SLIPP

Statistical Learning for Isotopic Paleoclimate Proxies

Team: Daniel Cairns, Jakob Thoms, Shirley Zhang (MDS, UBC)
Partner: Anais Orsi (EOAS department, Faculty of Science, UBC)
Mentor: Alexi Rodríguez-Arelis (MDS, UBC)

Image: Doug Clark, Univ. Washington

Image: James Yungel, NASA

Icebreaker

Outline

1. Background
2. Research Question
3. Data
4. Proposed Modelling Approaches
5. Final Data Product

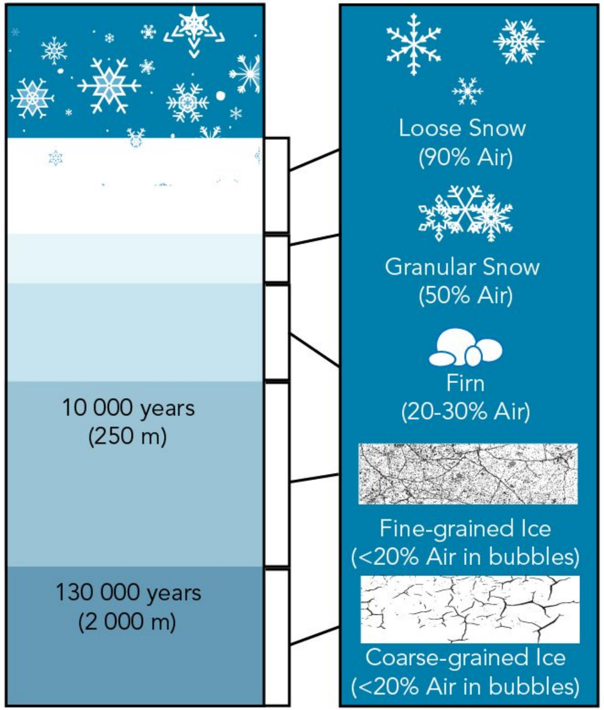
Background

Ice Cores

- Formed through snowfall (**precipitation**)
- Snow accumulates, gets compressed, becomes ice
- Layers = years
- Tell us the composition of snow during that year
  - **Affected by conditions in the atmosphere**

We are interested in:

Concentrations of isotopes in the snow/ice



Isotopes

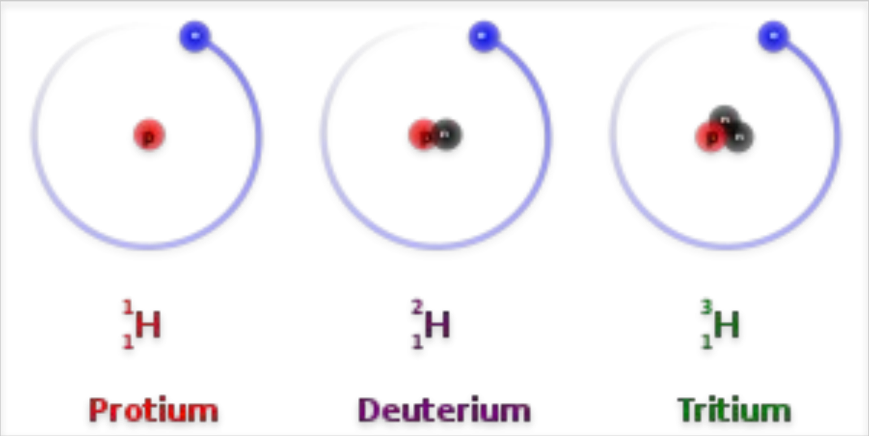


Image: Dirk Hünniger, 2016

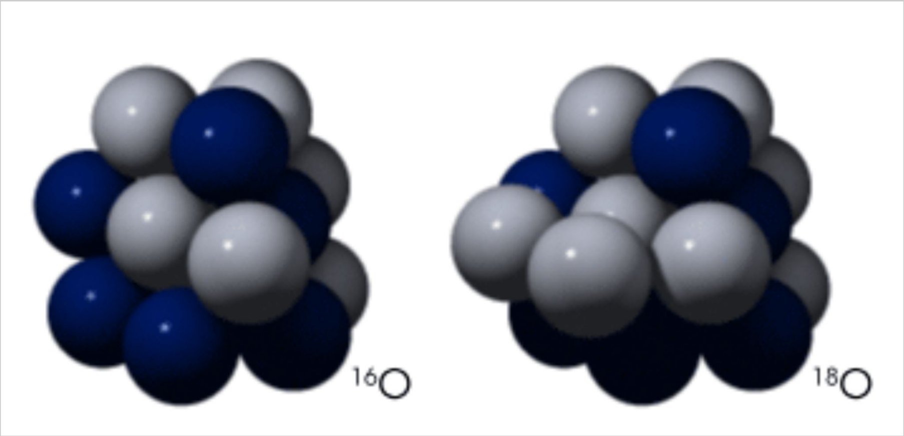


Image: NASA, 2005

Isotopes of H_2O:

