The University of Melbourne

Data Science Project

Emulation of dynamic simulation models of vegetation growth

Yuze Qu 1289458
Zening Zhang 1078374
Zhili Chen 980950
Ziqian Wu 1067683
Jiachen Huo 1293736

Team 25

A Report submitted for MAST90107

October 30, 2022

Abstract

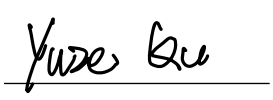
The original scientific theoretical model may be computationally expensive to construct and difficult to measure in some real-world engineering implementations. A different surrogate model is therefore desired; it operates as black-box modelling by ignoring the internal working theorem and concentrating just on the dataset. This data science capstone project is typically designed to develop such a surrogate model(s) to represent a complex terrestrial biosphere system where plant modulates their growth in response to the prevailing environmental conditions. When appropriate data is available, a surrogate model is typically expected to perform as well as the original emulation technique that has a lower computing cost.

In this report, we will start by introducing and describing the predictors and some of the outputs associated with the original JeDi-DGVM (Jena Diversity – Dynamic Global Vegetation Model) model which takes both the time dynamics of the growth process and multiple plants' functional trade-offs into account. Since multiple techniques and models are adopted in according with different tasks where the data process is required to various extents, thus the exploratory data analysis part, as well as any supported visualizations, are integrated within each main methodology section. The thorough explanation of a number of machine learning approaches and algorithms that we used to simulate our simulation process in various experimental conditions follows. We will next present and compare the predictions made by different algorithms, as well as their advantages and disadvantages.

Keywords— JeDi-DGVM, Surrogate Model, Random Forest, Clustering, ARIMA, LSTM

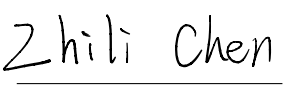
Signed Declaration

I certify that this report does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text. The report is 7418 words in length (excluding text in images, tables, bibliographies and appendices).

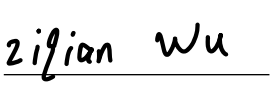


Yuze Qu

Zening Zhang



Zhili Chen



Ziqian Wu

Jiachn Huo

Acknowledgements

Throughout this year-long capstone data science project, we would want to offer our greatest appreciation to our supervisor, client, and educators for the Data Science Project. This initiative would not have been feasible without their tremendous assistance.

The client, Professor Peter Rayner, and the supervisors, Jeremey Silver and Joyce Zhang, have to be thanked first. Despite the arduous effort of monitoring and managing many capstone projects concurrently, your skills and expertise in offering a helpful data source and constructive advise really assisted us in staying on track. Your input on our weekly progress and our primary methodology is incredibly helpful, and we would like to convey our gratitude once more.

Then, we would like to express our gratitude to Yuze Qu, Zening Zhang, Zhili Chen, Ziqian Wu, and Jichen Huo. Everyone performed their responsibilities and displayed admirable abilities by executing flawless teamwork. During this culminating project, we forged a strong tie and connection with each other, which will be one of the most valuable things we harvested.

Contents

1 Introduction

1.1 Overview and Motivation

How to design a surrogate model so that it performs as well as the original scientific model while remaining computationally efficient remains a difficulty for the majority of machine learning experts. For a complex earth-system modeling procedure, vegetation typically adopts a variety of growing strategies and indicates distinct growth paths based on their specific prevailing environmental conditions. These environmental conditions are the most important input for our machine-learning algorithms, which we will discuss in greater detail later. We were tasked to develop several machine learning techniques and methods, as well as a benchmark model, to serve as a replacement benchmark surrogate model for the original JeDi-DGVM. Here, ”Simulation” and ”Simulators” is related to the original JeDi-DGVM, whereas ”Emulation” and ”Emulators” correspond to our surrogate model and techniques. We are using JeDi simulation results as a data source for the development and testing of our primary methodologies. Among the many valuable outputs that the JeDi simulation method generates, such as ’Biomass,’ ’Storage Carbon Pool,’ and ’Storage Carbon Pool,’ etc., our major prediction and emulation target in this capstone project is ‘NPP’ which stands for ‘Net Primary Productivity’ of vegetation. The aforementioned environmental factors can be further categorized as time-dependent variables, which will be used interchangeably with ’time-series data’, and static variables, which will be referred to as’static predictors’ or’static input’ for the remainder of this study.

There are a total of three working phases, which are based on three sets of simulation data sources from JeDi with distinct experimental settings, and each stage of development improves upon the preceding one by incorporating its strengths and addressing its weaknesses. Given that we constructed our emulator solutions in accordance with various datasets, each section of the Methodologies will explain in greater detail how we processed each dataset.

The structure of this report is as follows: The remainder of Section 1 concludes the introduction and high-level summarization of this project. The second section discusses relevant background knowledge and issues domain development. In Section 3, we present our framework(s) including data processing, techniques, and experiment conditions. The remainder of Section 4 consists of moderating findings and discussions. In Section 5, a conclusion and future directions are offered.

1.2 Challenges and Difficulties

From a data science perspective, working with large-scale datasets makes it more difficult to perform data-processing tasks such as decreasing noise and dealing with missing information, which takes additional computer resources. Given the nature of the original simulator that accepts both static and time-dependent input, how to properly combine or deal with them remains debatable throughout the development phase. Due to the possibility of geological interpolation of grid points, the potential overfitting issue must be resolved by establishing the appropriate limits or procedures. Then, from the perspective of candidate model selection, it comes to our agreement that no single surrogate model can handle

the entire problem, which shows that parallel coding and development requires a division of labour.

1.3 Dataset Description

As mentioned previously, a different version of the JeDi dataset with slightly modified simulation settings was provided to us at each working stage; these datasets are referred to as the ”South-America dataset,” the ”Global-Short-run dataset,” and the ”Global-Long-run dataset” for the remainder of this report. Each consists of identical static input and time-series input components:

Time-Series Input		Static Input	
’tas.nc’	Air Temperature	‘latitude.nc’	Latitude
‘rlns.nc’	Net Long Radiation Flux	‘longitude.nc’	Longitude
‘rsds.nc’	Shortwave Radiation	‘elevation.nc’	Elevation
‘pr.nc’	Precipitation	‘paw.nc’	Moisture
NPP	Net Primary Productivity		

Table 1: Input Variables

The South-America dataset is the very first version we were working with at our initial working stage (Fig. 1), it differs from the latter two in the sense of geological locations where it consists 51*60 grid points extracted from South-America area with 288 monthly time-series data (range from 2004 to 2029). It is composed of both land and sea grid points where we are considering them independent from each other neighbor grid points.

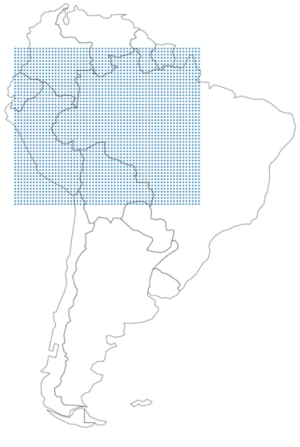


Figure 1: Full Grid Points of South-America Dataset

Both Global-Short-run and Global-Long-run dataset are consisting of a larger-scale grid points extracted globally, which interpolates the latitude-longitude pairs into 145*192 grid points. Figure 2 gives an illustration of grid points

of the global scale, where basically the whole globe is interpolated with the blue dots. The Global-Short-run dataset differs from the Global-Long-run one's in time dimension where the previous one consists of 288 monthly (24 years) data and the later one has a longer time dimension of 3516 months (293 years).

Figure 2: Grid Points of Global Scale Dataset

1.4 Research Questions

Given a complex climate and the earth's natural system, multiple tasks can be established during the modeling process. In this particular project, we mainly focused on the following two major tasks:

1. Short time-stamp prediction of NPP.
2. Simulated data (NPP)'s whole time-series emulation process with multiple period prediction.

The first task is supported by our benchmark solutions using Random Forest base classifiers. This task is when we have historical input of time-dependency variables and/or static input, we wanted to predict four-time ahead NPP's value. The second task is harder than the previous one, it requires emulating a whole dynamic growth process which potentially need large scale training data and more suitable machine learning strategies.

1.5 Methodology

As mentioned above that we are targeting two different tasks, we adopted four major techniques and methodologies throughout the whole project, and these will be introduced separately.

1. **Random Forest**: This is the major benchmark solution to task 1's short four-step-further prediction as Random Forest (RM for short) shows the advantage of being robust to noisy input data. The static predictors and time-

dependency predictors are fitted into RM regressors separately during each experiment, where a total of 15 combinations between dynamic features and NPP are carried out for time-dependency predictors.

2. **Clustering**: Given the large dimension of grid points with a potential change of latent distributions, one single regressor is insufficient of capturing all useful information, and having each grid point itself a regressor is too computationally expensive. Then a clustering method with square root of dataset-sized clusters is desired. Here a partitioning clustering with k-means on our target output NPP is developed as an essential preliminary data pre-process step for further combination with emulators.
3. **ARIMA**: Start with an emulator that only depends on endogenous variable itself which in this case, NPP, the auto-regressive integrated moving average method is adopted as an initial solution to our task 2. With the previous clustering result, each cluster specified ARIMA parameters are first determined (e.g., lags, degree of difference and moving-average of each cluster) then the normal training and emulation / forecasting are carried out consequently.
4. **LSTM**:Using LSTM as a black-box surrogate model, it has the ability of remembering some long-term information but is very data-hunger. Four exogenous predictors are taken into consideration when training the LSTM model with clustered training data taken from previous result.

1.6 Implication of Findings

To present a high-level summarization of our major results and findings here, the Random Forest regressor performs as a benchmark model which meets our expectations towards the first task of four-step-further prediction. Average model effects of RM achieve over 85% accuracy in terms of predict-score. Where on the other hand, the combination of ARIMA+Clustering and LSTM+Clustering gives good inspiration of how to emulate a given dynamic simulation process, but given the level of difficulty and limited working time period, further development and optimization are expected at a future time. We obtained acceptable result of the clustering method on NPP, on both South American dataset and the global-scale dataset with 77 clusters. Plus a slightly better emulation result which was given by ARIMA+clustering method.

2 Related Work

Adopting machine learning algorithms and techniques to construct a field-specific complex dynamic simulation model within various industries has been under discussion and gained lots of attention as the growing scale of big data and computational resources. As P. Stolfi et al. pointed out [SC21], original simulators generally come with non-linear ordinary or partial differential equations and multivariate output, they are capable of producing high-accuracy simulations and numerically describing the complex underlying scientific phenomena. However, it comes with expensive costs that it is only realistic to execute them on HPCs or workstations with huge parallel execution power.

P. Stolfi et al have then proposed a way of using the Random Forest model to 1) provide an approximate final prediction of their target variable (i.e. Presents of Type 2 Diabetes in their case), and 2) to recover the simulator's parameter value and identify most important predictors [SC21]. This gives us the major inspiration of also adopting Random Forest as our task 1's main model which is expected to perform at an acceptable benchmark level. Carine M. Rebello et al emphasized the existence of a conceptual difference between simulation and prediction by mentioning the term "Non-linear auto-regressive with exogenous input (NARX)", which gives a reminder that different techniques have to be developed for task 2's emulation process [RMC+22].

When it comes to the second task containing whole process emulation, the clustering technique is always preferred to effectively reduce the number of regressors or emulators (e.g. the partitioning k-means algorithm which was widely researched by Guha et al [GMM+03]). During a hydrological research work presented by Olutoyin A. Fashae et al, the prediction and simulation power of ARIMA outcomes ANN when they were forecasting the discharge of River Opeki from 2010 to 2020 [FONU18]. Thus, ARIMA as a powerful time-dependency forecasting model has been adopted for this capstone project. Take a step further than ARIMA, Manuel A. Rohrl et al indicated their experimental results that LSTM works pretty well as a surrogate model based on dynamic numerical simulation of a Heat Recovery Steam Generator (HRSG) [RLB+21].

3 Methodology

3.1 Random Forest-Benchmark

3.1.1 Overview

Considering the large scale of Global-Long-run dataset and the long time dimension which is 3516 months (293 years), we decided to use Random Forest as our benchmark model. It is known that Random Forest, as an ensemble of decision trees algorithms, can be used for both classification and regression tasks, , which can be formulated as:

$$\begin{aligned} y &= \psi(x) + \varepsilon, \varepsilon \sim \mathcal{N}(0, \Sigma) \\ &= \frac{1}{N} \sum_{i=1}^{N} \tau_i X + \varepsilon \end{aligned} \tag{1}$$

Where y is the vector of the target variable (here is NPP), x is the vector of correspondence predictors, N is the number of decision trees we need to tune for getting the best hyperparameter, and τ_i is the structure of the i-th tree. Compared to other traditional machine learning models, it has lower variance and can produce better results. The most important reason for us to choose Random Forest is that it is appropriate for time series prediction due to the ensemble of a series of weak learners(trees). We applied our Random Forest model to static features and time series features separately to figure out which features are most effective for the prediction of NPPs.

3.1.2 Random Forest with Static Data

For this part, we aimed to predict NPP using static features. To begin with, there are five features, such as location and elevation, of the static data and we integrated them to create a dataset with dimensions of 27840*5. Since the static features did not depend on time, we averaged the NPP for 293 years for each grid point on Earth. For the grid points of the ocean, NPP will be zero. Therefore, these instances with zero NPP were removed from the dataset. Then, the static features and NPP value are integrated, and a new dataset (Figure 3) with dimensions of 8093*6 will be used to fit the benchmark model.

	landSea	lantitude	longitude	elevation	moisture	NPP
2484	1.0	75.00	337.500	1293.0	0.060	-0.000164
2485	1.0	75.00	339.375	533.0	0.105	-0.000147
2486	1.0	75.00	341.250	-11.0	0.105	-0.000007
2623	1.0	73.75	238.125	276.0	0.075	-0.000017
2624	1.0	73.75	240.000	264.0	0.075	-0.002736
...	...	...	...	...	...	...
22427	1.0	-55.00	290.625	176.0	0.060	0.067544
22428	1.0	-55.00	292.500	56.0	0.105	0.071823
22433	0.0	-55.00	301.875	-3428.0	0.105	0.000624
22434	0.0	-55.00	303.750	-574.0	0.105	0.000271
22435	0.0	-55.00	305.625	-3139.0	0.105	0.000203

8093 rows × 6 columns

Figure 3: New dataset combining static features with NPP

Before model fitting, the data was split into training data and test data with the proportion of 80 percent and 20 percent along the geological dimension. Since NPP is a numeric variable, it could be known that the dataset can be also used to fit the classic regression model. In contrast to the Random Forest model, we subsequently trained a linear regression model whose accuracy was only about 0.325, significantly below our expectations.

