A Gaussian process state-space model for atmospheric CO_2 and sea surface temperature index reconstruction from boron isotope and planktonic $\delta^{18}O$ proxies

Taehee Lee
Harvard University
Cambridge, Massachusetts, USA
taehee_lee@fas.harvard.edu

Charles E. Lawrence
Brown University
Providence, Rhode Island, USA
charles_lawrence@brown.edu

ABSTRACT

It often occurs in practice that only a small number of observations are given for reconstructing past climate events in the field of paleoclimatology. State-space models can overcome such scarcity by giving priors to those hidden states to make them correlated to one another. Inferring multiple events simultaneously from various proxies to exploit their mutual dependency is another option. Here we present a Gaussian process state-space model to reconstruct both atmospheric CO_2 and sea surface temperature index from boron isotope and planktonic $\delta^{18}O$ proxies.

CCS CONCEPTS

• **Mathematics of computing → Nonparametric statistics**; **Variational methods**; • **Computing methodologies → Gaussian processes**; • **Applied computing → Environmental sciences**.

KEYWORDS

Gaussian process, state-space model, boron isotope, planktonic $\delta^{18}O$, atmospheric CO_2, sea surface temperature, paleoclimatology

ACM Reference Format:
Taehee Lee and Charles E. Lawrence. 2020. A Gaussian process state-space model for atmospheric CO_2 and sea surface temperature index reconstruction from boron isotope and planktonic $\delta^{18}O$ proxies. In *10th International Conference on Climate Informatics (CI2020), September 22–25, 2020, virtual, United Kingdom.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3429309.3429316

1 INTRODUCTION

In the field of paleoclimatology, the limited resolution of available proxy data often limits reconstruction of the past climate events over ages. For instance, boron isotope ($\delta^{11}B$) proxy is directly correlated to the atmospheric CO_2, but has low resolution and is unevenly spaced over ages [5, 7, 11]. The statistical learning that depends only on the individual inference is vulnerable to outliers and often inefficient to exploit information.

CI2020, September 22–25, 2020, virtual, United Kingdom

ACM ISBN 978-1-4503-8848-1/20/09.
https://doi.org/10.1145/3429309.3429316

One way is to give a more comprehensive prior on the past climate events over ages, just as how state-space models do. Traditional state-space models such as the particle filter/smoother [4, 14] often depend too much on the parametric transition models that miss their nonstationary aspects and model misspecification. Also, such models expect regularly spaced data over ages, which is problematic if the data are too scarce to keep information after rearranging the data regularly.

The Gaussian process state-space model (GPST) [6, 8] addresses these limitations. Gaussian processes [23] are nonparametric thus do not depend on parameters and can take the irregularly spaced data without the rearrangement. That GPSTs do not require the memoryless assumption is a bonus. [15] shows the reconstruction of atomospheric CO_2 from $\delta^{11}B$ by a GPST model.

Another way is to utilize the dependencies between a set of closely related climate events that have proxies of plentiful observations, such as the sea surface temperature (SST) *index* [25] for the atmospheric CO_2, for "borrowing" information from them indirectly: note that raw SSTs themselves are not global parameters.

Here we extend the GPST model in [15] to consider both atmospheric CO_2 and SST index simultaneously from two proxies, $\delta^{11}B$ and planktonic $\delta^{18}O$. Section 2 describes the modeling in detail and section 3 defines the data and how they are preprocessed. Section 4 shows the results and section 5 concludes the paper.

2 MODEL

We first define the following notations and symbols:

- $\mathrm{T} = \left(\mathrm{T}^{(1)}, \mathrm{T}^{(2)}\right)$: ages of the proxies.
  - $\mathrm{T}^{(1)} = T^{(1)}_{1:\mathrm{N}_1}$: ages of $\delta^{11}B$ proxy observations.
  - $\mathrm{T}^{(2)} = T^{(2)}_{1:\mathrm{N}_2}$: ages of the planktonic $\delta^{18}O$ proxy observations.
- $\mathrm{X} = \left(\mathrm{X}^{(1)}, \mathrm{X}^{(2)}\right)$: hidden paleoclimate events.
  - $\mathrm{X}^{(1)} = \mathrm{X}^{(1)}_{1:\mathrm{N}_1}$: atmospheric CO_2 at T.
  - $\mathrm{X}^{(2)} = \mathrm{X}^{(2)}_{1:\mathrm{N}_2}$: SST indices at T.
- $\mathrm{Y} = \left(\mathrm{Y}^{(1)}, \mathrm{Y}^{(2)}\right)$: observed proxies.
  - $\mathrm{Y}^{(1)} = \mathrm{Y}^{(1)}_{1:\mathrm{N}_1}$: $\delta^{11}B$ proxy observations.
  - $\mathrm{Y}^{(2)} = \mathrm{Y}^{(2)}_{1:\mathrm{N}_2}$: planktonic $\delta^{18}O$ proxy observations.