- $H_2$16-O - most common (99.7%)
- $H_2$18-O - heavier, less common (0.2%)

δ-Oxygen-18 ($\delta^{18}O$)

- We analyze the **δ-Oxygen-18** value:
  - A measure of the ratio of $^{18}O/^{16}O$ in a sample of water

Interpretation:

- Higher values/more positive = more ^{18}O
- Lower values/more negative = less ^{18}O

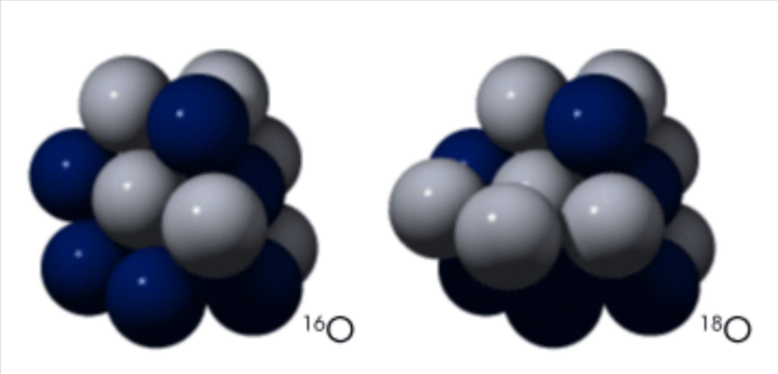


Image: NASA, 2005

Temperature and δ-Oxygen-18

Colder temperatures = more heavy isotopes in ice core samples

1. Takes more energy to evaporate heavier isotopes
2. Heavier isotopes preferentially removed in snowfall (precipitation)

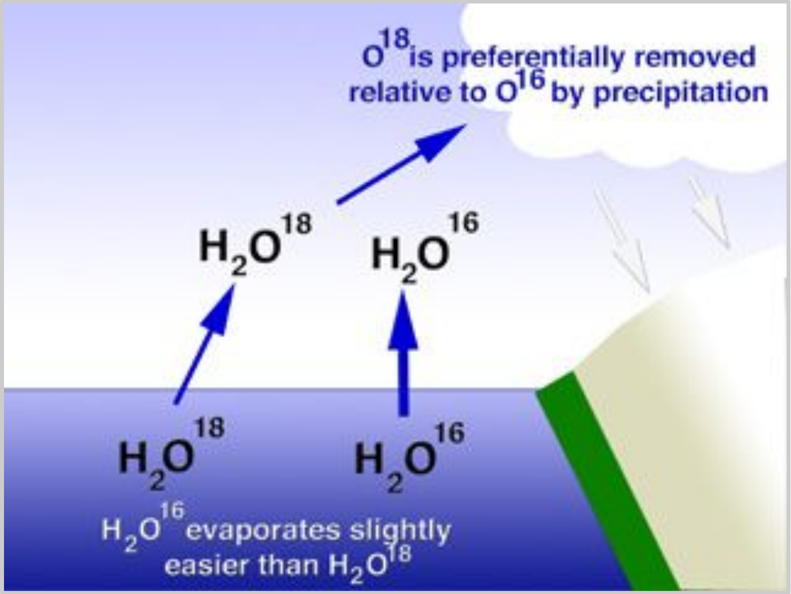


Image: University of Michigan

Circulation Mode, Precipitation, and δ-Oxygen-18

Circulation mode

- Heavier isotopes near the edge (higher $\delta^{18}O$)
- Lighter isotopes near the center (lower $\delta^{18}O$)
- **Depends on how air circulates**

