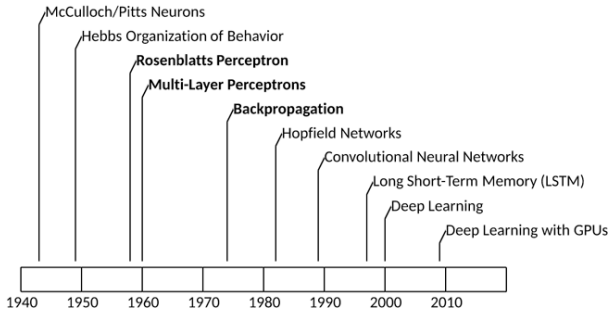

Figure 27: This figure shows a sequence of the history of DNN. [5]

3.3.3 Structure of DNN

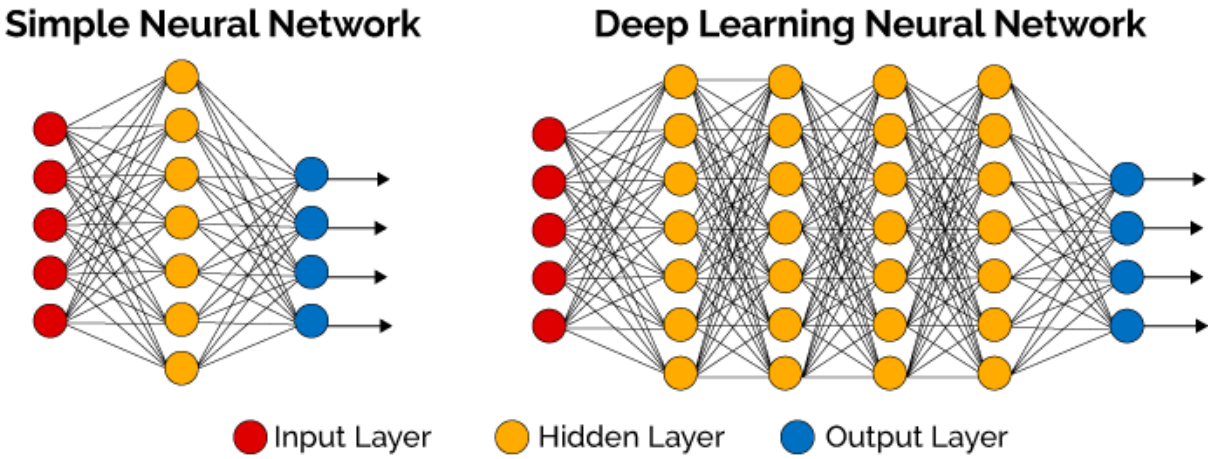

Figure 28: Above is the structure of a neural network compared to a deep neural network. Notice how the DNN contains more hidden layers in between the input and output layers. [2]

Introduction to Deep Neural Networks (DNN)

Deep Neural Networks (DNN) are machine learning models that consist of multiple layers of interconnected neurons that transform input data to produce predictions. DNNs can have different amounts of layers and neurons to predict the target variable and DNNs learn from data and make accurate predictions.

3.3.4 Elements of DNN

DNNs are used for tasks where a target variable is predicted from input features. Important elements are the model and parameters. The model makes predictions using the input features, which are the six factors. It also includes parameters (weights and biases) learned from training data. These parameters depend on the task, such as regression or classification. The structure also includes the objective function, which entails training loss and regularization. Training loss measures prediction accuracy, while regularization deals with more complex models to prevent overfitting, balancing model accuracy, and generalization (bias-variance trade-off).

DNNs also use a series of layers, including an input layer, multiple hidden layers, and an output layer, to build complex models. The input layer is the first layer in the network that receives the input data and each neuron in the input layer corresponds to one feature of the input data. [22] Hidden layers are intermediate layers between the input and output layers and you can choose how many hidden layers you need depending on what you are trying to do. Hidden layers extract and transform features from the input data through multiple processing steps. A DNN has multiple hidden layers, which allows it to learn hierarchical feature representations. The output layer is the final layer in the network that produces the output and generates predictions based on the transformed features from the hidden layers. The number of neurons typically matches the number of prediction targets.

Neurons are the basic units of a neural network and each neuron receives input, processes it through a weighted sum, applies an activation function, and outputs the result to the next layer. A DNN model transforms input data through layers using learned weights and biases: $\hat{y} = f_n(...f_2(f_1(x)))$ where f_i represents each layer's transformation. [1]

The training process also includes forward pass, which passes through the input through the network so that it can compute the output, loss calculation, which includes comparing the output to the target values using a loss function, backward pass, which consists of computing gradients of the loss and calculating using backpropagation, and parameter update, which is adjusting the weights and biases using an algorithm including optimization. [21]

Regularization controls the complexity of the model using: $\omega(\mathrm{w}) = \lambda \sum_{i=1}^{n} w_i^2$ where w_i are the weights and λ is the regularization parameter and this can improve the model fit and complexity. Another element, activation functions, allows neural networks to learn and model complex patterns in data, even if the data is not linear. [11]

$$\mathrm{f(z)} = \mathrm{ReLU(z)}$$
$$\mathrm{f(z)} = \sigma(\mathrm{z})$$

where z is the input to the activation function. ReLU is a common activation function that means a rectified linear unit, and it introduces non-linearity into a neural network.

3.4 Extreme Gradient Boosting (XGBoost)

3.4.1 Brief Description

Extreme Gradient Boosting (XGBoost) is an advanced implementation of gradient boosting designed for speed and performance. It is widely used in various machine learning tasks due to its efficiency and scalability. XGBoost builds an ensemble of decision trees sequentially, where each new tree corrects the errors of the previous ones. This iterative process continues until the model achieves optimal performance.

XGBoost's ability to handle large datasets and complex relationships makes it a suitable choice for data imputation tasks.

3.4.2 Background Information

XGBoost was developed by Tianqi Chen [9] and has become a popular choice for data scientists and researchers due to its robust performance in machine learning competitions and real-world applications. The algorithm incorporates several enhancements over traditional gradient boosting methods, such as regularization, which prevents overfitting, and the use of parallel processing, which speeds up computation. XGBoost's regularization techniques, including L1 (Lasso) and L2 (Ridge) regularization, help reduce model complexity and improve generalization.