Next, the dataset would be used to train a Random Forest model. Since the tree number of the model can affect the model performance, we set a range from 30 to 160 of the tree number parameter to figure out which number can make the model has the highest accuracy. Within the range, the model accuracy does not have significant differences. The results of different models are listed in part 4.2.2.

3.1.3 Random Forest with Time-Dependent Data

When it comes to time series data, since there are four time series features: airTemp, netLongRadi-antionFlux, shortwaveRadiation and precipitation, we experimented with total 15 different combinations of features plus previous 289 year's NPP to predict final 4 year's NPP among total 293 years of Global-Long-run dataset.

Considering the existence of the time dimension, we combined the four-feature data and the NPP value together to create a dataset with dimensions of 139200*3516. Each five rows represented four features and the corresponding NPP value of one grid point on Earth. Then, for the time dimension, we transformed the 3516 months which represents 3516 columns of dataset into 293 years which represents 293 columns of dataset. We removed ocean points and their corresponding features from the dataset to make further optimization. Finally, we got a new dataset with dimensions of 40465 * 293(Figure 4).

	0	1	2	3	4	5
0	259.569295	257.779251	261.612437	259.805583	259.517125	257.122696
1	-58.531250	-59.554688	-53.106120	-58.130208	-53.305990	-54.369792
2	121.298479	127.786171	105.556902	116.030153	119.126147	118.707997
3	0.000010	0.000012	0.000019	0.000015	0.000015	0.000011
4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
...	...	...	...	...	...	...
40460	271.042180	272.195791	271.917292	272.672750	271.172414	271.815928
40461	-36.747396	-37.533854	-37.666016	-34.920573	-38.606120	-38.881510
40462	113.727249	114.987738	104.334452	108.067668	119.969709	109.294378
40463	0.000026	0.000028	0.000037	0.000033	0.000033	0.000030
40464	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

40465 rows × 293 columns

Figure 4: New dataset combining time series features with NPP

Then we experimented with different Random Forest models with various combinations of features plus the previous 289 years' NPP as predictors. For each model, the data was split into training data and test data with the proportion of 80 percent and 20 percent. We also set a range from 80 to 120 of the tree number parameter to figure out which number can make the model has the highest accuracy. Besides, mean squared error and R-Squared score were viewed as criteria to evaluate these models as we did to the static data. The results of different models are listed in part 4.2.3.

3.2 Clustering

3.2.1 Overview

Due to the nature of the training data, computation complexity is raised as an issue; this section will discuss how clustering was used as a method to relieve the computation burden. For time series predictions, regressors such as the ARIMA family or random forest regressor always appear to be a popular solution. However, in order to preserve independence between instances of data, most of such solutions are focused on the scenario such that applying one regressor per example of times series. That means, with thousands of examples of NPP time series for each grid cell, considering features as exogenous variables, assuming the independence of the geological grid cells, and applying one regressor per cell will lead to an obvious problem: both time and space complexity will skyrocket.

To relieve the computation complexity while retaining most of the information, the idea of "grouping" the times series together is introduced to help reduce computation burdens. The idea of "grouping" time series is used when time series are parallel and independent, but time series are sharing some similar features or exogenous variables. For example, when features are categorical, one solution is using a hierarchical regression: regression with the child group first and then aggregating the results. The following figure shows an example of how the total time series disaggregate into 4 smaller series in 2 levels. Each node represents a child time series that holds similar features.

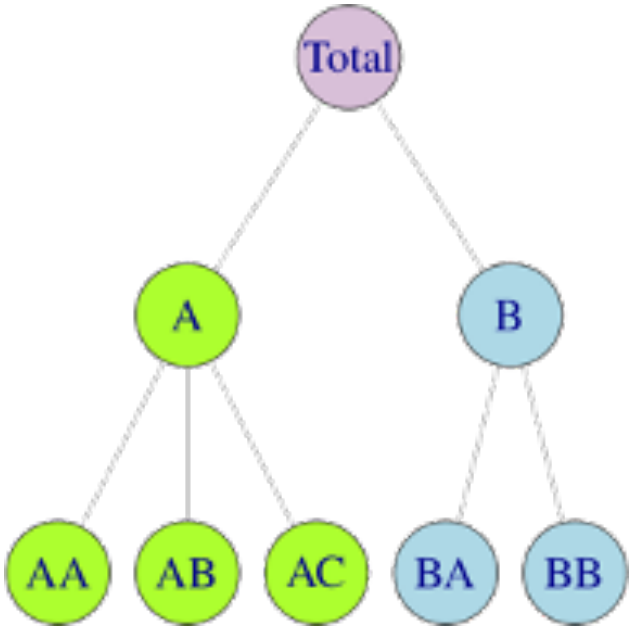


Figure 5: An example of hierarchical structure

However, hierarchical regression is mostly used when only the features are categorical, and the children can identify naturally based on such features. A more general form of hierarchical regression is simply grouping time series with similar features. Inspired by the idea of hierarchical regression and grouping, a more intuitive, suitable solution

for continuous data is to make use of clustering algorithms with a metric measuring the similarities between time series.

The primary idea is that vegetation growth may be similar to one another from time to time. That means it could be a viable method by using clusters to group similar time series together to reduce the number of regressors as well as sample size. Ideally, by applying clustering beforehand any "actual" regression or machine learning model, this method allows using fewer regressors or less complex models (smaller neural networks). By the rule of thumb of the clustering algorithm, the number of clusters is the square root of the sample size; this can reduce the dimension of data and thus reduce computation complexity.

3.2.2 K-means with Dynamic Time Warping

This project makes use of the package *tslearn*, which provides all ready-to-use tools for time series clustering. Within the solutions provided by the package, K means with Dynamic Time Warping (DTW) appears to be suitable for the project's scenario; after data preprocessing, K-Means with DTW is applied to the simulated NPP for clustering. Dynamic Time warping is a method of calculating the distance that is more accurate than Euclidean distance. The Following showed the general equation of DTW where X and Y represent two time series with m and n elements, respectively. DTW can be summarized as the square root of the sum of squared distances between each element of X corresponding to its nearest that of Y. The centroid of each cluster is computed in terms of DTW. K-means of *tslearn* package applied DTW Barycenter Averaging (DBA), an algorithm that searches for the minimum sum of squared DTW distance from the centroid of the cluster to another instance within the cluster.

$$DTW(x,y) = \min_{\pi} \sqrt{\sum_{(i,j\in\pi)} d(x_i,y_i)^2} \tag{2}$$

Where $\pi = [\pi_0, ..., \pi_k]$ is a path that satisfies:

- it is a list of index pairs with: $\pi_k = (i_k, j_k)$ with $0 \leq i_k < n$ and $0 \leq j_k < m$
- $\pi_0 = (0,0)$ and $\pi_k = (n-1, m-1)$
- for all k>0, $\pi_k = (i_k, j_k)$ is related to $\pi_{k-1} = (i_{k-1}, j_{k-1})$ as follows:
  - $i_{k-1} \leq i_k \leq i_{k-1} + 1$
  - $j_{k-1} \leq j_k \leq j_{k-1} + 1$

The following figure shows an example of clustering over NPP data. The red line represents the centroid or means of the cluster while the grey lines represent all the instances within the cluster. As a result of DTW and DBA from *tslearn* package, the centroid of the cluster appears to have an averaging shape that mimics all other instances of the same cluster.

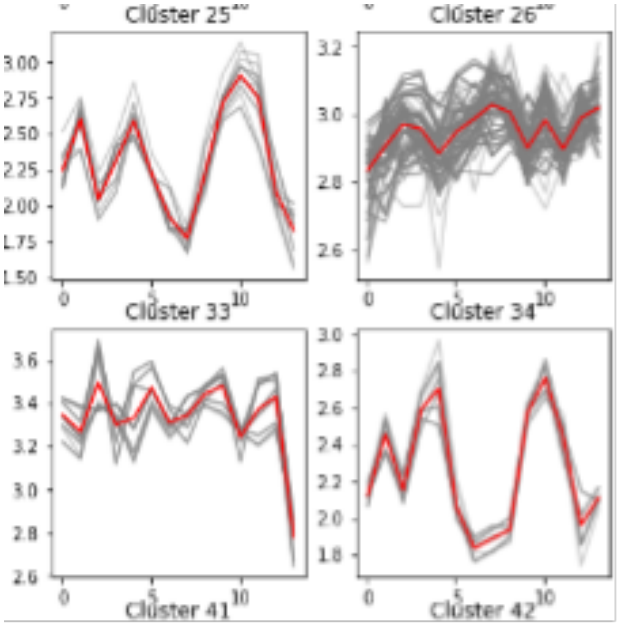


Figure 6: Example of cluster – a closer look at centroid shapes

DTW has an advantage over Euclidean if data points are shifted between each other. Instead of matching the positions of points, the focus of DTW is rather on its shape. More specifically, The Euclidean distance takes pairs of data points and compares them to each other; while DTW calculates the smallest distance between all points. For DTW, the matching between data points is considered in a one-to-many match strategy. The following figures showed the example match of Euclidean distance as well as DTW. The match of DTW appears to me more accurate and flexible.

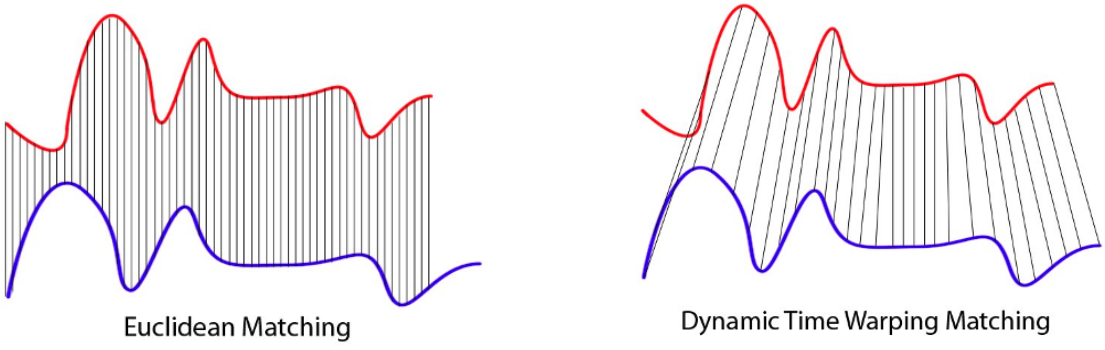


Figure 7: The comparison of DTW and Euclidean distance

When applying clustering algorithms on continental or even global-scaled data, the advantage of clustering starts to show up. The following figure is a heat map of clusters of NPP shown on the world map. Each dot represents an instance of time series of a geological grid point. The dots of similar color temperature represent that they belong to the same or similar cluster. As shown in the graph, geological grid points share similar static features, especially latitude and longitude, tending to belong to the same cluster. This observation meets the expectation since the time series of

vegetation growth is affected by static features such as spatial characteristics and climate distribution. The result also leads to a conclusion that clustering is a viable approach to be used beforehand or any other machine learning method in order to reduce computation complexity; although the clustering is obviously based on assumptions, it is able to capture most of the information of the original NPP time series. This will be discussed in more detail in the later section of this report; K-means with DTW will be used collaboratively with the ARIMA regressors as well LSTM neural networks and how such combination reflects on the result of the emulations.

Figure 8: Global Clusters in a Global view

3.3 ARIMA

Due to its specific statistical features and Box-Jenkins methodology, the autoregressive integrated moving average model (ARIMA) is a prominent technique for analysing time series data. ARIMA is widely used to predict the future value of a variable based on a number of historical observations and random mistakes, especially when dealing with data with linear relationships.[Zha03] Based on the properties of autoregression (AR), integrated(I), and moving average (MA), the ARIMA is more adaptable than some other autoregression models to the condition of temporal data kinds and structure. As the project's data spans 3516 months (293 years) in time dimension, and the NPP is not an unstable variable with few third-party impacts, the NPP is not a candidate for optimisation. The ARIMA model, which is basic and uses just endogenous variables, will be adequate for the project.
The ARIMA model's function may be represented as:

1. **as for the integrated part:**

$$y_t = \begin{cases} Y_t & \text{if d =0} \\ Y_t - Y_{t-1} & \text{if d =1} \\ (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) & \text{if d =2} \end{cases} \tag{3}$$

2. **as for ARIMA function:**

$$\hat{y}_t = \mu + \phi_1 * y_{t-1} + \cdots + \phi_p * y_{t-p} + \cdots + \theta_1 * y_{t-1} + \cdots + \theta_q * y_{t-q} \tag{4}$$

Where p stands for autoregression parameter, d stands for integrated parameter, and q stands for moving average parameter.

3.3.1 Data prepossessing

Similar to random forest's data preparation. As our response variable is NPP. Thus, any locations where the sea level is 0 will be eliminated (there are no NPPs).

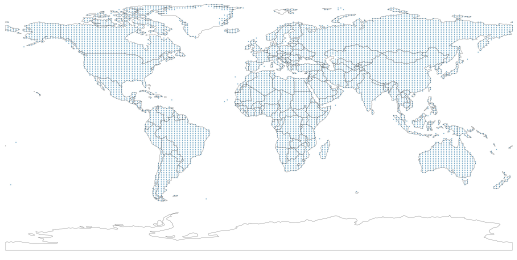


Figure 9: sea level 1 location

However, after pre-cleaning the data, we discovered that despite the fact that many places are above sea level, the majority of their NPP are zero owing to varying circumstances. During the modelling procedure, this data will also serve as interference terms. Consequently, these data must likewise be erased. Therefore, there will be a total of 5845 remaining places. Which is depicted in the diagram below as the filtered location in blue and the interference terms in red.

