The Bartlett School of Environment, Energy and Resources

MSc ESDA Title Page

UCL Candidate Code: Team Orange (KNCM2, KNHB8, NMHR0, KCSH2)

Module Code: BENV0091

Module Title: Energy Data Analysis

Coursework Title: Predicting Sea Ice in the Arctic with Supervised Learning

Module Leader: Prof. Aidan O'Sullivan

Date: 14.12.2020

Word Count: 5979

PREDICTING SEA ICE IN THE ARCTIC WITH SUPERVISED LEARNING

A BENV0091 ENERGY DATA ANALYSIS GROUP REPORT
SUBMITTED TO BARTLETT SCHOOL OF ENVIRONMENT, ENERGY AND RESOURCES
OF UNIVERSITY COLLEGE LONDON
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

UCL Candidate Code: KNCM2, KNHB8, NMHR0, KCSH2
December 2020

Abstract

Predicting future environmental information is one of the most elusive and long-standing challenges in Artificial Intelligence. This report tries to tackle the problem of predicting the Arctic sea ice: how to implement multiple Machine Learning algorithms. On the one hand, we think that Supervised Learning algorithm can predict the Arctic sea ice level. On the other hand, if we want to build an optimal predictor to ensure that our predictions on the Arctic sea ice are useful to the further future research, we would need an efficient and concise strategy to analyse the performance of different Supervised Learning algorithms. Besides, we need more strategies to analyse different factors related to the level of sea ice.

In this project, we firstly focus on Linear Regression, Penalised linear regression, Penalised Polynomial Regression, Random Forest, and Neural Networks: the most common and famous Supervised Learning algorithms by far. Then, according to the testing results, we will choose a better algorithm after testing the performance to be the algorithm that predicts the Arctic sea ice, both in a normal situation and the selected special situations.

This report consists of two parts. In the first part, we aim to understand the advantages of predicting sea ice of the Arctic and previous research on applying all kinds of algorithms on the related topic, especially predicting the Arctic sea ice or the Antarctic ice sheets. Then we present our efforts at implementing effective Supervised Learning algorithms.

In the second part of this project, we first get our performance results with different Supervised Learning algorithms by K-fold method. This is the so-called now-forecasting. Then, based on the performance results, we predict the Arctic sea ice for the further ten years with Random Forest with different specific scenarios.

Acknowledgements

A special thanks to Prof. Aidan O'Sullivan and Thomas Falconer. They always have a very insightful, high-level view of the field while they are also uncommonly detail oriented and understands the nature of the problems very well. More importantly, they are extremely caring and supportive for both the module BENV0091: Energy Data Analysis and the module BENV0094: Statistics for Energy Analytics that we could not have asked for more.

Collaboration is a valuable lesson that we have learnt, which is also a precious part of our graduate stage. We thank each group member for the hardworking and collaboration. We are glad that our project topic is different from any other group project in this academic year. But most importantly, we hope it can inspire future research in the area of energy, environment, and any other related discipline. It was a very unique and rewarding experience for us.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The motivation for this project includes attention to one of the most serious environmental risks on the planet in the early 21st century, the accelerating global temperature.

According to the paper from Nature, climate change happened in the past 70 years [1]. It has already had effects on the environment around us. Glaciers are shrinking and ices are breaking up earlier on the lakes and rivers.

The greenhouse effect has been working since the formation of the earth. Without the greenhouse effect, the surface of the earth would be extremely cold, the temperature would drop to minus 20°C, the ocean would freeze, and life would not form. Most climate scientists agree that it is the human expansion that causes global warming [2]. As we know, carbon dioxide (CO2) is a significant component of the atmosphere. Atmospheric CO2 concentration has been increased by more than a third since the Industrial Revolution began [2]. More importantly, as shown in Figure 1.1, atmospheric carbon dioxide has exceeded the highest level in the past 400,000 years.

Therefore, what we are facing is not the issue of whether there is a greenhouse effect, but the issue that humans emit a large number of greenhouse gases into the atmosphere through the burning of fossil fuels, causing the drastic greenhouse effect and the earth's climate.

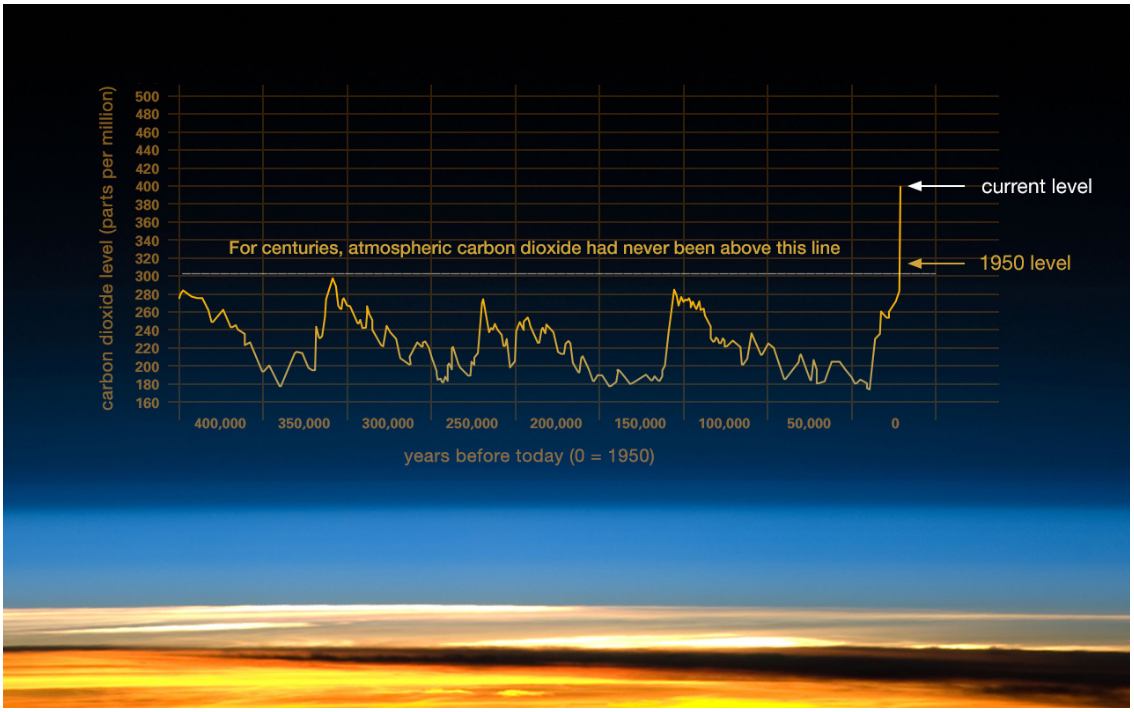


Figure 1.1: The evidence that atmospheric CO2 has increased since the Industrial Revolution began. Image courtesy: https://climate.nasa.gov/evidence

When the world's average temperature rises by 1°C, huge changes will occur: sea levels will rise, mountain glaciers will retreat, and snow-covered areas will shrink. As the global temperature rises, it will lead to uneven precipitation. In some areas, precipitation increases, while in others, precipitation decreases. For example, the Sahel region in West Africa has been severely arid since 1965, while in North China, precipitation has been decreasing year after year since 1965. Compared with the 1950s, precipitation in North China has been reduced by 1/3, and water resources have been reduced by 1/2. China's annual drought-affected area is about 400 million acres. In normal years, the national irrigation area lacks 30 billion cubic meters of water each year, and the cities lack 6 billion cubic meters of water.

When the average temperature of the world rises by 3°C, the world will suffer from food shortages. Due to rising temperatures, the global sea level has been rising at a rate of 1 to 2 millimetres per year in the past 100 years. It is expected that the sea level will continue to rise by 30 to 50 centimetres by 2050, which will flood a large amount of low-lying coastal land. Besides, due to climate changes have led to the aggravation of climatic disasters such as droughts, floods, and low temperatures, causing economic losses of more than tens of billions of dollars worldwide each year.

In this report, our team focuses on sea ice in the Arctic. We will look for the factors that we think affect the changes in Arctic sea ice, analyse their influence on the changes in sea ice, and select the two factors that we think are the most important. We will put these two influencing factors into our special situation model, and analyse the impact of significantly changing them on Arctic future sea ice.

1.2 Report Outline

This thesis consists of two parts — PART I ENVIRONMENTAL RISKS IN SUPERVISED LEARNING: FOUNDATIONS and PART II SCENARIO: PREDICTING ARCTIC SEA ICE IN SUPERVISED LEARNING.

PART I focuses on the task of understanding history development and the mathematical theory behind the Supervised Learning algorithms and building effective Supervised Learning algorithms so that we are able to predict Arctic sea ice in PART II.

In Chapter 2, we give an overview of the history and recent development of the field of environmental risks, especially:

(i) **why we choose to predict sea ice;**

(ii) **how to predict sea ice level using statistical models;**

(iii) **how to predict sea ice level using Machine Learning models;**

(iv) **how to predict sea ice level using non-statistical methods or non-machine learning methods.**

In Chapter 3, we mainly discuss five methodologies, which are:

(i) **feature description** that explains what these selected features are and why we choose these by referencing past researches;

(ii) **data** that simply introduces where our data comes from, and how to normalise our data sets to the range of zero to one;

(iii) **model accuracy** that introduces two methods that can evaluate the accuracy of our Machine Learning models: MEAN SQUARED ERROR and R-SQUARED;

(iv) **now forecasting methodology** that includes six different, important and suitable Supervised Learning algorithms that will be used to now-forecast the sea ice level with k-fold methods;

(v) **future forecasting methodology** that mainly explains why and how we designed three different scenarios to do future prediction.

PART II focuses on specific scenarios and we applied the Machine Learning algorithms and methodologies discussed above on these specific scenarios. Detailedly,

In Chapter 4, we get our performance results with different Supervised Learning algorithms via K-fold method.

In Chapter 5, based on the performance results above, we predict the Arctic sea ice for the further 10 years with Random Forest. In particular, 1) how we predict other environmental factors that influence the Arctic sea ice level; 2) how we analyse the impact of different environmental factors; 3) how we analyse the selected special situations with key factors.