Precipitation

- More precipitation in one region = lighter isotopes in the neighbouring region

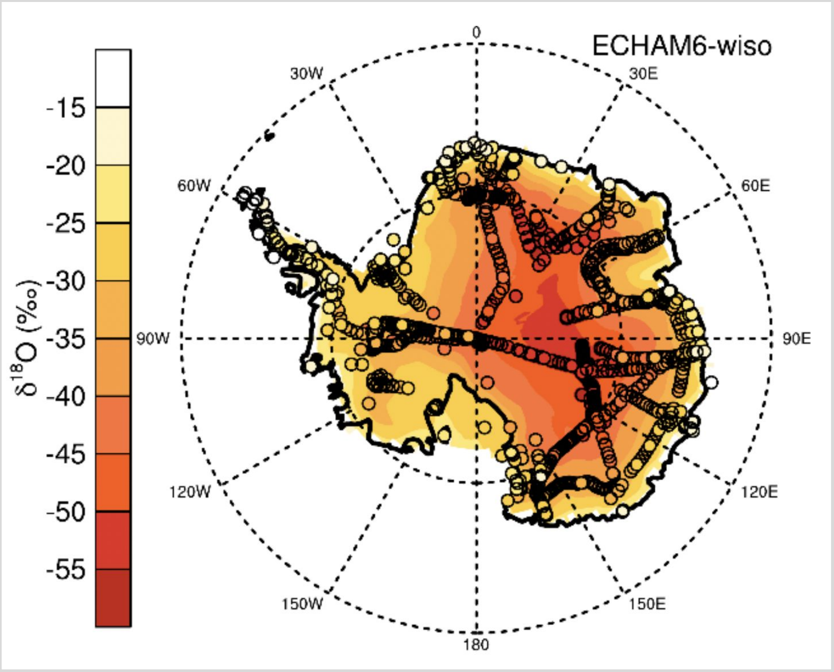
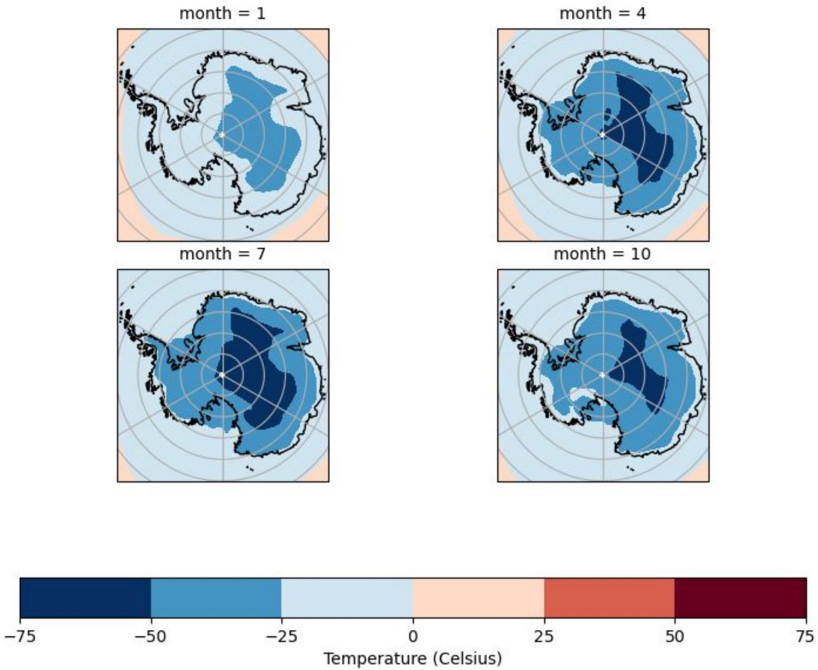


