
**School of Engineering and Computer Science**
Te Kura Mātai Pūkaha, Pūrorohiko

PO Box 600
Wellington
New Zealand

Tel: +64 4 463 5341
Internet: office@ecs.vuw.ac.nz

# Genetic Programming for Antarctic Ice Sheet Modelling

Samuel Mata

Supervisors: Dr. Bach Nguyen, Dr. Bing Xue

Submitted in partial fulfilment of the requirements for
Bachelor of Science with Honors in Artificial Intelligence.

**Abstract**

This project aims to investigate and evaluate the use of Genetic Programming *(GP)* and Evolutionary Learning techniques for the long-term modelling of Antarctic Ice Sheet measurements.

# Contents

# 1. Introduction

## 1.1 Problem Statement

## 1.2 Project Objectives

# 2. Background

- Do not focus on this for progress report - Machine learning - Read extra papers if have time
- classification / regression - How technology relates to problem

## 2.1 Antarctic Ice Sheet Modelling

## 2.2 Machine Learning

# 3. Exploratory Data Analysis

- Show understanding of dataset - Show original data - Show difference between original data and processed dataset - What findings have been found

## 3.1 Description of Initial Dataset

The dataset used in this project was obtained from Victoria University's Antarctic Research Center. It details the results of several physics-based simulations in effort to predict the future state of the Antarctic Ice Sheet. Specifically, the dataset consistutes 86 files, each representing 1 year of simulation data *(Ranging from 2015 to 2100 inclusive)*. Each of these files contains 2601 datapoints *(Totalling 223,686 across all files)*, with each datapoint representing one cell in a 51x51 grid of the Antarctic Ice Sheet. 8 measures are counted for each datapoint, which can be described in three forms:

**Positional Constants** These encode the constant positional data of each cell in a pair of `x_coordinate` and `y_coordinate` values. Both of these values range from -3,040,000 to 3,040,000 in discrete intervals of 121,600, with the grid being centered around point *(0,0)*, which lies on the South Pole. These are not directly useful for modelling without feature engineering, and primarily serve as a reference for the grid.

**Input Forcings** These are the primary inputs to be utilized in model prediction. These include 3 continuous features; `precipitation`, `air_temperature`, and `ocean_temperature`, which are the respective measurements for each cell provided by the physical simulation.

**Outputs** These are the target outputs for the model to predict. These include 2 continuous measurements; `ice_thickness` and `ice_velocity`, which represent the respective thickness and velocity of the ice in each cell. This also includes `ice_mask`, which is a discrete value representing whether a cell contains grounded ice, floating ice, or no ice at all *(i.e. open ocean)*.

## 3.2 Initial Data Cleaning

Initial analysis was performed without preprocessing to understand the nature of the initial dataset. This analysis excludes the positional constants `x_coordinate` and `y_coordinate`, as these are used for reference and not for prediction or evaluation.

Some minor data cleaning was performed before analysis, so as not to skew the results. Most prominently, several values in the dataset were found to be filler values original NaN *(Not a Number - used to represent empty or faulty data points)*.

The large proportion of NaN and filler values in the dataset is due to the relationship between ocean and ice cells. The majority of cells are in the ocean surrounding the ice sheet -

| Measure | Count | Proportion of Measure |
|---|---|---|
| ice_thickness | 143,819 | 64.29% |
| ice_velocity | 146,083 | 65.30% |
| ice_mask | 124 | <0.01% |
| ocean_temperature | 29,584 | 13.22% |
| precipitation | 385 | <0.01% |

Table 1: Counts of NaN and Filler Values in Dataset

and thus do not have ice measurements. This causes the ice measurements *(ice_thickness, ice_velocity, ice_mask)* to be NaN or filler values in these cells *(typically 0)*. Inversely, the ocean temperature measurements are NaN or filler values in cells containing ice *(typically 9.969...e+36)*, though this value is also used for cells without any measurement taken.

