**Using weather data and climate drivers to predict flooding in Victoria**

JD Chiang, Nicholas Nugent, Ryan Ngo, Bob Chen, William Mansfield

**Abstract**

Historical data of climate driver indices and weather were used alongside machine learning models to see if the occurrence of floods could be predicted.

No strong correlation could be found, and an accurate, reliable model could not be developed. There were some hopeful signs indicating the model was headed in the right direction, including increased predictive accuracy as more weather data was input into the machine learning models. The models and experimental findings are included in this report, and the code used is included in the Jupyter notebook file. The findings from this research did not have meaningful impact on the detection of floods, and further work needs to be done in order to achieve this. Numerous limitations were present that could be addressed, including time limitations and lack of manpower to handle the data exploration. Addition of more weather data may prove to be valuable as suggested by the models.

**Table of Contents**

[**Abstract** 2](#_Toc105009010)

[**Introduction** 4](#_Toc105009011)

[Contributions 6](#_Toc105009012)

[**Data Manipulation and Modelling** 7](#_Toc105009013)

[Data Preparation 7](#_Toc105009014)

[Data Processing and Manipulation 8](#_Toc105009015)

[Data Exploration and Analysis 10](#_Toc105009016)

[Data Modelling 13](#_Toc105009017)

[**Conclusion** 19](#_Toc105009018)

[References 20](#_Toc105009019)

**Introduction**

Australia is an extensive and diverse country, covering 7.692 million square km2, spanning three separate oceans and six distinct climate groups: equatorial, tropical, subtropical, desert, grassland and temperate (Australian Wildlife Journeys, n.d.). Characterised by numerous environmental and geographic factors such as latitude, proximity to the coast, presence of natural features (e.g., woods, hills and mountains), movement of ocean currents and etc, these climates each have varying levels of rainfall and temperature that creates varying degrees of difficulty for human settlement and utilisation. For instance, much of Australia is covered by desert and grassland biomes that are relatively drier and hotter, rendering them to be less productive and less inhabitable, causing 86.24% of Australians in 2020 to live in urban environments (Statista, 2021) which are situated mostly near the coast where the climate is wetter and cooler. This type of climate makes Melbourne and the rest of Victoria susceptible to flooding during periods of heavy rainfall, particularly in lower lying areas, with a large percentage of Australians being put at risk in such events.

The El Nino Southern Oscillation (ENSO) has a big influence on Australia’s climate over winter and spring. ENSO describes the position of warm and cool water, the strength of winds and atmospheric pressures in the Pacific Ocean regions. The Pacific Ocean winds blow from an easterly direction pushing moist air towards Australia. This tropical air is a big source of rain across eastern Australia including Victoria. During La Nina, the waters around Australia are warmer, meaning ENSO pushes greater amounts of this moist air, leading to more potential rainfall in Victoria. The converse is true for El Nino years, where the water surrounding Australia is cooled, resulting in drier conditions for Australia (Agriculture Victoria, 2021). These events occur irregularly, switching every 2 to 7 years and can last anywhere between 9 months and 2 years, creating long-lasting impact on the affected regions (Bureau of Meteorology, n.d.).

The Southern Oscillation Index (SOI) is one of the key indices used to predict and gauge how strong El Nino and La Nina events are, providing a rainfall outlook for Victoria. Specifically, the SOI represents the difference in air pressure, measured using a barometer, between Tahiti and Darwin, where strong positive values indicate pressure in Tahiti is lower than Darwin suggesting a La Nina event and strong negative values indicating that the pressure in Darwin is lower than Tahiti, suggesting an El Nino event.

The Dipole Mode Index (DMI) also sometimes referred to as the Indian Ocean Dipole (IOD), is an indicator of the east-west temperature gradient across the Tropical Indian Ocean. Like El Nino and La Nina in the pacific, the IOD gauges the relative temperatures in the East and West Indian Ocean. It is calculated as the difference between Western Tropical Indian Ocean surface temperature and the South-Eastern Tropical Indian Ocean surface temperature. A positive DMI value refers to the West Indian Ocean having warmer temperatures than the East Indian Ocean, and vice versa for negative DMI values.