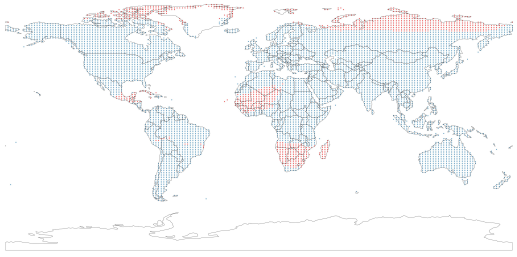


Figure 10: data after filtering

3.3.2 Stationary check

So far, a number of works on autoregressive models have shown that importance of the degree of data stationary. The degree of stationary determines whether the mean and variance are constant over time. In addition, it reflects that whether there is change in the statistical properties over time. The stationary of the data provides the autoregressive models and ARIMA ability to forecast a more accurate and credible future variables. To reach the stationary, the differencing can help to stabilize the statistical properties of the dataset by removing the trend or seasonal effect in time series. Besides, the differencing term can be indicated as the integrated(I) term in the ARIMA which is an important tuning parameter.

Normally, the Augmented Dickey-Fuller Test (ADF-test) will be applied to check whether the data is stationary. The hypothesis test about the unit root will be generated.

- **Null Hypothesis: H0** = If failed to be rejected,it suggests the time series has a unit root, meaning it is non-stationary
- **Alternative Hypothesis: H1**=The null hypothesis is rejected and suggests the time series does not have a unit root, meaning it is stationary

As the null hypothesis is subsequently rejected, there will be no unit root shown, and the stationary status of the time series will be demonstrated.

3.3.3 Hyper parameters tunning

Basically, the parameters in ARIMA are (p,d,q), which the

- **p** = the order of the autoregressive model (number of time delays).
- **d**=the degree of distinction (the number of times the data have had past values subtracted)
- **q**=is the order of the moving-average model

During the stationary check, the parameter d Integrated term was identified as being present. The values of p and q will be calculated using the approach of exhaustion from the to establish the minimum BIC. Due to the change in NPP, the transition is gradual and seamless. Therefore, the maximum value of autoregression and moving average cannot exceed nine terms.

3.3.4 Exogenous variable

In addition to NPP variables, the dataset also includes air-temperature, precipitation, shortwaveRadiation, and netLongRadiantFlux variables. They contribute differently to the NPP and should be considered exogenous factors to improve the accuracy of the model's NPP forecast. When exogenous variables are introduced, ARIMA becomes ARIMAX. Multiple variables are utilised by ARIMAX to include external data. The ARIMAX formula will then be

$$Y_t = \Sigma_{i=1}^{p}(\phi_i Y_{t-i}) + \Sigma_{j=1}^{q}(\theta_j \varepsilon_{t-j}) + \Sigma_{h=1}^{b}(\beta_h X_{t-h}) \tag{5}$$

as ε_t is the white noise, X_t is the exogenous variables

In addition, the technique of ARIMAX is identical to that of ARIMA in terms of checking stationary and tuning parameters.

3.3.5 Limitation on ARIMA/ARIMAX

As stated in the description, ARIMA/ARIMAX will function well with linear relationship data. Therefore, ARIMA/ARIMAX cannot address problems with nonlinear relations. As we cannot guarantee that the NPP is completely linear, the following nonlinear methods will assist to solve the NPP.

3.4 LSTM

3.4.1 Overview

In accordance with ARIMA/ARIMAX's limitations, a neural network model will be used to answer the query if the NPP has a nonlinear connection. Long Short-Term Memory (LSTM), is a kind of recurrent neural network (RNN). However, RNNs can only gaze back in time for around 10 timesteps [SM19]due to the disappearing or bursting fed-back signal. However, this issue was resolved by the construction of LSTMs.

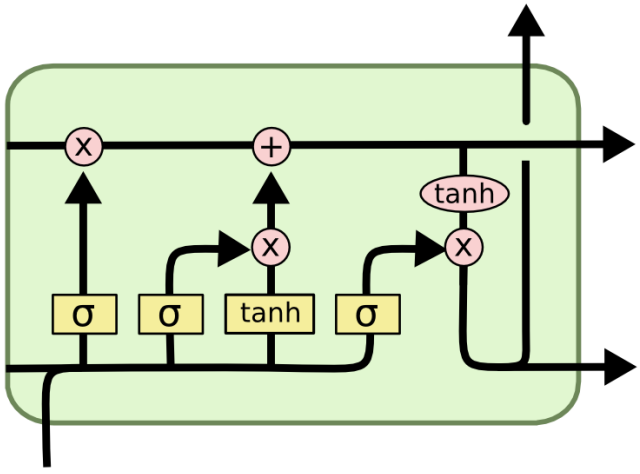


Figure 11: LSTM layer

- **input gate** = $I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i)$
- **forget gate**=$F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f)$
- **output gate**=$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o)$

From left to right along the yellow bar are the forget gate, input gate, fresh information gate, and output gate. And the forget gate determines whether a memory cell is retained or discarded during the current time step. This approach assists the LSTM with the disappearance of the fed-back signal and improves the capturing of time series dependencies with longer time step distances.

3.4.2 Data prepossessing

The method of data prepossessing in LSTM is as same as the ARIMA one.

3.4.3 Component and tunning parameter

3.4.3.1 Input layer

Our hope for LSTM is that it will determine the association between four factors and the NPP and attempt to predict the NPP based on these four variables. Therefore, for LSTM, the input shape is a 10-unit moving window.

- **X** = 4 variables from first to ninth time steps in the moving window
- **y**= NPP from last time step.

3.4.3.2 Hidden layer

The model's hidden layer was set to one layer. A numerous and complicated hidden layer will lead to model overfitting since the data size of 293 years is not sufficiently big.

3.4.3.3 Neurons

By comparing the different numbers of neurones (32, 64, 128) based on their mean square error, the number of neurones set to 128 was determined to have the lowest mean square error over the cluster.

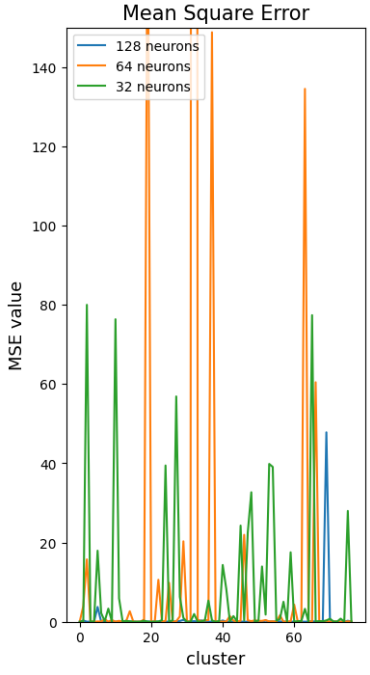


Figure 12: MSE for 32, 64, 128 Neurons

3.4.3.4 Recurrent activation function

The recurrent activation is the function that controls the forget, input, and output gates, whereas the hard sigmoid is

the fundamental activation for these gates.

3.4.3.5 Activation function

The activation function is used to regulate whether the previous data is stored in memory or lost. As the activation function, the Parametric Rectified Linear (PRelu) Activation Function was selected. PRelu, unlike regular Rectified Linear (Relu), enables negative values to show and does not convert them to zero. Adaptively learning the parameters of the corrected linear cell, the activation function is able to enhance accuracy with little additional computing cost. Second, we analyse the difficulty of model training and devise a theoretically valid initialization strategy that helps the deep network model converge.

$$f(y_i) = \begin{cases} y_i & \text{if } y_i > 0 \\ a_i y_i & \text{if } y_i \leq 0 \end{cases} \tag{6}$$

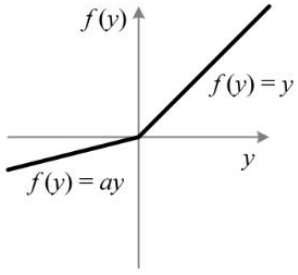


Figure 13: Parametric Rectified Linear

3.4.3.6 Early stop criteria

During the network sizing procedure, several decisions concerning the settings (hyperparameters) must be taken in order to generate a neural network with optimal performance. One of the hyperparameters is the number of epochs; too few or too many epochs will result in underfitting or overfitting issues, respectively. Consequently, the number of training rounds of a neural network is crucial. As a result, the early stop criterion will assist the model in stopping when optimisation is achieved.

monitor	min delta	patience	restore best weights
mean squared error	0.003	5	True

Table 2: Early stop criteria parameters

The model will stop if the mean square error loss decreases less than 0.003 for more than 5 round.

4 Result and Discussion

4.1 Lorenz Model

In the first part of the project, the group benefitted from trying to analyze and emulate Lorenz's time series. The group acquired a better understanding of the methodology of time series emulations in terms of preparation for the emulation of the actual targeting model which its data will be acquired in the second term of the project. The analysis of the Lorenz model showed that the study on the Lorenz model could be a good starting point before we begin the actual emulations of the targeting model. The following graph shows the following figure showed an example of prediction data vs. Testing data of the Lorenz time series.

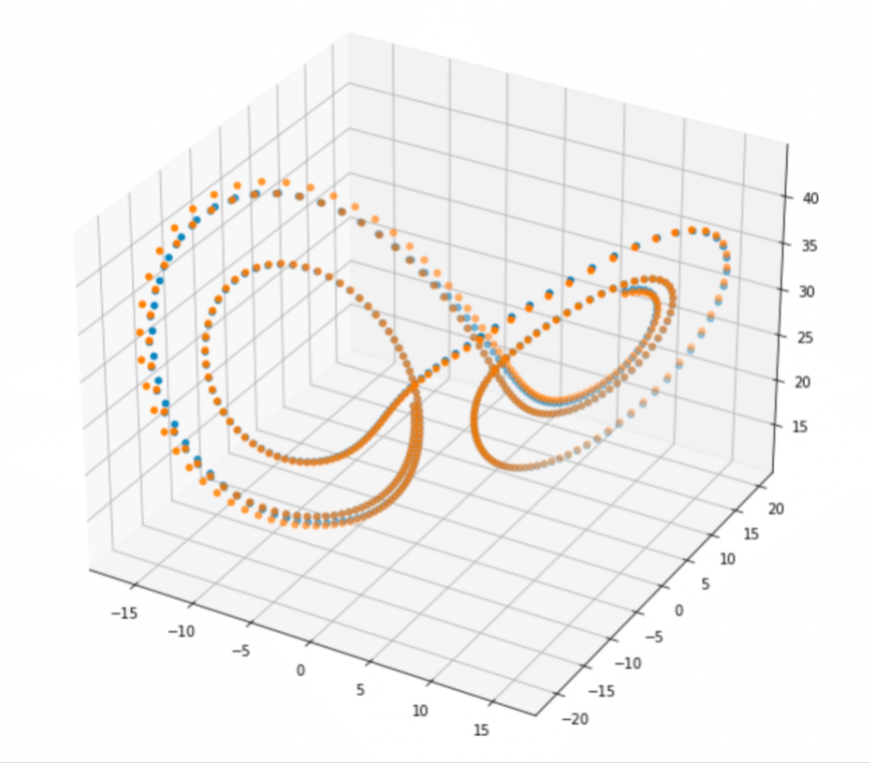


Figure 14: Lorenz model – testing data (blue) vs. Prediction data (orange)

4.2 Random Forest

4.2.1 Overview

We got the relevant results separately for static data and time series data. From the angle of evaluation criteria, the overall performance of Random Forest is quite ideal. The results of the static part are slightly better than the time series part which means the static features might have a stronger relationship with NPP than those of the time series. For the static part, we figured out the distribution of the predicted NPP values and thought about the reasons for its high accuracy. For the time series part, we finally found the best feature combination plus NPP with the highest model accuracy.

4.2.2 Static Data Result

For static data, from Figure 15, it can be known that the best tree number is 150 and the highest accuracy is around 0.9655. We decided to choose it as our best parameter for model evaluation.

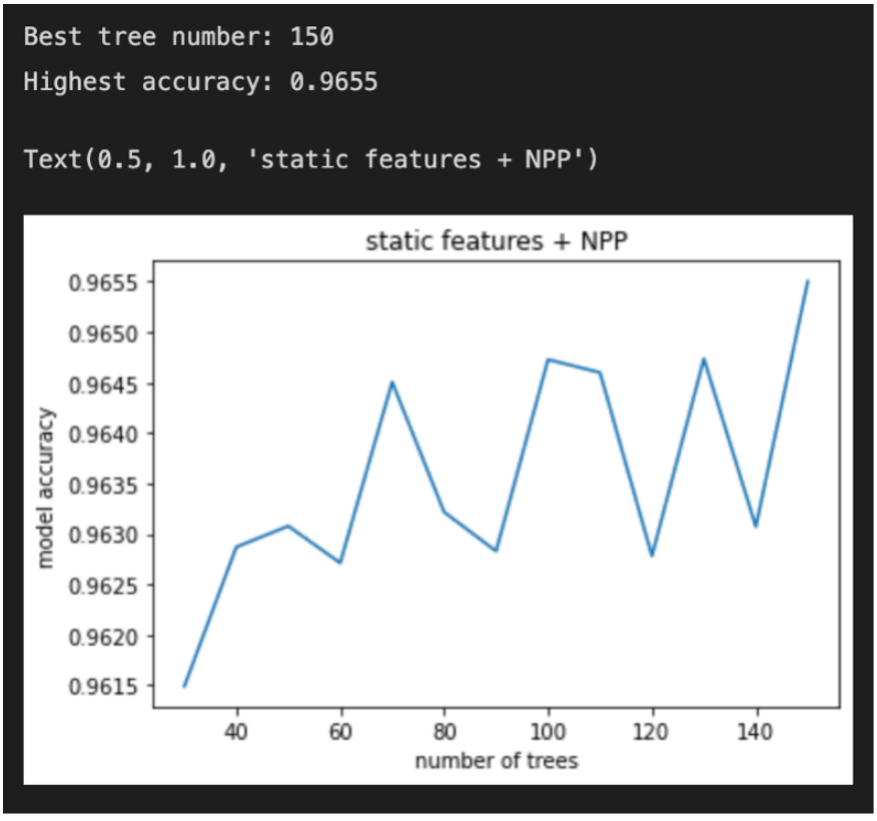


Figure 15: Best parameters (Tree Number) for the Random Forest model

Then, Mean Squared Error (MSE) and R-Squared score can be used as criteria to evaluate the model. These two criteria could measure how well a model fits the dataset. As is shown in the figure below, for the testing set, the MSE

between the predicted value and true NPP is around 0.026, which is close to zero. It reflects that the data points are dispersed closely around its central moment. It means that the regression is close to the line of best fit and indicates that the model is quite good. The R-Squared score is around 0.966, which means that around 97% of the observed variations can be explained by the model's inputs. It can be interpreted that it has a high level of correlation between static features and NPP values. What is more, from the figure below (Figure 16), it shows that at most of the training and testing residual points are around zero. It measures how well the regression line fits the individual data points. It appears to indicate that the model fits the data well.