Like the usual state-space models, our GPST model consists of emission and transition models. The emission model for $\delta^{11}B$ is given as follows, as in [15]: here we define $\mathcal{T}_\nu(\cdot|\alpha, \beta)$ as the generalized Student's t-distribution [2] with a degree ν and its

location and scale parameters α and β, respectively.

$$\begin{aligned} &p\left(\mathrm{Y}_n^{(1)}\middle|\mathrm{X}_n^{(1)}\right) \\ &= \mathcal{T}_6\left(\mathrm{Y}_n^{(1)}\middle| a_0 + a_1\mathrm{X}_n^{(1)} + a_2\log\left(a_3 + \mathrm{X}_n^{(1)}\right), \frac{4}{3}\sigma^2\right) \end{aligned} \tag{1}$$

where a_0, a_1, a_2 and a_3 are the coefficient parameters and σ is a standard deviation. Note that these parameters are given core-specifically.

The emission model for the planktonic δ^{18}O proxies is defined as follows:

$$p\left(\mathrm{Y}_n^{(2)}\middle|\mathrm{X}_n^{(2)}\right) = \mathcal{T}_6\left(\mathrm{Y}_n^{(2)}\middle| b_0 + b_1\mathrm{X}_n^{(2)}, \frac{4}{3}s^2\right) \tag{2}$$

where b_0 and b_1 are the coefficient parameters and s is a standard deviation.

The most distinctive feature of the GPST model is that the transition model does not assume the memoryless property. Instead, it is defined by the following Gaussian process:

$$\begin{aligned} p(\mathrm{X}|\mathrm{T}) &= \mathcal{N}\left(\mathrm{X}\middle|\vec{0}, \mathbb{K}_{\mathrm{TT}}\right) \\ \mathbb{K}_{\mathrm{TT}} &\triangleq \begin{pmatrix} \delta_1^2\mathbb{K}_{11}^{(0)} + \mathbb{K}_{11}^{(1)} & \rho\delta_1\delta_2\mathbb{K}_{12}^{(0)} \\ \rho\delta_1\delta_2\mathbb{K}_{21}^{(0)} & \delta_2^2\mathbb{K}_{22}^{(0)} + \mathbb{K}_{22}^{(2)} \end{pmatrix} \end{aligned} \tag{3}$$

where $\mathbb{K}_{ij}^{(k)} = \mathbb{K}_{\mathrm{T}^{(i)}\mathrm{T}^{(j)}}^{(k)}$ is an abbreviation and $\delta_1, \delta_2 > 0$ and correlation $-1 < \rho < 1$ are the kernel hyperparameters. To control ρ, we reparametrize it by $\rho = \tanh\rho_0$ for another parameter ρ_0 that takes any real values. The above covariance matrix consists of the following three kernels:

$$\begin{aligned} \mathbb{K}^{(0)}(u,v) &= \left(1 + \sqrt{3}\xi_0^2|u-v|\right)\cdot e^{-\sqrt{3}\xi_0^2|u-v|} \\ \mathbb{K}^{(1)}(u,v) &= \eta_1^2\cdot e^{-2\xi_1^2\sin^2(\pi|u-v|/r_1)} + \lambda_1^2\cdot 1_{\{u=v\}} \\ \mathbb{K}^{(2)}(u,v) &= \eta_2^2\cdot e^{-2\xi_2^2\sin^2(\pi|u-v|/r_2)} + \lambda_2^2\cdot 1_{\{u=v\}} \end{aligned} \tag{4}$$

where η_1, η_2, ξ_0, ξ_1, ξ_2, λ_1, λ_2 are also the kernel hyperparameters.