Negative DMI values tend to be less frequent and not as strong as positive DMI events, but can still bring severe climate conditions, such as heavy rainfall and flooding, to parts of Australia.

The Southern Annular Mode or SAM is the north-south contraction of the strong westerly winds that blow constantly in the southern hemisphere. This movement influences the strength and position of cold fronts, affecting rainfall across southern Australia. The SAM is divided into three phases: neutral, positive, and negative, each lasting one or two weeks. The index has its greatest influence during the winter. The belt of winds compresses towards Antarctica during a positive SAM phase, creating stable and dry conditions, while contrarily the winds intensify and move northwards during a negative SAM phase, leading to strong westerly winds and frontal activity, bringing more storms across southern Australia and triggering more intensive periods of rainfall.

With the effect these climate drivers have on the Australian climate and more specifically rainfall, it was believed they could be used to reliably predict floods utilising machine learning models. This information could be then put forth to the Australian authorities to help with earlier detection and put in place preventative and safety measures to minimise the overall social, economic and environmental impact of the disasters.

The supplied information was in the form of monthly data on the following climate driver indices that influence Australia: The Southern Oscillation Index, the Dipole Mode Index, and the Southern Annular Mode, as well as data on monthly rainfall, minimum and maximum temperatures for years ranging back to 1856. We supplemented this provided data with Victorian flood data containing the month and year of flood occurrence from 1971 to 2022, obtained from Flood Victoria.

# Contributions

The members of the group were JD Chiang, Nicholas Nugent, Ryan Ngo, Bob Chen and William Mansfield. JD oversaw the completion of the project and was specifically in charge of the data handling and modelling, the final writeup of the report, and the presentation speech and keynote slides. Nicholas, Ryan, and William gathered some background information on the weather indices for the presentation. Additionally, Nicholas provided a summary of the exploratory data analysis undertaken, and Ryan analysed the processing of the data. Bob contributed to the introduction of the speech and report.

**Data Manipulation and Modelling**

# Data Preparation

The data was first analysed using Pandas describe and info functions. The year 2022 was dropped due to having incomplete data and thus containing NaN values for all the datasets. Following an overview of the minimum temperature, maximum temperature and rainfall data, the unnecessary columns ‘Product Code’ and ‘Station Number’ were dropped. For the SOI index, the March month was found to be string values instead of float, and so this was converted using Pandas to\_numeric function. In the rainfall dataset, the October month was found to have a NaN value for one of the years. This was filled using the mean for October across all years as that would most likely be closer to the real value than 0, using the fillna function. The ‘Unnamed: 0’ column of the DMI and SAM datasets were renamed to ‘YEAR’ to keep the same format as the SOI dataset and allow for easier manipulation in the future. The SAM dataset needed to have the new ‘YEAR’ column values converted into integer format, rather than float to match the other datasets. This was done by using the astype(‘int’) function on the filtered ‘YEAR’ data series from the SAM dataset and then merging it with the original dataset using pd.merge. The datasets were all filtered so only data from 1971-2021 were used, in order to match the Victorian Flood data, which was limited to the year 1971 onwards. The index for the climate driver datasets, previously indexed by the ‘YEAR’ column, was reset using reset\_index, and the years were instead added to the columns. All datasets were given a ‘name’, as part of the Pandas package, to allow for easier referencing later. These steps transformed the datasets to share the same format, which would allow for easier manipulation. A function called clean\_load\_data was defined that would automate all these processes, in order to ‘reset’ the data to new prior to a model being applied or exploration undertaken.

# Data Processing and Manipulation

There were 2 ways to format the cleaned data for exploration: by season and by month. For future manipulation, transforming the datasets from wide to long was necessary. This was unable to be done in the format wanted by any of the included Pandas’ functions, including pivot\_table, so a function needed to be created. To process the data into seasonal format, through the get\_seasonal\_data function, each year was split up into summer and winter halves: spring and summer for the summer half, and autumn and winter for the winter half. This was initially done due to the seasonality of the climate drivers, which are more in effect during certain seasons, as well as making it easier to work with the data due to the data being handled by a single team member. The monthly data for all datasets was then also manipulated to represent the average for the corresponding season halves of every year. The result was a merged Pandas dataframe, containing each climate driver index and the temperature and rainfall data for Summer and Winter, for every year. A categorical column could be added to the seasonal dataset when necessary, using the add\_flood\_col\_seasonal function, which indicated if a flood occurred in the seasonal half for every year between 1971 and 2021.

