NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

MH6212-ANALYTICS SOFTWARE II

Climate Factors Forecast

Group Member Name	Matriculation Number
Ang Shu Wei	G2302794C
Goel Ishita	G2302904C
Khoo Chun Yun	G2302847F
Yu Qing	G2303017K
Zhao Jiayi	G2302767L

1. Introduction

Climate change is a long-lasting change of the temperature and weather patterns. Many human activities contribute to climate change: for example, industrial production and transportation emit large amounts of greenhouse gases; deforestation and urbanisation cause a reduction in forest cover and a decrease in the diversity of species, above activities exacerbating the problem of climate change. Climate change has become a very challenging global issue, which leads to many serious consequences, such as extreme weather and sea level rise.

At the same time, global climate change has also impacted Singapore's temperature and precipitation patterns. As a coastal country, Singapore is also experiencing more frequent extreme weather events, including heavy rainfall, droughts, and sea level rise, etc.. This leads to significant challenges to Singapore's environment and society. Research on climate change in Singapore is very necessary because this provides insights into climate change trends, maintains ecological balance and helps to formulate effective policies to ensure sustainable development.

Singapore is located near the equator, in the tropics. Also, it is close to the ocean, with frequent precipitation. Due to these features, we wish to study the changes in precipitation, maximum temperature and minimum temperature in Singapore from 1901 to 2022. Because these factors are core indicators of the climate, they are critical to understand the trends of climate change. Temperature changes are the main manifestation of climate change. Changes in temperature can lead to extreme weather, such as droughts and floods. Climate change can also lead to changes in precipitation patterns, especially the quantity of rainfall. In the end, we hope to predict future trends in precipitation, maximum temperature and minimum temperature by using past years data.

In this report, we aim to conduct two types of analysis based on the information we have. One is machine learning model analysis (linear regression and SVR) and visualisation. Another one is time series analysis (using the ARIMA model). They are used to predict how temperature and precipitation will change over time. Both types of analytics can help to understand the data better, show trends and patterns, and provide useful information for forecasting.

2.1 Data description

There are 4 variables - Category, temp_min, temp_max, and rainfall.

Variables	description	type	e.g.
Category	Year	integer	1901, 1902,..., 2022
temp_min	Minimum temperature for the year	numeric (float)	23.03, 22.98,..., 23.89
temp_max	Maximum temperature for the year	numeric (float)	30.4, 30.34,…,31.58
rainfall	Total amount of rain collected for the year (mm)	numeric (float)	2106.77, 2334.47,..., 2894.53