Note that:

$$\begin{pmatrix} \delta_1^2 & \rho\delta_1\delta_2 \\ \rho\delta_1\delta_2 & \delta_2^2 \end{pmatrix} \otimes \mathbb{K}_{\mathrm{TT}}^{(0)} = \begin{pmatrix} \delta_1^2\mathbb{K}_{\mathrm{TT}}^{(0)} & \rho\delta_1\delta_2\mathbb{K}_{\mathrm{TT}}^{(0)} \\ \rho\delta_1\delta_2\mathbb{K}_{\mathrm{TT}}^{(0)} & \delta_2^2\mathbb{K}_{\mathrm{TT}}^{(0)} \end{pmatrix} \tag{5}$$

$$\begin{aligned} &\begin{pmatrix} \delta_1^2\mathbb{K}_{11}^{(0)} + \mathbb{K}_{11}^{(1)} & \rho\delta_1\delta_2\mathbb{K}_{12}^{(0)} \\ \rho\delta_1\delta_2\mathbb{K}_{21}^{(0)} & \delta_2^2\mathbb{K}_{22}^{(0)} + \mathbb{K}_{22}^{(2)} \end{pmatrix} \\ &= \begin{pmatrix} \delta_1^2\mathbb{K}_{11}^{(0)} & \rho\delta_1\delta_2\mathbb{K}_{12}^{(0)} \\ \rho\delta_1\delta_2\mathbb{K}_{21}^{(0)} & \delta_2^2\mathbb{K}_{22}^{(0)} \end{pmatrix} + \begin{pmatrix} \mathbb{K}_{11}^{(1)} & 0 \\ 0 & \mathbb{K}_{22}^{(2)} \end{pmatrix} \end{aligned} \tag{6}$$

and the first term of (6) is obtained by removing some rows and columns symmetrically from (5), thus $\mathbb{K}_{\mathrm{TT}}$ is a positive semi-definite symmetric matrix to become a covariance matrix.

It is straightforward to show that (3) is consistently extendable, i.e., for a query age pair $t = (t_1, t_2)$ and the associated hidden events $x = (x_1, x_2)$, we have:

$$\begin{aligned} p(x, \mathrm{X}|t, \mathrm{T}) &= \mathcal{N}\left(\begin{pmatrix} \mathrm{X}^{(1)} \\ x_1 \\ \mathrm{X}^{(2)} \\ x_2 \end{pmatrix}\middle|\middle|\begin{pmatrix} \vec{0} \\ 0 \\ \vec{0} \\ 0 \end{pmatrix}, \mathbb{K}_{\mathrm{T}t,\mathrm{T}t}\right) \\ &\mathbb{K}_{\mathrm{T}t,\mathrm{T}t} \\ &\triangleq \begin{pmatrix} \delta_1^2\mathbb{K}_{1t_1,1t_1}^{(0)} + \mathbb{K}_{1t_1,1t_1}^{(1)} & \rho\delta_1\delta_2\mathbb{K}_{1t_1,2t_2}^{(0)} \\ \rho\delta_1\delta_2\mathbb{K}_{2t_2,1t_1}^{(0)} & \delta_2^2\mathbb{K}_{2t_2,2t_2}^{(0)} + \mathbb{K}_{2t_2,2t_2}^{(2)} \end{pmatrix} \end{aligned} \tag{7}$$

The motive of $\mathbb{K}_{\mathrm{TT}}$ comes from the following hierarchical prior on X:

$$\begin{aligned} &p\left(\mathrm{X}\middle|\underline{\mu}, \mathrm{T}\right) \\ &= \mathcal{N}\left(\begin{pmatrix} \mathrm{X}^{(1)} \\ \mathrm{X}^{(2)} \end{pmatrix}\middle|\middle|\begin{pmatrix} \underline{\mu}^{(1)} \\ \underline{\mu}^{(2)} \end{pmatrix}, \begin{pmatrix} \delta_1^2 & \rho\delta_1\delta_2 \\ \rho\delta_1\delta_2 & \delta_2^2 \end{pmatrix} \otimes \mathbb{K}_{\mathrm{TT}}^{(0)}\right) \end{aligned} \tag{8}$$

$$\begin{aligned} p\left(\underline{\mu}^{(1)}\middle|\vec{0}, \mathbb{K}_{11}^{(1)}\right) &= \mathcal{N}\left(\underline{\mu}^{(1)}\middle|\vec{0}, \mathbb{K}_{11}^{(1)}\right) \\ p\left(\underline{\mu}^{(2)}\middle|\vec{0}, \mathbb{K}_{22}^{(2)}\right) &= \mathcal{N}\left(\underline{\mu}^{(2)}\middle|\vec{0}, \mathbb{K}_{22}^{(2)}\right) \end{aligned} \tag{9}$$