Table

Description automatically generated

Figure 1: Clip of the finished seasonal data format for the year 1971, W representing winter half and S representing summer half data.

A picture containing text

Description automatically generated

Figure 2: The same dataframe processed through the add\_flood\_col\_seasonal function; a 0 in the ‘Flood’ column indicating no flood for that season/year and 1 indicating flood occurred.

The get\_monthly\_data was later added to process the data into months for every year, instead of seasons, to assist in data exploration. This was done in a similar fashion to the seasonal data, converting the raw wide format datasets into long datasets, with 12 rows for every year corresponding to each month.

**Table

Description automatically generated**

Figure 3: Clip showing part of the dataframe produced by the get\_monthly\_data function, for the same year (1971) as the seasonal figures above.

Table

Description automatically generated

Figure 4: The same dataframe as Figure 3 with ‘Flood’ column added, indicating if a flood occurred in a particular month and year.

These two formats formed the basis of the data exploration and modelling undertaken, outlined in the following sections.

# Data Exploration and Analysis

The first step in the data analysis was to create annual means for each of the provided climate driver indices, being DMI, SOI and SAM respectively. This was created using the code shown below in Figure 5.

Graphical user interface, text, application, chat or text message

Description automatically generated

Figure 5: Code for annual means

A general correlation table was then created using the annual means for each of DMI, SOI, SAM as well as rainfall, minimum and maximum temperature. The code and table produced can be seen below in Figure 6.

A picture containing graphical user interface

Description automatically generated

Figure 6: Annual correlation table

As can be seen in the table above, there was a strong correlation between all three of the climate drivers: 0.95 between DMI and SAM, 0.87 between SOI and DMI and 0.83 between SOI and SAM. Furthermore, there were also strong correlations between all the climate drivers and minimum temperature, and a slightly weaker correlations between them and the maximum temperature. However, the objective of this project was to create a model which can predict flooding, which in essence is extreme rainfall. Thus, none of these correlations were beneficial to this report. The one row which is beneficial is that of annual rainfall, and as is illustrated in the table, had a very low correlation with all the climate drivers, thus the correlations and annual columns were discarded and not considered for the remainder of the exploration.

Following this, scatterplots were created using the seasonal format of the data, illustrating the relationship between each of the different climate drivers and rainfall. Shown below is the code and scatterplots created for the three different climate drivers.

Chart, scatter chart

Description automatically generated

Figure 7: Rainfall vs DMI

As visualised in the graph above, DMI had a negative correlation with rainfall, although the data is widely spread, there still is a clear direction in the data. This corresponds with the background information on DMI suggesting negative values indicate warmer temperatures around Australian waters, leading to increased rainfall.

Chart, scatter chart

Description automatically generated

Figure 8: Rainfall vs SOI

As expected from the background information, SOI had a positive correlation with rainfall, as positive SOI values indicate the occurrence of a La Nina event. The variations in the data were larger than that of DMI.

Chart, scatter chart

Description automatically generated

Figure 9: Rainfall vs SAM

The SAM index seemed to have no comprehensive correlation with rainfall, with datapoints spread out inexplicably.

This initial data analysis illustrated that while the SAM index seemed to have no correlation with rainfall, both DMI and SOI indices were correlated with rainfall. This correlation was a hopeful sign that floods may be able to be predicted using climate driver data.

# Data Modelling

Due to the correlations observed in the exploratory data analysis, a linear regression model seemed to be a valid approach in gaining insight into the strength of the correlation of rainfall and the indices. To use this model, the data was split into testing and training segments, an 80:20 split. The model uses the training data to place a line of best fit through data points by minimising measured distance between the data points and the line, and a training score is calculated through predicting rainfall for a given training data point on the index, effectively calculating the ratio of the difference in points from the line of best fit, and the total difference between points. The higher the score the better, meaning the model represented the data used for modelling accurately. The same score was calculated for the testing data, representing the testing score, which effectively applies the same model for training to see if it can also explain and potentially predict new datapoints which it had not learned on.

