Data Science for Polar Ice Core Climate Reconstructions

Capstone Project Proposal Report

May 12, 2023

Authors

Daniel Cairns, Jakob Thoms, and Shirley Zhang

Master of Data Science Program, University of British Columbia

Mentor

G. Alexi Rodríguez-Arelis, PhD

Department of Statistics, University of British Columbia

Partner

Anaïs Orsi, PhD

Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia

Contents

1. Executive Summary . . . 2
2. Introduction . . . 3
    Background . . . 3
    Research Question . . . 4
    Objectives . . . 4
    Data Product . . . 5
3. Data Science Techniques . . . 6
    Dataset Description . . . 6
    Proposed Modelling Approaches . . . 9
    Success Criteria . . . 12
4. Timeline . . . 13
5. Appendix . . . 14
    Equations . . . 14
    Comparison of OLS and GP models . . . 14
6. References . . . 16

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 simulated data obtained from global climate models. By the project's conclusion, 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. This work will support climate science research in Antarctica by developing tools to help reconstruct polar climates going back thousands of years.

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, scientists study water-stable isotopic records in ice cores dating back thousands of years (Stenni et al. 2017). One such measure is the isotopic composition of Oxygen in precipitation, expressed as $\delta^{18}O$ (delta Oxygen-18). This measure can act as a proxy to estimate key climate variables such as temperature, precipitation, and geopotential height (see Table 1).

$\delta^{18}O$ reflects a ratio of the heavy oxygen isotope ^{18}O to the light isotope ^{16}O in a sample of water (see Appendix for equation). Broadly speaking, warmer temperatures result in more ^{18}O in the ice cores, since the heavier ^{18}O isotopes require more energy than ^{16}O to evaporate (Mulvaney 2004). Precipitation processes also affect $\delta^{18}O$, as the heavier isotope preferentially precipitates before the lighter one. Finally, air circulation guides the travel paths of temperature and moisture, therefore also affecting $\delta^{18}O$ across Antarctica (Noone and Simmonds 2002). We use a variable called "geopotential height" to measure air circulation.

Real ice core data is scarce, but $\delta^{18}O$ estimates can be generated uniformly over large areas using global climate models (Stevens et al. 2013). Climate models simulate natural processes with computer codes that implement complex mathematical models. These climate models can be extremely computationally intensive and require significant computing time and resources to run (Bastos and O'Hagan 2009), limiting their flexibility.

Previous research has used linear regression with ordinary least squares (OLS) to model the relationship between $\delta^{18}O$ and temperature in data simulated from a climate model (Stenni et al. 2017). Building upon this research, our project will use more powerful data science techniques on data obtained from the *IsoGSM* climate model (Yoshimura et al. 2008) to model the relationships between the isotopic composition of precipitation and key climate variables (outlined in Table 1 below).

Table 1: Definitions of key climate variables for this project.

Variable	Definition
$\delta^{18}O$	Delta Oxygen-18 in precipitation (‰)
Temperature	Air temperature 2 metres above the surface (K)
Precipitation rate	Rate of precipitation reaching the Earth's surface (mm/s)
Geopotential height	A vertical coordinate relative to Earth's mean sea level at 500 milibars (1/2 the atmosphere) (m)

Research Question

The question which will guide our project is as follows:

> *Using simulated data obtained from a global climate model, how can we model the relationship between **isotopic proxies** ($\delta^{18}O$) and weather conditions in Antarctica such as **temperature**, **precipitation**, and **geopotential height**?*

Objectives

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

1. Implement machine learning (ML) models:

   Using data from the IsoGSM climate model, implement Gaussian Process (GP) and neural network (NN) models to predict temperature, precipitation, and geopotential height (Y) from values of $\delta^{18}O$ (X).

   (see Proposed Modelling Approaches for more details on GP and NN models)

2. Examine the performance of the ML models:

   After training our models, use heatmap visualizations to examine their predictive performance for (1) different regions of Antarctica and for (2) predicting different climate variables.

