SLIPP Project Proposal

Daniel Cairns, Jakob Thoms, Shirley Zhang

2023-05-10

Contents

1. Executive Summary . . . 2
2. Introduction . . . 3
   Background . . . 3
   Research Question . . . 3
   Objectives . . . 3
   Data Product . . . 3
3. Data Science Techniques . . . 5
   Description of the Data . . . 5
   Proposed Modelling Approaches . . . 7
   Success Criteria . . . 8
4. Timeline . . . 9
5. References . . . 10

1. Executive Summary

The isotopic composition of ice cores (i.e. $\delta^{18}O$) is a proxy for understanding historical climate and its variability in Antarctica. Our proposal aims to model the relationship between $\delta^{18}O$ and three climate variables: temperature, precipitation, and geopotential height. We will build Gaussian Process and Deep Learning models to predict these outcomes across space and time, using data simulated from global climate models. We will deliver a reproducible and well-documented workflow to train our models and a Python package to apply our trained models on new data.

2. Introduction

Background

The earliest climate observations in Antarctica date back to 1958, when the first weather stations were set up (Bromwich et al. 2013). To characterize the climate before this time, water stable isotopic records in ice cores dating back thousands of years can be studied (Stenni et al. 2017). Water stable isotopes are known to be related to variables such as temperature, precipitation, and air circulation/geopotential height (Sodemann, Aemisegger, and Risi 2022; Servettaz et al. 2020). Thus, they can be used as proxies to estimate these key climate variables.

Beyond direct observations collected from ice-cores, water stable isotope estimates can also be generated from global climate models (Stevens et al. 2013). Climate models are large data sets created by complex and computationally intensive supercomputer simulations. They predict Earth's climate over time by modeling interactions between many physical, chemical, and biological processes. The complexity means these are effectively black-box models, making it difficult to directly parse the relationship between isotopes and climate variables.

Previous research has utilized the ECHAM5-wiso climate model (Werner 2019) to reconstruct temperature in Antarctica using the isotopic composite measure $\delta^{18}O$. Linear regression with ordinary least squares (OLS) was used to find a significant trend of cooling across Antarctic regions throughout 0 to 1900 CE (Stenni et al. 2017). Despite this result, there remains a need for more complex models in the literature to characterize this relationship. This need could potentially be addressed through a partnership between the Antarctic ice-core and data science research fields.

To this end, our project will aim to use data science techniques, exploring machine learning models to characterize the relationship between isotopic composites and climate variables in a chosen climate model, IsoGSM (Yoshimura et al. 2008).

Research Question

The question which will guide our project is as follows:

> *How can we interpret the relationship between isotopic proxies ($\delta^{18}O$) and weather conditions in Antarctica such as temperature, precipitation, and air circulation/geopotential height within a climate model?*

Objectives

Our research question will be broken down into the following objectives:

1. Using data from the IsoGSM climate model, implement a Gaussian Process (GP) model to predict temperature, precipitation, and air circulation (Y) from values of $\delta^{18}O$ (X).
2. Interpret the "skill" of the model in various regions of Antarctica, and for various climate variables (i.e. is the relationship found by the model stronger in one area than another?).
3. Using IsoGSM data, implement a deep learning (DL) model to use as a comparison of the performance of the GP model.

Data Product

The deliverables addressing our objectives will be the following:

1. Workflow notebook

   A well-documented Jupyter notebook containing a workflow how the models were implemented, where our model has skill, and evaluations and visualizations of output metrics.

2. Python package

A ready-to-use and documented package containing functions that allow the user to reproduce everything from the workflow notebook using their own dataset. A toy dataset and example use cases will be included.

3. Data Science Techniques

Description of the Data

We will use data from the IsoGSM climate model specifically, since it is the most recent model and stored entirely in one file. Its data is 4-dimensional; each variable (i.e. temperature) has a value, latitude, longitude, and time axis. To get a full picture of climate, we must predict three variables: temperature, precipitation, and geopotential height. These form our response variables, which we will predict using isotope ratios measured as $\delta^{18}O$ values. See the table below.