Thus, each mean prior function $\underline{\mu}^{(i)}$ is assumed to follow a Gaussian process with the zero mean function and periodic kernel [17] and the hidden climate event X takes those mean prior functions together with the covariance function that is defined by a Kronecker product of a scaling covariance matrix and Matérn covariance matrix [9, 19, 27] with degree 3/2: note that a Gaussian process adopting Matérn kernel with degree ν is connected to a particular form of a continuous autoregressive (AR) model of order $\nu + 0.5$ [23], which means that our model implicitly assumes an AR(2) model.

The idea of coupling multiple hidden events in the framework of the Gaussian process with the Kronecker product is inspired by [1]. Regarding (8) and (9) as a likelihood and priors respectively and marginalizing $\underline{\mu}$ out bring $\mathbb{K}_{\mathrm{TT}}$ in (3).

The primary goal is to compute the posterior of X given T and Y, $p(\mathrm{X}|\mathrm{T},\mathrm{Y}) \propto p(\mathrm{X}|\mathrm{T})\,p(\mathrm{Y}|\mathrm{X})$. Then it follows $p(x|t,\mathrm{T},\mathrm{Y}) = \int p(x|\mathrm{X},t,\mathrm{T})\,p(\mathrm{X}|\mathrm{T},\mathrm{Y})\,d\mathrm{X}$.

Because our emission models (1) and (2) are not Gaussian, expressing $p(\mathrm{X}|\mathrm{T},\mathrm{Y})$ in a known form is not possible. Instead, we consider a variational method. Let $q(\mathrm{X}|\Theta)$ be another Gaussian distribution defined as below:

$$\begin{aligned} q(\mathrm{X}|\Theta) &= \mathcal{N}(\mathrm{X}|\mu,\Sigma) \\ &= \mathcal{N}\left(\mathrm{X}^{(1)}\middle|\mu^{(1)},\Sigma^{(1)}\right)\mathcal{N}\left(\mathrm{X}^{(2)}\middle|\mu^{(2)},\Sigma^{(2)}\right) \\ &= \prod_{n=1}^{\mathrm{N}_1}\mathcal{N}\left(\mathrm{X}_n^{(1)}\middle|\mu_{1n},\sigma_{1n}^2\right)\prod_{n=1}^{\mathrm{N}_2}\mathcal{N}\left(\mathrm{X}_n^{(2)}\middle|\mu_{2n},\sigma_{2n}^2\right) \end{aligned} \tag{10}$$

The learning procedure consists of tuning kernel hyperparameters and inferring the variational parameters $\Theta = \{\mu_{1n},\sigma_{1n}\}_{n=1}^{\mathrm{N}_1} \cup \{\mu_{2n},\sigma_{2n}\}_{n=1}^{\mathrm{N}_2}$ with the following evidence lower bound (ELBO) as the objective function to maximize:

$$\begin{aligned} &\log p(\mathrm{Y}|\mathrm{T}) \geq \mathcal{L}(\Theta) \\ &= \int q(\mathrm{X}|\Theta)\log p(\mathrm{Y}|\mathrm{X})d\mathrm{X} - \mathbb{D}_{\mathrm{KL}}(q(\cdot|\Theta)||p(\cdot|\mathrm{T})) \end{aligned} \tag{11}$$

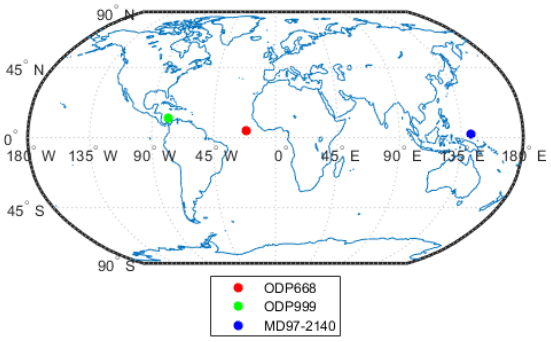

Figure 1: Core locations on map.