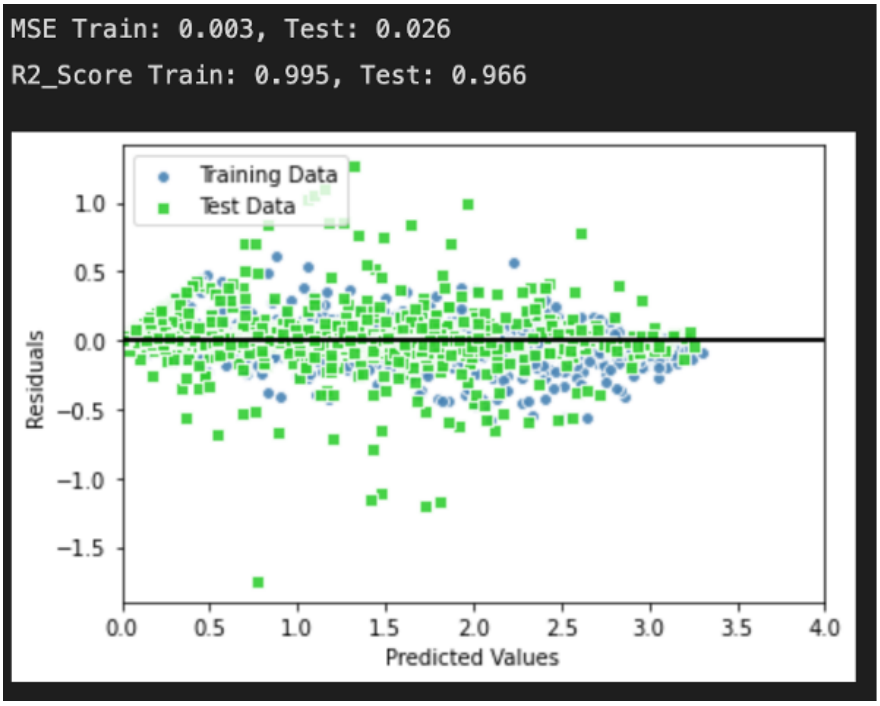


Figure 16: Criteria result of the model

For the high accuracy of the model, we can analyze that the difference between the predicted feature, NPP, is small. The values are similar. The static features also have this issue. It may lead to the issue of gradient exploding. The error gradient is accumulated in the updating process to obtain a very large gradient, which will greatly update the network parameters and thus lead to network instability. In addition, the static features may have a high correlation with NPP. It helps us deeply understand the relationship between the features. It also provides some guidance for analyzing time series data later. Compared to the time series data, static data has some limitations. For example, it does not have a time dimension, and it can not predict future data.

4.2.3 Time Seris Data Result

From the results of 15 different combinations of features plus NPP as follows, we got the average accuracy of the different number of features plus NPP. It can be seen that the interval of accuracy is from 0.829 to 0.865. Among these

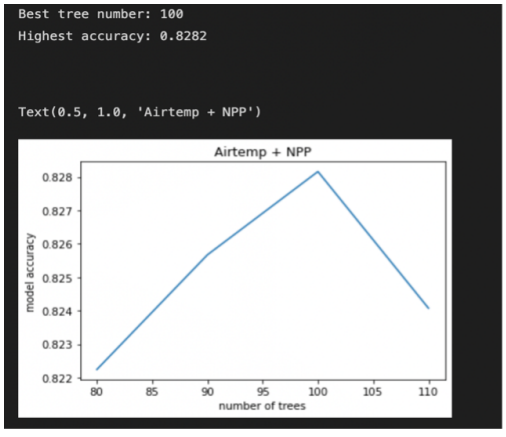
combinations, all four features plus NPP had the highest accuracy. The accuracy increases with the number of features increases. The increasing trend is obvious, and it implies that all 4 features which are Airtemp, Netlong, Shortwave and Precipitation contribute to the improvement of prediction which means all of them are related with NPP value.

The performance of Random Forest on time series data is quite good. The mean squared error and R-Squared score is stable among all 15 different combinations. Its overall accuracy is appropriate as the benchmark model, neither overfitting nor underfitting. Therefore, it meets our expectations and reveals the correlation between time series features and NPP value in the time dimension.

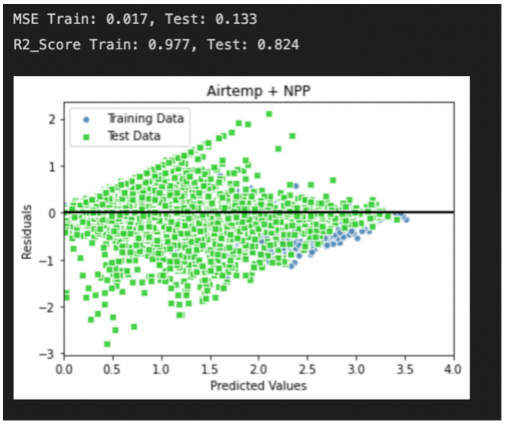
1. **Single Feature + NPP:**

   To conclude, in this experiment, we trained four Random Forest models using the dataset with one feature and NPP. From the plots below, the best-predicted accuracy of the models is around 82% to 83%. Besides, the MSE of the models is close to 0, and the R-Squared score is close to 1. Most of the training and testing residual points are around zero. It seems that the models fit the data well.

   (a) AirTem + NPP: This experiment test the model effect of one time-series predictor of Air Temperature on NPP



(a) Airtemp+NPP's accuracy



(b) Airtemp+NPP's criteria

Figure 17: Accuracy and Criteria Result of the Model Airtemp + NPP

   (b) netLong + NPP: This experiment tests the model effect of one time-series predictor of Net Long Radiation Flux on NPP

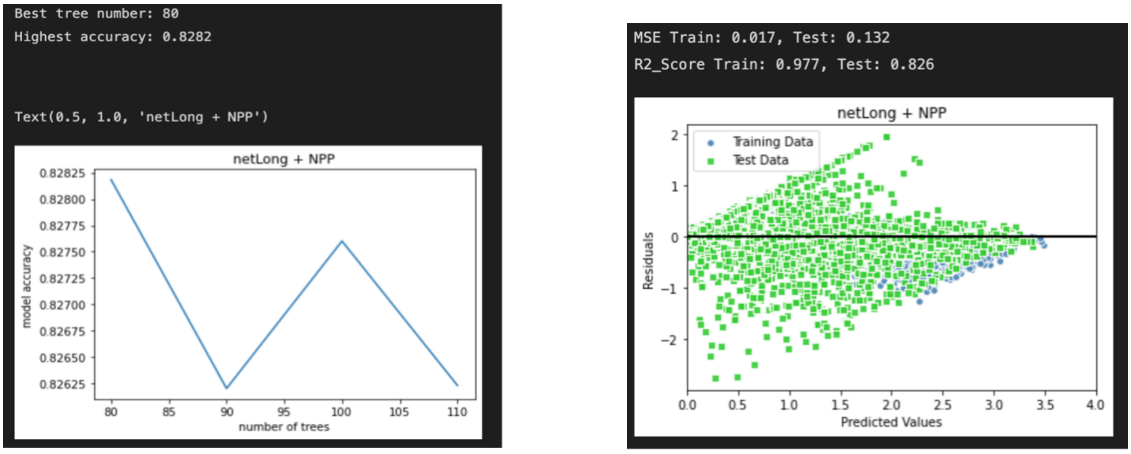


Figure 18: Accuracy and Criteria Result of the Model Netlong + NPP

(c) Shortwave + NPP: This experiment tests the model effect of one time-series predictor of Shortwave Radiation on NPP

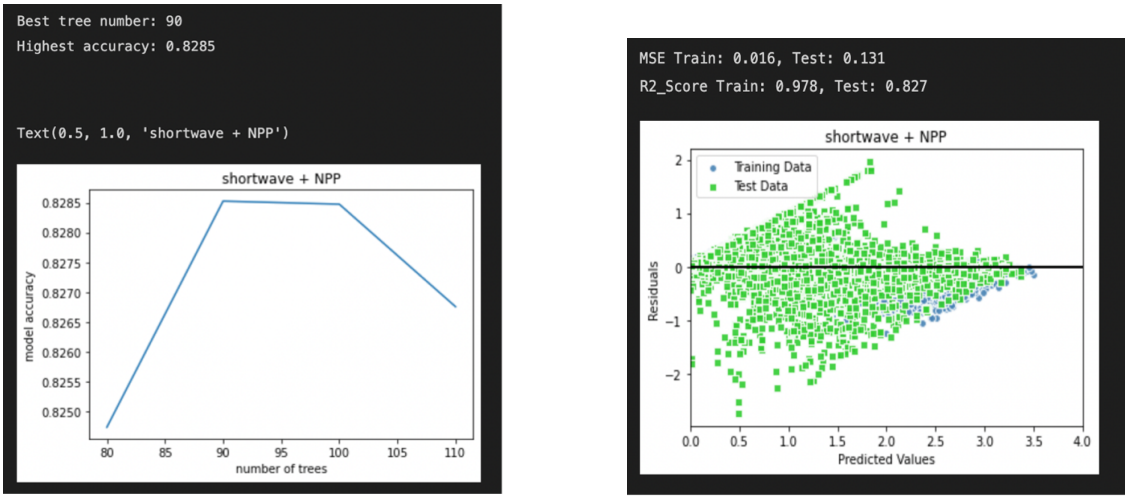


Figure 19: Accuracy and Criteria Result of the Model Shortwave + NPP

(d) Precipitation + NPP: This experiment tests the model effect of one time-series predictor of Precipitation on NPP

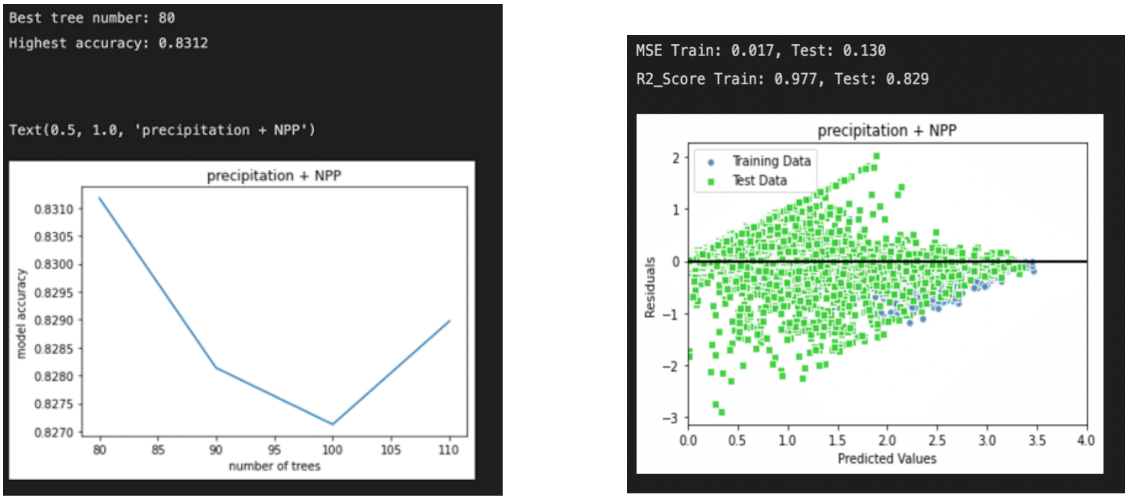


Figure 20: Accuracy and Criteria Result of the Model Precipitation + NPP

2. **Two Features + NPP:**

   Among all the two-feature experiments, we trained a total of six Random Forest models. From the plots below, the best-predicted accuracy of the models is around 83% to 86%. Similar to the above, the MSE of the models is close to 0, and the R-Squared score is close to 1. Most of the training and testing residual points are around zero. But the combination of Air Temperature + Precipitation with NPP seems to work the best with the highest predict-score of 0.8544. It seems that the models fit the data well.