For ocean temperature values, the datapoints containing NaN or filler values were removed from the dataset. This was done as these values are exclusively around the perimeter of the collected data, and only used to provide a square grid despite the circular nature of the collected data. This can be viewed with a spatial heatmap of the ocean temperature values, which shows a clear circular border around the data where these values are not present.

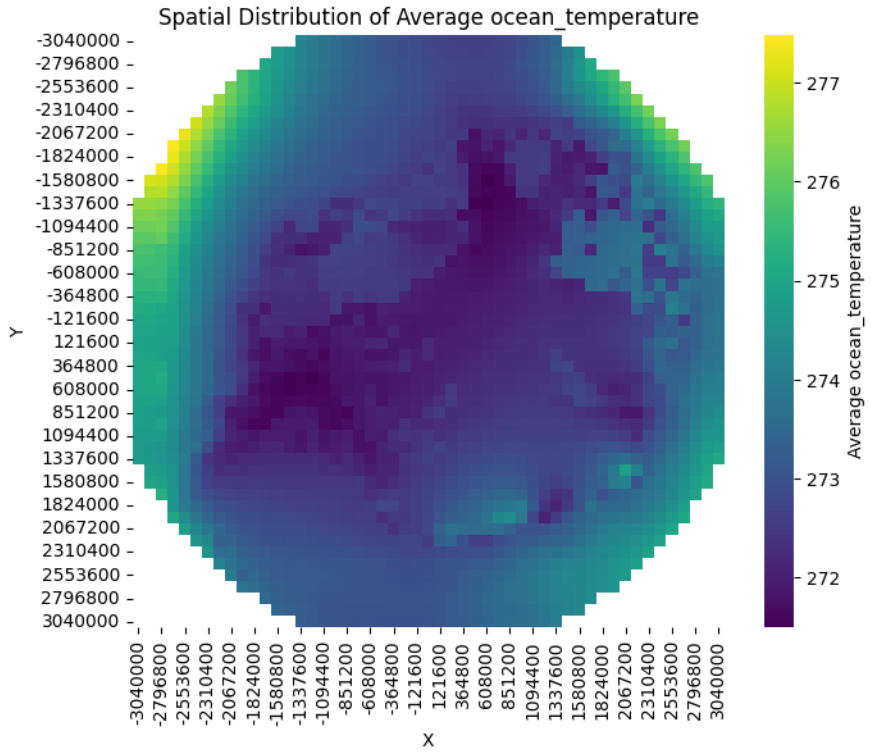


Figure 1: Spatial Heatmap of Average Ocean Temperatue Values

Filler values for ice mask were set to 4, as this value is used to represent open ocean. Additionally, all measures were quantised into integers to reflect the discrete nature of the variable, and to discard some minor variations in values. Ice velocity and ice thickness do not have set values for ocean cells, so these were left as NaN values for the initial analysis.

Finally, the few NaN values in the precipitation variable were filled with the mean of the variable, as these appeared randomly distributed throughout the dataset.

Table 2: Summary Statistics for Input Forcings

| | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| **Precipitation** | 223686 | 575.61 | 364.76 | -522.47 | 204.15 | 638.85 | 871.18 | 2459.46 |
| **Air Temperature** | 223686 | 258.14 | 16.93 | 214.14 | 245.16 | 265.52 | 271.24 | 297.43 |
| **Ocean Temperature** | 194102 | 273.12 | 1.09 | 271.19 | 272.31 | 272.89 | 273.69 | 277.98 |