2.2 Data exploration

A. Overview of numeric variables

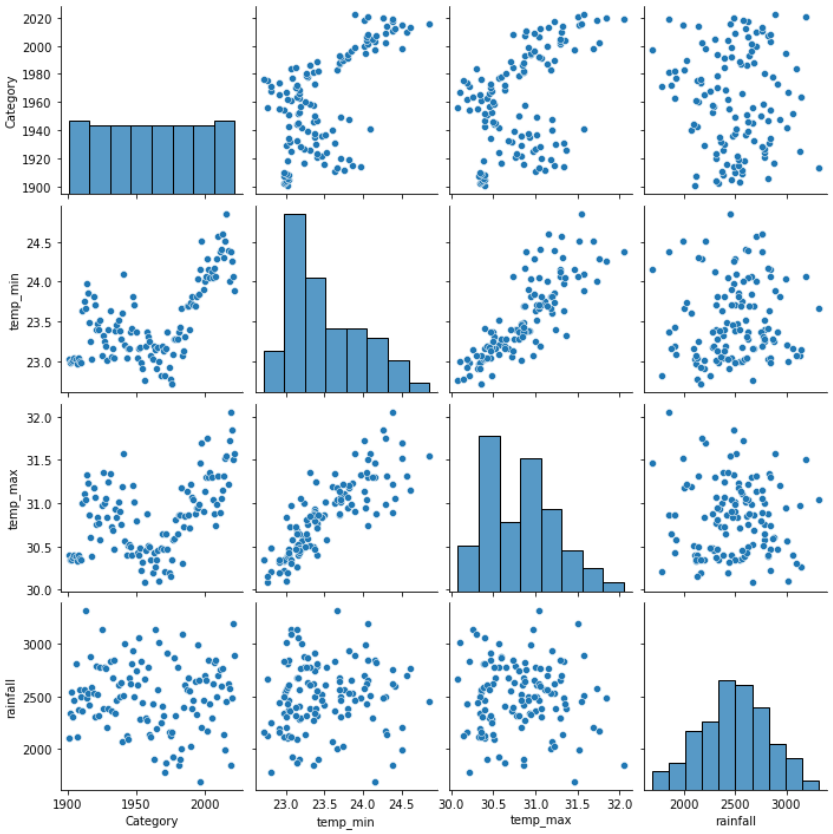

Figure 1: Exploration of variables using a pairplot

The distribution for the minimum temperature is more right-skewed as compared to the maximum temperature and rainfall which resemble more of a normal distribution.

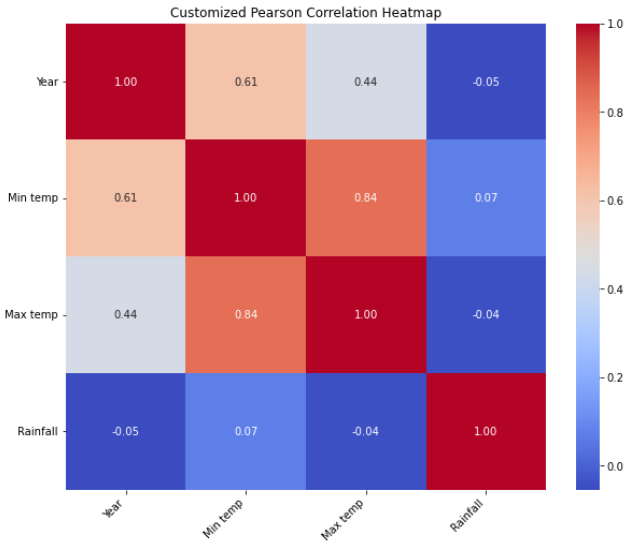

Figure 2: Correlation Heatmap on the variables

There is a strong correlation between minimum temperature and maximum temperature. A relatively strong correlation can also be observed between rainfall and minimum temperature as seen by correlation score of 0.61. Also, there is a weak/no correlation between rainfall and the other variables since the correlation figures are less than 0.1

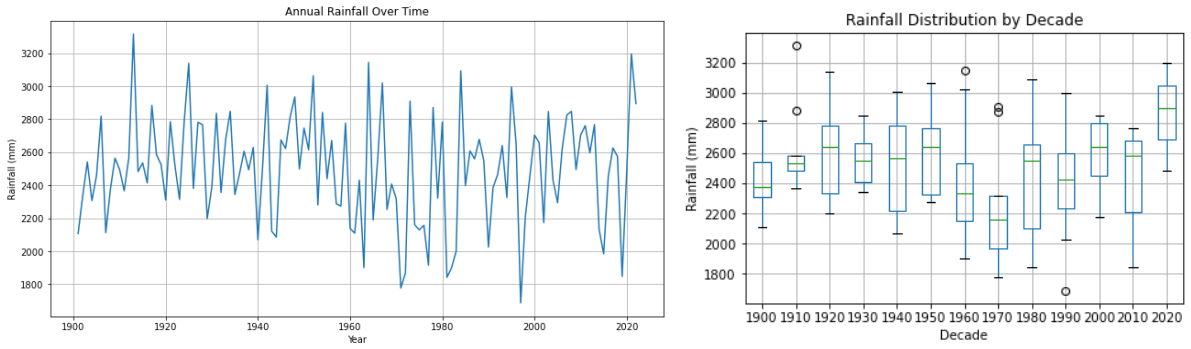

Figure 3: Distribution of rainfall over time

Unlike the temperatures which have an increasing trend from year 1960 onwards, the rainfall over time hovers in the range of 2200 to 2800mm. There are exceptional years (outliers) as seen in the boxplot above for year 1913 we have unusually high rainfall exceeding 3300mm while for year 1997 we have very low rainfall of less than 1800mm.