   (a) AirTem + Netlong + NPP: This experiment tests the model effect of two time-series predictors of Air Temperature and Net Long Radiation Flux on NPP

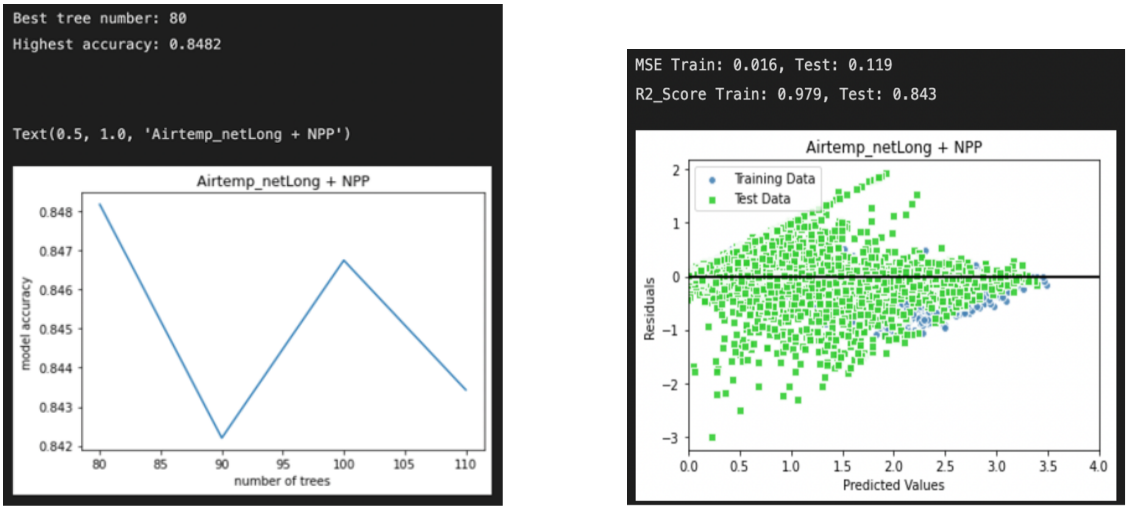


Figure 21: Accuracy and Criteria Result of the Model Air Temperature and Net Long Radiation Flux + NPP

   (b) AirTem + Shortwave + NPP: This experiment tests the model effect of two time-series predictors of Air Temperature and Shortwave Radiation on NPP

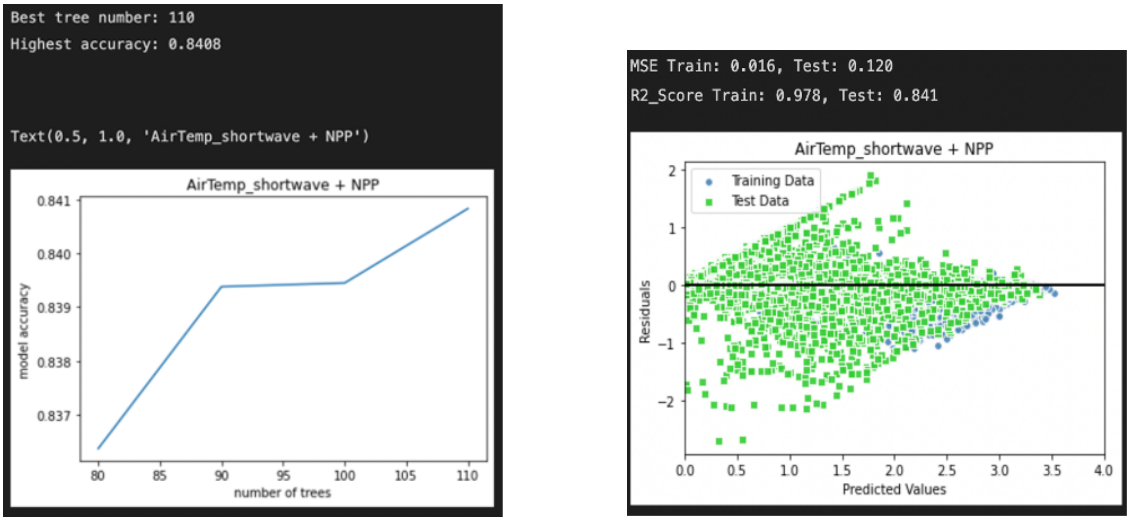


Figure 22: Accuracy and Criteria Result of the Model Air Temperature and Shortwave Radiation + NPP

(c) AirTem + Precipitation + NPP: This experiment tests the model effect of two time-series predictors of Air Temperature and Precipitation on NPP

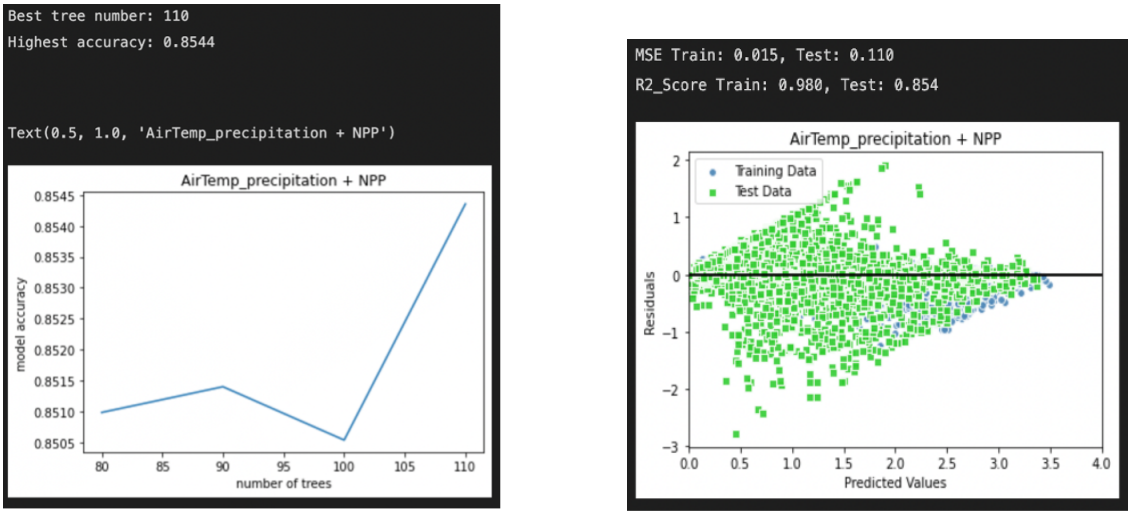


Figure 23: Accuracy and Criteria Result of the Model Air Temperature and Precipitation + NPP

(d) Netlong + Shortwave + NPP: This experiment tests the model effect of two time-series predictors of Net Long Radiation Flux and Shortwave Radiation on NPP

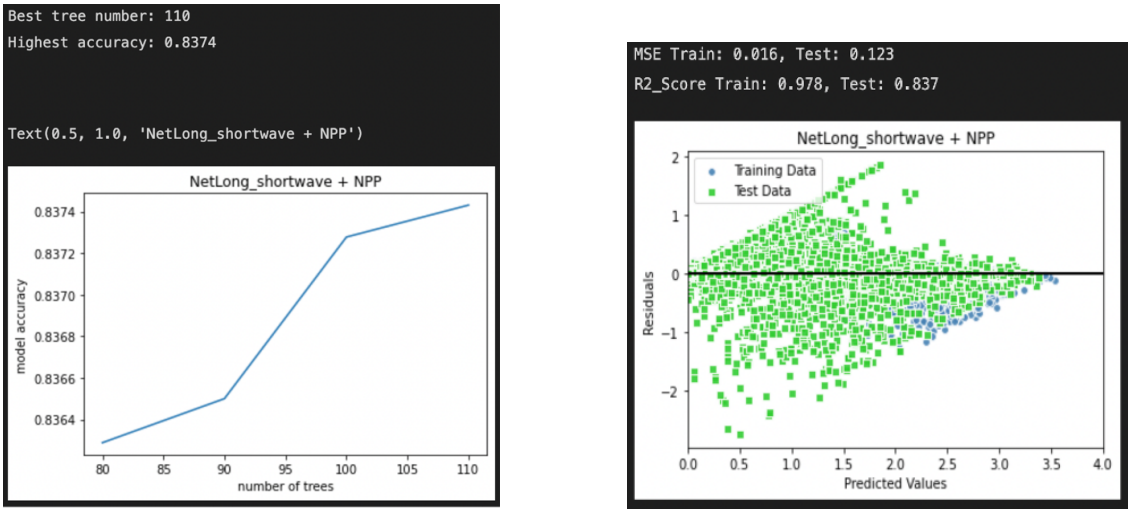


Figure 24: Accuracy and Criteria Result of the Model Net Long Radiation Flux and Shortwave Radiation + NPP

(e) Netlong + Precipitation + NPP: This experiment tests the model effect of two time-series predictors of Net Long Radiation Flux and Precipitation on NPP

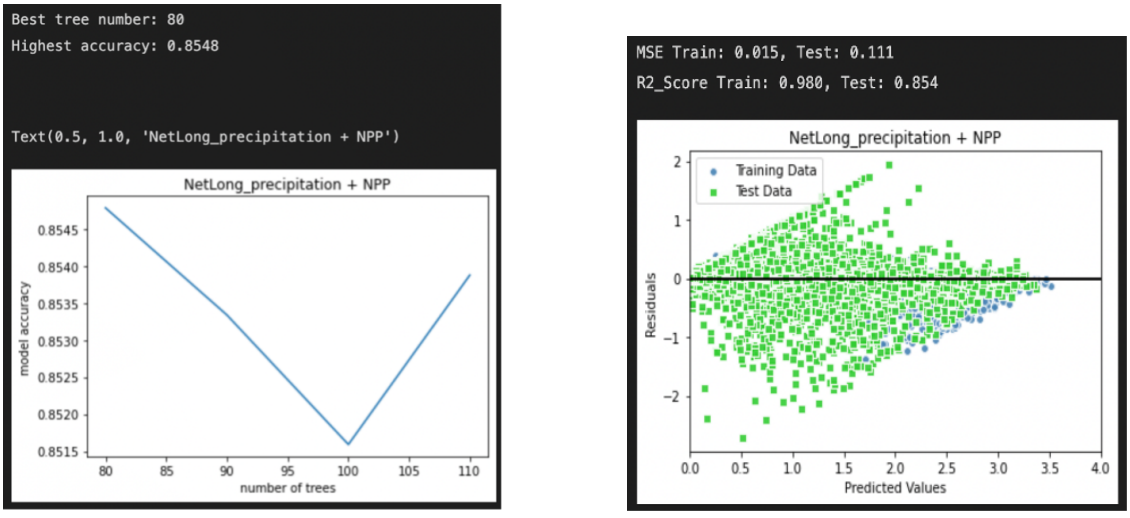


Figure 25: Accuracy and Criteria Result of the Model Net Long Radiation Flux and Precipitation + NPP

(f) Shortwave + Precipitation + NPP: This experiment tests the model effect of two time-series predictors of Shortwave Radiation and Precipitation on NPP

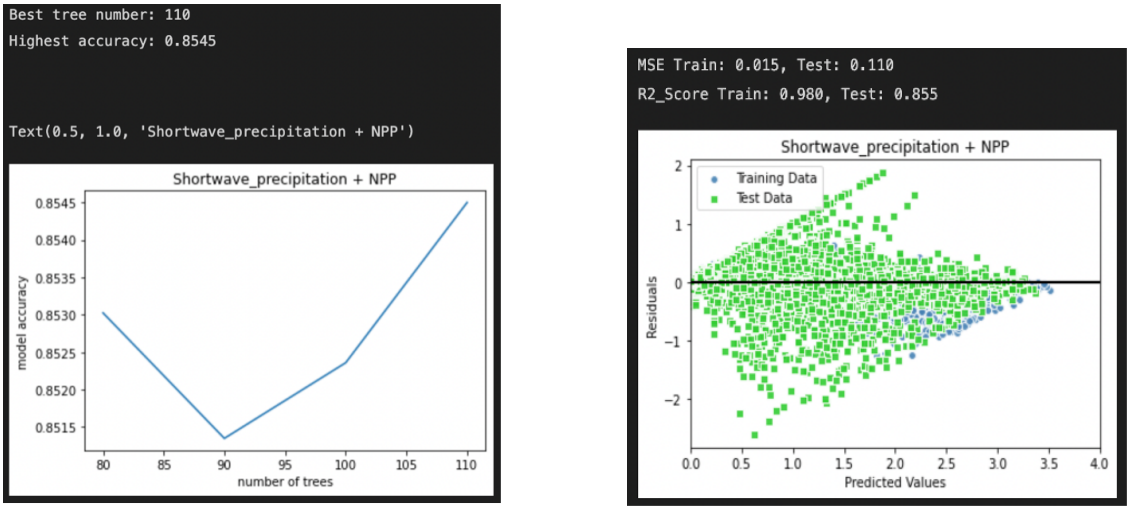


Figure 26: Accuracy and Criteria Result of the Model Shortwave Radiation and Precipitation + NPP

3. **Three Features + NPP:**

   During the thrid experiment setting, we trained four Random Forest models using the dataset with three features and NPP. As the resulting plot indicated, the best-predicted accuracy of the models increased to around 85% to 86%. There's the trend of increasing relative predictors may result in better model effects. The combination of Net Long Radiation Flux, Shortwave Radiation, and Precipitation has the best-resulted predict-score of 0.8611.