Image: Wang et al., 2022

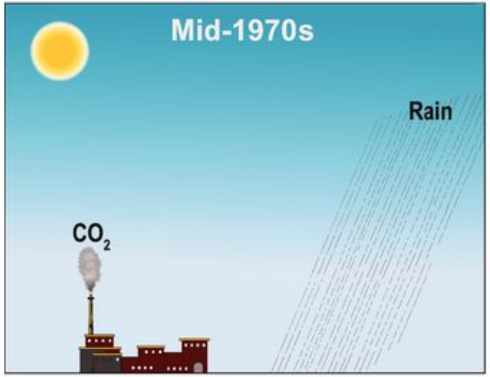
Research Question

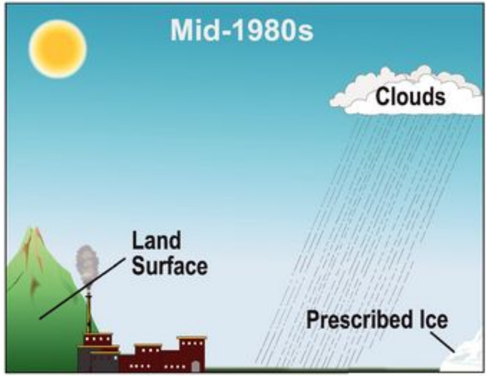
How can we interpret the **relationship** between **isotopic proxies (i.e $\delta^{18}O$)** and weather conditions in Antarctica such as **temperature, precipitation, and circulation mode**?


month = 1
month = 4
month = 7
month = 10
−75
−50
−25
0
25
50
75
Temperature (Celsius)

Data

Global climate models have become more complex over time





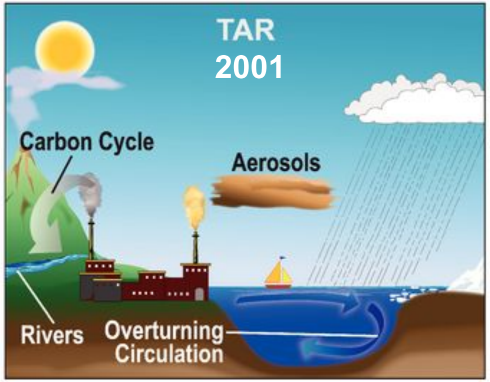
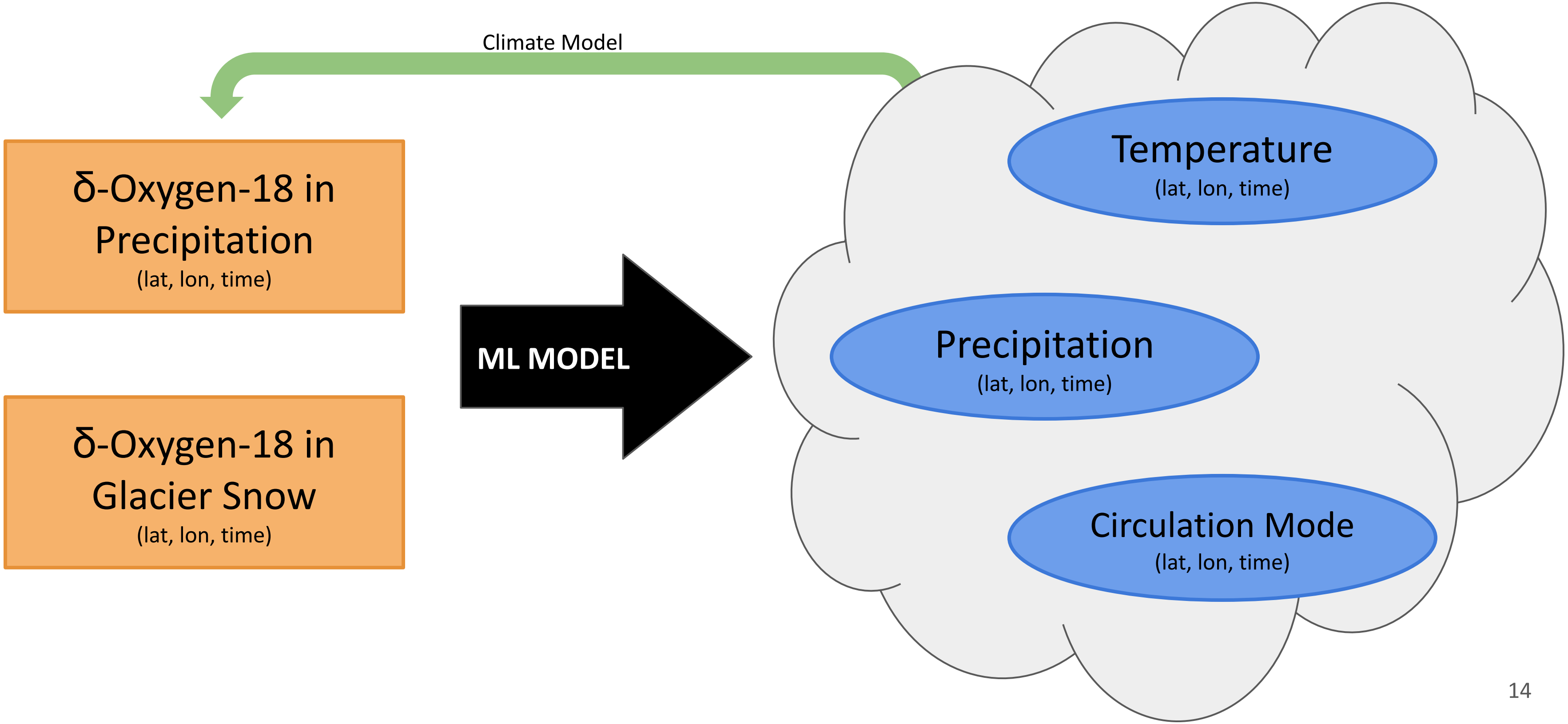