The XGBoost algorithm works by constructing an ensemble of weak learners, typically decision trees [23]. Each tree is built to minimize the residual errors of the previous trees, gradually improving the model's accuracy. The model uses a learning rate to control the contribution of each tree, preventing the model from fitting the training data too closely and improving its ability to generalize to new data.

3.4.3 Structure

XGBoost, short for "Extreme Gradient Boosting," is based on Gradient Boosting, introduced by Friedman. This method refines the model using supervised learning principles, providing a clear and structured approach to boosted trees [6]. XGBoost is used for tasks where a target variable is predicted from input features. Key elements include the model and parameters, which make predictions based on input features, with parameters (θ) learned from training data. These parameters vary by task, such as regression or classification. The objective function consists of training loss and regularization. Training loss measures prediction accuracy, while regularization penalizes complex models to prevent overfitting, balancing model accuracy and generalization (bias-variance trade-off). XGBoost uses decision tree ensembles, combining multiple trees to enhance performance. A CART model assigns scores to its leaves for various tasks. XGBoost improves accuracy by summing predictions of multiple trees:

$$\hat{y} = \sum_{k=1}^{K} f_k(x)$$

where K is the number of trees, f_k represents each tree, and $\mathcal{F}$ is the space of all CARTs [3].

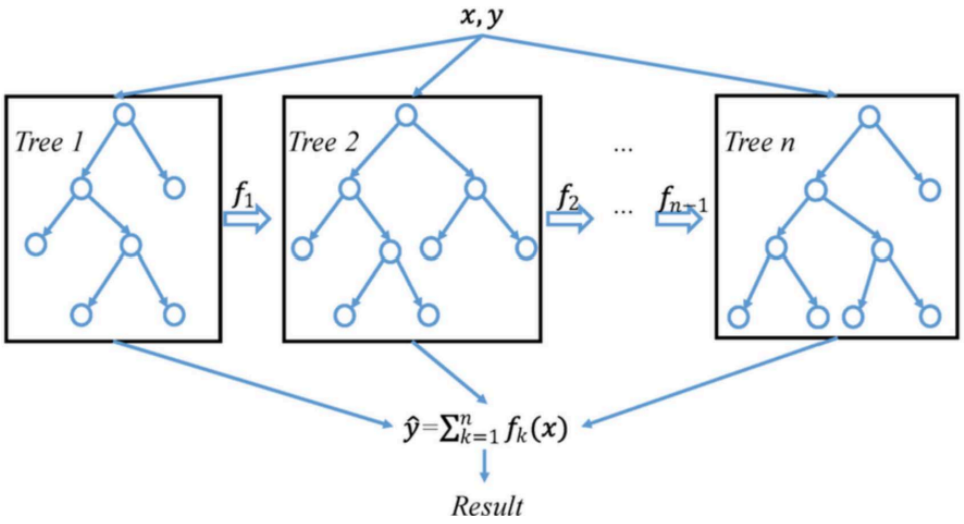

Figure 29: Illustration of the XGBoost process, where an ensemble of trees combines their predictions to produce the final result. Each tree is sequentially added and trained to correct the errors of the previous trees, leading to improved accuracy and robustness. [10]

Training involves optimizing an objective function iteratively. Parameters include tree structures and leaf scores. Trees are added sequentially to improve performance:

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + f_t(x)$$

A second-order Taylor expansion simplifies optimization for general loss functions. Regularization controls tree complexity:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

where T is the number of leaves, w_j are leaf scores, and γ and λ are regularization parameters. This helps balance model fit and complexity. The structure score evaluates tree quality based on data fit and complexity, summing contributions from all leaves. Optimizing tree structure involves splitting leaves to improve the score. Gains from splits are calculated, and insufficient gains are pruned to avoid overfitting. For real-valued data, optimal splits are found by sorting instances and scanning split points efficiently.

4 Performance metrics

4.1 Mean Squared Error (MSE)

To evaluate the models we tested, we measured and calculated for the MSE. The MSE is the error that tells us how far our data points are from the regression line. [19] MSE is a common metric used to evaluate the performance of a regression model. It measures how close a regression line is to a set of data points. The MSE is the variance - bias^2, and it is calculated by taking the average (mean) of errors

squared from data as it relates to a function.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where n is the number of observations, y_i is the actual value and $\hat{y}_i$ is the predicted value. [20] A larger MSE means that the data points are dispersed widely around its central moment (mean), whereas a smaller MSE suggests the opposite. A smaller MSE means there is a smaller error which means the method is a better estimator. We also know that the MSE will always be positive no matter what since the differences are squared, and we know that it is sensitive to outliers because it squares the errors, which would give a larger value to a larger error. Also, hyperparameter tuning can assist in minimizing the MSE because tuning chooses the best hyperparameters for the model with the data we are testing. This attempts to decrease the MSE as much as possible and is an essential tool in lowering the MSE and making our predictions more accurate. In summary, MSE is an important and common metric for testing the accuracy of regression models, and it is a valuable tool in statistical analysis and machine learning.