Data Product

To meet our objectives, we will deliver the following data products:

1. Workflow notebook:

   A well-documented Jupyter notebook containing a workflow detailing how our models were implemented with evaluations and visualizations of output metrics. It will be contained within a private GitHub repository accessible by our partner.

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 data set. A toy data set and example use cases will be included. It will be contained within a public GitHub repository accessible by anyone interested in our project.

3. Data Science Techniques

Dataset Description

We will build models using simulation-generated data from the *IsoGSM* climate model. Its data is 4-dimensional; each variable has a value, latitude, longitude, and time axis. See Figure 1, which illustrates temperature values across space and time. 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 $\delta^{18}O$ values. Table 2 shows a sample of our data set with these relevant columns.

Table 2: Sample of relevant *IsoGSM* climate model data. Variables include the monthly average $\delta^{18}O$ (delta Oxygen-18, per mille), GPH (geopotential height at 500 mbar, meters), precipitation rate (millimeters per second), and temperature (degrees Kelvin) values across longitude, latitude, and time dimensions.

month	lat	lon	$\delta^{18}O$ (‰)	GPH (m)	Precip. (mm/s)	Temp. (K)
Jul 2005	-84.75	50.62	-50.5	4855	5.0e-07	220
Jan 2006	-86.65	43.12	-29.6	5126	2.7e-06	245
Feb 2007	-84.75	11.25	-32.5	5009	1.9e-06	241
Feb 2007	-88.54	69.38	-51.4	4999	7.0e-07	234
Jun 2007	-88.54	30.00	-54.4	4997	7.0e-07	230
Oct 2007	-82.85	16.88	-39.2	4939	3.0e-06	235
Sep 2008	-80.95	7.50	-33.8	4894	1.0e-06	229
May 2009	-86.65	35.62	-39.3	5031	5.2e-06	233
Jun 2009	-82.85	54.38	-47.0	5038	8.0e-07	228
Nov 2009	-80.95	35.62	-38.5	5091	6.0e-07	240