Image: Carbon Brief, 2018

Key Variables from Global Climate Models



NetCDF and xArray Format

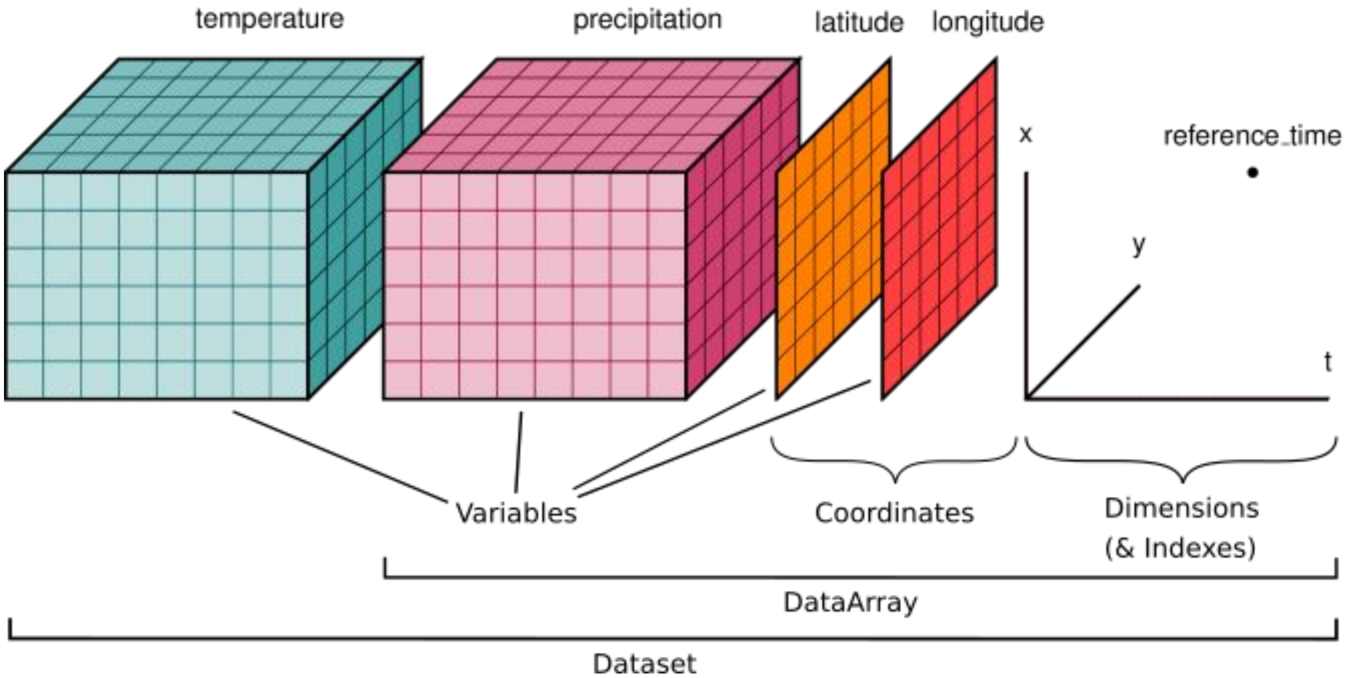


Image: xarray Developers, 2023

Challenge 1: Compatibility

ML packages expect 2d arrays, but our data has 4 dimensions (x, lat, lon, t)