We finally conclude and discuss the future work in this area in Chapter 6.

1.3 Contributions

The contributions of this thesis are summarised as follows:

- We analysed multiple factors influencing the sea ice level and gave a result of their correlation relationships, and used these relationships to implement our selected Supervised Learning models. We generated clear figures and summarised clear tables to compare the performance of different models.

- We designed some special situation models where policy-makers could have a clear insight that now is the time to take action to control the global temperature and reduce the excessive greenhouse effect.
- We create open source web page including Shiny App [1] and GitHub [2].

[1] Shiny App: https://grouporangearcticiceextent.shinyapps.io/shiny/?$_{g}a$ $=$ $2.114418575.940438834.1607903962 - 2063170584.1607903962$

[2] GitHub: https://github.com/UCL-BENV0091-Antarctic

Part I

Environmental Risks in Supervised Learning: Foundations

Chapter 2

Literature Review

2.1 Why We Choose to Predict Sea Ice

The change of Arctic sea ice is a sensitive indicator of climate change. The rate of Arctic sea ice disappearance exceeds even the most pessimistic climate model predictions. The current Arctic summer ice conditions are 30 years earlier than model predictions on average, and seasonal ice-free conditions are happened early [3]. Therefore, the prediction of sea ice is essential for understanding the future Arctic environment and global changes. Global warming has led to a reduction in sea ice and aggravated the deterioration of the Arctic environment, while the reduction in sea ice, in turn, has accelerated global warming. Increasing concentrations of greenhouse gases are playing an increasingly important role in the disappearance of the Arctic ice cap [4]. Many studies have also linked the loss of sea ice to atmospheric circulation patterns.

2.2 How to Predict Sea Ice Level Using Statistical Models

There are many statistical models that study the relationship between the Arctic sea ice concentration and climate factors. Stroeve et al. pioneered the use of singular value decomposition SVD, empirical orthogonal function EOF and other multivariate analysis techniques, focusing on the relationship between the decline of sea ice concentration in summer and winter and the warming trend and AO driving [3]. TIvy et al. used the multivariate analysis technique of Code Correlation Analysis (CCA) to compare the SIC value of Hudson Bay in July from 1971 to 2005 by comparing sea surface temperature, position altitude, sea level pressure, and regional surface air temperature. Among them, in the 6-month forecast period, the forecast results are the most accurate. Surface temperature in autumn is the most influential predictor [5]. Ahn et al. introduced the automatic regression integrated moving average (ARIMA) method for the first time into the sea ice concentration statistical model, using seven climatic factors (skin temperature, sea surface temperature, total column liquid water, total column water vapour, and instantaneous moisture flux). It is better at predicting large data sets than the single equation of the ordinary minimum hours (OLS) regression method and has higher accuracy. The average improvement of RMSE is 0.076 [6]. The forward stepwise regression model is used to predict summer sea ice conditions several months in advance based on four predictors which are winter multi-year combined concentration, total ice concentration in spring, North Atlantic Oscillation Index and East Atlantic Index where R^2 value is 90.7% and MAE value is 34 [7]. In 2003, it was expanded with multiple linear regression models [8]. Finally, in 2007, on this basis, Drobot and others created a multiple linear regression model (MLR) to predict the annual minimum Arctic sea ice range for the monthly interval from February to August. The forecast data is based on the average monthly weighted index of sea ice concentration (WIC), surface skin temperature (WST), surface illuminance (WAL) and surface long-pass volume (WDL). Each MLR model is better than the climatology model [9], but due to insufficient statistical time series modelling, the performance of machine learning models is often better than statistical models.

2.3 How to Predict Sea Ice Level Using Machine Learning Models

Machine learning models are often used to predict Arctic sea ice concentration and Arctic sea ice classification. Chi et al. used a large Arctic sea ice dataset to train a neural network to predict the Arctic sea ice concentration [10]. The neural network prediction results that use long and short-term memory (LSTM) are better than traditional autoregressive (AR) models, and can successfully adapt to long-term data sets. The average monthly forecast error is less than 9%, but the predictability in summer is low [10]. Choi et al. used artificial neural networks (ANN) to make short-term predictions of the Arctic Sea Ice Concentration (SIC) [11]. Using global SIC data for training will result in higher prediction accuracy than using only Arctic SIC data for training [11]. Shu et al. proposed an object-based random forest (ORF) to classify the ice map types of the Arctic, and the overall classification accuracy was 90.1%. When providing surface-related values of sea ice density and snow cover, the estimation of ice thickness can be improved [12].

2.4 How to Predict Sea Ice Level Using Non-Statistical Methods or Non-Machine Learning Methods

The field of sea ice prediction is very extensive. Since 1979, the satellite-based multi-channel passive microwave imaging system has continuously monitored the Arctic sea ice concentration. Common monitoring systems are Scanning Multichannel Microwave Radiometer (SMMR), Special Sensor Microwave/Imager (SSM/I) and Advanced Microblog Scanning Radiometer (AMSR) [13]. Numerical models predict interactions based on physical equations. In the short term, The prediction is usually better than the statistical model. However, it is difficult and expensive to obtain data by physical models [10]. Sea ice retrieval algorithms usually process satellite data, and various sea ice parameters have been determined, such as age, concentration, range, thickness, etc. [10], but statistical models are usually better than dynamic models [5].

Chapter 3

Methodology

3.1 Feature Description

The extent of Arctic sea ice in the is deemed relevant to the following nine features. The content of carbon dioxide, the area of the ozone hole over the Arctic, the land and ocean temperature in the northern hemisphere, the Max/Ave/Min temperature of North Slope Alaska, the rainfall in the Arctic, daylight of Arctic and the population of the world.

Changes in the area of sea ice in the Arctic are believed to be related to the content of carbon dioxide. As shown in Figure 3.1, Carbon dioxide accounts for 81% of greenhouse gases. Thus the main component of greenhouse gases is carbon dioxide, which is also the main cause of the greenhouse effect. According to J.H.Mercer, the greenhouse effect will have a catastrophic impact on the ice sheets [14]. Thus the content of carbon dioxide is chosen to be one of the features.

In addition, according to the results of the common correlated effects mean group (CCEMG) estimator, the GDP growth and the population size influence CO2 emission levels positively and significantly, at both the global and regional levels [15]. In that case, the GDP growth and the population size would have an impact on the ice sheets.

The ozone hole area is also considered to be related to the area of sea ice in the Arctic. In 2010, Sigmond and Fyfe found that the ozone depletion leads to a positive SAM response in austral summer, which induces sea ice melt [16]. In that case, changes in the size of the ozone hole would also affect the changes in sea ice area. Thus the ozone hole area is chosen

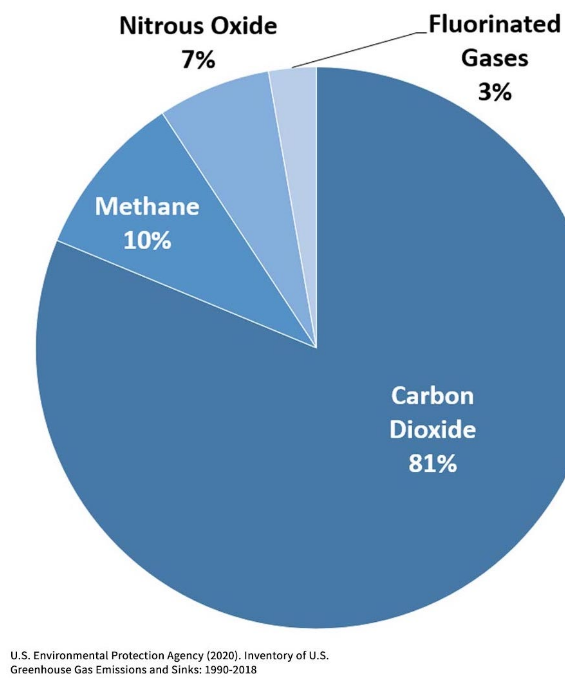


Figure 3.1: Overview of US Greenhouse Gases in 2018.
Image courtesy: https://www.epa.gov/ghgemissions/overview-greenhouse-gases

to be one of the features.

Furthermore, the temperature is deemed to be related to the area of sea ice in the Arctic. According to Ditlevsen and Grinsted's research, by considering a minimal model of an ice sheet, it shows that fluctuating temperatures have an effect on the mass balance and thus on the steady-state volume of the ice sheet [17]. Thus the temperature is considered to be a feature. In our project, we initially selected five different temperatures related to the Arctic area which are the ocean temperature in the Northern Hemisphere, the land temperature in the Northern Hemisphere, the maximum temperature of North Slope Alaska, the average temperature of North Slope Alaska, and the minimum temperature of North Slope Alaska.

Changes in the area of sea ice in the Arctic are also believed to be related to the rainfall. According to Bromwich and Robasky, the increased precipitation may reduce the net loss of ice sheets area or even turn it into net gain [18]. Thus rainfall is considered to be a feature.

The amount of daylight in the Arctic is also believed to be related to the area of ice sheets. The Arctic has polar day and polar night phenomena, so there is periodicity in the phenomenon of daylight. As shown in Figure 3.2, due to the sunshine could produce heat, thus different amount of daylight time would affect the temperature. Thus the daylight is considered to be a feature.

3.2 Data

3.2.1 Data Sources

For the dependent variable in the project, data for Arctic sea ice could be found in National Snow & Ice Data Centre (NSIDC). For independent variables, data for global CO2 content and Arctic ozone hole area are made available to the public via the National Aeronautics and Space Administration (NASA) website. Data for the global population were accessed through the Our World in Data website. In addition, data for the different temperatures and rainfall and average daylight of Arctic are provided by National Oceanic and Atmospheric Administration (NOAA) and Weather Atlas website respectively. Furthermore, the data for the global GDP could be found in the world bank website.