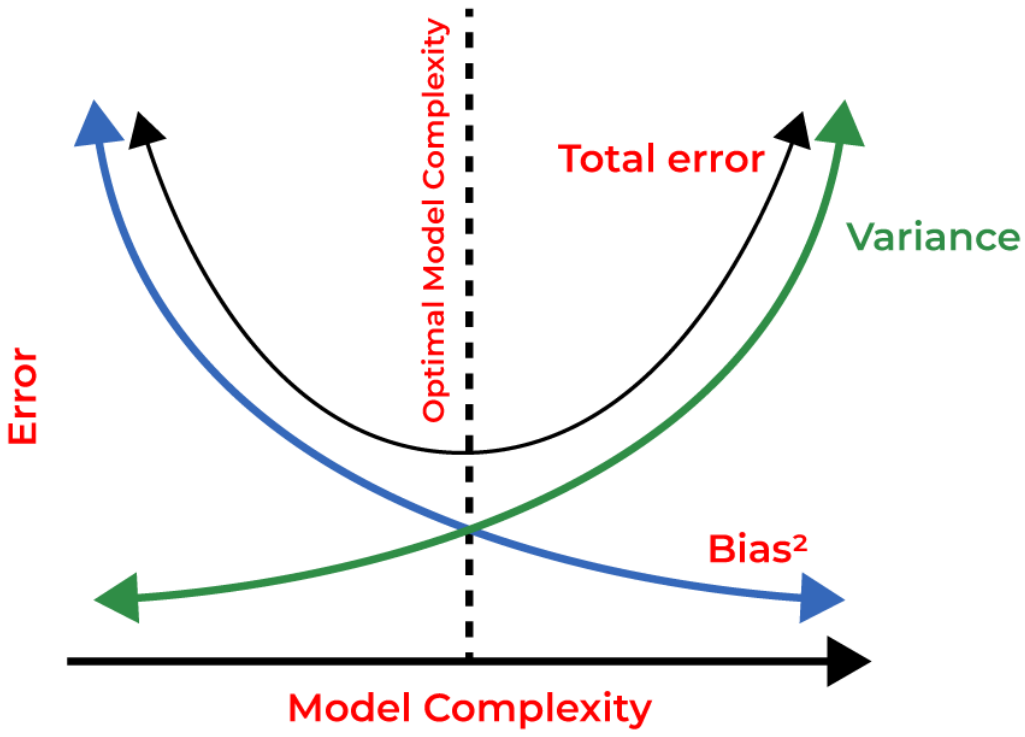

Figure 30: This image shows how to get the optimal model and how to reduce error as much as possible. The graph shows the error versus the model complexity and shows how bias and variance affect the error. [7]

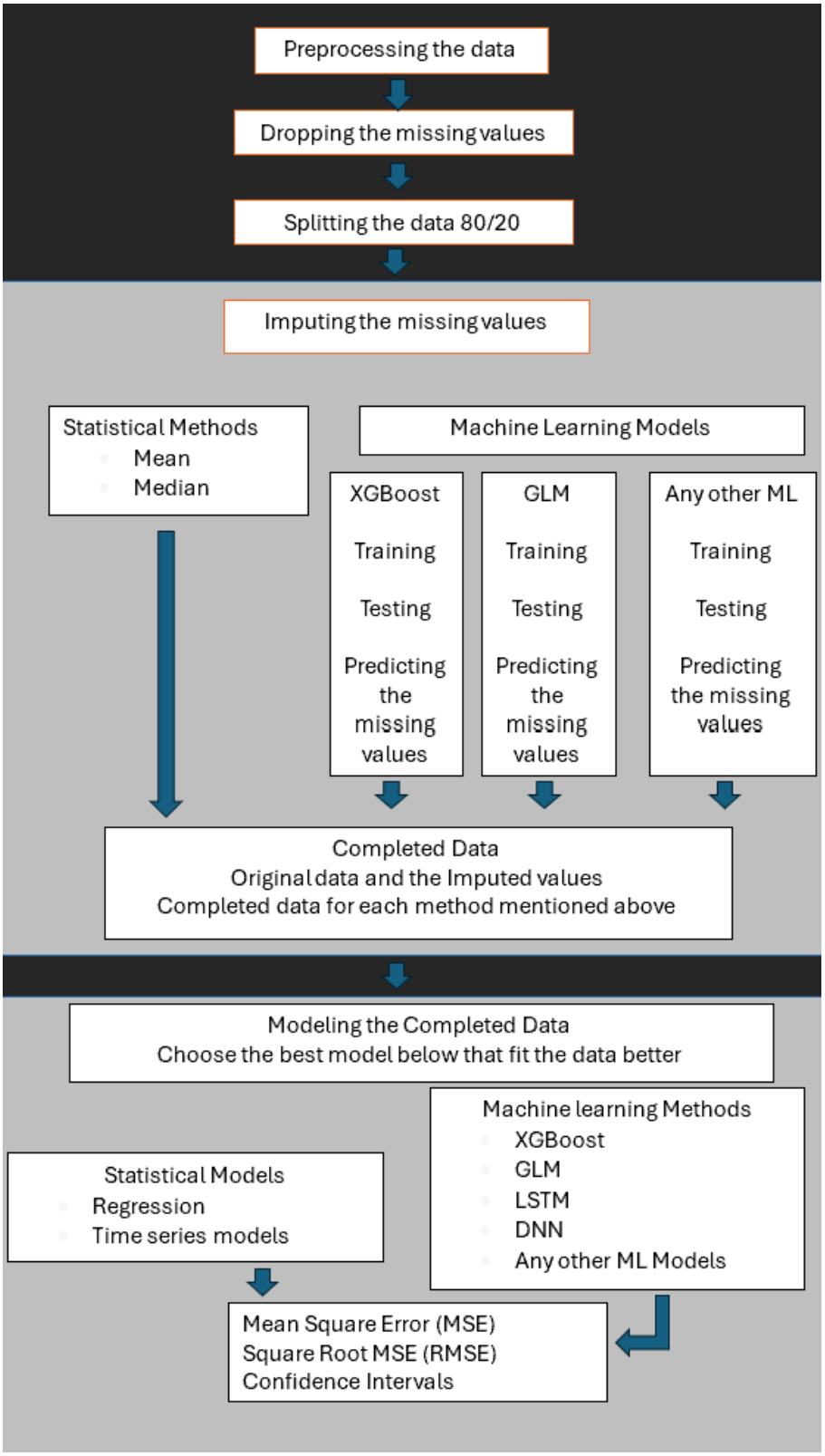

Figure 31: This flowchart shows our steps for analyzing the two basins.

Our goal for this project was to get the lowest MSE we could using different machine-learning models. Analyzing the two basins that we tested, we learned that the MSE for Sao Francisco will be significantly higher since the GRACE values range much more than Saudi Arabia's GRACE values. Saudi Arabia's

values range from approximately -8 to 3, and Sao Francisco's range from approximately -45 to 19. These differences in scale between the two basins will always result in varying MSE outcomes. We also wanted to test how the MSE changes from the exogenous factors we described. We imputed missing values using different models with the factors as our input variables and used the model predictions to find MSE. The goal was to find the model with the lowest MSE because we wanted to decrease it as much as possible.

5 Results

To achieve our results, we first tested these four models that we described above. We obtained the MSE and a graph showing how close the predictions looked to the actual GRACE data for each model and basin. We were able to see which models worked well, and which models did not quite give us ideal results. We also worked on learning about and implementing hyperparameters to some of our models, and we tried to get better results with these hyperparameters. We worked through these models one at a time to see which one would produce the best results.