Naive approach: xArray ➡️ Pandas ➡️ sklearn

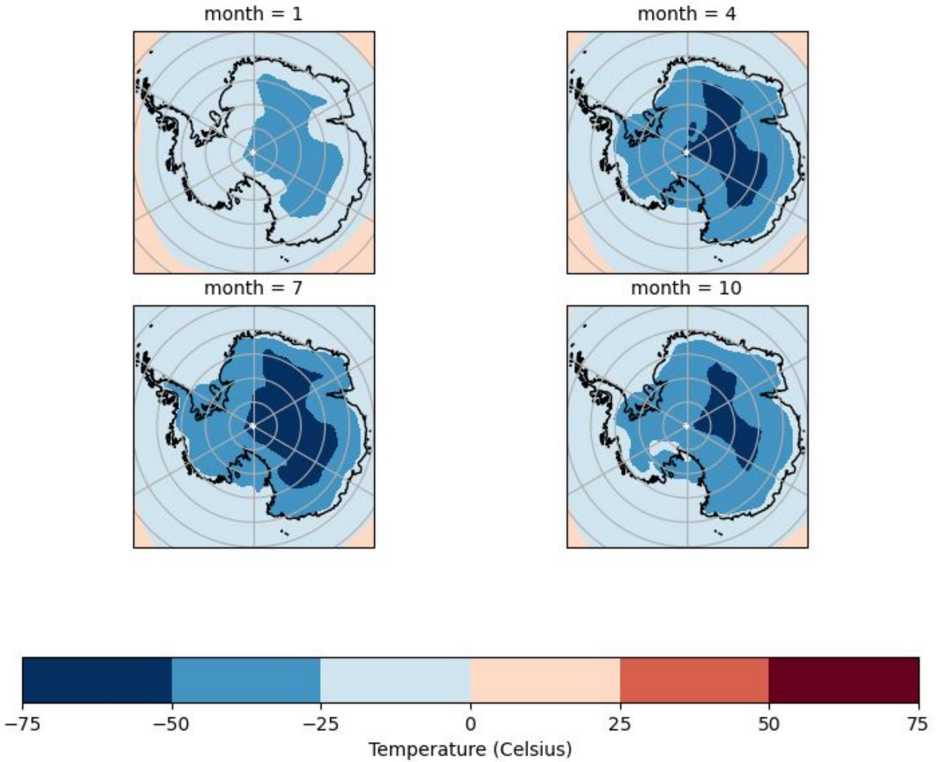
Better plan: NumPy mapping & reshape functions

Challenge 2: Volume

- Climate model outputs 53 x 320 grid over Earth per time slice
- One variable’s data over 120 months contains 2 million data points
- Need **parallelization** and **cloud computing**

Challenge 3: Time & Space Autocorrelation

- Significant Yearly Seasonality
- Global Warming Trends
- Spatial Interpolation

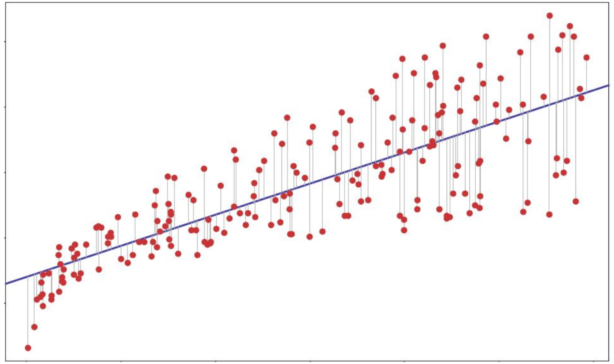


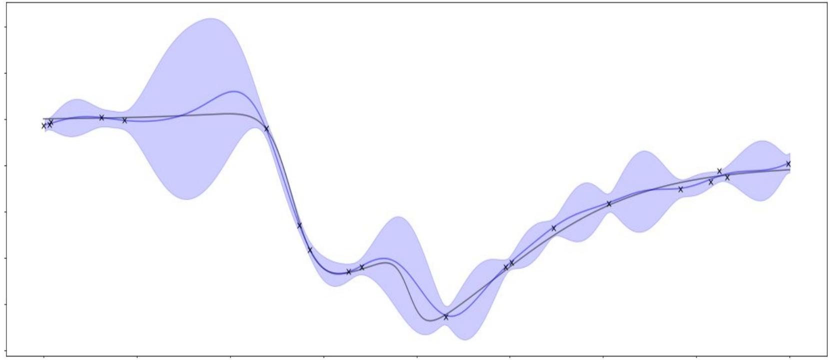
Modelling Approaches

Previously: Linear Regression with Ordinary Least Squares (OLS)