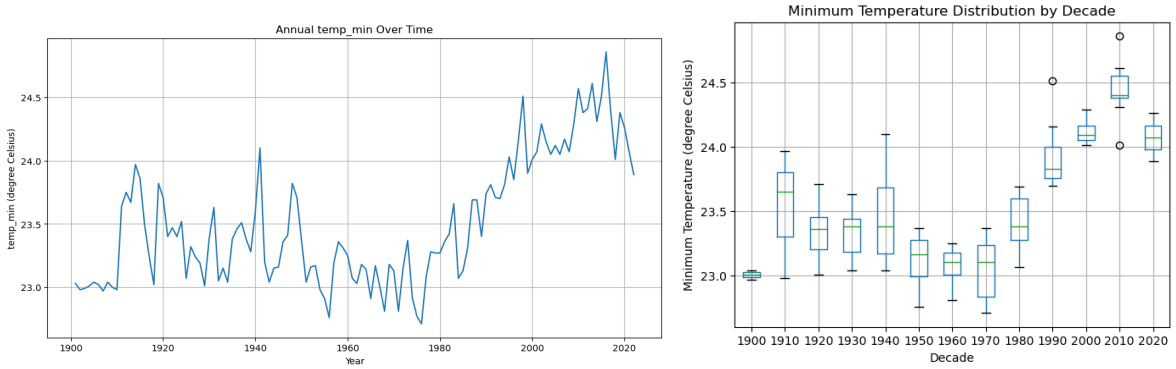

Figure 4: Distribution of minimum temperature over time

The range of the temperatures seems relatively big in these decades experienced in 1910-1920, 1940-1950, and 1970-1980, on the other hand, there is a very short range of temperatures in the decade experienced in 1900-1910. There are more outliers in the recent decades such as from year 1990 onwards as compared to the older decades. Generally, there are higher minimum temperatures experienced in the recent decades.

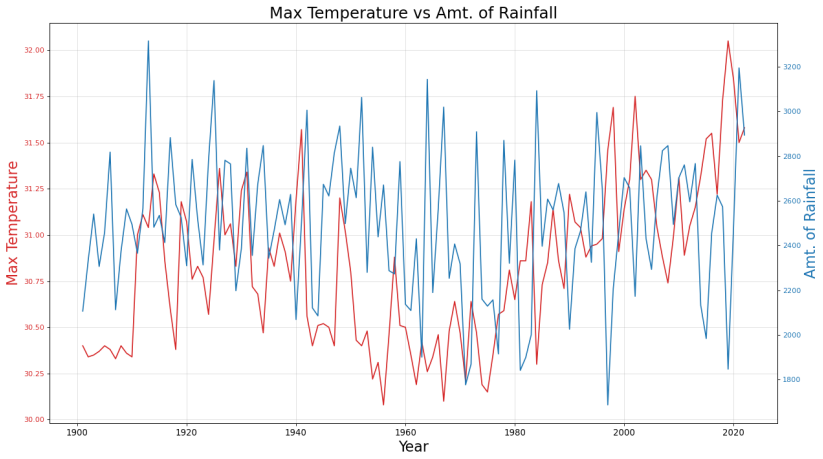

Figure 5: Trend between maximum temperature and rainfall throughout the years

From the years 1940 to 1960, there is a decreasing trend for both maximum temperature and rainfall. A few years before leading to 1934, the temperatures are sliding while the amount of rainfall is increasing. Also in the year 1934, it had a high rainfall of around 2800 mm but a low maximum temperature of 30.47 degrees celsius.

3. Methodology

3.1 Analysis Explored

The data collected, about the change of minimum, maximum temperature over the years as well as the precipitation/rainfall over the years was analyzed using two methods.