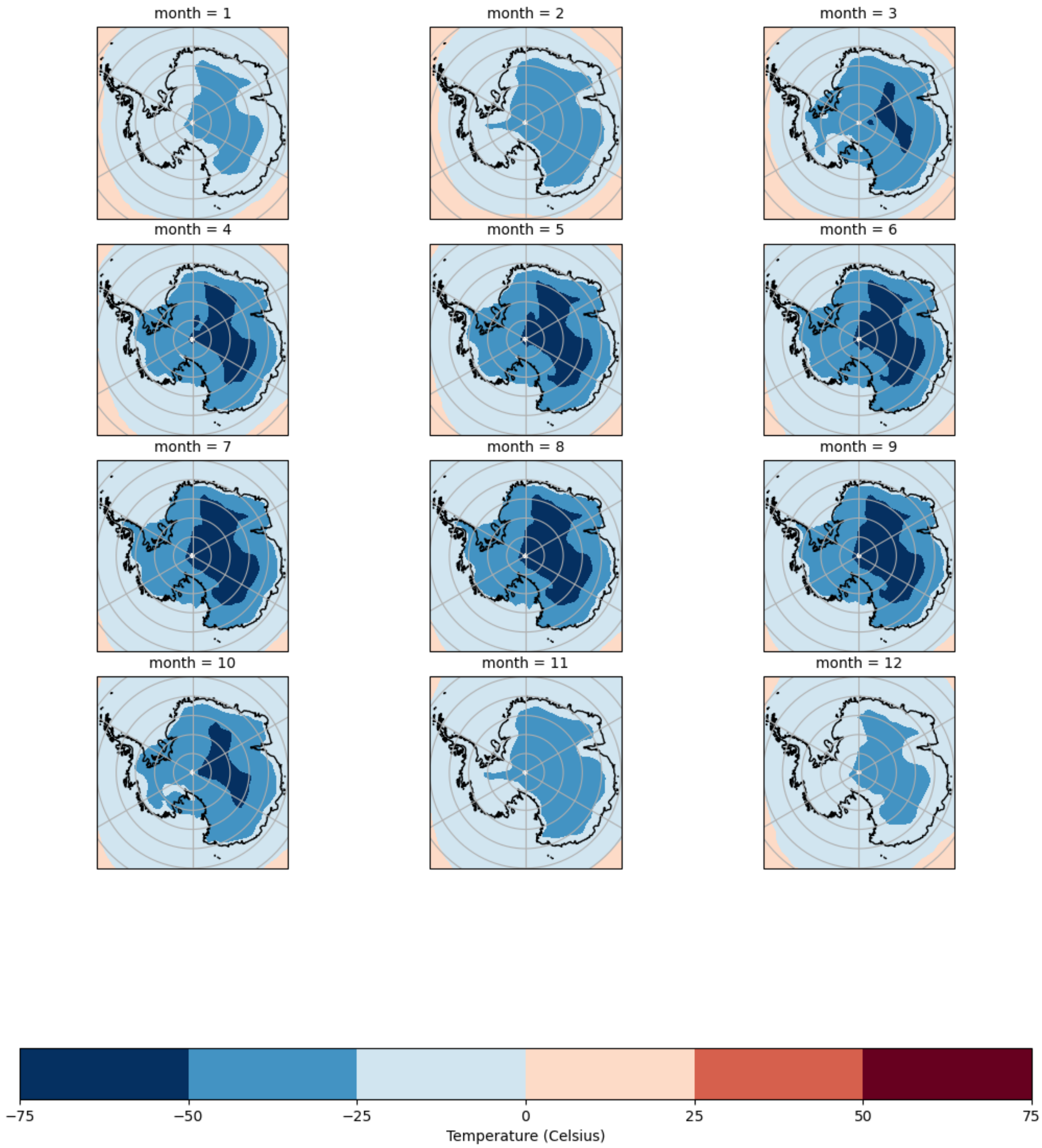

Figure 1: Sample of monthly average air temperature in Antarctica calculated from our data set.

Data Challenges There are two significant data challenges we must overcome:

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 format is more space efficient (see Figure 2), but does not natively integrate with some machine learning packages like `sklearn` and `PyTorch` (Pedregosa et al. 2011; Paszke et al. 2019). We need to wrap the base functions so they work with our data.

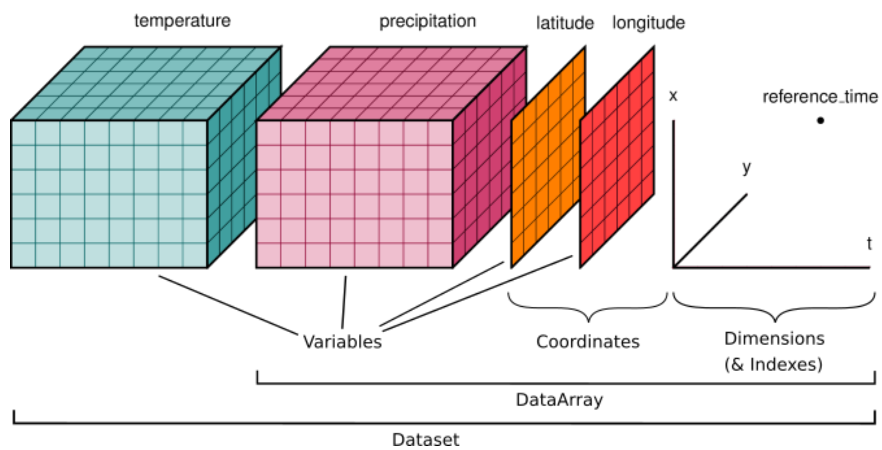

Figure 2: xArray Data Structure (Hoyer and Hamman 2017)

Proposed Modelling Approaches

Previous Efforts The OLS model assumes **linearity**, i.e. there is a linear relationship between the model's input and output variables. It also assumes **independence**, i.e. that distinct observations are uncorrelated. The independence assumption does not hold for spatial-temporal climate data, and the linearity assumption likely does not hold between input isotope measurements and the output climate variables. In light of these challenges, we will propose two alternative modelling approaches.