Note that:

$$\begin{aligned}&\mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\Theta\right)||p\left(\cdot|\mathrm{T}\right)\right)\\&=\frac{1}{2}\mathrm{tr}\left(\mathbb{K}_{\mathrm{TT}}^{-1}\Sigma\right)+\frac{1}{2}\gamma^{\mathbb{T}}\mathbb{K}_{\mathrm{TT}}^{-1}\gamma+\frac{1}{2}\log\left|\mathbb{K}_{\mathrm{TT}}\right|\\&-\sum_{n=1}^{\mathrm{N}_1}\log\sigma_{1n}-\sum_{n=1}^{\mathrm{N}_2}\log\sigma_{2n}-\frac{1}{2}\left(\mathrm{N}_1+\mathrm{N}_2\right)\end{aligned}\tag{12}$$

$$\begin{aligned}&\int q\left(\mathrm{X}|\Theta\right)\log p\left(\mathrm{Y}|\mathrm{X}\right)d\mathrm{X}\\&=\sum_{n=1}^{\mathrm{N}_1}\int\mathcal{N}\left(\mathrm{X}_n^{(1)}\middle|\mu_{1n},\sigma_{1n}^2\right)\log p\left(\mathrm{Y}_n^{(1)}\middle|\mathrm{X}_n^{(1)}\right)d\mathrm{X}_n^{(1)}\\&+\sum_{n=1}^{\mathrm{N}_2}\int\mathcal{N}\left(\mathrm{X}_n^{(2)}\middle|\mu_{2n},\sigma_{2n}^2\right)\log p\left(\mathrm{Y}_n^{(2)}\middle|\mathrm{X}_n^{(2)}\right)d\mathrm{X}_n^{(2)}\end{aligned}\tag{13}$$

Consequently, the partial derivative of (12) and (13) with respect to each variational parameter θ is given as follows:

$$\begin{aligned}&\frac{\partial}{\partial\theta}\mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\Theta\right)||p\left(\cdot|\mathrm{T}\right)\right)=\frac{1}{2}\mathrm{tr}\left(\mathbb{K}_{\mathrm{TT}}^{-1}\frac{\partial\Sigma}{\partial\theta}\right)\\&+\mu^{\mathbb{T}}\mathbb{K}_{\mathrm{TT}}^{-1}\frac{\partial\mu}{\partial\theta}-\sum_{n=1}^{\mathrm{N}_1}\frac{1}{\sigma_{1n}}\frac{\partial\sigma_{1n}}{\partial\theta}-\sum_{n=1}^{\mathrm{N}_2}\frac{1}{\sigma_{2n}}\frac{\partial\sigma_{2n}}{\partial\theta}\end{aligned}\tag{14}$$

$$\begin{aligned}&\frac{\partial}{\partial\theta}\int q\left(\mathrm{X}|\Theta\right)\log p\left(\mathrm{Y}|\mathrm{X}\right)d\mathrm{X}\\&=\sum_{n=1}^{\mathrm{N}_1}\int\frac{\partial}{\partial\theta}\log p\left(\mathrm{Y}_n^{(1)}\middle|\mu_{1n}+\sigma_{1n}\epsilon\right)\mathrm{N}\left(\epsilon|0,1\right)\mathrm{d}\epsilon\\&+\sum_{n=1}^{\mathrm{N}_2}\int\frac{\partial}{\partial\theta}\log p\left(\mathrm{Y}_n^{(2)}\middle|\mu_{2n}+\sigma_{2n}\epsilon\right)\mathrm{N}\left(\epsilon|0,1\right)\mathrm{d}\epsilon\\&\approx\frac{1}{\mathrm{K}}\sum_{n=1}^{\mathrm{N}_1}\sum_{k=1}^{\mathrm{K}}\frac{\partial}{\partial\theta}\log p\left(\mathrm{Y}_n^{(1)}\middle|\mu_{1n}+\sigma_{1n}\epsilon_{1k}\right)\\&+\frac{1}{\mathrm{K}}\sum_{n=1}^{\mathrm{N}_2}\sum_{k=1}^{\mathrm{K}}\frac{\partial}{\partial\theta}\log p\left(\mathrm{Y}_n^{(2)}\middle|\mu_{2n}+\sigma_{2n}\epsilon_{2k}\right)\end{aligned}\tag{15}$$

where $\epsilon_{k1},\epsilon_{k2}\sim_{i.i.d.}\mathcal{N}\left(0,1\right)$ and K is a large integer. Note that the reparameterization trick [13] is applied to (15).