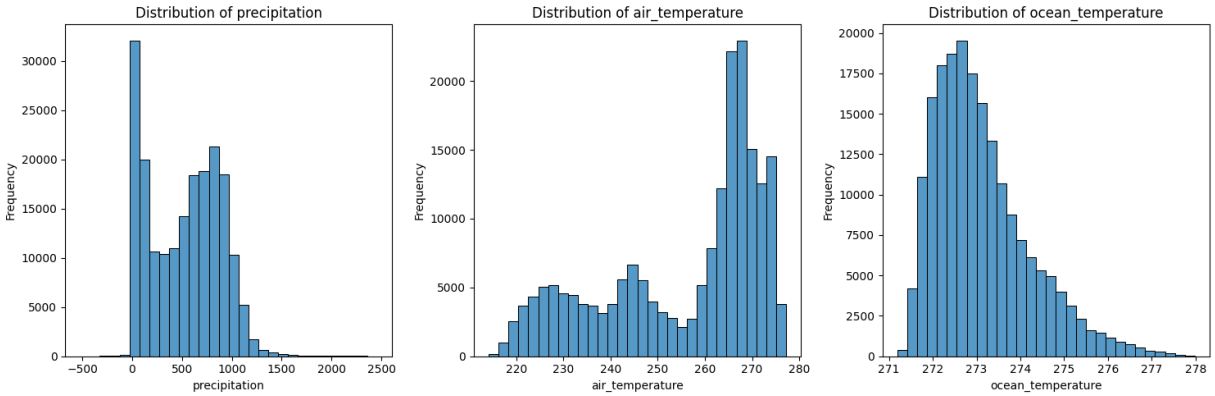


Figure 2: Distribution of Input Forcings in Dataset

## 3.3 Univariate Analysis

With initial data cleaning complete, univariate analysis can be completed on each variable. This was first done with

ANALYSIS
ANALYSIS

## 3.4 Spatial Analysis

## 3.5 Temporal Analysis

## 3.6 Correlation Analysis

## 3.7 Preprocessing

Several stages of preprocessing were performed on the dataset as directed by the analysis. This included several

- Talk about feature before -¿ why a change is justified -¿ after

Table 3: Summary Statistics for Target Outputs

| | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| **Ice Thickness** | 79867 | 1901.61 | 1084.26 | 0.00 | 922.46 | 2061.50 | 2823.69 | 4614.76 |
| **Ice Velocity** | 77603 | 86.49 | 298.34 | 0.00 | 2.78 | 8.63 | 29.53 | 12527.31 |
| **Ice Mask** | 223562 | 3.33 | 0.93 | 0.51 | 2.00 | 4.00 | 4.00 | 4.36 |

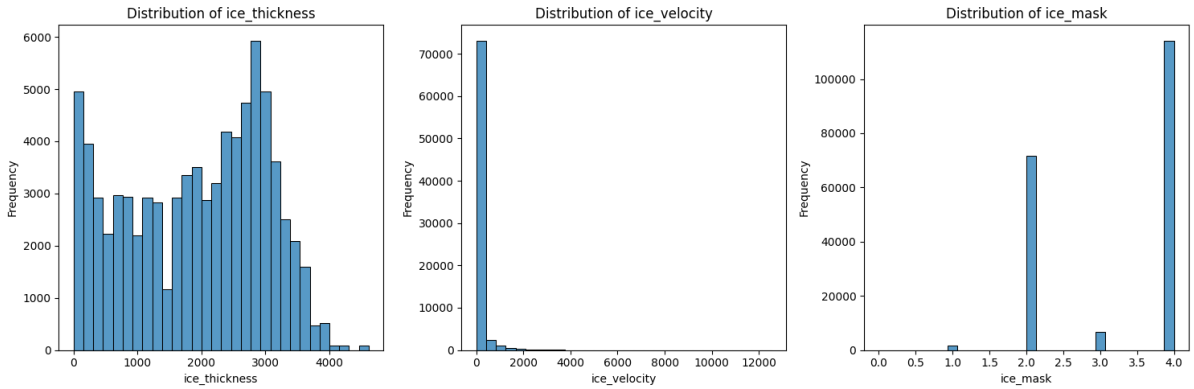


Figure 3: Distribution of Target Outputs in Dataset

### 3.8 Feature Engineering

## 4. Conclusions

- What I have learnt from this - Identify problems in solution - what are potential future works

# Bibliography