In the first method we made use of machine learning techniques like linear regression and support vector machine (SVR to be precise) to predict these climate factors with years as the predictor. In the second method, we implemented time series analysis to obtain deeper insights into the data collected. The time series analysis/model was implemented using ARIMA (Autoregressive Integrated Moving Average). ARIMA is a statistical model that makes use of the time series data to forecast future trends.

3.2 Training Procedure

For implementing machine learning for the climate factors over the years, we first split the dataset into training and testing sets. We used both random splitting and time sequenced based splitting. Three different models were built one for each of the climate factors as the response variable and year as the predictor variable. We evaluate the model on the test dataset. In addition, we use the chosen model (linear regression or support vector machine) to further predict the change of the climate factor over the future years.

For the time series analysis using ARIMA model, first the structure of data was tested to find out whether the data obtained is stationary or not. After this ACF and PACF graphs were plotted to get the initial values of the parameters for the ARIMA model. The parameters are, p, the number of autoregressive terms in the model, d, is the number of differentiations

applied on the time series values and q refers to the number of moving average terms in the model.

3.3 Evaluation Procedure

To evaluate the machine learning models we make use of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R-Squared. The model that explains the data most accurately is used to predict the climate change of the future.

For the time series analysis, we make use of Mean Squared Error (MSE) and also evaluate the residual plots for each of the climate factors we have chosen.

4. Results/ Discussion

4.1 Machine learning models:

a. Random splitting for train and test set

In the case of minimum temperature, the linear regression model demonstrates a high level of accuracy and a good fit. It yields the following performance metrics: Mean Squared Error: 0.134 and R-squared score: 0.398. The results suggest that the minimum temperature is predicted to steadily increase over the years and is expected to surpass 28 degrees Celsius by the year 2500. We can observe the fitted lines and the predicted outcomes in Figure 6.

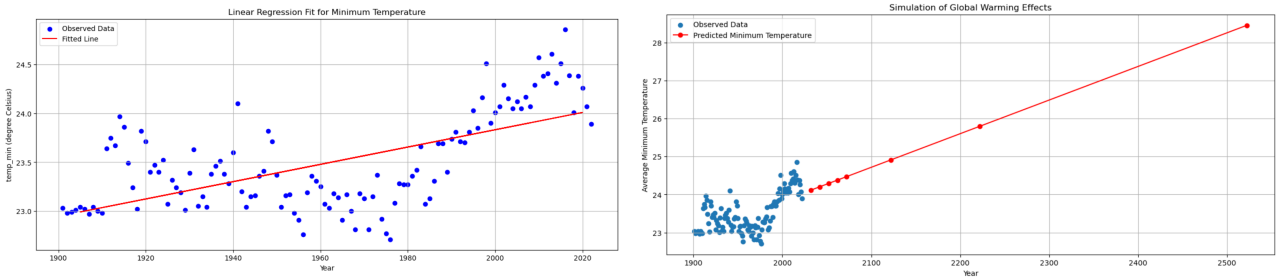

Figure 6: Fitting and Forecast of linear regression in minimum temperature

Similarly, for the maximum temperature, the model provides reliable predictions and a strong fit. The outcomes indicate that the maximum temperature will progressively rise year by year, with expectations of exceeding 34 degrees Celsius by the year 2500.The results are shown in Figure 7.