Gaussian Processes GP models were first introduced by Sacks et al. (1989). A GP model extends the OLS model in a way that removes the assumption of uncorrelated observations while also introducing some non-linearity within the model. Assuming that we have a training data set with n observations, let y_i denote the ith observed value of a specific climate variable (e.g. temperature or precipitation), and let $\mathbf{x}_i = \begin{pmatrix} \delta^{18}\text{O} & \text{lat} & \text{lon} & \text{time} \end{pmatrix}^\intercal$ denote the corresponding observed values of the model's input variables. Analogous to OLS, a GP model has a linear regression component of the form

$$y_i = \beta_0 + \beta^\intercal \mathbf{x}_i, \qquad (1)$$

where $\beta = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{pmatrix}^\intercal$ is a vector of coefficients. The main difference between OLS models and GP models is in their correlation structure (for a more detailed comparison, see the Appendix). Whereas an OLS model assumes that $\text{Cor}(y_i, y_j) = 0$ for $i \neq j$, a GP model assumes that observations have a correlation structure which is determined by a kernel function $R(\cdot\ ,\ \cdot)$:

$$\text{Cor}(y_i, y_j) = \underbrace{R(\mathbf{x}_i, \mathbf{x}_j)}_{\text{kernel function}} \in [0, 1] \qquad (2)$$

The kernel function models the correlation between observations 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. There are many kernel functions available in the literature (Duvenaud 2014). For example, there are kernel functions which are able to directly model seasonal temporal correlation structures (Roberts et al. 2013). Thus, we anticipate dedicating a large amount of effort to selecting the most appropriate kernel functions.

Note that GP models have some drawbacks. Like OLS, they assume normally distributed error terms (see Appendix), and this assumption may not hold for our data. GP models are also computationally intensive because they compute pairwise distances between all training data points, resulting in exploding space and time complexities on big data. This matrix must be stored as part of the trained model to generate predictions.

Neural Networks A Neural Network (NN) model features a collection of connected nodes, called "neurons". Any NN can be represented as a graph of nodes and edges (see Figure 3). Neurons are usually organized into "layers", and any NN has at least two layers: an "input layer" whose nodes represent the model's input variables, and an "output layer" whose nodes represent the model's output variables. Thus NN models are capable of handling multivariate outputs with ease, which is an important advantage in the context of climate modelling. Besides the input and output layers, NN models can have intermediate "hidden layers". The term "deep learning" (DL) refers to training NN models with more than one hidden layer.

There are many different ways to organize the neurons into layers. The "architecture" of an NN refers to the structure of the model's neurons and the connections between them. We plan to leverage the existing literature on NN architectures (Murphy 2016). For example, there are architectures designed for modelling data with both spatial and temporal correlation (Serifi, Günther, and Ban 2021). Picking a neural network architecture is an art, and we anticipate testing out a variety of architectures to find one that fits best.

DL models are extremely flexible "black box" models. They are effective regardless of the distribution of the data, and are more efficient to train and to store compared to GP models. The main drawback of DL models compared to GP models is that they sacrifice interpretability in favour of increased flexibility.