Therefore, we have the following:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\theta}&\approx\frac{1}{\mathrm{K}}\sum_{n=1}^{\mathrm{N}_1}\sum_{k=1}^{\mathrm{K}}\frac{\partial}{\partial\theta}\log p\left(\mathrm{Y}_n^{(1)}\middle|\mu_{1n}+\sigma_{1n}\epsilon_{1k}\right)\\&+\frac{1}{\mathrm{K}}\sum_{n=1}^{\mathrm{N}_2}\sum_{k=1}^{\mathrm{K}}\frac{\partial}{\partial\theta}\log p\left(\mathrm{Y}_n^{(2)}\middle|\mu_{2n}+\sigma_{2n}\epsilon_{2k}\right)\\&-\frac{1}{2}\mathrm{tr}\left(\mathbb{K}_{\mathrm{TT}}^{-1}\frac{\partial\Sigma}{\partial\theta}\right)-\mu^{\mathbb{T}}\mathbb{K}_{\mathrm{TT}}^{-1}\frac{\partial\mu}{\partial\theta}\\&+\sum_{n=1}^{\mathrm{N}_1}\frac{1}{\sigma_{1n}}\frac{\partial\sigma_{1n}}{\partial\theta}+\sum_{n=1}^{\mathrm{N}_2}\frac{1}{\sigma_{2n}}\frac{\partial\sigma_{2n}}{\partial\theta}\end{aligned}\tag{16}$$

Because (13) is not a function of kernel hyperparameters, the partial derivatives of $\mathcal{L}$ with respect to the kernel hyperparameters are given as follows:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\vartheta}&=\frac{\partial}{\partial\vartheta}\mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\Theta\right)||p\left(\cdot|\mathrm{T}\right)\right)\\&=\frac{1}{2}\mathrm{tr}\left(\mathbb{K}_{\mathrm{TT}}^{-1}\left(\Sigma+\mu\mu^{\mathbb{T}}-\mathbb{K}_{\mathrm{TT}}\right)\mathbb{K}_{\mathrm{TT}}^{-1}\frac{\partial\mathbb{K}_{\mathrm{TT}}}{\partial\vartheta}\right)\end{aligned}\tag{17}$$

Once the kernel hyperparameters and variational parameters are learned, we can explicitly approximate the distribution of hidden climate event x at a continuous query age t as follows:

$$\begin{aligned}p\left(x|t,\mathrm{T},\mathrm{Y}\right)&=\int p\left(x|\mathrm{X},t,\mathrm{T}\right)p\left(\mathrm{X}|\mathrm{T},\mathrm{Y}\right)d\mathrm{X}\\&\approx\int p\left(x|\mathrm{X},t,\mathrm{T}\right)q\left(\mathrm{X}|\Theta\right)d\mathrm{X}\\&=\int p\left(x|\mathrm{X},t,\mathrm{T}\right)\mathcal{N}\left(\mathrm{X}|\mu,\Sigma\right)d\mathrm{X}\\&=\mathcal{N}\left(x|\overline{\mu}\left(x\right),\overline{\Sigma}\left(x\right)\right)\end{aligned}\tag{18}$$

where:

$$\begin{aligned}\overline{\mu}\left(x\right)&=\mathbb{K}_{t\mathrm{T}}\mathbb{K}_{\mathrm{TT}}^{-1}\mu\\\overline{\Sigma}\left(x\right)&=\mathbb{K}_{tt}-\mathbb{K}_{t\mathrm{T}}\left(\mathbb{K}_{\mathrm{TT}}^{-1}-\mathbb{K}_{\mathrm{TT}}^{-1}\Sigma\mathbb{K}_{\mathrm{TT}}^{-1}\right)\mathbb{K}_{\mathrm{T}t}\end{aligned}\tag{19}$$

3 DATA AND PREPROCESSING

For δ^{11}B data, we chose the cores ODP668 and ODP999 [5] as the sources. Data overlapping over ages are replaced with their average. Each δ^{11}B observation is standardized by $y\longrightarrow(y-20.5)/1.5$. To construct the emission model, we used the pairs of the standardized proxy observations (δ^{11}B indices) and the associated published atmospheric CO_2 inferences (CO_2 indices) after standardizing to $x\longrightarrow(x-300)/150$.