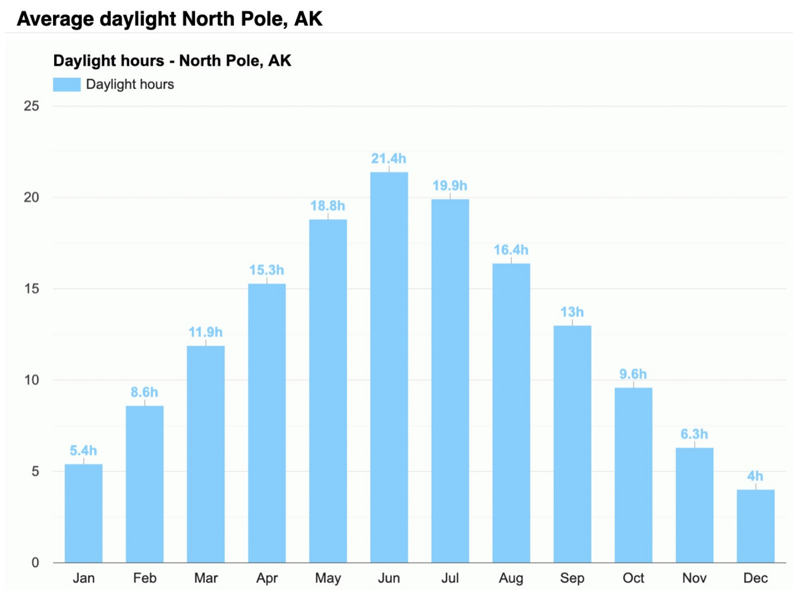


Figure 3.2: Average daylight in North Pole, Alaska. Image courtesy: https://www.weather-us.com/en/alaska-usa/north-pole-climate

All independent variables data were based on monthly records during the period from January 1980 to October 2020. In the process of collecting the data, the group found that some data such as population and GDP contain only annual data set. Therefore, it is necessary to convert these data into monthly format applying Mean Reversion. Moreover, there is a small amount of data missing from the original data (due to the failure of instrumentation). Thus, Mean Reversion can be applied to supplement missing data as well.

3.2.2 Data Normalisation

Normalisation was applied on the whole data sets for lower noise generated by the different magnitude of data during the training process [19]. For the process of normalisation, the minimum value in the data set was placed to 0, while the maximum was placed to 1. For processing the whole data set, Function 3.1 was applied.

$$\text{Normalised Value} = \frac{\text{Value} - \text{Minimum Value}}{\text{Maximum Value} - \text{Minimum Value}} \tag{3.1}$$

3.3 Model Accuracy

3.3.1 Mean Squared Error

Mean Squared Error (MSE) is generally used to observe the deviation between the predicted and true value of the model. MSE (Function 3.2) is used to measure the accuracy between testing results from models. MSE is the expectation of the sum of squared residuals between the true value and predicted value. The smaller the MSE value, the better the performance of the model in the testing process.

$$\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y_n})^2 \tag{3.2}$$

3.3.2 R-squared

R-squared (R^2) (Function 3.3) is the sum of squares of residuals over the total sum of squares, where the value lies between 0 and 1. It is used to compare the performance of the training results from each training data set. The greater the R^2, the better the performance of the model in the training process.

$$R^2 = 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}} \tag{3.3}$$

3.4 Now Forecasting Methodology

3.4.1 K-Fold Cross-validation

Cross-validation is a data splitting and testing method. In this project, k-fold splits the data set into 10 parts. Each time, 9 parts represented by the white grid in Figure 3.3 are used as training data, and the remaining part represented by the red grid in Figure 3.3 is used as testing data for model validation. Since k-fold cross-validation is suitable for small data scenarios, it was applied in this project where only 490 observations are contained in the data set.

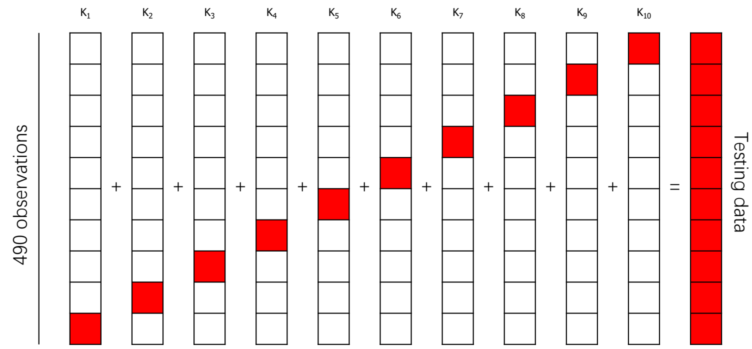


Figure 3.3: K-fold Cross-validation

3.4.2 Linear Regression

Linear regression uses statistical analysis to determine the quantitative relationships between multiple variables [20]. By a given training set, one reasonable method to pick or learn the parameters θ is to make the hypothesis $h(x)$ close to y. To formalize this, we will define a function that measures, for each value of the θ's, how close the $h(x^{(i)})$'s are to the corresponding $y^{(i)}$'s. We define the cost function or ordinary least squares:

$$J_\theta = \frac{1}{2}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2. \tag{3.4}$$

After minimising the sum of the residual, the weights vector for each independent variable will be generated. In the result, P-value of each feature is also notable and can be applied to express the magnitude of the significant influence of the independent variable on the dependent variable. However, owing to the low deviation of linear regression, as the number of features increases, it is susceptible to high variance or over-fitting.

3.4.3 Penalised Linear Regression (Lasso, min/1se)

Comparing with linear regression, Penalised regression is a highly computationally efficient prediction method that can reduce a large number of features into a manageable set and make good predictions on various large data sets, especially when the features are correlated. [21]. As Function 3.5 shown, a penalty term with coefficient Lambda is added to cost function to penalise model complexity by limiting the weight magnitude for reducing the variance.

$$J_\theta := J_\theta + \lambda \sum_{K=1}^{k} b_k. \tag{3.5}$$

In the project, Lasso regression was applied. Comparing with ridge regression, which is also a form of Penalised Regression, Lasso regression is able to cancel unimportant features based on the choice of Lambda. Generally, the choice of Lambda is based on the MSE. In this algorithm, the k-fold method was applied for calculating MSE with a different value of Lambda.

Based on MSE, there are two methods to choose coefficient Lambda which are MIN and 1SE. These two methods have different preferences. Applying MIN is intended to obtain the model with the minimum error and the most accurate, while 1SE tends to simplify the model as much as possible with the loss of a small amount of accuracy.

3.4.4 Penalised Polynomial Regression (Lasso, min/1se)

For linear regression and penalised linear regression, due to the lack of flexibility, the fitting result is often poor. Facing this problem, features with higher-order were applied in them for training. Similarly, when the features of higher-order were added to the model for raising flexibility, the polynomial regression may also have the problem of over-fitting. Therefore, similar to linear regression, Lasso penalisation was also added, simplifying models while addressing the problem of over-fitting.

3.4.5 Random Forest

Random Forest is a model that uses the main vote of multiple decision trees to achieve prediction. It can be applied both on regression and classification algorithm [22]. For Random Forest, n training samples with i features (less than the total number of features) are randomly selected from the original data sets using Bootstrapping method that is a type of reaping sampling method. After k rounds of random selection, k independent training sets are selected and k decision trees are generated.

Due to the randomness of observation selection, non-selected observations are defined as out-of-bag data. Use these observations as labelled testing data, out-of-bag errors (OOB) can be calculated. As k increases, OOB will decrease and tend to be stable. Selecting the appropriate k value based on the OOB is critical. Small k value causes OBB instability while great k value brings high Computational cost.

Also, the feature number applied in each decision tree i can be also important. Small i value causes low flexibility problem, while great i value reduce the diversity of the "forest". Generally, with the increase of i, OOB value drops first and then rises. Choosing the i value corresponding to the minimum OBB helps to achieve better performance of the model.

3.4.6 Neural Networks

A neural network is a massively parallel processor composed of simple processing units. Neurons are formed the basic structure of the network. With the structure of multiple hidden layers, Neural networks algorithm is highly flexible for fitting but also suffers from over-fitting and high computational cost problem.

Forward propagation is a process that multiple inputs enter a hidden layer and time corresponding weights. After that, a nonlinear activation function is applied to convert the input signal into an output signal. For training Neural Networks, backpropagation method is applied. Through gradient descent of loss function, the weights in the whole network can be updated. Since the application of differential operation, the computational cost of training can be mainly attributed to backpropagation.

3.5 Future Forecasting Methodology

3.5.1 Normal Situation

In order to predict the Arctic sea ice extent variation in the next 10 years, the existing features will be processed in three ways to obtain the value of each feature from the year of 2021 to 2030.

(a) **Pure Periodical Feature (Daylight, Rainfall)**. Theses features are monthly-average values over years since 20th century. They would be applied on the predictions for the next 10 years.

(b) **Data with Accessible Future Scenario (Population, GDP, CO2)**. Since the future variation scenarios are given, the monthly record in scenarios would be applied directly.

(c) **Other data (Ozone, Temperature related features)**. As no future variation scenario was found, the values of future 10 years would be generated by applying existing trends via linear regression.

The way of generating future data is based entirely on recent trends, so the predicted result would be treated as NORMAL SITUATION. It would be compared as criteria against forecast based on SPECIAL SITUATIONS which would be discussed in the following part.

3.5.2 Special Situation

The SPECIAL SITUATIONS was defined as the prediction of the Arctic sea ice variation under the circumstance that the fluctuation of a feature does not follow the trend of previous years. Firstly, two feature that has the most significant impact on extent variation would be obtained by variable importance analysis as key factors. For the key factors, two different scenarios would be generated, correspondingly representing sharper and smoother variation comparing with NORMAL SITUATION.