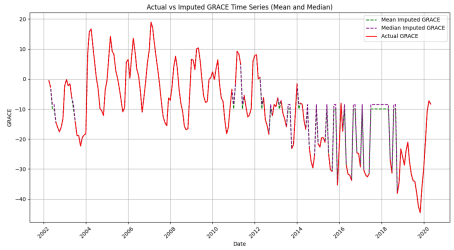

Figure 32: mean/median imputation Sao Francisco

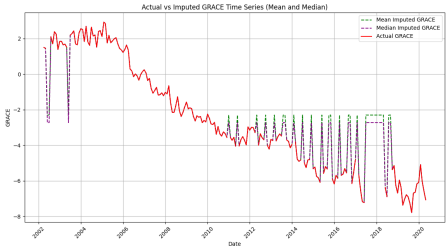

Figure 33: mean/median imputation Saudi Arabia

The GRACE TWS data visualizations for both basins highlight the limitations of mean and median imputation methods. In the graphs, the red line represents actual GRACE values, while the green and purple dashed lines represent mean-imputed and median-imputed values, respectively.

Mean and median imputation result in a loss of variability, as they replace missing values with a single static value. This smoothing effect is evident in the imputed lines, which appear unnaturally flat compared to the actual data. This obscures important trends and patterns, particularly those associated with seasonal and annual variations.

Furthermore, these methods do not account for temporal dependencies in the GRACE data, resulting in imputed values that often do not align with surrounding data points. This misalignment creates abrupt and unrealistic jumps in the time series, as seen in the imputed segments of the graphs.

Additionally, mean and median imputation introduce bias by pushing imputed values towards the central tendency, which is problematic for datasets with skewed distributions or non-random missingness. This bias reduces the ability to accurately model and predict specific events. The imputed lines do not

accurately reflect the peaks and troughs of the actual GRACE values.

Given these limitations, it became clear that more sophisticated imputation techniques were necessary to handle the complexity and variability of the GRACE data. We began exploring Generalized Linear Models (GLM) as a potential solution. GLMs offer a more flexible approach, allowing for the incorporation of various predictors and accounting for the inherent structure of the data. By leveraging the strengths of GLMs, we aimed to achieve more accurate and realistic imputations that better reflect the underlying dynamics of the TWS data.

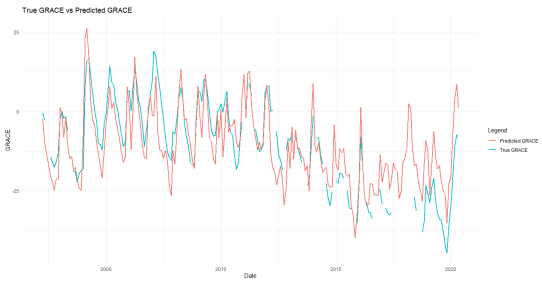

Figure 34: GLM test Sao Francisco

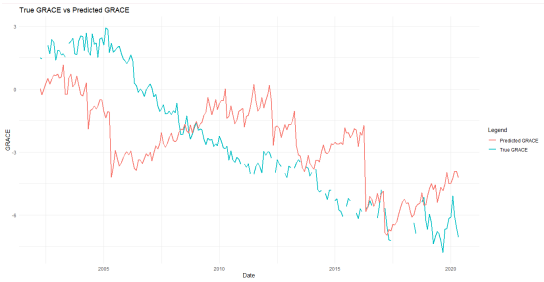

Figure 35: GLM test Saudi Arabia

These visualizations for both basins show the results of the GLM model, which was the next model we tested. In the graphs, the blue line represents actual GRACE values, while the red line represents the GLM model.

We can see from the results that GLM did not give us a great prediction, mostly because our graphs are not linear. We also can see that as we thought, Sao Francisco's predicted values were closer than Saudi Arabia's. The MSE for Sao Francisco was 65.11246, and the MSE for Saudi Arabia was 6.109285. The GLM model can be effective, but in our case, it was not, and this could have been for many reasons. We believe that the biggest problem is that our data was not linear, and therefore the results were not what we expected. Also, our graphs, especially Saudi Arabia, did not have any specific patterns it was difficult for this model to be accurate. This led Sao Francisco's to be a little better since it had a bit of a pattern, but we believed that these results could improve. Another problem was that our data was too complex, and GLM just was not complex enough to find these patterns and come up with better predictions. From here, we decided to do some more elaborate machine learning models, and a popular one we thought would be good was DNN, which was the next model we tested.

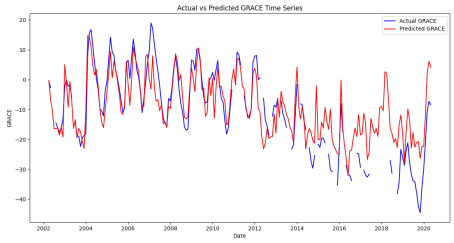

Figure 36: DNN test Sao Francisco

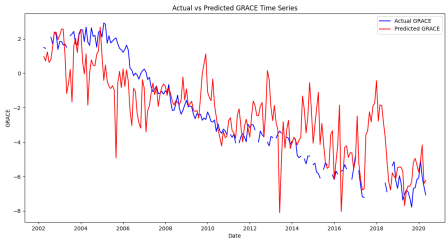

Figure 37: DNN test Saudi Arabia

The output of our DNN model is displayed in the visualizations for the basins of Sao Francisco and Saudi Arabia. The DNN model predictions are represented by the red line in these graphs, while the actual GRACE data are shown by the blue line. Saudi Arabia's MSE was 4.9318. The graph illustrates how the projected values of the DNN model frequently differed greatly from the actual values. This discrepancy suggests that the model has trouble adequately representing the intricate patterns found in the Saudi Arabian GRACE data. With an MSE of 87.0267, the DNN model's performance for the South American basin was also poor. The model's bad generalization for this dataset was evident from the predicted values' significant departures from the real GRACE readings.