Table 1: Relevant columns sampled from our dataset

lon	lat	time	temp_C	precip_mm	air.press_kPA	d.18O_per.mille
151.88	-82.43	2008-05-02	-38.83	8.01	99056.47	-44.26
88.88	-80.19	2008-05-02	-63.90	0.99	100211.88	-59.84
113.62	-80.19	2008-09-02	-58.55	1.51	100362.44	-56.70
127.12	-88.03	2009-05-02	-46.39	6.29	100446.34	-46.60
95.62	-79.06	2009-07-02	-57.39	2.43	100937.20	-54.27
159.75	-81.31	2009-07-02	-34.89	0.95	99601.26	-38.83
117.00	-79.06	2009-08-02	-58.54	0.50	101404.05	-56.06
56.25	-81.31	2009-08-02	-63.80	1.22	100265.41	-61.01
178.88	-79.06	2010-05-02	-34.12	5.76	98662.49	-32.37
39.38	-82.43	2010-11-02	-39.96	1.47	99195.65	-52.61

Data Challenges:

1. **Volume**. Our data has 17,000 grid points per variable per time slice. Small subsets of monthly data can easily exceed one million rows. We need parallelization and cloud computing resources to handle this big data problem.
2. **Compatibility**. The data is in NetCDF format, and handled in Python using the `xArray` package (Hoyer and Hamman 2017). This is efficient, but does not natively integrate with some machine learning packages like `sklearn` and `PyTorch`. We need to wrap the base functions so they work with our data.

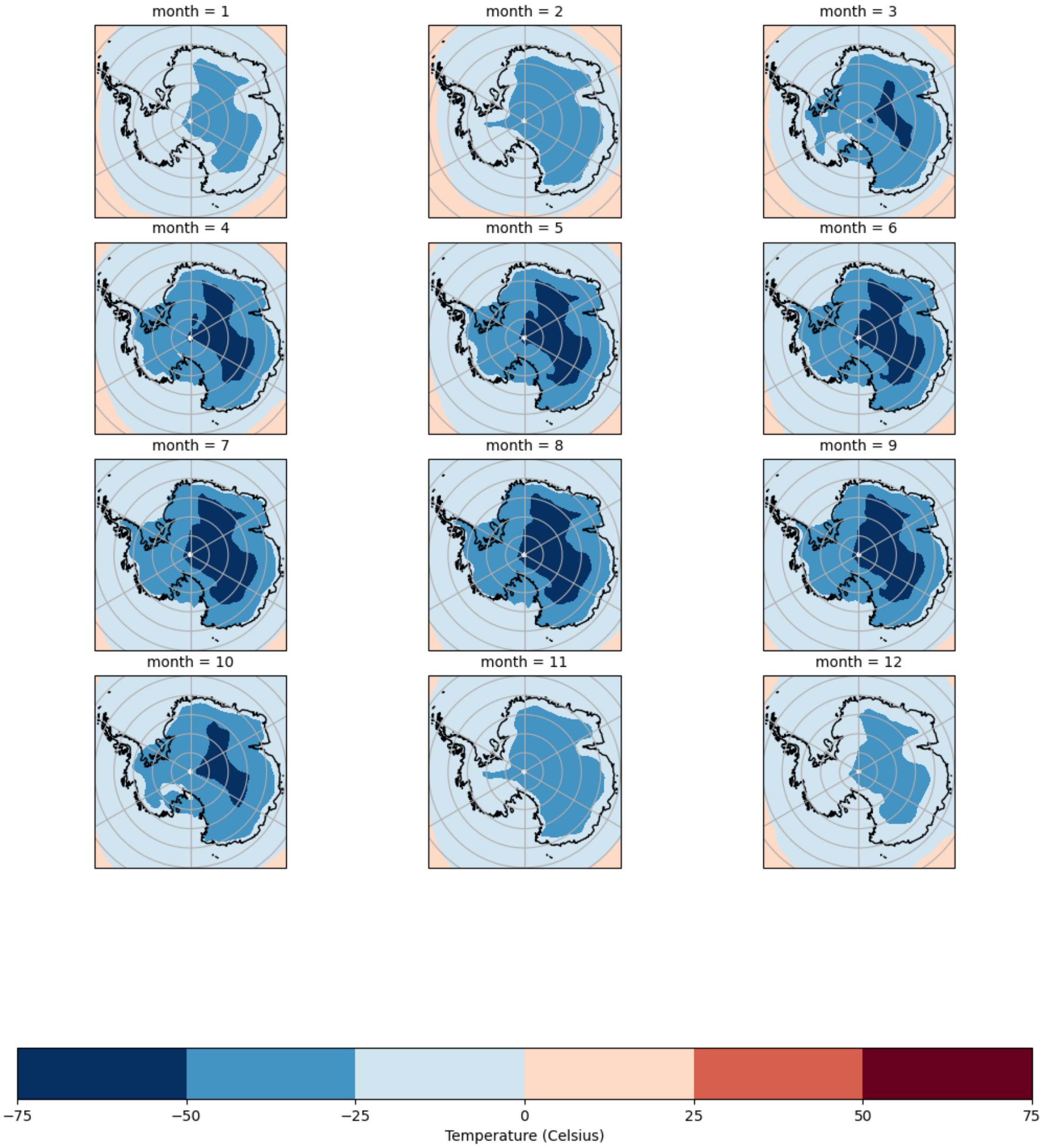

Figure 1: Sample of Antarctica temperature data over time pulled from our dataset