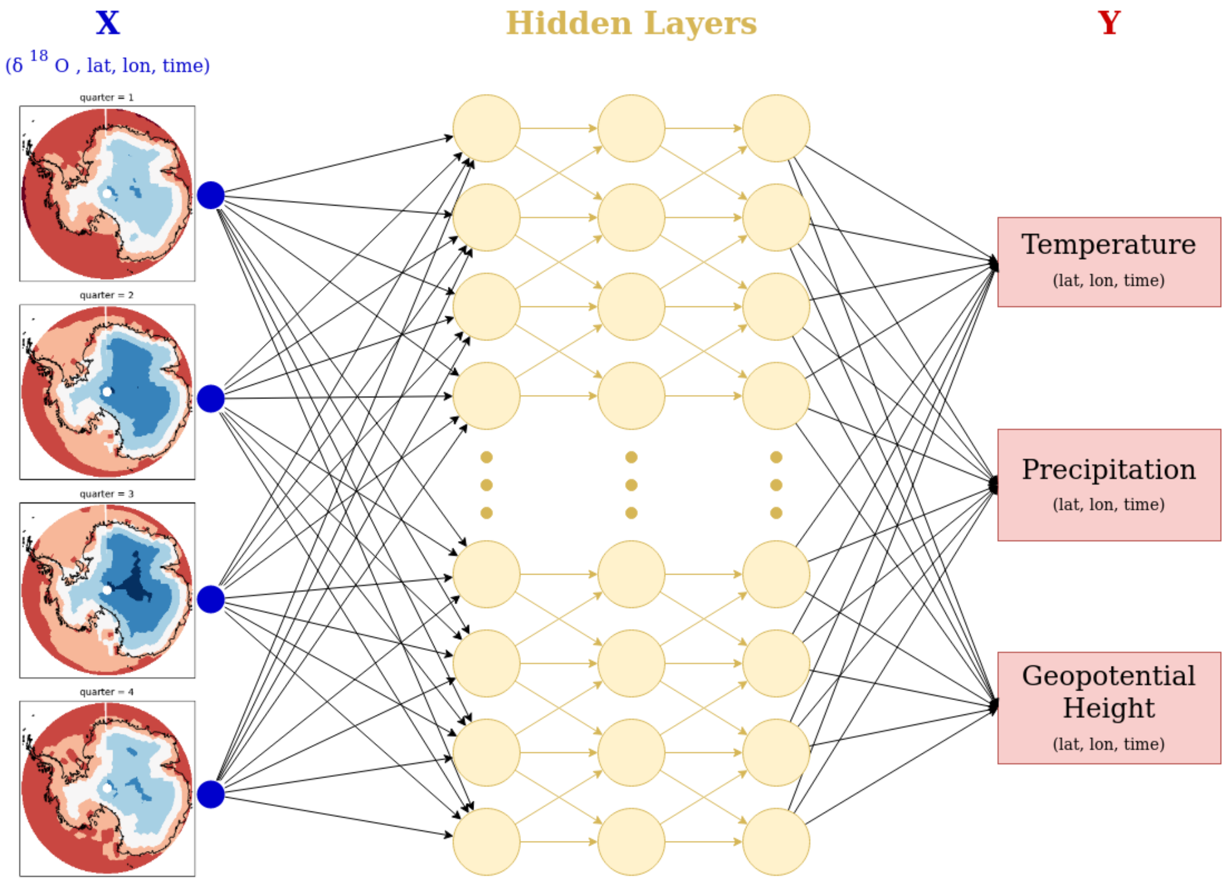

Figure 3: Abstract representation of our proposed Deep Learning model. Hidden layers model the relationships between isotopic composition and three output climate variables (temperature, precipitation, and geopotential height) across space and time dimensions.

Success Criteria

We will evaluate our models' success using RMSE (Root Mean Square Error) on validation data (see Appendix for equation). 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 communicating to our partner where our model performs well in terms of prediction accuracy. For example, we could do this through heat map visualizations of RMSE scores across Antarctica. Our final models will impact Antarctic ice-core research by indicating which modeling approaches (e.g. GP kernel functions, NN architectures) may be most promising to continue pursuing further.

4. Timeline

The following table outlines the milestones and objectives we will aim to achieve throughout the 8 weeks of the capstone project.

Table 3: Milestones and Objectives.

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 resultsto the Master of Data Science cohort; Draft final report
8	June 19-23	Final report and data product	Complete final report; Complete reproducible notebook deliverable; Publish Python package deliverable

5. Appendix

Equations

Equation for calculation of $\delta^{18}O$

$$\delta^{18}O = \left(\frac{(^{18}O/^{16}O)_{sample}}{(^{18}O/^{16}O)_{VSMOW}} - 1 \right) \times 1000\,‰ \quad (3)$$

Where $(^{18}O/^{16}O)_{sample}$ is the ratio of the heavy to light isotope in a sample, and $(^{18}O/^{16}O)_{VSMOW}$ is the ratio in the Vienna Standard Mean Ocean Water (Wet, West, and Harris 2020; Stenni et al. 2017).