Chart, scatter chart

Description automatically generated

Figure 10: Visualisation of minimising squared difference

As foreseen by the extremely low correlation scores, the linear regression model did not produce good results and produced extremely low training and testing scores, which could be explained largely by the high variability in data points of the indices.

From this result, linear regression did not seem like a sensible approach, and was abandoned.

The nature of the outcome variable was categorical: a flood occurring in a certain time period, and so a logistic regression model was chosen as the next best approach. Since there was only two outcomes for the outcome variable- the outcome measure is either 1 or 0, it follows that drawing a straight line between points with limits to infinity/negative infinity would not be viable. In the logistic regression model, the straight line is replaced by a ‘S’ shaped curved, which is fitted to the data. The curve is chosen through a process that maximises the total likelihood, given the data, rather than the measured distance. The output of the logistic regression model gives the likelihood of a datapoint being either of the categorical outcomes, adding up to 1. This model was first applied to the seasonal dataframe, using DMI, SOI, SAM, minimum temperature and maximum temperature as inputs for the model. The inputs were experimented with, but as suspected from background research on the drivers of rainfall, the combination of all the climate drivers and both minimum and maximum temperatures yielded the best results, for reasons that could not be fully qualified. The seasonal dataset needed to be modelled twice, one for winter and one for summer. Rainfall was omitted from the modelling process, due to the high similarity with the outcome variable, flooding, which is essentially extremely heavy, prolonged rainfall.

Chart, diagram, line chart

Description automatically generated

Figure 11: Logistic regression model curve with categorical outcomes at each limit of the curve. The classification point is highlighted as a point on the curve, the probability after which it was determined a flood did or did not occur

With the output variables being a probability scale for a flood occurring in winter of a particular year, a classification point needed to be determined, defined as the probability at which a flood was decided to have occurred or not. This point was a large focal point of the model as the probabilities were relatively low for a flood occurring, naturally, due to the rare occurrence of such an event compared to a floodless season: 25 seasons out of 102. The point was experimented with using statistical concepts such as the mean and interquartile ranges

Table

Description automatically generated

Figure 12: Output of the logistic regression model for the Winter seasons, ‘class’ column indicates predicted flood occurrences, and ‘true’ column indicates the actual occurrences of floods

for the probability distribution of the flood, to try to maximise the correct prediction of floods. This was found to be a double-edged sword; as the classification point was lowered to increase the number of floods accurately predicted, the accuracy of the model reflected in training and testing scores, which compares all the predictions against the actual values, headed lower, as the model raised a greater number of false flags for floods occurring. There also seemed to be no interpretable point at which the floods occurred as seen in Figure 12, Floods occurred even when the model determined the probability of the event to be extremely low, such as index 34 where the probability was 0.11 and the flood still occurred. This raised the possibility that either there was other data that was not considered which could determine a flood occurring, such as wind direction or speed, which from background information was known to influence rainfall, or the fact that a natural occurrence such as floods were by nature harder to predict due to the randomness of the events at times.

The classification point for this particular model was set as the 75th percentile, defined as the top 25% of the probabilities of flood occurring in the data set. This was done through experimentation in maximising the number of cases correctly predicted and minimising the number of false predictions. From a mathematical standpoint, this seemed reasonable too, as the point was far away enough from the mean to isolate the outliers in the probability distribution. The training and testing scores with this model were 0.75 and 0.909 respectively, with the testing model predicting every single case and only falsely predicting 1 event, as seen in Figure 13.

Table

Description automatically generated

Figure 13: Testing outcome for the Winter model. ‘test\_class’ is the predicted flood occurrences, ‘true’ column is the actual flood occurrences. Row 5 was falsely predicted

The exact same model was applied to the summer seasons, where the model suffered in accuracy with training and testing scores of 0.75 and 0.636 respectively. The classification point was a large part of this reduction in accuracy, as a greater number of floods were occurring in summer with extremely low probabilities compared to winter.