Part II

Scenario: Predicting Arctic Sea Ice in Supervised Learning

Chapter 4

Now Forecasting Results

4.1 Correlation matrix

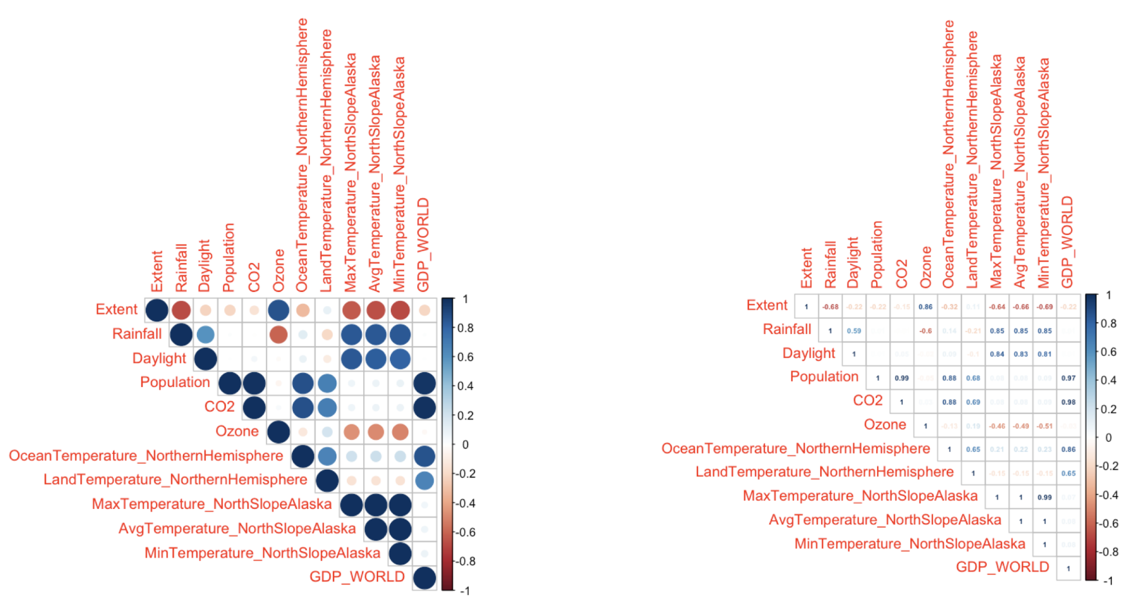


Figure 4.1: Left: The correlation matrix with "circle" method. Right: The correlation matrix with "number" method.

The correlation matrix is a symmetric matrix. Based on the results of the Correlation Matrix, it shows that the extent of Arctic sea ice has a strong correlation with ozone, temperature gained from North Slope Alaska, and rainfall. The temperature measured in the

North Pole Alaska is divided into maximum, average, and minimum temperatures. These three measured temperature values are not only highly correlated with extent but also highly correlated with each other. Therefore, it is unreasonable to add all three variables into the model. Here, the feature with the greatest correlation (min temperature) was selected by Correlation Matrix through number but not through symbol. Hence, the minimum temperature gained from North Slope Alaska was selected.

4.2 Linear Regression

In the first algorithm, Linear Regression model provides a basic fitting results. Without the addition of penalty term, R^2 reached 0.8982 (as Figure 4.1 shown), the mean square error (MSE) reached 0.00946 when the test set was used for evaluation.

Coefficients:	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	0.72626	0.02602	27.912	< 2e-16	***
Rainfall	0.04478	0.02495	1.799	0.0727	.
Daylight	0.18852	0.03376	5.583	3.95e-08	***
Population	-0.98423	0.13192	-7.461	4.07e-13	***
CO2	1.86621	0.19379	9.630	< 2e-16	***
Ozone	0.39157	0.03262	12.003	< 2e-16	***
OceanTemperature_NorthernHemisphere	-0.23312	0.04408	-5.289	< 1.87e-07	***
LandTemperature_NorthernHemisphere	0.08561	0.03910	2.190	0.0290	*
MinTemperature_NorthSlopeAlaska	-0.63691	0.05038	-12.642	< 2e-16	***
GDP_WORLD	-69403	0.07102	-9.772	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1
Residual standard error: 0.08382 on 480 degrees of freedom
15894 observations deleted due to missingness
Multiple R-squared: 0.8982, Adjusted R-squared: 0.8963
F-statistic: 470.5 on 9 and 480 DF, P-value: < 2e-16

Table 4.1: Linear Regression training results.

Fitting diagram was generated as Figure 4.2. It shows that the fitting points are not concentrated around the straight line $y = x$. In fact, this prediction was over-fitting. Comparing to other models that will be discussed later, the MSE of Linear Regression is relatively bigger, which means Linear Regression cannot predict the sea ice very well.

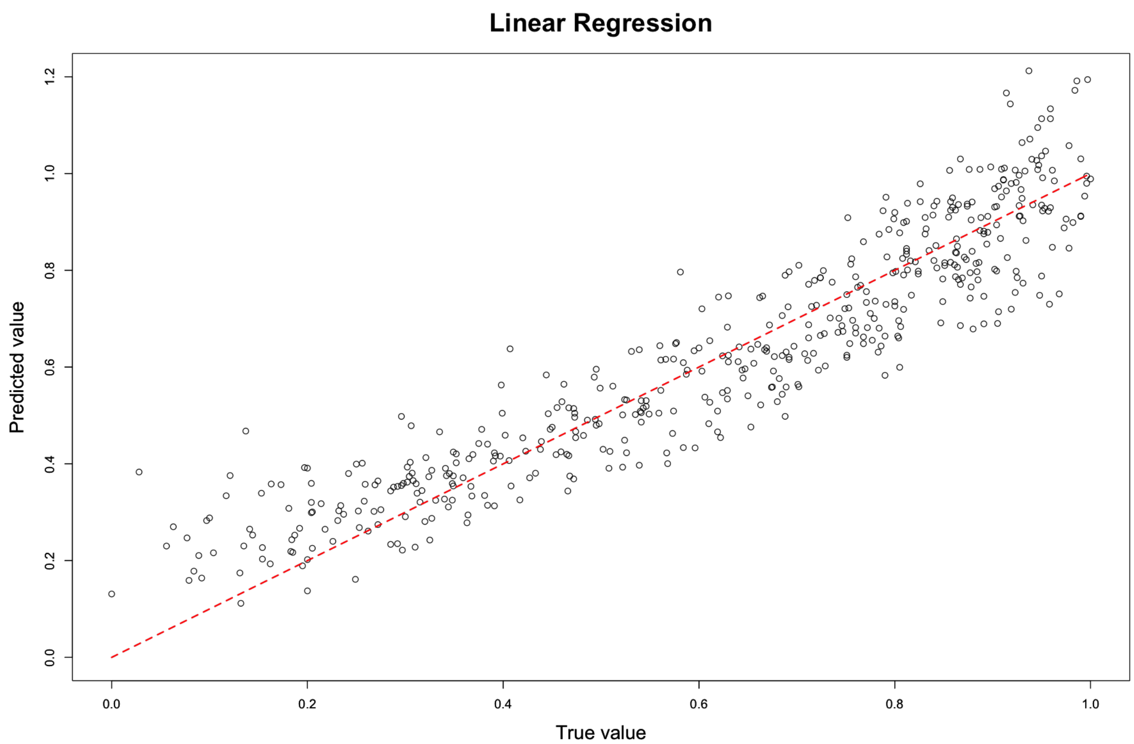


Figure 4.2: The predicted Arctic sea ice extent value vs the real Arctic sea ice extent value with **Linear Regression**. The red referenced dotted line represents the straight line y=x. Mean Square Error (MSE) is **0.00946**.

4.3 Penalised Linear Regression (Lasso, min/1se)

In order to simplify the model by reducing the number of features whilst increasing the accuracy, Lasso Penalisation was applied. Using the GLMNET package, the weights of all features were presented as the change of Lambda value.

The ridge-trance figure (Figure 4.3) shows that more and more coefficients become zero as the value of log lambda increases. In Lasso Penalisation, the features with poor performance are removed to reduce the complexity of the model.

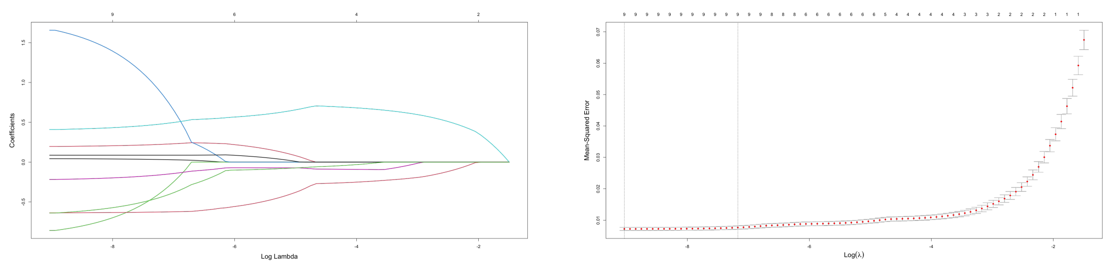


Figure 4.3: Left: Ridge trace diagram of Penalised Linear Regression. Right: Log Lambda vs Testing Error diagram of Penalised Linear Regression.

At this point, appropriate Lambda values were selected to output the weights of all features. In the selection, MIN and 1SE were applied respectively (Figure 4.3).

Applying different options for the Lambda value, two different sets of results were obtained. When Lambda option was MIN, the MSE value of the model was 0.00690, which was slightly less than the 0.00734 when Lambda option was 1SE.

The output result was demonstrated in Table 4.2. In this case, using 1SE did not reduce the complexity of the model but the number of features. Therefore, MIN was better than 1SE in Lasso Penalisation in our project.

In Linear Regression, the performance of test fitting was not ideal even it was optimised because of the lack of flexibility of the model. By adding a penalty term to the Linear Regression, the performance of test fitting was improved by both MIN and 1SE options.

Coefficients:	min	1se
	1	1
(Intercept)	0.71559335	0.66666336
Rainfall	0.04299020	0.03049067
Daylight	0.19686230	0.22725958
Population	-0.85828586	-0.30290499
CO2	1.65605218	0.74378027
Ozone	0.40941811	0.48953790
OceanTemperature_NorthernHemisphere	-0.21778615	-0.14950798
LandTemperature_NorthernHemisphere	0.08669526	0.08684475
MinTemperature_NorthSlopeAlaska	-0.63667606	-0.62486026
GDP_WORLD	-0.63894945	-0.40706240

Table 4.2: Hyper-parameters of Penalised Linear Regression.