Equation for calculation of RMSE

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (4)$$

Where n is the number of samples, $\hat{y}_i$ is the $i-th$ predicted value, and y_i is the actual value (for $i = 1, 2, ...n$) (Chai and Draxler 2014).

Comparison of OLS and GP models

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

$$\mathbf{x}_i = \begin{pmatrix} \delta^{18}\mathrm{O} & \mathrm{lat} & \mathrm{lon} & \mathrm{time} \end{pmatrix}^{\intercal} \quad (5)$$

denote the corresponding observed values of the model's input variables.

OLS regression assumes the following structure:

$$y_i = \beta_0 + \beta_1\delta^{18}\mathrm{O}_i + \beta_2\mathrm{lat}_i + \beta_3\mathrm{lon}_i + \beta_4\mathrm{time}_i + \epsilon_i \quad (6)$$

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.,

$$\begin{aligned}
&\epsilon_i \sim \mathcal{N}(0, \sigma^2), \\
&\mathrm{Cor}(\epsilon_i, \epsilon_j) = 0 \quad \text{ for } i \neq j \quad (7)
\end{aligned}$$

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:

$$y_i = \beta_0 + \beta_1 \delta^{18}\mathrm{O}_i + \beta_2 \mathrm{lat}_i + \beta_3 \mathrm{lon}_i + \beta_4 \mathrm{time}_i + Z(\mathbf{x}_i) \quad (8)$$

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

$$Z(\mathbf{x}_i) \overset{marginal}{\sim} \mathcal{N}(0, \sigma^2) \quad (9)$$

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

$$\mathrm{Cor}(Z(\mathbf{x}_i), Z(\mathbf{x}_j)) = \underbrace{R(\mathbf{x}_i, \mathbf{x}_j)}_{\text{kernel function}} \in [0, 1] \quad (10)$$

6. References

Bastos, Leonardo S., and Anthony O'Hagan. 2009. "Diagnostics for Gaussian Process Emulators." *Technometrics* 51 (4): 425–38. https://doi.org/10.1198/TECH.2009.08019.

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.

Chai, Tianfeng, and Roland R Draxler. 2014. "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)?–Arguments Against Avoiding RMSE in the Literature." *Geoscientific Model Development* 7 (3): 1247–50.

Duvenaud, David. 2014. "Automatic Model Construction with Gaussian Processes." PhD thesis.

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.

Mulvaney, Robert. 2004. "How Are Past Temperatures Determined from an Ice Core?" *Scientific American.*

Murphy, John H. 2016. "An Overview of Convolutional Neural Network Architectures for Deep Learning." In.

Noone, David, and I. Simmonds. 2002. "Associations Between Delta-18O of Water and Climate Parameters in a Simulation of Atmospheric Circulation for 1979–95." *Journal of Climate* 15 (22): 3150–69. https://doi.org/https://doi.org/10.1175/1520-0442(2002)015%3C3150:ABOOWA%3E2.0.CO;2.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In *Advances in Neural Information Processing Systems 32*, 8024–35. Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Roberts, S., M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain. 2013. "Gaussian Processes for Time-Series Modelling." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371 (1984): 20110550. https://doi.org/10.1098/rsta.2011.0550.

Sacks, Jerome, William J. Welch, Toby J. Mitchell, and Henry P. Wynn. 1989. “Design and Analysis of Computer Experiments.” *Statistical Science* 4 (4): 409–23. https://doi.org/10.1214/ss/1177012413.

Serifi, Agon, Tobias Günther, and Nikolina Ban. 2021. “Spatio-Temporal Downscaling of Climate Data Using Convolutional and Error-Predicting Neural Networks.” *Frontiers in Climate* 3. https://doi.org/10.3389/fclim.2021.656479.

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.

Wet, Ruan F de, Adam G West, and Chris Harris. 2020. “Seasonal Variation in Tap Water δ2H and δ18O Isotopes Reveals Two Tap Water Worlds.” *Scientific Reports* 10 (1): 13544.

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