The complex temporal dependencies and noise in the GRACE data presented a significant challenge for the DNN model, which may have underfitted or overfitted due to improper tuning or model architecture. Additionally, the model's subpar performance could have been caused by a lack of optimal hyperparameter settings. These factors combined to make the DNN model's performance subpar. Having acknowledged the shortcomings of the DNN model, we made the decision to investigate alternate machine learning strategies that might be more adept at handling the unique qualities of our data. One strong and effective gradient boosting framework that has shown promise is XGBoost. It was an appealing alternative for raising the GRACE datasets' prediction accuracy because of its capacity to manage missing data, recognize intricate patterns, and perform better in a variety of regression tasks. As a result, we tested the XGBoost model next.

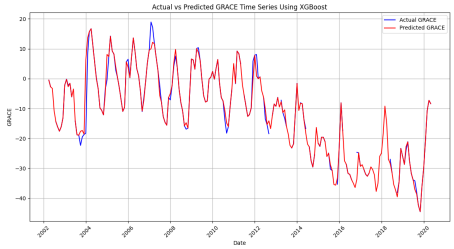

Figure 38: XGBoost test Sao Francisco

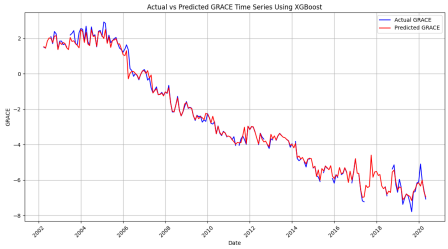

Figure 39: XGBoost test Saudi Arabia

After conducting some research, we discovered that XGBoost is highly suitable for complex datasets like the ones we have. In the graphs above, the red line represents the predicted values while the blue line represents the original values. The visualizations clearly demonstrate the efficacy of XGBoost in

handling the intricate patterns present in our data. Given the trends in our datasets, we required a model capable of managing such complexity effectively.

Upon examining the MSE for both the Saudi Arabia basin and the Sao Francisco basin, it is evident that XGBoost performs exceptionally well. The MSE for the Saudi Arabia basin was 0.263359484, indicating a very close match between the predicted and actual values. For the Sao Francisco basin, the MSE was higher at 28.56557523, but this is justifiable due to the much larger range of values in this dataset. Despite the higher MSE, the model's predictions still closely follow the actual data trends, demonstrating XGBoost's robustness.

One of the key factors contributing to the superior performance of XGBoost was our success in identifying strong hyperparameters. These optimized hyperparameters played a crucial role in mitigating issues of overfitting and underfitting, which are common challenges when dealing with complex datasets. By fine-tuning the model parameters, we were able to enhance the model's ability to generalize well across both basins, thereby achieving higher predictive accuracy.

In summary, the transition to XGBoost marked a significant improvement in our predictive modeling efforts. Its capability to handle complex data, coupled with effective hyperparameter tuning, resulted in better alignment of predicted values with actual GRACE data. This made XGBoost a highly effective choice for our analysis, addressing the limitations we faced with previous models and providing more reliable predictions for both the Saudi Arabia and Sao Francisco basins.

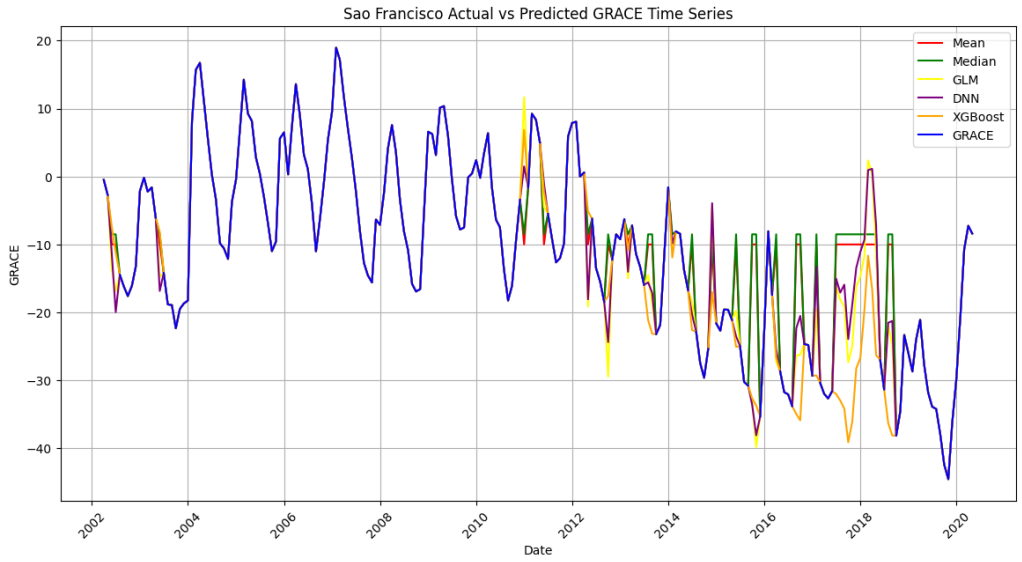

Figure 40: All Sao Francisco models

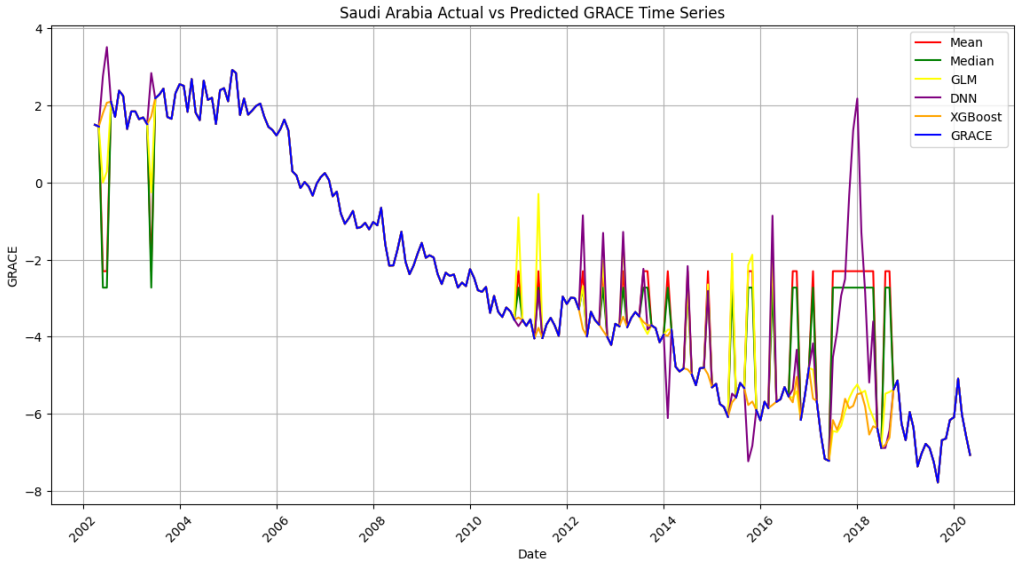

Figure 41: All Saudi Arabia models

Model	Sao Francisco MSE	Saudi Arabia MSE
GLM	65.11246	6.109285
DNN	87.0267	4.9318
XGBoost	28.56557523	0.263359484

Table 3: MSE values for different models in Sao Francisco and Saudi Arabia

Initially, we dropped the NAs when we first ran every model to evaluate their performance without any imputation. After identifying that XGBoost yielded the best results, we decided to explore various imputation methods to see if we could further enhance the model's performance.