- A fundamental assumption of OLS is **independence** of data points.
- Climate data has strong spatial/temporal correlation

- We will use **gaussian processes (GPs)**, an extension of the OLS model
    - Remove the assumption of independence





Gaussian Processes (GPs)

- Like OLS, a GP has a linear regression component

  + interpretability

- Like OLS, a GP assumes normally distributed error terms

  + confidence intervals for predictions

- Unlike OLS, a GP assumes a correlation structure

  + model spatial/temporal correlation structure of data

- Basic idea: points that are closer together should be more correlated

OLS:

$$y_i = \beta_0 + \beta_1 \vec{x}_{i,1} + \cdots + \beta_d \vec{x}_{i,d} + \underline{\epsilon_i}$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$\mathrm{Cor}(\epsilon_i, \epsilon_j) = 0 \text{ for } i \neq j$$

GP:

$$y_i = \beta_0 + \beta_1 \vec{x}_{i,1} + \cdots + \beta_d \vec{x}_{i,d} + \underline{Z(\vec{x}_i)}$$

$$Z(\vec{x}_i) \overset{marginal}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\mathrm{Cor}(Z(\vec{x}_i), Z(\vec{x}_j)) = \underbrace{R(\vec{x}_i, \vec{x}_j)}_{\text{kernel function}} \quad \text{for } i \neq j$$

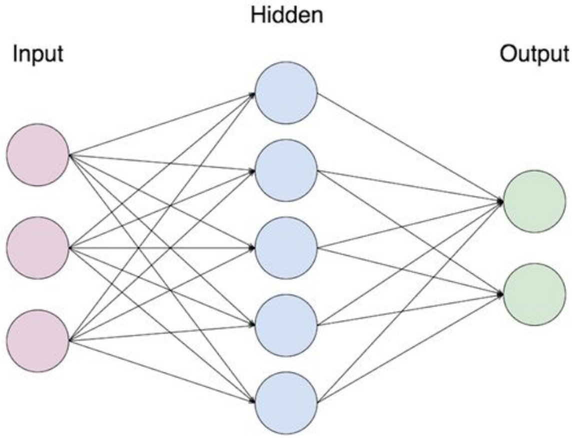
Deep Learning

Advantages

- Flexible, numerous architectures available
- Can handle multivariate outputs
- Don’t need to store training data to make predictions

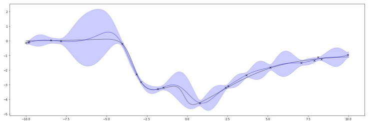
Disadvantages

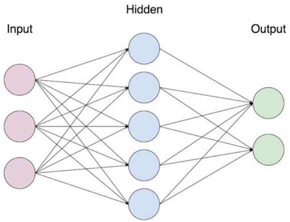
- No confidence intervals
- No interpretability

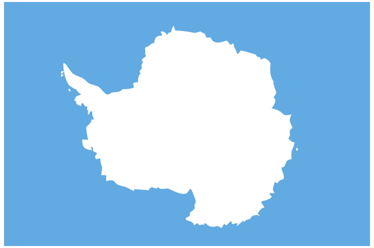


Final Data Product

1. A well documented workflow or notebook exploring:







a. Different **statistical and machine learning (ML) models**

b. How ML models perform for different **regions of Antarctica**, or different **weather conditions**

2. A package with our final models

Where future climate scientists can train and get predictions with their own data

github.com/UBC-MDS/SLIPP

Significance

Antarctic Ice Core Research Community

- Help move towards finding more complex models that would work well with ice core data
- Help compare and contrast different climate models

Generally

- Help characterize the Antarctic climate and understand anthropogenic climate change

Timeline and Milestones

Timeline and Milestones (unfinalized)

Week 1
Hackathon

Week 2
Data preprocessing
Decide how to deseasonalize and/or detrend the time series

Week 3
Initial models
Implement Gaussian Processes and Deep Learning models on very small datasets, fine-tune models

Week 4-5
"Final" models
Leverage cloud computing, train models on a much larger dataset, fine-tune, and interpret the coefficients and output metrics

Week 6
Polish documentation and create package
Finish the workflow notebook, and wrap our code in a package

Week 7
Presentation

Test on other climate models
Test our workflow and models on other climate models to see if results are consistent

Week 8
Draft of report
Write the report, create the presentation

Thank you!