The findings from this initial modelling raised some hopeful signs, and it was decided to apply the logistic regression model to monthly data, as there was a large variation month to month in the data, particularly with indices. Monthly predictions would also be more useful to the aim of the research, in providing a more specific timeframe of when a flood would occur. The probabilities of flood occurring were significantly lower for the monthly model, due to only 29 flood months out of 612 from 1971-2021. The flood months were even more unpredictable than the seasonal model, as floods occurred when the probability of such occurrence was as low as 0.02% as shown in Figure 14.

Table

Description automatically generated

Figure 14: Row 95 highlights a month where the likelihood of a flood occurring was 0.027 according to the model, yet was a flood month

The classification point again needed experimenting, which was difficult due to the extremely low probabilities couples with a large variance in data points. It was settled to be at the mean of the predicted probability of a flood occurring. This resulted in training and testing accuracy scores of 0.615 and 0.65, highlighting the impact that the lower classification point had on the accuracy of the model, raising a lot of false flags. Despite the large number of false flags, the model was able to determine 18/27 flood cases correctly for the training dataset, and 2/2 flood cases correctly for the testing dataset. Due to time and team member limitations further modelling could not be successfully completed.

**Conclusion**

The research and modelling proved to be largely inconclusive as to whether climate driver indices were able to be utilised in flood detection in Victoria.

There were numerous limitations that hindered project completion, including a small data size, lack of knowledge in other machine learning models, as well as the random element that is present when dealing with the weather. Additionally, the data handling being completed by a single team member meant further avenues for potential research were not able to be utilised, such as the addition of more weather data as suggested in the modelling section, or additional geographical regions being studied to gather a larger dataset. The large variability of drivers on the Australian climate may also have been a contributing factor for the lack of reliable evidence. These limitations may have meant that key areas were overlooked, or the wrong questions were asked of the data itself. Some of these limitations could be rectified in the future, such as implementing the use of more weather variables such as wind and humidity, as well as using a larger set of data to make these calculations and seeing the results. There is further work to be done in order to limit the number of false flags the system raised, so that it can serve as a more consistent predictor of floods, as well as improving the accuracy through a different approach that has yet to be determined from this study. A broader approach through modelling the seasonal dataset seemed to bear more accurate results compared to the smaller scope of the monthly models, however due to the limited data size this could not be reliably extracted from the results. The only conclusive finding from this research was that there was additional elements present that impacted the occurrence of floods, that were not quantified or accounted for in our models or research, which could be a focus point in the future.

In its current state, while showing some hopeful signs, floods could not be reliably or accurately predicted using the supplied climate drivers and weather data.

**References**

*Australia: Urbanisation from 2010 to 2020*. (December 2021). Statista. <https://www.statista.com/statistics/260498/degree-of-urbanization-in-australia/#:~:text=Degree%20of%20urbanization%20in%20Australia%202020&text=This%20statistic%20shows%20the%20degree,Australia's%20population%20lived%20in%20cities>.

*Climate & Seasons*. (n.d.). Australian Wildlife Journeys. <https://australianwildlifejourneys.com/plan-your-trip/climate-and-seasons#:~:text=There%20are%20six%20distinct%20climate,Kangaroo%20Island%2C%20Eyre%20Peninsula>).

*Observation of Rainfall*. (n.d.). Bureau of Meteorology. <http://www.bom.gov.au/climate/cdo/about/rain-measure.shtml#:~:text=Introduction-,Instruments,up%20to%2025mm%20of%20precipitation>

*What is La Nina and how does it impact Australia?* (n.d.). Bureau of Meteorology. <http://www.bom.gov.au/climate/updates/articles/a020.shtml#:~:text=The%20presence%20of%20La%20Ni%C3%B1a,than%20El%20Ni%C3%B1o%20years%201>.

*The Climatedogs — The six drivers that influence Victoria's climate (2022)* Agriculture Victoria

<https://agriculture.vic.gov.au/climate-and-weather/understanding-weather-climate-and-forecasting/the-climatedogs-the-six-drivers-that-influence-victorias-climate/enso>