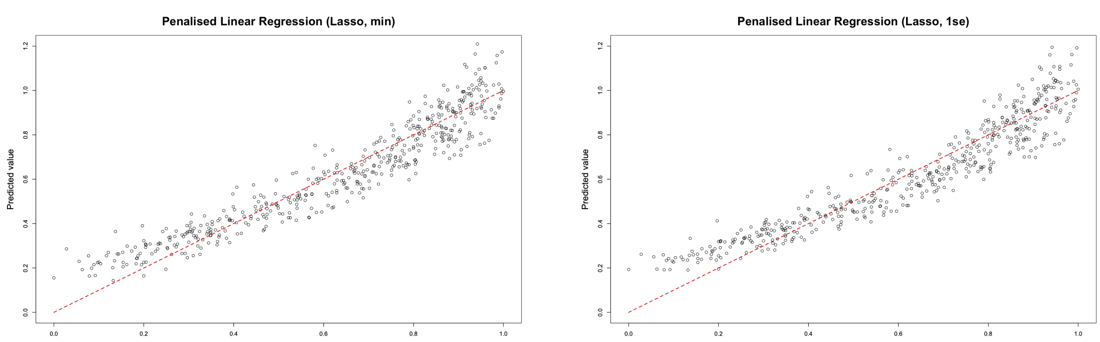


Figure 4.4: **Left**: The predicted Arctic sea ice extent value vs the real Arctic sea ice extent value with **Penalised Linear Regression** (Lasso, **min**). The red referenced dotted line represents the straight line y=x. Mean Square Error (MSE) is **0.00690**. **Right**: The predicted Arctic sea ice extent value vs the real Arctic sea ice extent value with **Penalised Linear Regression** (Lasso, **1se**). The red referenced dotted line represents the straight line y=x. Mean Square Error (MSE) is **0.00732**.

4.4 Penalised Polynomial Regression (Lasso, min/1se)

To increase the flexibility of the model, polynomial regression was applied.

Based on the correlation matrix, the four features with the highest correlation with the Arctic sea ice were selected for this model. The 1st, 2nd, 3rd, and 4th polynomials of each feature were added to the training sets. As polynomial regression still experience over-fitting, a penalty term was applied in this model.

A ridge-trance graph was generated. By applying log Lambda vs testing error figure (left of Figure 4.5), the Lambda value was chosen.

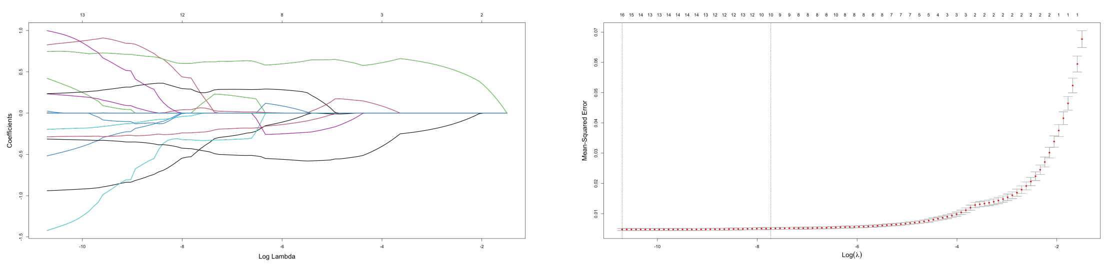


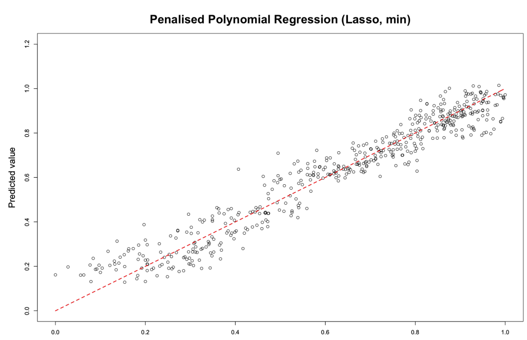
Figure 4.5: Left: Ridge trace diagram of Penalised Polynomial Regression. Right: Log Lambda vs Testing Error diagram of Penalised Polynomial Regression.

According to the following results in Table 4.3, no features are removed in this model with min, and its MSE is 0.00448. In 1se, 6 features are removed and its MSE is 0.00483. At the cost of small precision loss, the model can be greatly simplified, which is exactly the purpose of the 1se option of Lambda.

By comparing the fitting diagram (Figure 4.6) of two results, it was found that the results were similar, which further proved that the influence of 1SE value on the accuracy of the model was almost negligible. At the same time, compared with Linear Regression, the fitting results were significantly better as polynomial regression brought more flexibility to the model.

Coefficients:	min	1se
	1	1
(Intercept)	7.817628e-01	0.7495621617
Rainfall	-9.399989e-01	-0.5338097486
Rainfall.2	8.267722e-01	0.4307466064
Rainfall.3	4.211873e-01	.
Rainfall.4	-5.176367e-01	.
Daylight	-1.975953e-01	-0.0364294056
Daylight.2	2.334835e-01	.
Daylight.3	2.349987e-01	0.2928271727
Daylight.4	1.117847e-05	0.0415480304
Ozone	7.446573e-01	0.6009477148
Ozone.2	2.420938e-02	-0.0002389955
Ozone.3	-1.422105e+00	-0.3234189416
Ozone.4	9.982037e-01	.
MinTemp	-3.142767e-01	-0.4278954907
MinTemp.2	-2.878996e-01	-0.2404459826
MinTemp.3	1.936775e-05	.
MinTemp.4	3.213799e-03	.

Table 4.3: Hyper-parameters of Penalised Polynomial Regression.



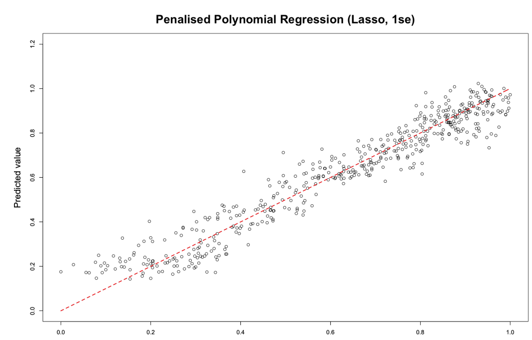


Figure 4.6: **Left**: The predicted Arctic sea ice extent value vs the real Arctic sea ice extent value with **Penalised Polynomial Regression** (Lasso, **min**). The red referenced dotted line represents the straight line y=x. Mean Square Error (MSE) is **0.00448**. **Right**: The predicted Arctic sea ice extent value vs the real Arctic sea ice extent value with **Penalised Polynomial Regression** (Lasso, **1se**). The red referenced dotted line represents the straight line y=x. Mean Square Error (MSE) is **0.00484**.

4.5 Random Forest

To use Random Forest, three parameters had to be determined which are the number of total features, the number of trees, and the number of features that randomly applied in each tree (MTRY).

Firstly, nine out of eleven features were applied on this algorithm based on correlation matrix. Then, the ”tree number vs out-of-bag error” graph (the right of Figure 4.7) was plotted. Based on different MTRY values, when the tree number approaches 200, the out-of-bag error tends to be stable. Hence, 200 was chosen as the tree number of this model.

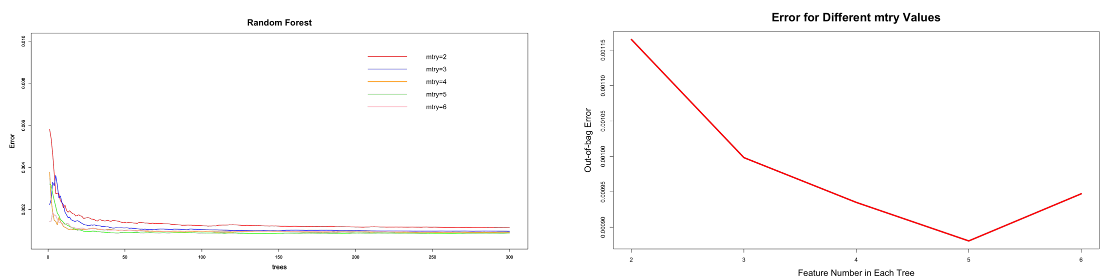


Figure 4.7: Left: Check the error stability of random forest with different number of trees. Right: Check the out-of-bag error of random forest with different number of features in each tree when three number is 200.

Moreover, the choice of the MTRY value is also important. Based on the selected tree number, different out-of-bag errors would be obtained by using different MTRY values. Through comparison, feature number in each tree was selected as 5 which corresponds with the minimum out-of-bag error.

After parameter was selected, Random Forest model was trained and tested. The MSE value is 0.00105 which is much smaller than the previous models. Observing the fitting diagram (bottom of Figure 4.8), it was found that dots were distributed in a narrow area fitting the straight line $y = x$, which indicates a robust fitting performance.