We used several imputation methods, including Mean, Median, GLM, and DNN, and compiled the results into a data frame as seen in the below tables. This data frame included one column with the original GRACE data (including the NAs) and separate columns for each imputation method, where the missing values in the GRACE dataset were filled using the respective imputation techniques.

After constructing this comprehensive data frame, we used it to test each imputation method by feeding the imputed data into what we believed to be our best model based on the initial Mean Squared Error (MSE) comparison—XGBoost. Our aim was to see if the imputed data could further improve the model's performance.

DATE	GRACE	Mean	Median	DNN	XGBoost	GLM
4/1/2002	-0.48764	-0.48764	-0.48764	-0.48764	-0.48764	-0.48764
5/1/2002	-2.71737	-2.71737	-2.71737	-2.71737	-2.71737	-2.71737
6/1/2002	NA	-9.99033	-8.5002	-10.7795	-7.39412	-13.5669
7/1/2002	NA	-9.99033	-8.5002	-18.3155	-11.1649	-16.8717
8/1/2002	-14.4403	-14.4403	-14.4403	-14.4403	-14.4403	-14.4403
9/1/2002	-16.1875	-16.1875	-16.1875	-16.1875	-16.1875	-16.1875
10/1/2002	-17.5875	-17.5875	-17.5875	-17.5875	-17.5875	-17.5875
11/1/2002	-16.0834	-16.0834	-16.0834	-16.0834	-16.0834	-16.0834

Table 4: Sao Francisco Data

DATE	GRACE	Mean	Median	DNN	XGBoost	GLM
4/1/2002	1.500754	1.500754	1.500754	1.500754	1.500754	1.500754
5/1/2002	1.454773	1.454773	1.454773	1.454773	1.454773	1.454773
6/1/2002	NA	-2.2995	-2.72643	2.260744	1.799514	-0.000073
7/1/2002	NA	-2.2995	-2.72643	2.712308	2.073102	0.282646
8/1/2002	2.100225	2.100225	2.100225	2.100225	2.100225	2.100225
9/1/2002	1.704547	1.704547	1.704547	1.704547	1.704547	1.704547
10/1/2002	2.38778	2.38778	2.38778	2.38778	2.38778	2.38778
11/1/2002	2.250339	2.250339	2.250339	2.250339	2.250339	2.250339

Table 5: Saudi Arabia Data

Initially, we used Mean and Median imputation methods for handling the missing values in the GRACE dataset. As expected, these methods did not perform well. The Mean imputation resulted in the Mean Squared Error (MSE) increasing from 28.56557523 to approximately 55.4 for the São Francisco basin, and from 0.263359484 to roughly 1.9 for the Saudi Arabia basin. Similarly, the Median imputation caused the MSE to rise to 61.6 for São Francisco and to 1.5 for Saudi Arabia.

Next, we imputed the missing values using a Generalized Linear Model (GLM). This approach showed improvement for the São Francisco basin, where the MSE decreased from 28.56557523 to 22.57039881, indicating a strong enhancement. However, for the Saudi Arabia basin, the MSE increased slightly from 0.263359484 to approximately 0.7, which, while not as drastic as the increases seen with the Mean and Median imputations, still represented a bad imputation.

When we used a Deep Neural Network (DNN) for imputation, the results were less favorable. The MSE for São Francisco increased to 29.4, and for Saudi Arabia, it rose to 1.2. This considerable difference highlighted the poor performance of the DNN on Saudi Arabia's GRACE time series, likely due to the heavy downward trend in the data.

Finally, we imputed the missing data using XGBoost. This method yielded the lowest MSE for both basins, with the São Francisco basin achieving an MSE of 13.0 and the Saudi Arabia basin reaching 0.22. Although the improvement for Saudi Arabia was modest compared to its initial test, the fact that the MSE decreased aligned with our expectation that XGBoost would be the most effective model. Notably, Saudi Arabia only showed an improvement in MSE when using XGBoost, whereas São Francisco showed improvements with both GLM and XGBoost and had results close to DNN. This indicates that a basin

with a strong trend, like São Francisco, performs significantly better with XGBoost.

Model	Sao Francisco MSE	Saudi Arabia MSE
Mean	55.39412772	1.92191265
Median	61.60424332	1.530585255
GLM	22.57039881	0.711753973
DNN	29.41633735	1.2212988
XGBoost	13.00034159	0.224926562

Table 6: MSE values for different models in Sao Francisco and Saudi Arabia

5.1 Confidence Intervals

After we got our results, we looked at the confidence intervals for each of our basins and tested them to see if our models were in our confidence intervals. The confidence interval is the mean of your estimate plus and minus the variation in that estimate. Our confidence level is 95%, meaning that we are 95% confident that this is the confidence level range. With this region of confidence, we can see how much of it is being mapped within the confidence range.