Figure 2 gives core-specific patterns of the pairs, so their emission models are given core-specifically. However, these models do not reflect the uncertainty along with CO_2 indices. To resolve it, we instead consider a generalized Student's t-distribution that has the mean and standard deviation functions of each emission model in figure 2 as the location and scale parameters, just as (1). In the reconstruction, we use the observations up to 800 kiloyears only, i.e., 25 observations in ODP668 and 58 in ODP999. Core-specifically learned parameters are given in figure 3.

For the planktonic δ^{18}O, we chose the core MD97-2140 [3] that has the 202 observations up to 800 kiloyears. The observations are constantly translated to fit to the planktonic δ^{18}O stack [25]:

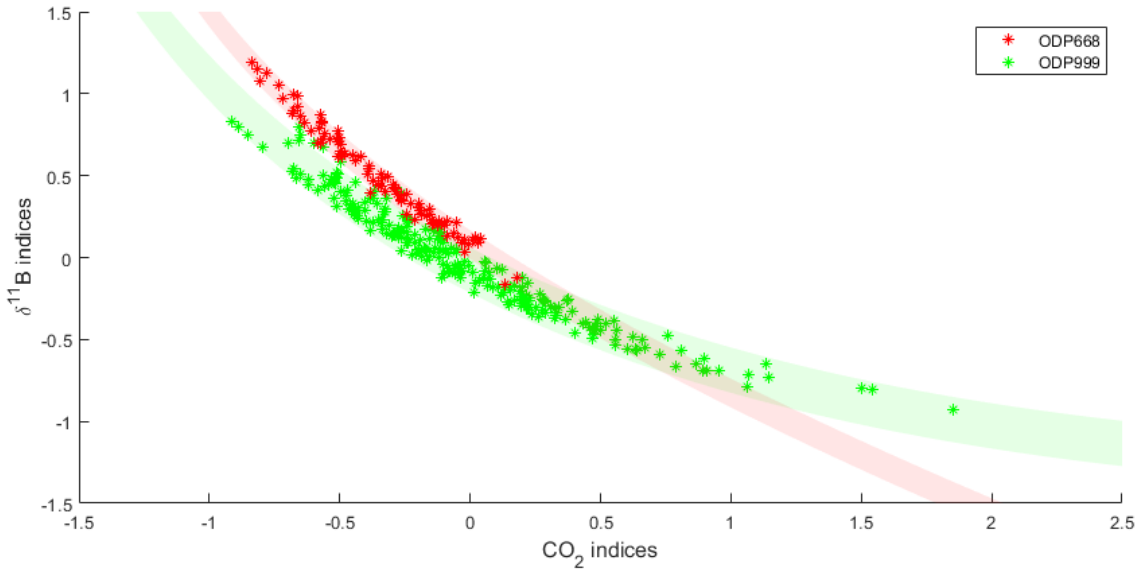

Figure 2: Red and green stars are pairs of the published CO_2 indices and δ^{11}B indices of ODP668 and ODP999, respectively. The shaded regions represent 95% confidence bands of the Gaussian emission models.

	a_0	a_1	a_2	a_3	σ
ODP668	0.6973	-0.3576	-1.1799	1.6880	0.0447
ODP999	2.5659	0.4296	-2.9984	2.4150	0.0704

Figure 3: A table of the inferred core-specific coefficients in (1).