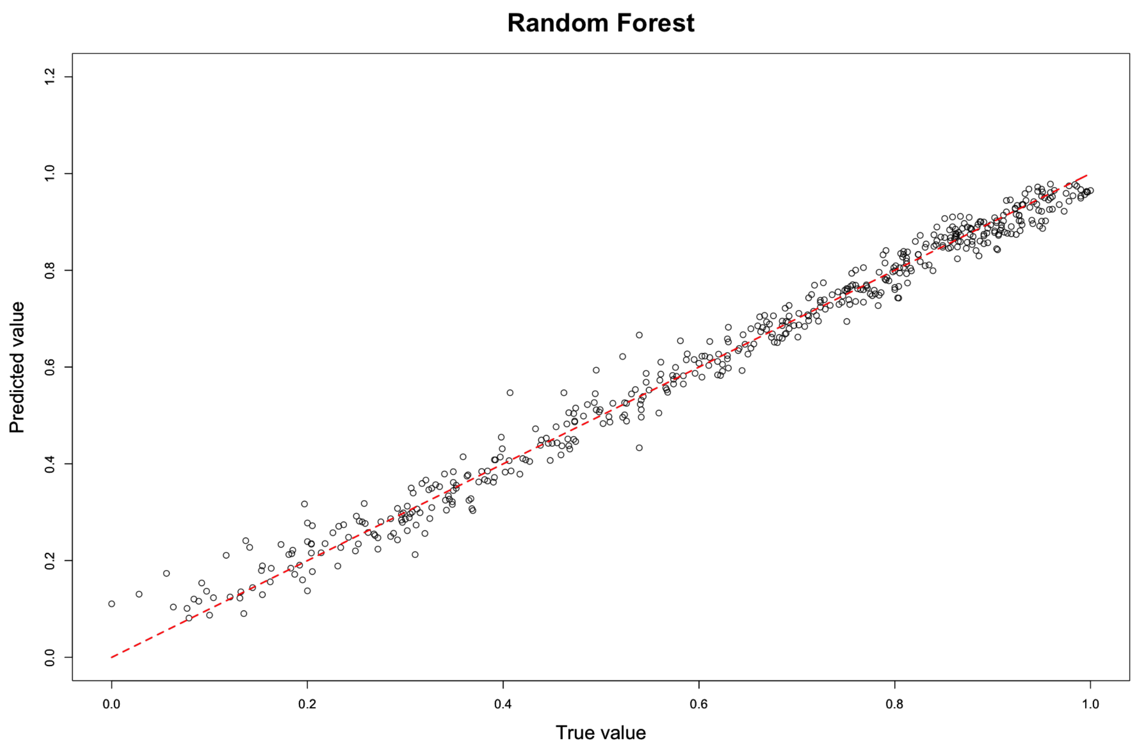


Figure 4.8: The predicted Arctic sea ice extent value vs the real Arctic sea ice extent value with **Random Forest** (200 trees, 5 features). The red referenced dotted line represents the straight line y=x. Mean Square Error (MSE) is **0.00105**.

4.6 Neural Network

Unlike random forests, parameters are selected based on out-of-pocket errors in Neural Networks (NN). During the parameter tuning process, k-fold was applied on output testing errors for reference. After multiple tuning, 5 hidden layers were selected. Nine, seven, five, three and one neurons were used in each hidden layer from front to back as Figure 4.9 shown.

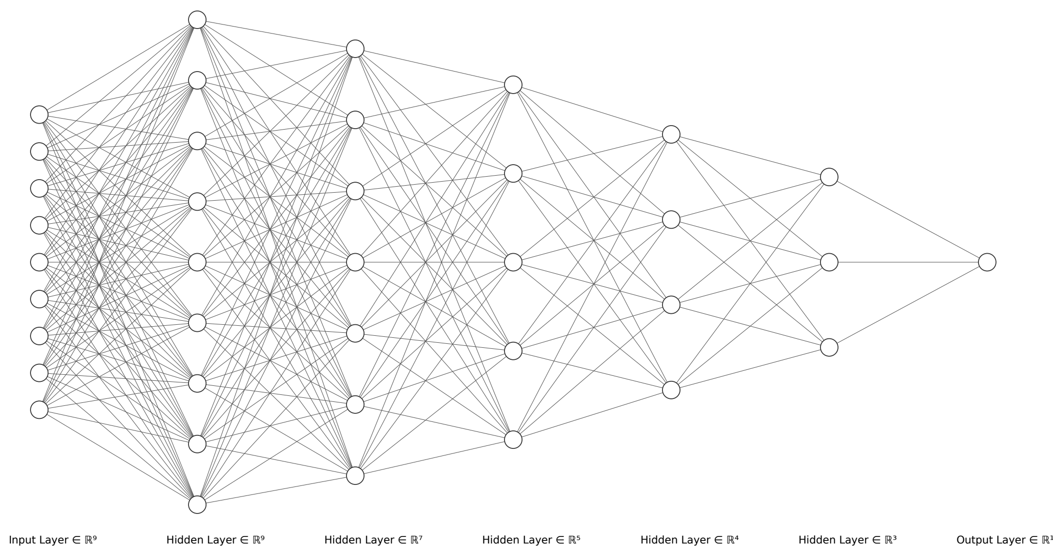


Figure 4.9: Neural Networks Structure.

Similar to Random Forest, the Neural Network algorithm achieved excellent performance. The MSE value of Neural Network was slightly higher than the Random Forest, reaching 0.00106. The fitting diagram (right of Figure 4.10) also proves that the variance of the Neural Networks model is very low comparing with previous Supervised Learning algorithms excluding Random Forest.

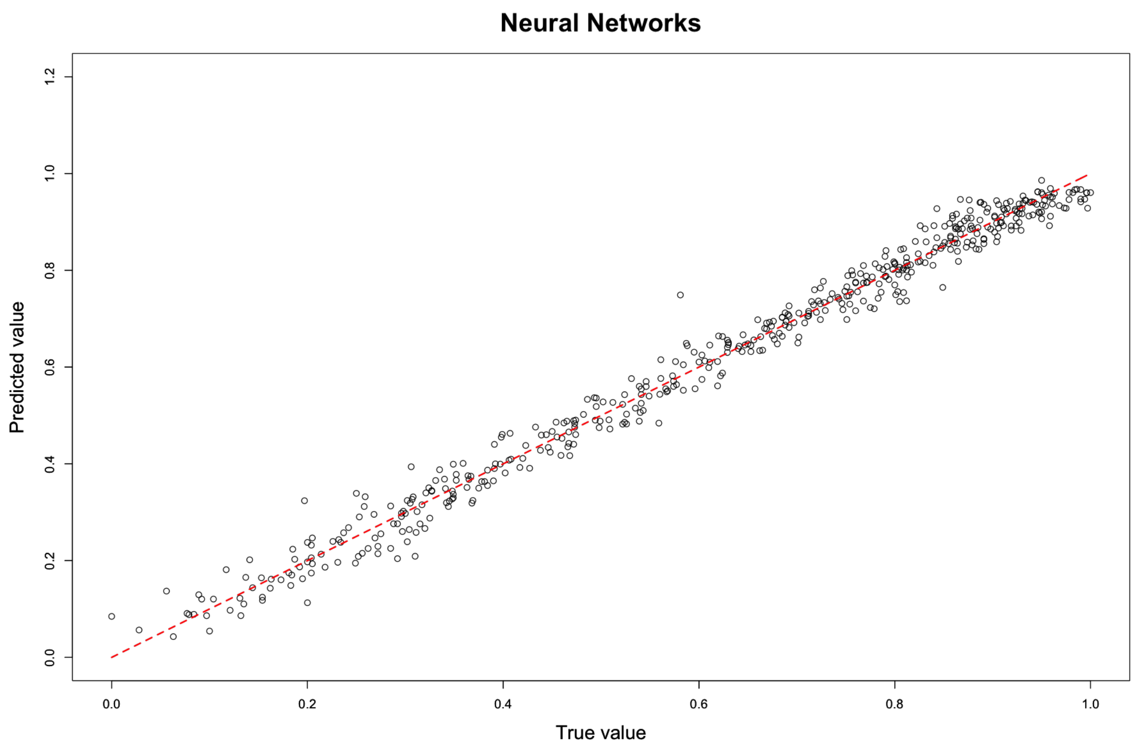


Figure 4.10: The predicted Arctic sea ice extent value vs the real Arctic sea ice extent value with **Neural Networks** (9 neurons, 5 hidden layers with (9,7,5,4,3) nodes respectively). The red referenced dotted line represents the straight line y=x. Mean Square Error (MSE) is **0.00106**.

4.7 Comparison

Based on the above four models and seven prediction results, all the results of R-squared and MSE are listed and compared in Table 4.4.

Model	R^2	MSE
Linear Regression	0.898	0.00946
Penalized Linear Regression (Lasso, min)	0.863	0.00690
Penalized Linear Regression (Lasso, 1se)	0.857	0.00734
Penalized Polynomial Regression (Lasso, min)	0.931	0.00448
Penalized Polynomial Regression (Lasso, 1se)	0.917	0.00483
Random Forest	**0.983**	**0.00105**
Neural Networks	0.927	0.00106

Table 4.4: Comparison of performance metrics.

Random Forest was selected for further future predictions as it was the best-performed model. As we can see from Table 4.4, the model with low R^2 might not perform well, which indirectly proves the existence of over-fitting phenomenon and the necessity of penalization as well.

Chapter 5

Future Forecasting Results

5.1 Normal Situation

Applying data sets which contains features for prediction (mentioned in Section 3.5.1), the Arctic Sea Ice Extent variation in 10 years (from 2021 to 2030) was plotted (Figure 5.1).

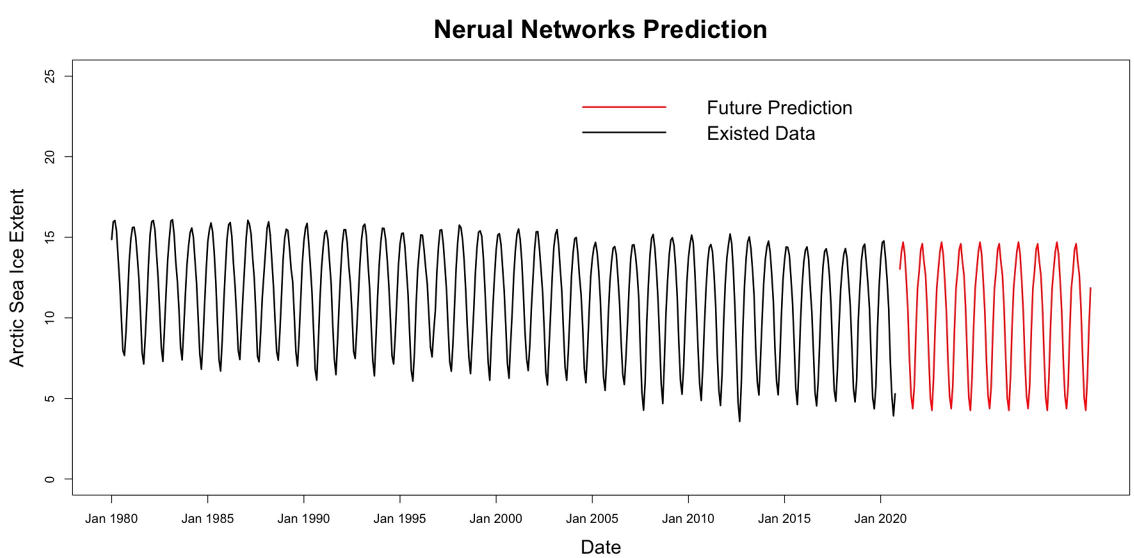


Figure 5.1: Future prediction of Arctic sea ice: 10 years' normal situation prediction —- monthly.