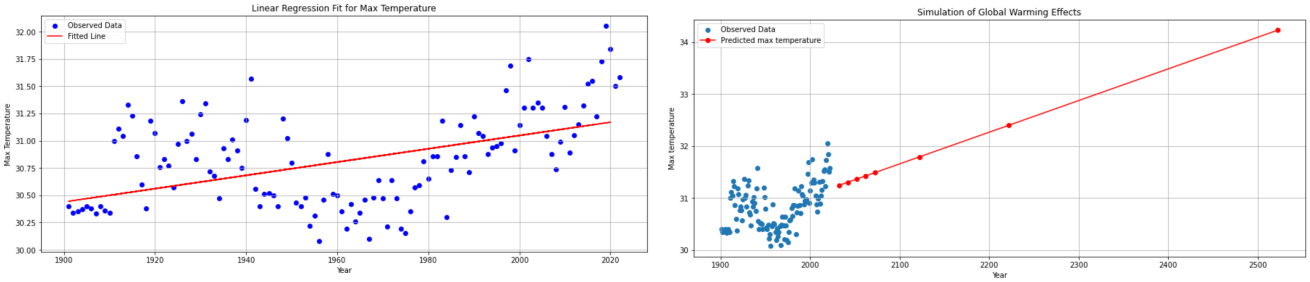

Figure 7: Fitting and Forecast of linear regression in maximum temperature

However, as for the rainfall, we have more confidence in the results obtained from the SVR model, as shown in Figure 8. It is well-known that minimum temperature and maximum temperature exhibit a strong correlation, and their future trends are similar, both showing a year-by-year upward trend. However, in the case of rainfall, the predictions indicate that it will remain relatively constant in the future. Additionally, the model evaluation metrics for rainfall are not as favourable, with results such as Mean Squared Error: 84098.146 and R-squared score: 0.002.

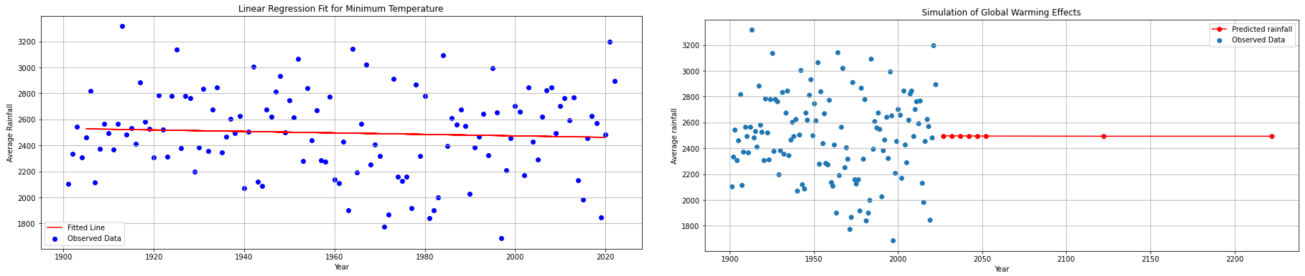

Figure 8: Fitting and Forecast of linear regression in rainfall

b. Time sequence based Splitting

To prevent leak of data to the machine learning model, we made use of time sequence based splitting, where we split the data from 1901-2010 as the train data and from 2010-2022 as train data.

Minimum temperature:

Models	**R-squared**	**MSE**
Linear regression	**-5.943**	**0.440**
Linear SVR	**-5.627**	**0.420**
Non-linear SVR (using RBF kernel)	**-0.065**	**0.067**

Maximum temperature:

Models	R-squared	MSE
Linear regression	**-2.650**	**0.358**
Linear SVR	**-2.640**	**0.357**
Non-linear SVR (using RBF kernel)	**-2.481**	**0.342**

Rainfall:

Models	R-squared	MSE
Linear regression	**-0.102**	**140704.223**
Linear SVR	**-0.062**	**135581.193**
Non-linear SVR (using RBF kernel)	**-0.017**	**129825.557**

4.2 statistical model: ARIMA

The time series for minimum temperature, maximum temperature and rainfall are all non-stationary which means there is non-randomness in the structure in nature. This is ascertained after obtaining p-value of more than 0.05 (assuming 5% level of significance) upon implementation of Augmented Dickey-Fuller Test. So, our model could only be ARIMA among various time series models. By visualizing raw data, rolling average, rolling standard deviation and comparing them year by year, it is found that there is no significant trend. By testing the autocorrelation and seasonality utilizing ACF and PACF, we discovered that:

- Due to the annual nature of the data, there should not be any seasonality.
- The ACF plot for temp_min has a 9th-order trailing and a 2nd-order truncation of the PACF plot, which indicates AutoRegressive Order=2, Differencing Order=1, and Moving Average Order=9 for ARIMA model
- The ACF plot for temp_max has a 6th-order trailing and a 2nd-order truncation of the PACF plot, which indicates AutoRegressive Order=2, Differencing Order=1, and Moving Average Order=6 for ARIMA model.
- The ACF plot for rainfall has a 1st-order truncation and a 1st-order truncation of the PACF plot, which indicates AutoRegressive Order=1, Differencing Order=1, and Moving Average Order=1 for ARIMA model. Noted that due to the bad model performance ARIMA (1,1,1) has got, we tuned the parameters by iterating all possible values of parameters with the criteria of MSE. The final model is ARIMA (1,1,2).

indicators	ACF/PACF	ARIMA parameters
min_temp	ACF: 9th-order trailing PACF: 2nd-order truncation	ARIMA (p=2, d=1, q=9)
max_temp	ACF: 6th-order trailing PACF: 2nd-order truncation	ARIMA (p=2, d=1, q=6)
rainfall	ACF: 1st-order truncation PACF: 1st-order truncation	ARIMA (p=1, d=1, q=2)

For minimum and maximum temperatures as well as rainfall, the ARIMA models predict relatively stable future trends. Specifically, the minimum temperature is projected to remain around 24.1 degrees Celsius, the maximum temperature is expected to hover around 31.6 degrees Celsius, and the rainfall is anticipated to stay at approximately 2500 mm for an extended period. See more details in Figure 9.

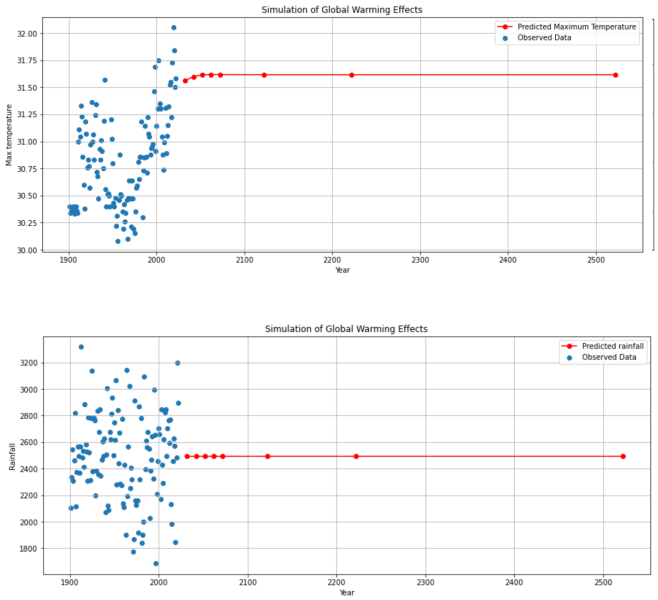

Figure 9: Forecast of ARIMA in different indicators

Meanwhile, the residual plots of all indicators' ARIMA models exhibit no discernible patterns or trends, show no autocorrelation, and remain stationary. The average Mean Squared Error (MSE) for the minimum and maximum temperature models are 0.051 and 0.070, respectively, indicating that the models have relatively small average prediction errors and can be considered reliable. However, it is worth noting that despite parameter adjustments, the

ARIMA model for rainfall still yields a high average MSE of 116805, suggesting a relatively large average prediction error and reduced reliability.

4.3 Comparison across models

a. Comparison across the machine learning models

For the machine learning models we focus on the time sequence based splitting, since it is more appropriate, as random splitting is inadvisable due to it causing data leakage.

In general, R-squared scores are negative with the use of linear regression, linear SVR, non-linear(rbf) SVR to predict existing data. This suggests that these models fit the data very poorly and may not be suitable for predicting climate change data. However, it is notable that non-linear SVR performs least poorly in predicting minimum temperature, maximum temperature, and rainfall.