   (a) Airtemp + Netlong + Shortwave +NPP: This experiment tests the model effect of three time-series predictors of Air Temperature, Net Long Radiation Flux and Shortwave Radiation on NPP

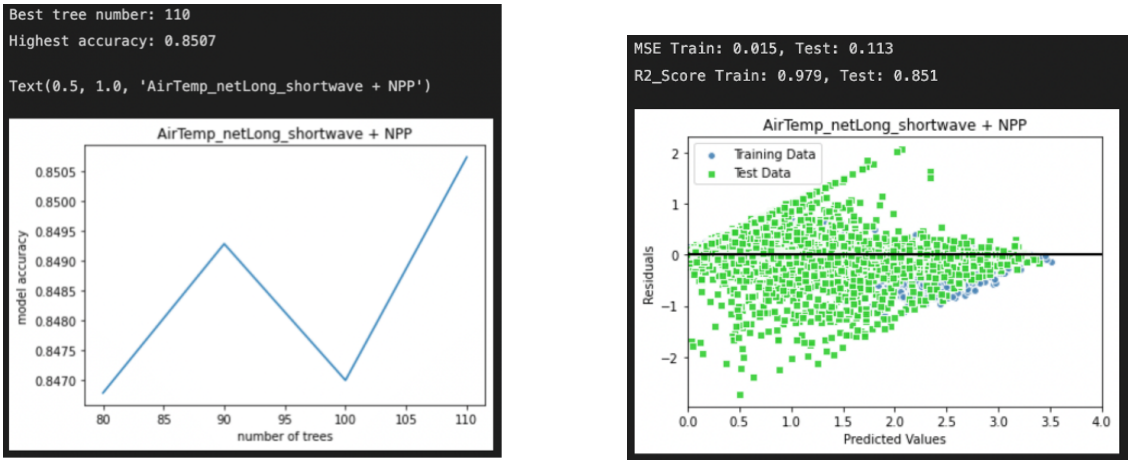


Figure 27: Accuracy and Criteria Result of the Model Air Temperature, Net Long Radiation Flux and Shortwave Radiation + NPP

   (b) Airtemp + Netlong + Precipitation +NPP: This experiment tests the model effect of three time-series predictors of Air Temperature, Net Long Radiation Flux and Precipitation on NPP

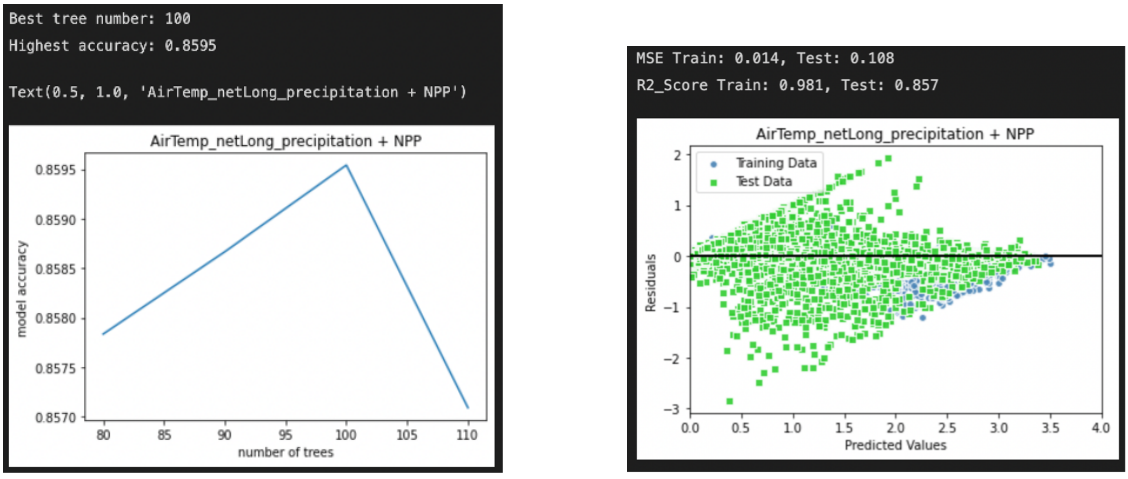


Figure 28: Accuracy and Criteria Result of the Model Air Temperature, Net Long Radiation Flux and Precipitation + NPP

(c) Airtemp + Shortwave + Precipitation +NPP: This experiment tests the model effect of three time-series predictors of Air Temperature, Shortwave Radiation and Precipitation on NPP

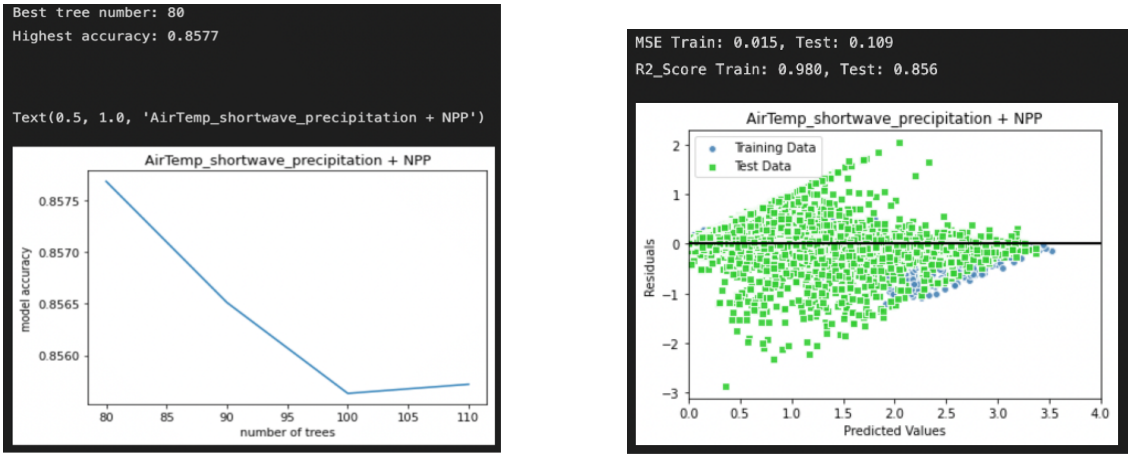


Figure 29: Accuracy and Criteria Result of the Model Air Temperature, Shortwave Radiation and Precipitation + NPP

(d) Netlong + Shortwave + Precipitation +NPP: This experiment tests the model effect of three time-series predictors of Net Long Radiation Flux, Shortwave Radiation, and Precipitation on NPP

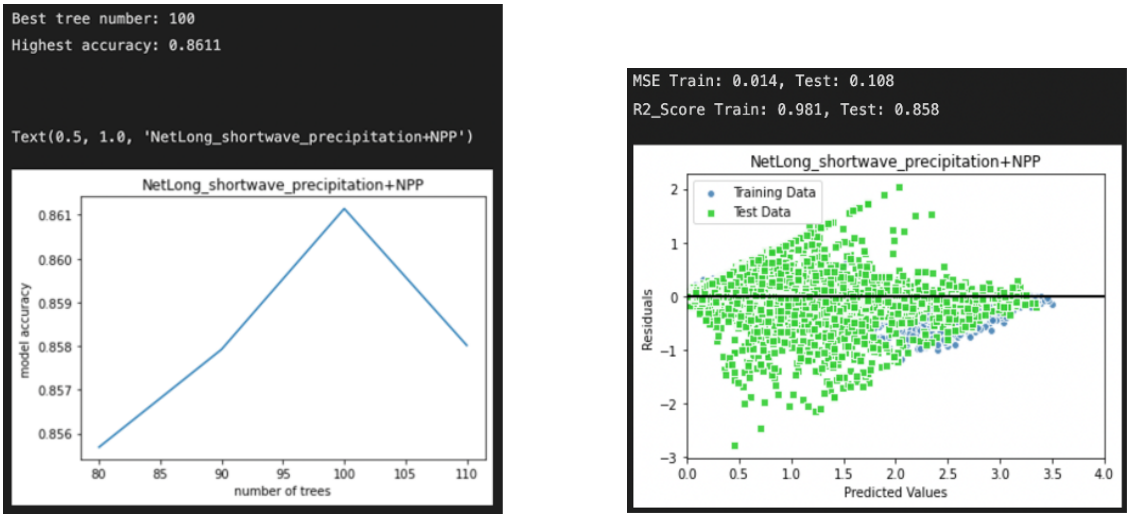


Figure 30: Accuracy and Criteria Result of the Model Net Long Radiation Flux, Shortwave Radiation, and Precipitation + NPP

4. **All Four Features + NPP:**

   This experiment tests the model effect of all four time-series predictors of Air Temperature, Net Long Radiation Flux, Shortwave Radiation, and Precipitation on NPP. It maks sense that using all predictors with NPP gives the best benchmark model with a predict-score of 0.8645.

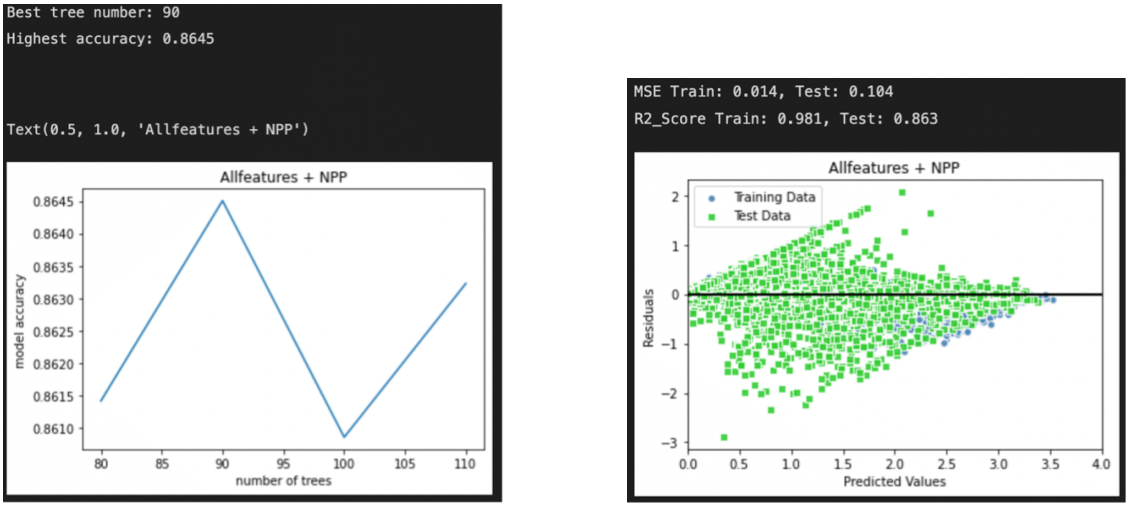


Figure 31: Accuracy and Criteria Result of the Model all four features + NPP

4.3 Clustering-K-means with DTW

The following graphs show the clustering result with South America as well as global NPP. For each instance of NPP time series, the first ten years are removed to prevent the initial drastically increasing effect to influence the model accuracy. Meaningless data such as NPP in the ocean are also taken out from consideration. The grey lines represent each instance of the time series while the red lines represent the means of each cluster. The graphs showed that the cluster captured the majority of the information.

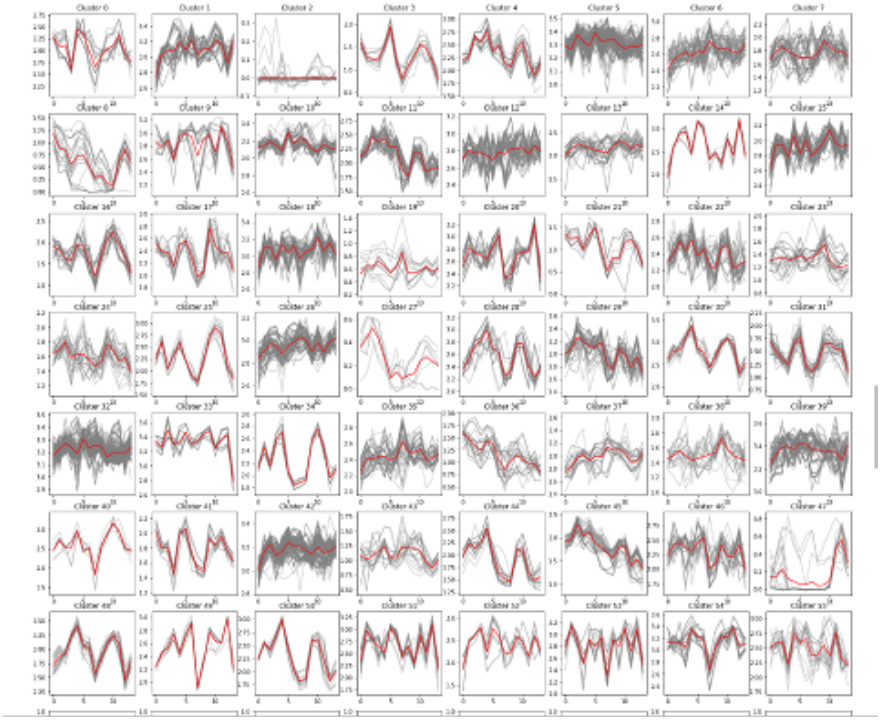


Figure 32: South America Clusters Result

Some of the clusters have darker gray background lines which indicate more geological grid points are classified into this cluster (e.g. cluster 5,32 and 42 etc.), Whereas lighter ones have fewer classified grid points' NPP data (e.g. cluster 27, 40, and 49, etc.). Larger clusters suffer from dataset's noise as we can see the value range of larger clusters is greater than smaller ones. Whereas smaller clusters could be somehow affected by the existence of outliers as cluster 3 as we could see that the fitting result of cluster 3 is not as good as expected. This can be caused by the fact that the partition-based clustering method works really well on natural spherical-shaped clusters but lacks the ability to capture arbitrary-shaped clusters. And in our case with the illustration of Figure 8 from above, we know that there are geological properties of these NPP clusters which don't necessarily follow a spherical-shaped cluster, thus this clustering method could be further improved by density-based clustering techniques such as DBSCAN.