As the monthly figure (Figure 5.2) shown, the sea ice extent fluctuation conforms to the expected seasonality and continues the main trend. Therefore, we can conclude that the prediction based on the neural network and the prediction data set is reasonable.

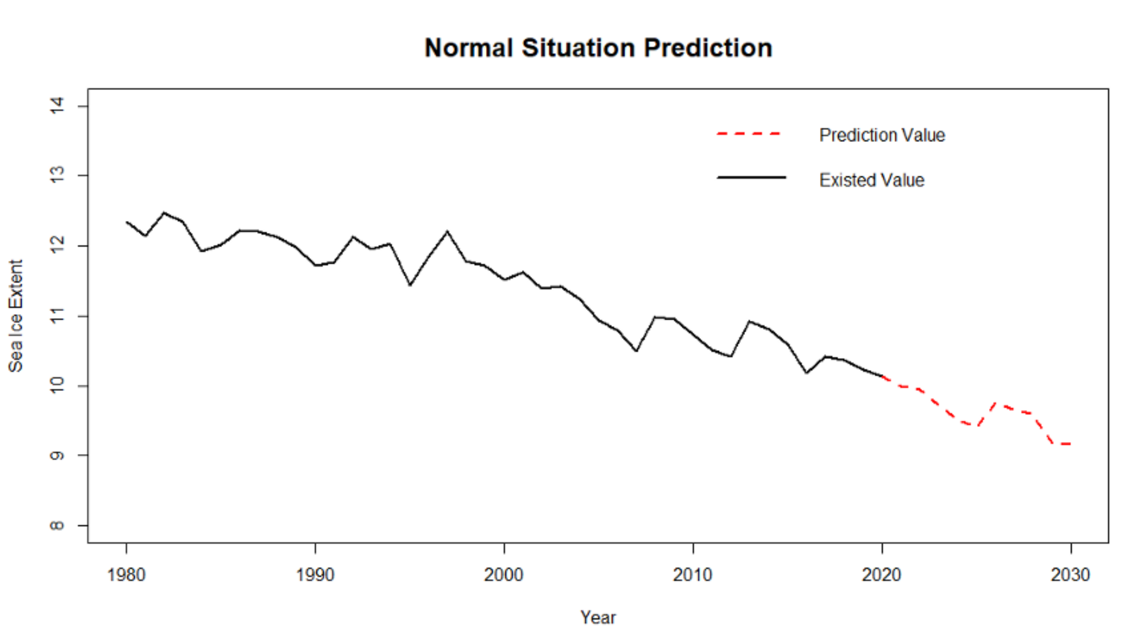


Figure 5.2: Future prediction of Arctic sea ice: 50 years' normal situation prediction.

To predict a long-term trend, it is more intuitive to visualize using annual average. Based on Random Forest model, the Arctic sea ice extent decreased from 10.00 Mkm^2 in 2020 to 9.41 Mkm^2 in 2030, reducing by 5.9%.

5.2 Special Situation

First, visualizations were performed using VARIABLE IMPORTANCE function. According to Figure 5.3, it can be seen that ozone, rainfall, and temperature are the most important features. However, rainfall is not as closely related to climate change as temperature, so we chose ozone and temperature as two key factors.

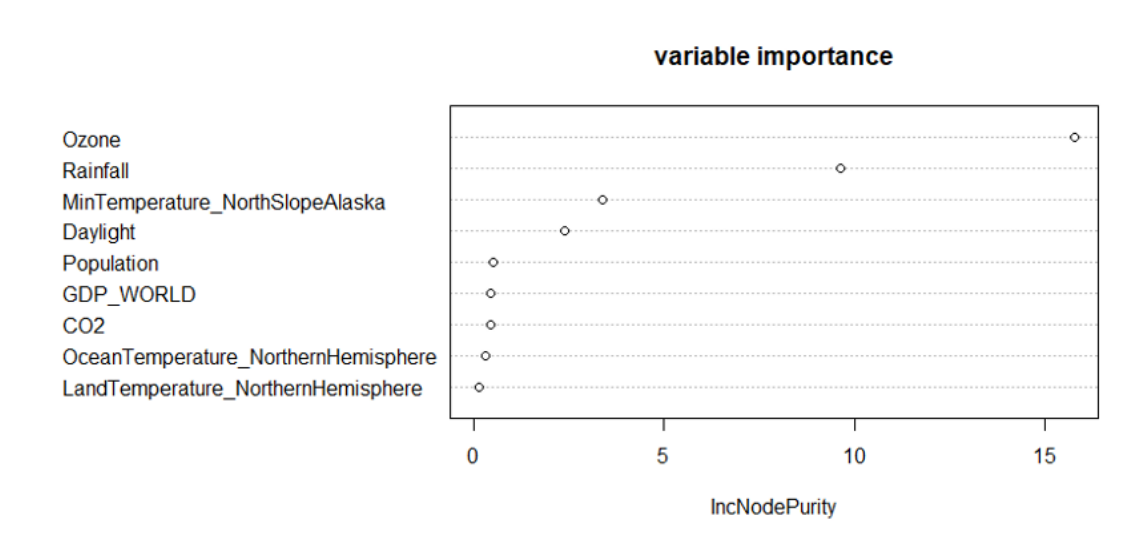


Figure 5.3: Comparison of variance importance.

Linear Regression was applied on annul average ozone value only, a straight line was plotted and the ozone level decreased by 4.6% in last 40 years. Based on the NORMAL SITUATION data processing (Section 3.5.1), the ozone level was predicted to decrease about 1.2% in 10 years. For generating ozone data of SPECIAL SITUATION, a ramp was applied on NORMAL SITUATION data for simulating ozone variation in high level where it was decreased by 2.5% in 10 years and low level where it was set no variation in 10 years).

As Figure 5.4 shown, applying high level ozone, the predicted extent of Arctic sea ice was 0.92% lower than applying Normal Situation, which was 9.414 Mkm^2. Moreover, applying low level ozone level, the predicted result was 0.87% higher than that of NORMAL SITUATION. There was a negative correlation between ozone level and Arctic ice extent, which further confirmed the relationship shown in correlation matrix.

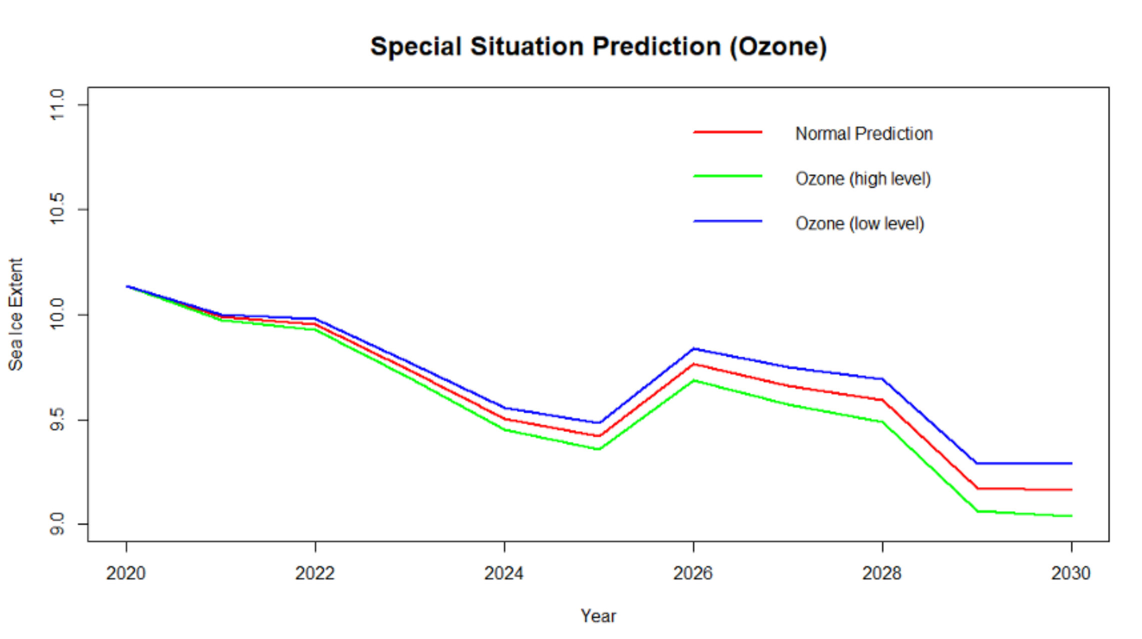


Figure 5.4: Special situation prediction based on different Ozone level.

Similarly, applying the same method on *MinTemperature*, in high level, normal situation and low level, the temperature was set to rise by 7,4,1 degree Celsius respectively.

According to the results above (Figure 5.5), applying high level temperature, the predicted extent of Arctic sea ice was 0.13% higher than applying Normal Situation. While applying low level ozone level, the predicted result was 0.16% lower than NORMAL SITUATION. It was found that the temperature is positively correlated with the Arctic sea ice extent from the correlation matrix. In addition, the influence of temperature change on the sea ice extent is not as sensitive as ozone, which also proved the feature importance of temperature is lower than ozone.