This suggests that the relationship between minimum temperature, maximum temperature and rainfall and time is definitely not linear, and may be more complex than what is reflected within these three models. This mirrors the multi-faceted nature of climate change in reality.

b. Overall comparison

It is evident that the ARIMA model gives more accurate predictions of minimum temperature, maximum temperature and rainfall.

5. Conclusion

The following analysis is based on results from the ARIMA model.

a. Analysis in the short run:

5-year analysis

	Minimum Temperature	Maximum Temperature	Rainfall
Previous 5 year Average	24.122	31.74	2598.942
Future 5 year Average	24.16	31.64	2493.5863

In 5 years, the average minimum temperature is projected to rise by 0.038 degree celsius, while the average maximum temperature is projected to fall by 0.10 degree celsius. The decrease in maximum temperature could be attributed to the La Niña effect present in 2020-2022, which may lead to temporarily lower temperature. Overall, it is conceivable that with the rise in average minimum temperature, there is likely to be an increase in temperature in the short run. Given the high level of humidity in Singapore, the temperature increase will likely be acutely felt and heat-related illness may become a growing concern. Also, research

has shown that an increase in temperature may negatively impact cognitive performance(Brink et al., 2020), which may affect productivity and economic growth. Thus, there is an urgent need for the Singapore government to come up with stop-gap measures to help residents deal with rising temperatures in their day-to-day activities. For example, there will likely be a need for greater heat insulation and flexibility in dressing in schools and workplaces.

The average rainfall is projected to fall by 4% in 5 years, which may pose a risk to Singapore's already limited water supply. Thus, the Singapore government may have to find ways to diversify its water sources, including importing more from other countries. The possible decrease in Singapore's water supply may increase the risk of inflationary pressures in the future, which may affect standard of living for residents in Singapore in the near future.

Overall, it is clear that increased government expenditure is required to help residents manage the effects of climate change in the short term.

b. analysis in the long run

10-year analysis

	Minimum Temperature	Maximum Temperature	Rainfall
Previous 10 year Average	24.329	31.546	2495.919
Future 10 year Average	24.13	31.62	2492.955

Average minimum temperature is projected to decrease by 0.20 degree celsius in the next decade, while average maximum temperature is projected to increase by 0.07 degree celsius in the next decade. This suggests not only a rise in temperature in the next decade, but increased temperature volatility in the next decade. Thus the temperature changes projected for the next decade are likely to be more adverse than the 5-year projections. The average rainfall decrease projected in the next decade is more moderate compared to the 5-year projections; with a decrease of 0.12% in the next decade. However, weather changes may become more pronounced with the increasing environmental damage brought on by unexpected increasing military efforts worldwide, such as in the Russia-Ukraine War and Israel-Palestinian conflict.

Thus, it is clear that stop-gap measures will not be sufficient given that the increase in temperature will continue to persist in the long run. Therefore, the Singapore government needs to make systematic changes to mitigate the increase in temperature and decrease in rainfall in the long run. Singapore has recently embarked on the Green Plan 2030 in its bid for sustainable development to mitigate climate change, in which one of the proposed actions

is to restore nature in its landscape. If this is brought to fruition, increased vegetation could help to reduce surface temperature and mitigate temperature changes in the next decade.

6. Future Work

There is scope to reflect greater complexity in how climate change data evolves over time in the linear regression and SVR models. For example, engineered features such as moving average, lagged predictors in SVR, and the El Niño and La Niña effects can be incorporated into our models to improve predictions.

Also, there is merit in performing cross-sectional multivariate analysis, for e.g. by using measures of pollution as predictors to rainfall and maximum and minimum temperature. This can help to improve the performance of machine learning models.

References

Brink HW, Loomans MGLC, Mobach MP, Kort HSM. Classrooms' indoor environmental conditions affecting the academic achievement of students and teachers in higher education: A systematic literature review. Indoor Air. 2021 Mar;31(2):405-425. doi: 10.1111/ina.12745. Epub 2020 Oct 21. PMID: 32969550; PMCID: PMC7983931.