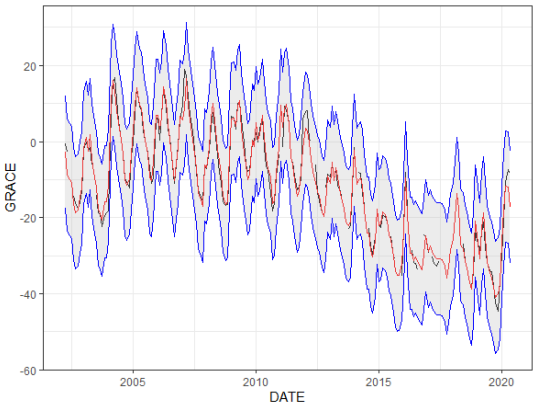

Figure 42: Sao Francisco confidence interval

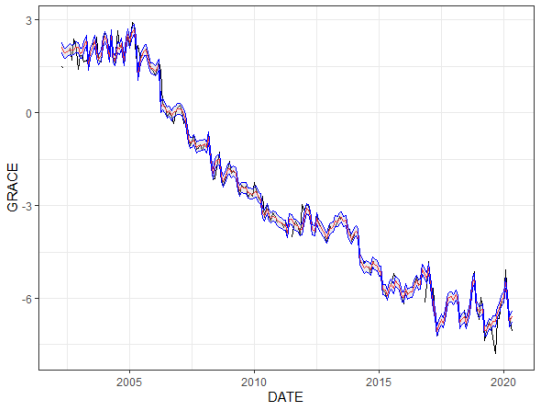

Figure 43: Saudi Arabia confidence interval

The figures above show the confidence intervals for each basin. Saudi Arabia's confidence intervals are very small because of the low variance and the lower mean. On the other hand, Sao Francisco's confidence intervals are bigger because of the higher variance and higher mean. Saudi Arabia is tight and close with its data and Sao Francisco has a larger range so everything is in the confidence interval. We tested how well our models work and how they fit with the confidence interval. Below you can see that XGBoost is within a 95% confidence interval the green dots are where all of the predicted XGBoost values. As you can see in figures 40 and 41 if you were to put all of the predicted values in the confidence interval for Sao Francisco both DNN and GLM are either in or nearly in 95% confidence while Saudi Arabia didn't have any other models anywhere close.

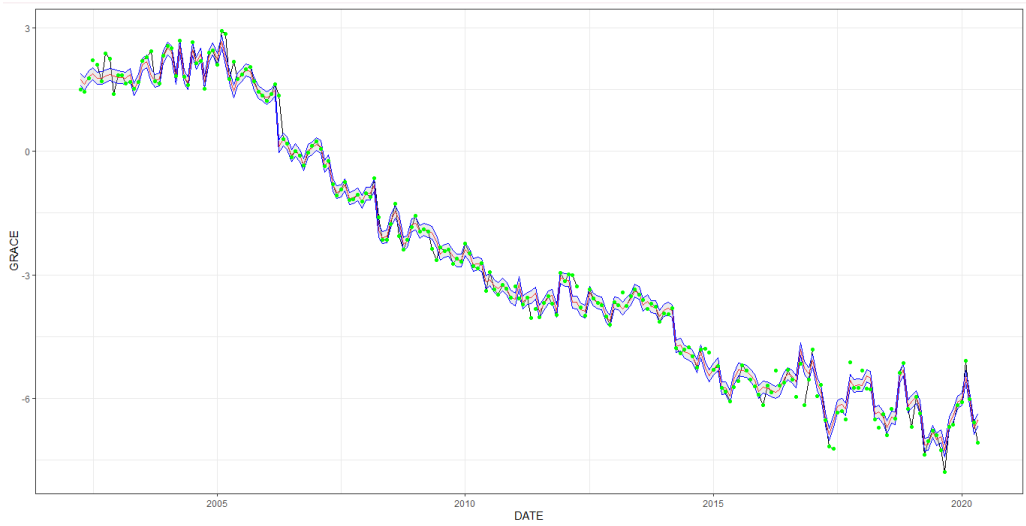

Figure 44: Saudi Arabia confidence interval for XGBoost

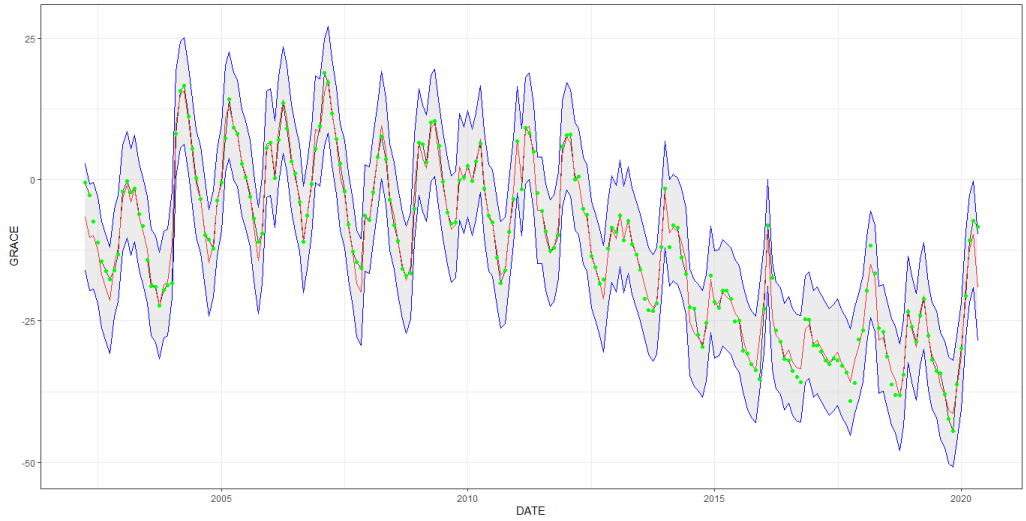

Figure 45: Sao Francisco confidence interval XGBoost

6 Conclusion

In this study, we have explored various statistical and machine-learning techniques to address the challenge of imputing missing GRACE data. We used two specific basins, Sao Francisco and Saudi Arabia, and our goal was to develop models that could accurately predict and impute missing GRACE data using associated factors such as TWS, temperature, precipitation, evapotranspiration, NDVI, and ENSO. We evaluated multiple imputation methods, including Mean/Median Imputations, GLM, DNN, and XGBoost. Each method was tested as we attempted to minimize the MSE, providing a measure of the model's performance and accuracy. Our findings show that while mean and median imputations are simple, they do not fully capture the variability and correlations within the data, leading to potential biases. GLMs were also found to be less effective in basins with complex, nonlinear patterns. In contrast,

DNN and XGBoost demonstrated superior performance due to their ability to model more complicated relationships within the data. Among the methods tested, XGBoost consistently provided the lowest MSE, making it the best approach for imputing missing GRACE data in our study.