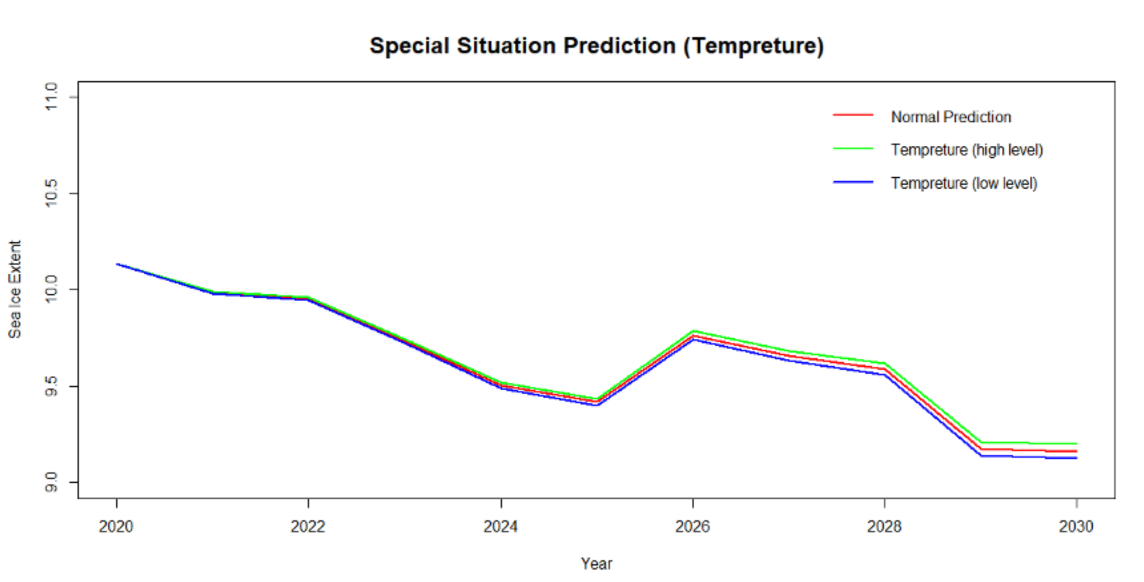


Figure 5.5: Special situation prediction based on different minimum temperature level in North Slope Alaska.

Chapter 6

Conclusion

In this report, we gave readers a thorough overview of predicting the Arctic sea ice in Supervised Learning: the foundations (PART I) and the scenario (PART II), as well as how we contributed to the analytical strategy to connect the theories and testings.

In Chapter 2, we first explain how important the sea ice to the climate and environment research, especially the global warm. Then, we walked through the overview of the field of environmental risks in terms of multiple kinds of algorithms, including predicting sea ice level using statistical models, predicting sea ice level using Machine Learning models, and predicting sea ice level using non-statistical methods or non-machine learning methods.

In Chapter 3, we covered the analytical strategies of our project. We explained the features we used and where we can obtain these data. Model accuracy is one of the most important steps in our project. We used R-squared and mean square error (MSE) to analyse the accuracy of our Supervised Learning algorithms. K-fold cross-validation was used in this project since small data were used to predict the sea ice level in this project. Then, in mathematically, we introduced the cost function for Linear Regression algorithm and explained how it worked. On the basis of Linear Regression, we introduced penalised linear regression by adding a penalised term. Furthermore, penalised polynomial regression can be derived by applying higher-order parameters on features. Last but not least, we introduced the famous Random Forest and Neural Networks.

In PART II, the key questions we want to answer are: which factor(s) is/are the most important factor to the change of the sea ice, and more importantly, how the sea ice will be melt if the situations go worse?

In Chapter 4, we gave readers clear visualisations on correlation relationships and the fitting diagram of each Supervised Learning model. At the end of chapter, in Section 4.7, we summarised that Random Forest had the best fitting performance in our case.

In Chapter 5, we designed three specific scenarios for our future prediction. The first one is NORMAL SITUATION where all factors/features are followed the current trends. Then we designed two SPECIAL SITUATIONS scenarios for the future prediction. We first selected two factors that influence the sea ice level the most, then these factors are applied on SPECIAL SITUATIONS respectively. In a result, for the worst situation, the extent of the Arctic sea ice will lose around 1% more compared to the NORMAL SITUATION. It is time for policy-makers to take action to control the global temperature and reduce the excessive greenhouse effect.

We am really excited about the progress that has been made for this semester. At the same time, we also deeply believe that there is still a long way to go towards predicting the Arctic sea ice or other environmental information, and we are still facing challenges and a lot of open questions that need to be addressed in the future. For instance, we need spend more time on more scenario analysis. For instance, we need to find out what the ”high-level scenario” means for Ozone. Is increasing 2.5% in 10 years for ”high-level scenario” in our case reasonable? Secondary, we need to find our more possible features. Thirdly, a large and tidy data set is important for the scientific research since big data are helpful to show the advantage of Neural Networks or other Artificial Intelligence algorithms.

Bibliography

[1] C. Parmesan and G. Yohe, "A globally coherent fingerprint of climate change impacts across natural systems," *Nature*, vol. 421, no. 6918, p. 37, 2003.

[2] R. K. Pachauri, M. R. Allen, V. R. Barros, J. Broome, W. Cramer, R. Christ, J. A. Church, L. Clarke, Q. Dahe, P. Dasgupta, N. K. Dubash, O. Edenhofer, I. Elgizouli, C. B. Field, P. Forster, P. Friedlingstein, J. Fuglestvedt, L. Gomez-Echeverri, S. Hallegatte, G. Hegerl, M. Howden, K. Jiang, B. J. Cisneroz, V. Kattsov, H. Lee, K. J. Mach, J. Marotzke, M. D. Mastrandrea, L. Meyer, J. Minx, Y. Mulugetta, K. O'Brien, M. Oppenheimer, J. J. Pereira, R. Pichs-Madruga, G.-K. Plattner, H.-O. Pörtner, S. B. Power, B. Preston, N. H. Ravindranath, A. Reisinger, K. Riahi, M. Rusticucci, R. Scholes, K. Seyboth, Y. Sokona, R. Stavins, T. F. Stocker, P. Tschakert, D. van Vuuren, and J.-P. van Ypserle, *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, R. Pachauri and L. Meyer, Eds. Geneva, Switzerland: IPCC, 2014.

[3] J. Stroeve, A. Frei, J. McCreight, and D. Ghatak, "Arctic sea-ice variability revisited," *Annals of Glaciology*, vol. 48, p. 71–81, 2008.

[4] J. Kim, K. Kim, J. Cho, Y. Q. Kang, H.-J. Yoon, and Y.-W. Lee, "Satellite-based prediction of arctic sea ice concentration using a deep neural network with multi-model ensemble," *Remote Sensing*, vol. 11, no. 1, p. 19, 2019.

[5] A. Tivy, S. E. Howell, B. Alt, J. J. Yackel, and T. Carrieres, "Origins and levels of seasonal forecast skill for sea ice in hudson bay using canonical correlation analysis," *Journal of Climate*, vol. 24, no. 5, pp. 1378–1395, 2011.

[6] J. Ahn, S. Hong, J. Cho, Y.-W. Lee, and H. Lee, "Statistical modeling of sea ice concentration using satellite imagery and climate reanalysis data in the barents and kara seas, 1979–2012," *Remote Sensing*, vol. 6, no. 6, pp. 5520–5540, 2014.

[7] S. D. Drobot and J. A. Maslanik, "A practical method for long-range forecasting of ice severity in the beaufort sea," *Geophysical Research Letters*, vol. 29, no. 8, pp. 54–1, 2002.

[8] S. Drobot, "Long-range statistical forecasting of ice severity in the beaufort–chukchi sea," *Weather and Forecasting*, vol. 18, no. 6, pp. 1161–1176, 2003.

[9] S. D. Drobot, J. A. Maslanik, and C. Fowler, "A long-range forecast of arctic summer sea-ice minimum extent," *Geophysical Research Letters*, vol. 33, no. 10, 2006.

[10] J. Chi and H.-c. Kim, "Prediction of arctic sea ice concentration using a fully data driven deep neural network," *Remote Sensing*, vol. 9, no. 12, p. 1305, 2017.

[11] M. Choi, L. W. A. De Silva, and H. Yamaguchi, "Artificial neural network for the short-term prediction of arctic sea ice concentration," *Remote Sensing*, vol. 11, no. 9, p. 1071, 2019.

[12] S. Shu, X. Zhou, X. Shen, Z. Liu, Q. Tang, H. Li, C. Ke, and J. Li, "Discrimination of different sea ice types from cryosat-2 satellite data using an object-based random forest (orf)," *Marine Geodesy*, vol. 43, no. 3, pp. 213–233, 2020.

[13] K. Fennig, M. Schröder, A. Andersson, and R. Hollmann, "A fundamental climate data record of smmr, ssm/i, and ssmis brightness temperatures." *Earth System Science Data*, vol. 12, no. 1, 2020.

[14] J. H. Mercer, "West antarctic ice sheet and co2 greenhouse effect: a threat of disaster," *Nature*, vol. 271, no. 5643, pp. 321–325, 1978.

[15] K. Dong, G. Hochman, Y. Zhang, R. Sun, H. Li, and H. Liao, "Co2 emissions, economic and population growth, and renewable energy: Empirical evidence across regions," *Energy Economics*, vol. 75, pp. 180 – 192, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140988318303256

[16] M. Sigmond and J. Fyfe, "Has the ozone hole contributed to increased antarctic sea ice extent?" *Geophysical Research Letters*, vol. 37, no. 18, 2010.

[17] T. B. Mikkelsen, A. Grinsted, and P. Ditlevsen, "Influence of temperature fluctuations on equilibrium ice sheet volume." *Cryosphere*, vol. 12, no. 1, 2018.

[18] D. Bromwich and F. Robasky, "Recent precipitation trends over the polar ice sheets," *Meteorology and Atmospheric Physics*, vol. 51, pp. 259–274, 1993.

[19] H. Bessembinder, J. F. Coughenour, P. J. Seguin, and M. M. Smoller, "Mean reversion in equilibrium asset prices: Evidence from the futures term structure," *The Journal of Finance*, vol. 50, no. 1, pp. 361–375, 1995.

[20] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2012, vol. 821.

[21] M. Bowles, *Machine learning in Python: essential techniques for predictive analysis*. John Wiley & Sons, 2015.

[22] J. Tan, H. Liu, M. Li, and J. Wang, "A prediction scheme of tropical cyclone frequency based on lasso and random forest," *Theoretical and Applied Climatology*, vol. 133, no. 3-4, pp. 973–983, 2018.