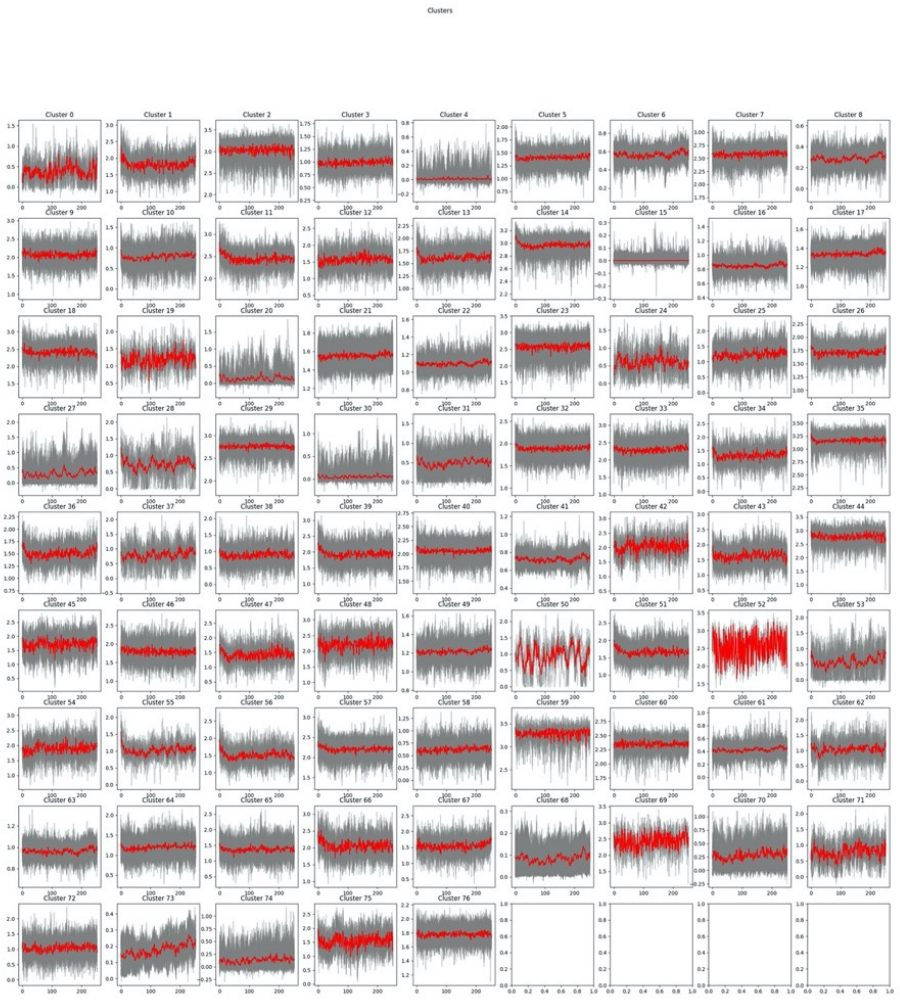


Figure 33: Global clusters Result

The clustering task for the global dataset turns to be more computationally expensive than South America one. We see more darker background lines from the global dataset's clustering result as we constrained the number of clusters. We use the same size of clusters as when clustering South America dataset which is 77. With now the effectively reduced dimensions, we pave the path for combining clustering with following ARIMA and LSTM.

4.4 ARIMA+Clustering

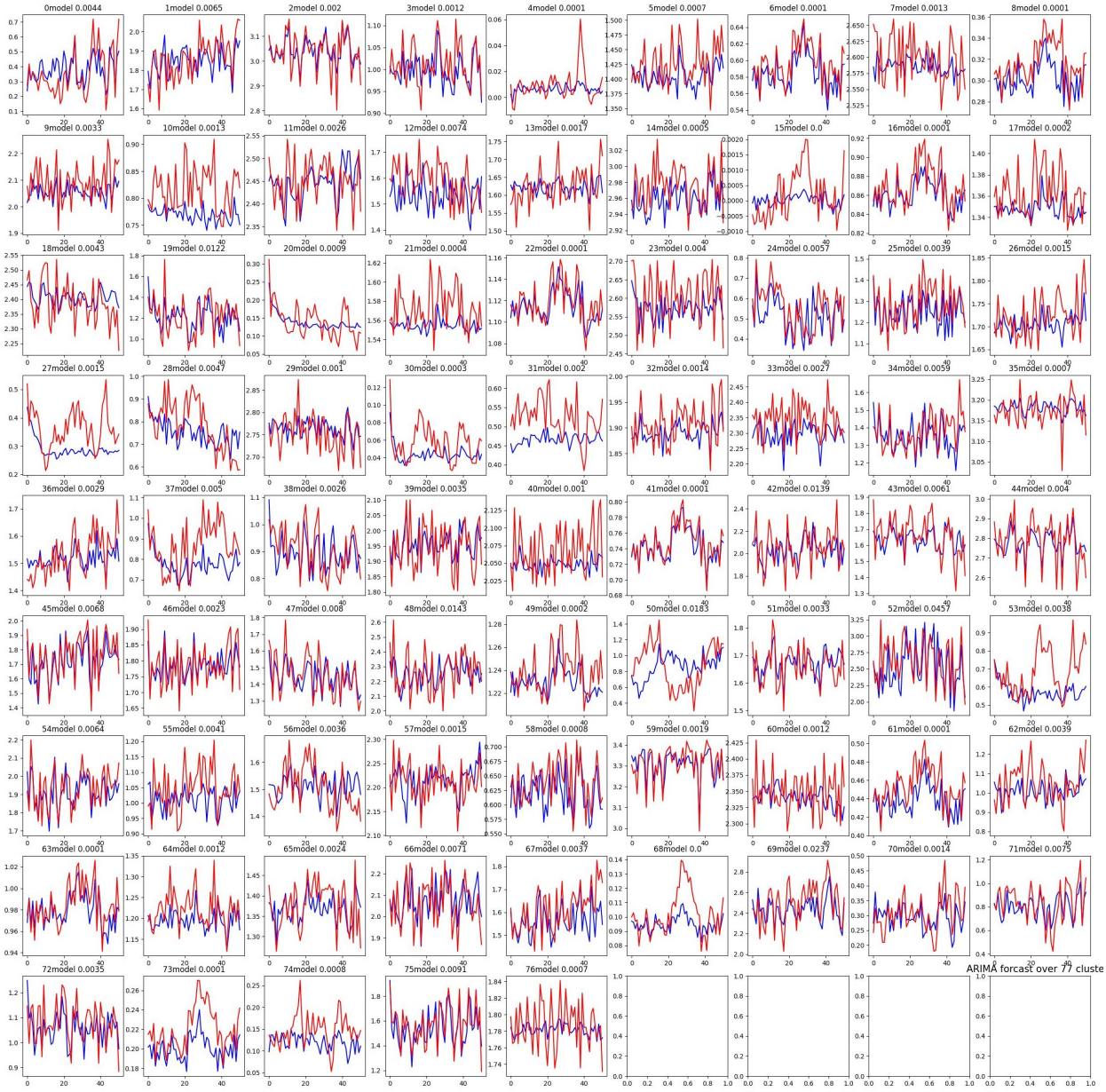


Figure 34: ARIMA + Cluster result

For each cluster, 77 models are created based on 77 clusters. Moreover, the above graph demonstrates that the prediction on 51 time steps NPP with the same time steps exogenous factors. As the red line represents the real NPP and the blue line represents the projection NPP, we determined that the majority of the models had the capacity to accurately explain and forecast the data.

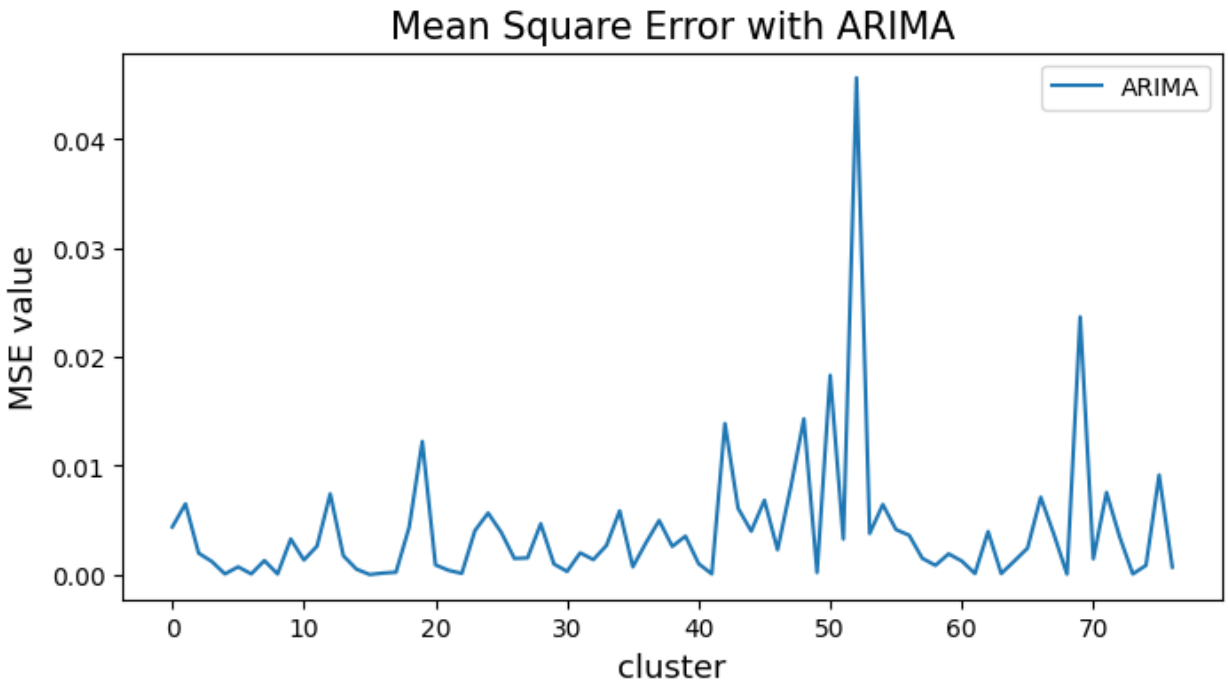


Figure 35: ARIMA + Cluster MSE

In addition, the MSE plot reveals that the majority of models have MSE values between 0.00 and 0.02. Comparing the ARIMA to the random forest baseline model, the ARIMA has a comparable MSE, but the ARIMA in cluster has a greater capacity to explain more subtleties owing to its classification of each location into a cluster based on Dynamic Time Warping. In addition, the cluster ARIMA is able to forecast numerous predictions, achieving the second objective of the research.

4.5 LSTM+Clustering

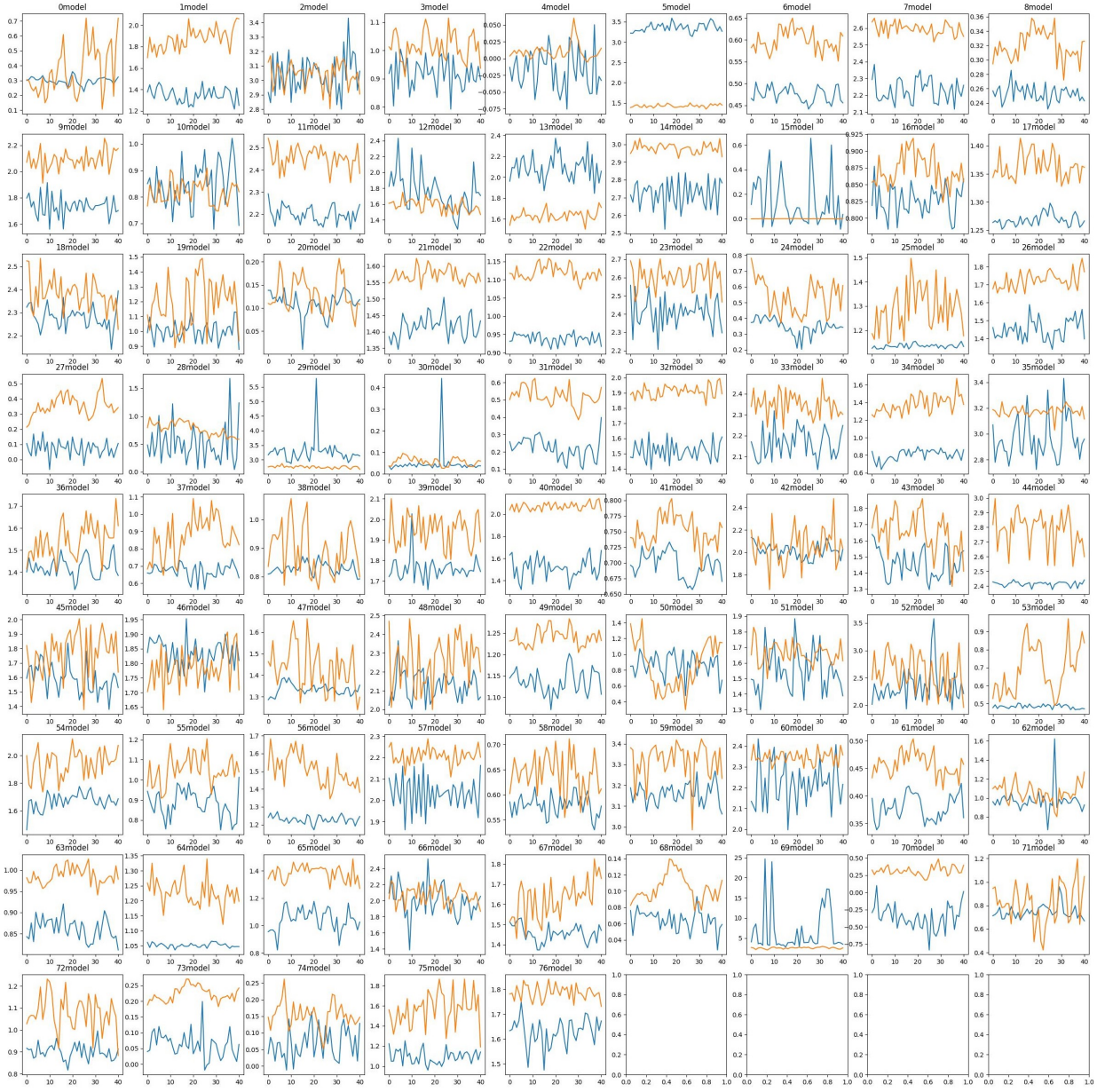


Figure 36: LSTM + Cluster result

LSTM, like ARIMA in cluster, applies each model to each of the 77 clusters. AS the LSTM use previous 9 variable to predict future 1 value which is the first goal of our project. As the blue line represents the prediction and the yellow line represents the origin, we discovered that a number of models fail to adequately explain and forecast the data under a variety of scenarios.