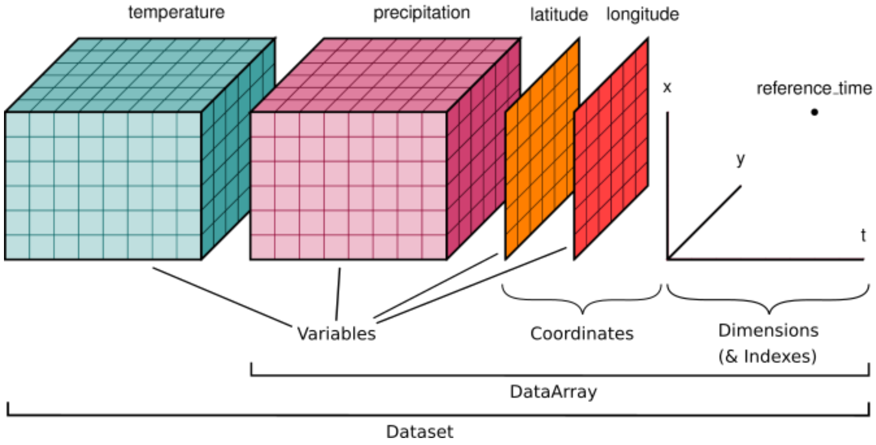

Figure 2: xArray data structure

Proposed Modelling Approaches

Assuming that we have a training dataset with n observations, let y_i denote the ith observed value of a specific climate variable (e.g. temperature or precipitation), and let $\mathbf{x}_i = \begin{pmatrix}\delta^{18}\mathrm{O} & \mathrm{lat} & \mathrm{lon} & \mathrm{time}\end{pmatrix}^\intercal$ denote the corresponding observed values of the model's input variables.

Previous Efforts Previous efforts have made use of OLS regression when attempting to model the relationship between $\delta^{18}O$ and Antarctic climate. OLS regression assumes the following structure:

$$y_i = \beta_0 + \beta_1\delta^{18}\mathrm{O}_i + \beta_2\mathrm{lat}_i + \beta_3\mathrm{lon}_i + \beta_4\mathrm{time}_i + \epsilon_i$$

The ϵ_i in (1) denotes the ith random **error term** corresponding to observation i. The error term accounts for deviations in the data from the assumed linear relationship (i.e. it accounts for the fact that the model is not perfect). The error term is assumed to be normally distributed with no correlation between distinct observations, i.e.

$$\begin{aligned}&\epsilon_i \sim \mathcal{N}(0, \sigma^2),\\ &\mathrm{Cor}(\epsilon_i, \epsilon_j) = 0 \quad \text{for } i \neq j.\end{aligned}$$

The assumption of uncorrelated observations is restrictive and does not hold for spatial-temporal climate data. Similarly, the assumption of a linear relationship between the input isotope measurements and the output climate variables likely fails to hold for this problem. A suitable model for this problem should not assume observations are uncorrelated and should be capable of modelling non-linear relationships.

Gaussian Processes (OLS Extension) Gaussian Processes (GPs) build on OLS by removing the assumption of uncorrelated training data while also introducing some non-linearity within the model. (**ADD REFERENCE FOR GPS**)

A GP model replaces ϵ_i from OLS with a stochastic term $Z(\mathbf{x}_i)$. It models correlation structures between distinct observations. The GP analog of the previous OLS model assumes the following model structure:

$$y_i = \beta_0 + \beta_1\delta^{18}\mathrm{O}_i + \beta_2\mathrm{lat}_i + \beta_3\mathrm{lon}_i + \beta_4\mathrm{time}_i + Z(\mathbf{x}_i)$$

The stochastic term is assumed to be (marginally) normally distributed, i.e.

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

Further, a GP model assumes that distinct random error terms have a correlation structure which is determined by a kernel function $R(\cdot\ ,\ \cdot)$:

$$\mathrm{Cor}(Z(\mathbf{x}_i), Z(\mathbf{x}_j)) = \underbrace{R(\mathbf{x}_i, \mathbf{x}_j)}_{\text{kernel function}} \in [0, 1]$$

The kernel function models the correlation between data points as a function of their distance. Although a GP has a linear regression component, a kernel function can be non-linear, and as such GP models are capable of modelling non-linear relationships. Further, there are kernel functions which are able to directly model seasonal temporal correlation structures.

GPs are well-suited for this problem for a variety of reasons: they extend previous methods (OLS), capture spatial/temporal correlations, provide confidence intervals for predictions, maintain interpretability via a linear regression component, and accommodate non-linear relationships using versatile kernel functions.