In conclusion, we achieved our goal of imputing these missing values and finding the best option for them. These different statistical and machine learning techniques, particularly XGBoost, offer a powerful solution for imputing missing values in GRACE data. This approach not only strengthens the completeness and reliability of the dataset but also supports more accurate forecasting and analysis of TWS. In other previous findings we looked at, most said that DNN would work the best, and we thought that this was interesting that our findings differed from those. We believe that it is because there is more dynamic change in our basins that DNN did not produce as good of results as we thought. Since we only tested two basins, and most papers tested using all of the basins, our results were not the same as other prior research. Overall, our study emphasizes the potential of advanced machine learning techniques in imputing missing data in certain datasets, and this sets a foundation for further research about GRACE data and the best approach for filling in missing gaps in the data.

7 Future Research

We intend to investigate how our data might be used with the Autoregressive Integrated Moving Average (ARIMA) model in future work. Because ARIMA can represent the autocorrelations within the data and takes into consideration the relationships between observations, it is a good choice for time series forecasting. Our goal in applying ARIMA is to enhance the forecasting of missing GRACE data points and better capture the temporal patterns.

Furthermore, a large portion of our upcoming work will include broadening the scope of our analysis to encompass basins other than the São Francisco and Saudi Arabia basins. We can assess the robustness and generalizability of our results by testing these identical models—XGBoost, DNN, GLM, and imputation methods—on a range of basins with varying climatic and hydrological features. By doing this, we may ascertain which models and imputation techniques work best in a variety of geographical and environmental circumstances, which will ultimately strengthen the imputation of GRACE data and advance our knowledge of the dynamics of terrestrial water storage.

References

[1] Mariette Awad and Rahul Khanna. *Deep Neural Networks*, pages 127–147. Apress, Berkeley, CA, 2015.

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2016. Accessed: 2024-07-19.

[3] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, and Yuan Tang. Xgboost documentation - model training. `https://xgboost.readthedocs.io/en/stable/tutorials/model.html`, 2021.

[4] National Research Council. *A Review of the Landscape Conservation Cooperatives*. The National Academies Press, 2009. Accessed: 2024-07-19.

[5] IBM Developer. Cognitive neural networks: A deep dive. `https://developer.ibm.com/articles/cc-cognitive-neural-networks-deep-dive/`, 2024. Accessed: 2024-07-19.

[6] J.S. Famiglietti. Remote sensing of terrestrial water storage, soil moisture and surface waters. *State of the Planet: Frontiers and Challenges in Geophysics*, pages 197–207, 2004.

[7] GeeksforGeeks. Bias vs variance in machine learning. `https://www.geeksforgeeks.org/bias-vs-variance-in-machine-learning/`, 2024. Accessed: 2024-07-19.

[8] Bimal Gyawali, Mohamed Ahmed, Dorina Murgulet, and David N. Wiese. Filling temporal gaps within and between grace and grace-fo terrestrial water storage records: An innovative approach. *Remote Sensing*, 14(7), 2022.

[9] Latinx in AI. Xgboost: The king of machine learning algorithms. `https://medium.com/latinxinai/xgboost-the-king-of-machine-learning-algorithms-6b5c0d4acd87`, 2024. Accessed: 2024-07-19.

[10] Muhammad Faisal Javed, Muhammad Fawad, Rida Lodhi, Taoufik Najeh, and Yaser Gamil. Forecasting the strength of preplaced aggregate concrete using interpretable machine learning approaches. *Scientific Reports*, 14(1):1234–1238, 2024. Accessed: 2024-07-19.

[11] Osval Antonio Montesinos López, Abelardo Montesinos López, and Jose Crossa. Fundamentals of artificial neural networks and deep learning. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, pages 379–425. Springer International Publishing, Cham, 2022.

[12] NASA. Grace-fo mission. `https://grace.jpl.nasa.gov/mission/grace-fo/`, 2024. Accessed: 2024-07-19.

[13] Algorithm Hall of Fame. Backpropagation in neural networks. https://www.algorithmhalloffame.org/algorithms/neural-networks/backpropagation/, 2024. Accessed: 2024-07-19.

[14] M. Rodell, B.F. Chao, A.Y. Au, J.S. Kimball, and K.C. McDonald. Global biomass variation and its geodynamic effects: 1982–98. *Earth Interactions*, 9(2):1–19, 2005.

[15] M. Rodell and J.S. Famiglietti. An analysis of terrestrial water storage variations in illinois with implications for the gravity recovery and climate experiment (grace). *Water Resources Research*, 37(5):1327–1339, 2001.

[16] T.H. Syed, J.S. Famiglietti, M. Rodell, J. Chen, and C.R. Wilson. Analysis of terrestrial water storage changes from grace and gldas. *Water Resources Research*, 44(2), 2008.

[17] Unknown. A quick history of generalized linear models. https://blog.revolutionanalytics.com/2014/05/quick-history-glm.html, 2014. Accessed: 2024-07-23.

[18] Unknown. Generalized linear models. http://stratigrafia.org/8370/lecturenotes/generalizedLinearModels.html#:~:text=Generalized%20linear%20models%20use%20likelihood,how%20the%20errors%20are%20distributed., 2024. Accessed: 2024-07-23.

[19] Unknown. Mean squared error. https://www.simplilearn.com/tutorials/statistics-tutorial/mean-squarederror#:~:text=Mean%20square%20error%20is%20calculated,it%20relates%20to%20a%20function.&text=A%20larger%20MSE%20indicates%20that,smaller%20MSE%20suggests%20the%20opposite., 2024. Accessed: 2024-07-23.

[20] Unknown. Mean squared error (mse). https://statisticsbyjim.com/regression/mean-squared-error-mse/, 2024. Accessed: 2024-07-23.

[21] Unknown. Neural networks: Forward pass and backpropagation. https://towardsdatascience.com/neural-networks-forward-pass-and-backpropagation-be3b75a1cfcc, 2024. Accessed: 2024-07-23.

[22] Analytics Vidhya. New to neural networks? https://medium.com/analytics-vidhya/new-to-neural-networks-544ca4d6dacd, 2024. Accessed: 2024-07-19.

[23] YouTube. When to use xgboost. YouTube, https://www.youtube.com/watch?v=YgUcwTAmgpc&t=10s, 2024.