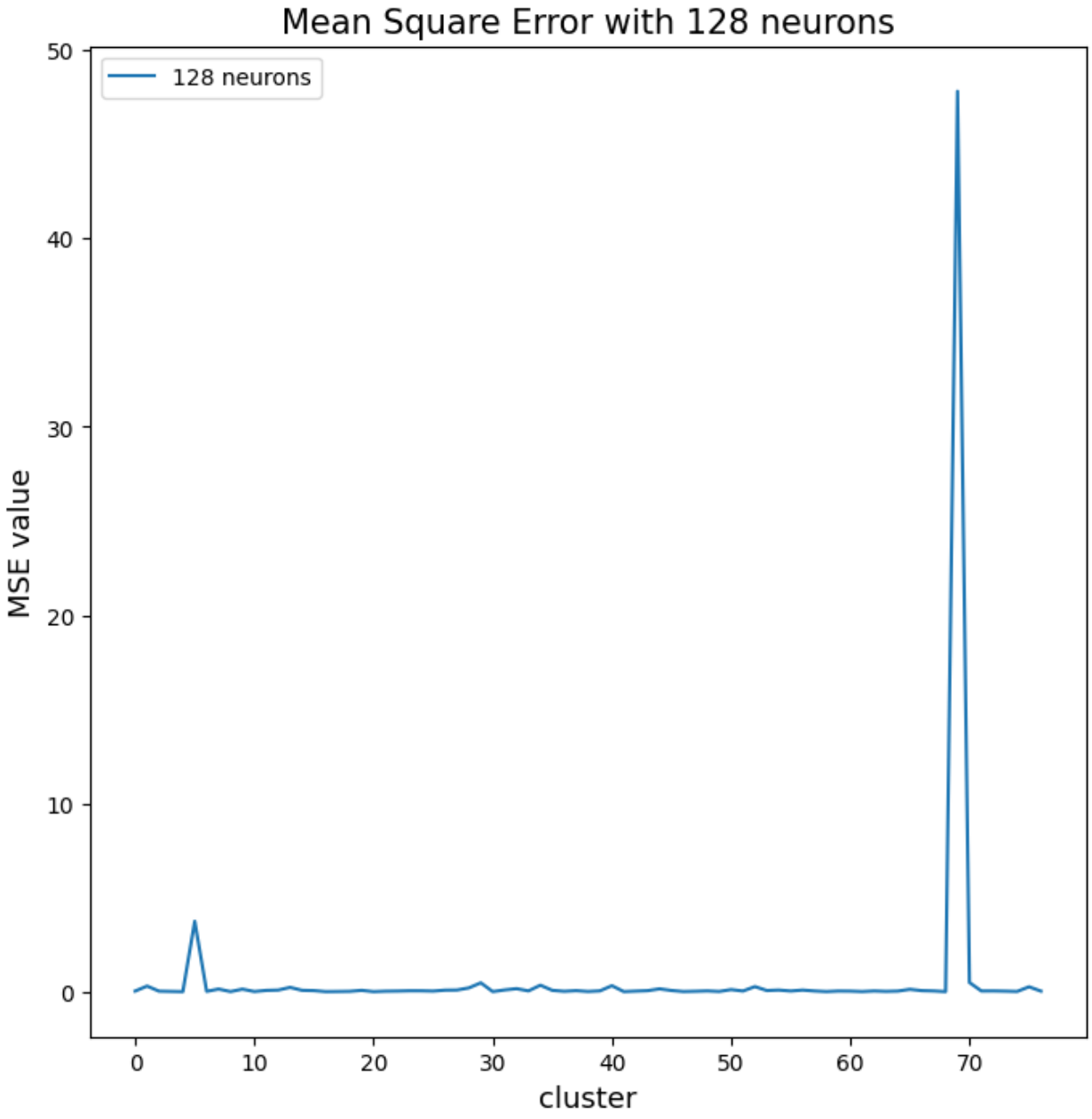


Figure 37: LSTM + Cluster MSE

Also, the MSE graph indicates that LSTM's performance is inferior to that of ARIMA and Random forest. However, LSTM also offers benefits over black box modelling, which is the model that simulates JEDI data the most accurately. As neural networks are data-hungry, the performance of LSTM will improve as more data is collected or as the season mean rather than the yearly mean based on 3516 months is utilised.

5 Conclusion and Achievements

5.1 Conslusion

Throughout this project, we have developed multiple machine-learning techniques and algorithms on a large-scale Jedi-DGVM dataset, and also performed two different tasks related to a simulation process. We adopted Random Forest as the main regressor for task 1's short four-step-further prediction of NPP. Two experimental settings are carried out using only static predictors or time-dynamic predictors respectively, as we were intended to test should NPP be more determined by static factors (e.g. latitude, longitude, moisture, or elevation), or by time-dependency factors (e.g. air temperature, net long radiation flux, shortwave radiation, or precipitation). Both static and tie-dependency group indicated good prediction result, but since the group of static experiments achieved the highest predict-score of 0.9655 we conclude that when both static and dynamic predictors are available, the choice of static variables may lead to a better prediction result.

Then when it comes to the second task of whole process emulation, a primary clustering method is applied to the global-scale dataset which aims to reduce the number of regressors at a lower computational cost. Then ARIMA and LSTM are carried out on the resultant 77-clustered dataset for doing the emulation task. The majority ARIMA models have MSE between 0.00 and 0.02 which seems to be a pretty good result, whereas most of the LSTM models have a result MSE within the range 0 to 5 which has slightly inferior performance.

5.2 Achievement

To present a deliverable result, we attached our main application development result as a GitHub repository link as well as our presentation link as follows:

1. Github link: `https://github.com/QuYuze/VegeSimulation`

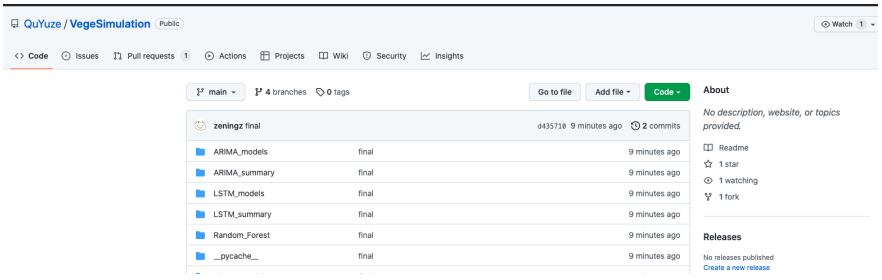


Figure 38: Git Repository

2. Presentation link: `https://youtu.be/0DoxLSOES4Q`

Figure 39: Youtube Presentation

5.3 Future Direction

Considering the fact that we are using RM as a benchmark model, such a high-accuracy result may be caused by some overfitting problem. As mentioned above we adopted the global-scale dataset due to the geological interpolation effect on the South-American dataset which could give overfitting problems. However, one more factor that may lead to this problem is we were actually using the whole 293-year time sequence to do short-step prediction as the model could learn existence patterns over the whole history. Thus one improvement on the Random Forest algorithm is to use shorter historical training data to predict the longer future sequences.

Further improvements can also be discussed related to ARIMA and LSTM. ARIMA has a restriction of mainly capturing the linear relationship of its targets and exogenous predictors. This can be slightly improved by doing logistic interpolation when sampling the historical data, or adopting seasonal data in our case. LSTM, on the other hand, requires much longer historical data as the general deep learning method depends on vast amount of training data.

Taking a step further, an emulator is expected to perform as a well-functioning black-box model, where given an initial state of some factors it has the ability to generate the original simulation process and finally comes to a resultant value of the target within a user-defined time-horizon. Our data science project only paves the path for further exploration of this exciting research field pursued by a great amount of machine learning researchers, and more complex or suitable models are under discovering for us in the future.

6 Appendix

6.1 Contribution

Group Member	Contributions
Zhili Chen	Random Forrest Regressor, model Benchmark, feature engineering, research on the Jedi model, RP-model simulation, and preliminary analysis
Jiachen Huo	LSTM research, early iteration frameworks, project documentations and coordination, RP model research,
Yuze Qu	High-level coordination and progress tracking, clustering algorithm, simulation, and analysis of Lorenz models
Ziqian Wu	Random Forrest Regressor, model benchmark, feature engineering and data cleaning, RP model preliminary analysis and simulation
Zening Zhang	Research on ARIMA Family Model selection Research. Simulation and model development with ARIMA and LSTM in the collaboration with clustering

Table 3: Group Contributions

6.2 Metting Logs

Meeting Time & Date	Meeting Contents
April 1st, 2022, 11:00am - 12:00pm	-First Client meeting -Introductions -Communication and version control Setup (email, github, slack etc.)
April 8th, 2022, 2:00 - 3:00pm	-Decision on the targeting model (JeDi-DGVM) of emulation -Discussion of the project objectives at a high-level.
April 14th, 2022, 2:00 - 3:00pm	-Discussion on targeting model required input output as well as parameters estimates
April 29th, 2022, 2:00 - 3:00pm	-Discussion on model for preliminary analysis to get familiar with the final model - Discussion on Lorenz model and p-model
May 9th, 2022, 2:00 - 3:00pm	-Discussion on preliminary analysis with Lorenz model and p-model: regression for time series prediction as well as static data predictions. -Propose the method for emulations: random forest regressor, XG-boost. - Questions on model understanding and relation to the target model
May 18th, 2022, 2:00 - 3:00pm	-Preparation for presentation and report for Semester 1 -Discussion on current progress - Requirement clarifications -Project report discussion

Table 4: Data Science Project-Part 1 Meeting Logs

Meeting Time & Date	Meeting Contents
August 4th, 2022, 1:00 - 2:00pm	-Discussion on simulated data output by JeDi model -Emulation objectives clarifications -Discussion on JeDi model input-output reinstated
August 11th, 2022, 1:00 - 2:00pm	-Discussion on analysis of JeDi input and output
August 18th, 2022, 1:00 - 2:00pm	-Dicussion on ARIMA as potential solution
August 25th, 2022, 1:00 - 2:00pm	-Potential solution: LSTM, ARIMA, VARIMA -Model benchmarking: Random Forest Regressor, XGBoost
September 5th, 2022, 1:00 - 2:00pm	-Consult the client/supervisor about the unexpected result of the benchmark -Ask for the possibility of getting longer time series for the data-hungry models (eg. LSTM) -Bring up hierarchical regression
September 12th, 2022, 1:00 - 2:00pm	-Bring up that clustering is a more realistic approach rather than hierarchical regression -Discussion on the high accuracy of random forest regressor
September 19th, 2022, 1:00 - 2:00pm	-LSTM recent result and ask for supervisor/clients' feedback -results for clustering time series with DTW -Random Forest Regressor recent results
September 20th, 2022, 10:00 -11:00am	-Discussion with Course coordinator on the current progress
October 13th, 2022, 1:00 - 2:00pm	-Show the recent emulation result to the client and the supervisor -prepare for the presentation
October 19th, 2022, 1:00 - 2:00pm	-Discussion on the final report and presentation -Finalize the model and emulation results
October 23th, 2022, 1:00 - 2:00pm	-Discussion of the report write up -Distribute the work needed for the final report
October 26th, 2022, 1:00 - 2:00pm	-Check the current progress of the final report -Adjustments on work done so far
October 29th, 2022, 1:00 - 2:00pm	-Report finalized -Work done checklist -Peer review reminder

Table 5: Data Science Project-Part 2 Meeting Logs

References

[ADSC13] Bruce H Andrews, Matthew D Dean, Robert Swain, and Caroline Cole. Building arima and arimax models for predicting long-term disability benefit application rates in the public/private sectors. *Society of Actuaries*, pages 1–54, 2013.

[BAF+19] Helizani Couto Bazame, Daniel Althoff, Roberto Filgueiras, Maria Lúcia Calijuri, and Julio Cesar de Oliveira. Modeling the net primary productivity: A study case in the brazilian territory. *Journal of the Indian Society of Remote Sensing*, 47(10):1727–1735, 2019.

[FMH19] Amir Farzad, Hoda Mashayekhi, and Hamid Hassanpour. A comparative performance analysis of different activation functions in lstm networks for classification. *Neural Computing and Applications*, 31(7):2507–2521, 2019.

[FONU18] Olutoyin A. Fashae, Adeyemi O. Olusola, Ijeoma Ndubuisi, and Christopher Godwin Udomboso. Comparing ann and arima model in predicting the discharge of river opeki from 2010 to 2020. *River Research and Applications*, 35(2):169–177, 2018.

[GMM+03] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: Theory and practice. *IEEE Transactions on Knowledge and Data Engineering*, 15(3):515–528, 2003.

[LZX+22] Song Li, Rui Zhang, Lingxiao Xie, Junyu Zhan, Yunfan Song, Runqing Zhan, Age Shama, and Ting Wang. A factor analysis backpropagation neural network model for vegetation net primary productivity time series estimation in western sichuan. *Remote Sensing*, 14(16):3961, 2022.

[PDB+13] Ryan Pavlick, Darren T Drewry, Kristin Bohn, Björn Reu, and Axel Kleidon. The jena diversity-dynamic global vegetation model (jedi-dgvm): a diverse approach to representing terrestrial biogeography and biogeochemistry based on plant functional trade-offs. *Biogeosciences*, 10(6):4137–4177, 2013.

[RLB+21] M. Röhrl, F. Listl, V. Brandstetter, T. Schulze, and T. Runkler. Surrogate modeling based on dynamic numerical simulation and measurements for fast emulation. *14th WCCM-ECCOMAS Congress*, 2021.

[RMC+22] Carine M. Rebello, Paulo H. Marrocos, Erbet A. Costa, Vinicius V. Santana, Alírio E. Rodrigues, Ana M. Ribeiro, and Idelfonso B. Nogueira. Machine learning-based dynamic modeling for process engineering applications: A guideline for simulation and prediction from perceptron to deep learning. *Processes*, 10(2):250, 2022.

[SC21] Paola Stolfi and Filippo Castiglione. Emulating complex simulations by machine learning methods. *BMC Bioinformatics*, 22(S14), 2021.

[SM19] Ralf C Staudemeyer and Eric Rothstein Morris. Understanding lstm–a tutorial into long short-

term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*, 2019.

[Whi53] Peter Whittle. Estimation and information in stationary time series. *Arkiv för matematik*, 2(5):423–434, 1953.

[ZGM+21] Rui Zhang, Zhen Guo, Yujie Meng, Songwang Wang, Shaoqiong Li, Ran Niu, Yu Wang, Qing Guo, and Yonghong Li. Comparison of arima and lstm in forecasting the incidence of hfmd combined and uncombined with exogenous meteorological variables in ningbo, china. *International journal of environmental research and public health*, 18(11):6174, 2021.

[Zha03] G Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.