Note that GPs have drawbacks. They still contain an assumption of normally distrbuted error terms, which may not hold for our data. GP models are also computationally intensive because they compare each point to each other point, resulting in exploding space and time complexities on big data. This matrix must be stored as part of the trained model to generate predictions.

Deep Learning (Neural Networks) Our second modeling approach is Deep Learning (DL) using neural networks. DL models are effective regardless of the distribution of the data, and are more efficient to train and to store compared to GPs. We can also leverage the existing literature on neural network architectures. For example, there are architectures designed for modeling data with both spatial and temporal correlation. Picking an neural network architecture is an art, and we anticipate testing out a variety of architectures to find one that fits best.

Neural Networks are extremely flexible "black box" models. They sacrifice interpretability in favour of increased prediction accuracy. Further, NNs cannot generate only point predictions without confidence intervals. For these reasons, we anticipate that our GP models will be our primary models for the sake of inference, and that our NN models will mainly be used for the purpose of benchmarking the accuracy of our GP models.

Success Criteria

We will evaluate our models' success using RMSE (Root Mean Square Error) on validation data. Our project prioritizes exploration and interpretation over prediction accuracy; thus there is no benchmark RMSE we must achieve for success. Instead, RMSE scores will be used to compare the effectiveness of our approaches.

Success means interpreting and conveying to our partner where our model has skill, for example through map visualizations of RMSE scores across Antarctica.

Our final models will impact Antarctic ice-core research by indicating which research directions are most promising based on the RMSE scores.

4. Timeline

Week	Dates	Milestone	Objectives
1	May 1-5	Hackathon	Understand the problem; Become familiar with the dataset; Brainstorm modeling approaches
2	May 8-12	Data wrangling	Learn how to use `xArray` with machine learning models; Create small lightweight dataset; Implement a baseline dummy model
3	May 15-19	Finalize models with small dataset	Implement GP and NN model on small dataset; Build reproducible modeling workflows
4 and 5	May 22-June 2	Finalize models with full dataset	Utilize cloud computing; Implement GP amd NN models on full dataset
6	June 5-9	Evaluate models	Evaluate model results; Create visualizations
7	June 12-16	Final presentation	Present results to MDS; Draft final report
8	June 19-23	Final report and data product	Complete final report; Complete reproducible notebook deliverable; Publish Python package deliverable

5. References

Bromwich, David H, Julien P Nicolas, Andrew J Monaghan, Matthew A Lazzara, Linda M Keller, George A Weidner, and Aaron B Wilson. 2013. "Central West Antarctica Among the Most Rapidly Warming Regions on Earth." *Nature Geoscience* 6 (2): 139–45.

Hoyer, Stephan, and Joe Hamman. 2017. "Xarray: N-d Labeled Arrays and Datasets in Python." *Journal of Open Research Software*, April. https://doi.org/10.5334/jors.148.

Servettaz, Aymeric PM, Anais J Orsi, Mark AJ Curran, Andrew D Moy, Amaelle Landais, Cécile Agosta, V Holly L Winton, et al. 2020. "Snowfall and Water Stable Isotope Variability in East Antarctica Controlled by Warm Synoptic Events." *Journal of Geophysical Research: Atmospheres* 125 (17): e2020JD032863.

Sodemann, Harald, Franziska Aemisegger, and Camille Risi. 2022. "How Stable Water Isotope Measurements and Modeling Can Help Bridge the Gap Between Research on Weather and Climate Time Scales." *Bulletin of the American Meteorological Society* 103 (8): E1886–93.

Stenni, Barbara, Mark AJ Curran, Nerilie J Abram, Anais Orsi, Sentia Goursaud, Valerie Masson-Delmotte, Raphael Neukom, et al. 2017. "Antarctic Climate Variability on Regional and Continental Scales over the Last 2000 Years." *Climate of the Past* 13 (11): 1609–34.

Stevens, Bjorn, Marco Giorgetta, Monika Esch, Thorsten Mauritsen, Traute Crueger, Sebastian Rast, Marc Salzmann, et al. 2013. "Atmospheric Component of the MPI-m Earth System Model: ECHAM6." *Journal of Advances in Modeling Earth Systems* 5 (2): 146–72.

Werner, M. 2019. "ECHAM5-Wiso Simulation Data–Present-Day, Mid-Holocene, and Last Glacial Maximum, PANGAEA."

Yoshimura, K, M Kanamitsu, D Noone, and T Oki. 2008. "Historical Isotope Simulation Using Reanalysis Atmospheric Data." *Journal of Geophysical Research: Atmospheres* 113 (D19).