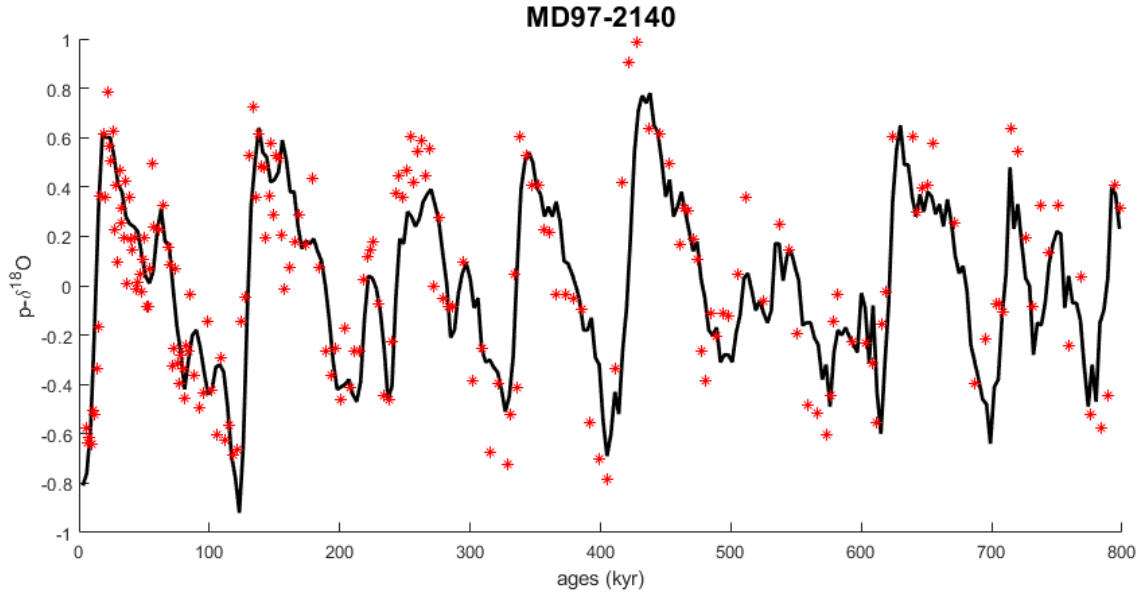

Figure 4: The planktonic δ^{18}O stack (black curve) and the translated δ^{18}O observations of MD97-2140 (red stars).

figure 4 visualizes the stack and records. The emission model of the planktonic δ^{18}O proxy given SST index is derived from their pairs of the above stack and the SST stack of [25], shown in figure 5. The inferred values are $b_0 = 0.0051$, $b_1 = -0.2831$ and $s = 0.2213$. Note that the model uncertainty is larger than those of δ^{11}B because planktonic δ^{18}O is not a direct proxy of SST index whereas δ^{11}B is of atmospheric CO_2. For the same reason of δ^{11}B, our emission model

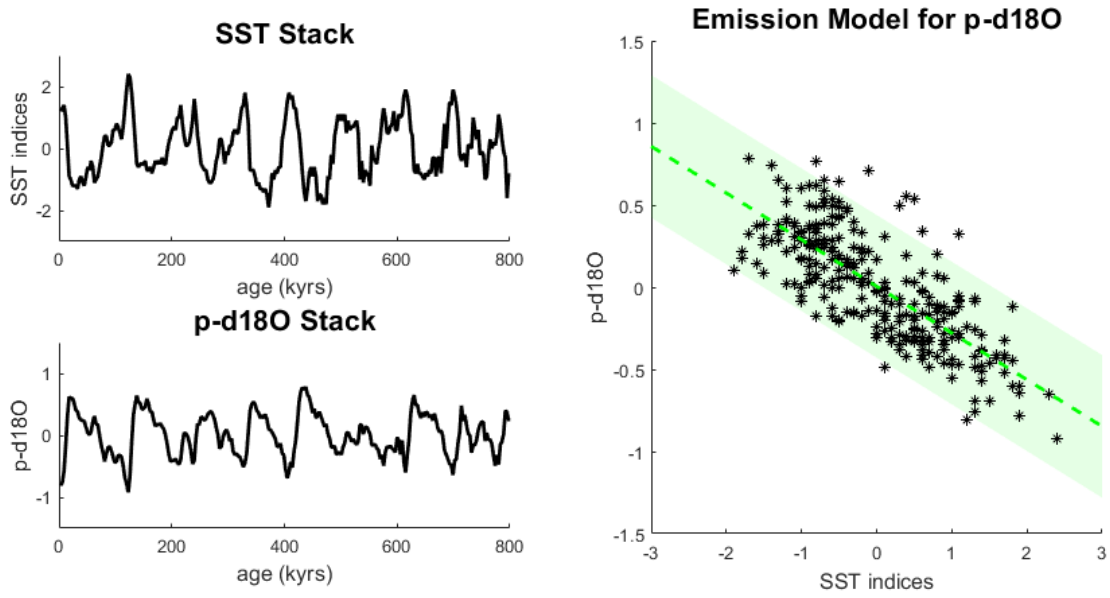

Figure 5: The SST index and planktonic $\delta^{18}O$ stack (left) and the plots of their pairs (stars) on the emission model (right). The shaded region indicates the 95% confidence band and the dashed line is the mean function.

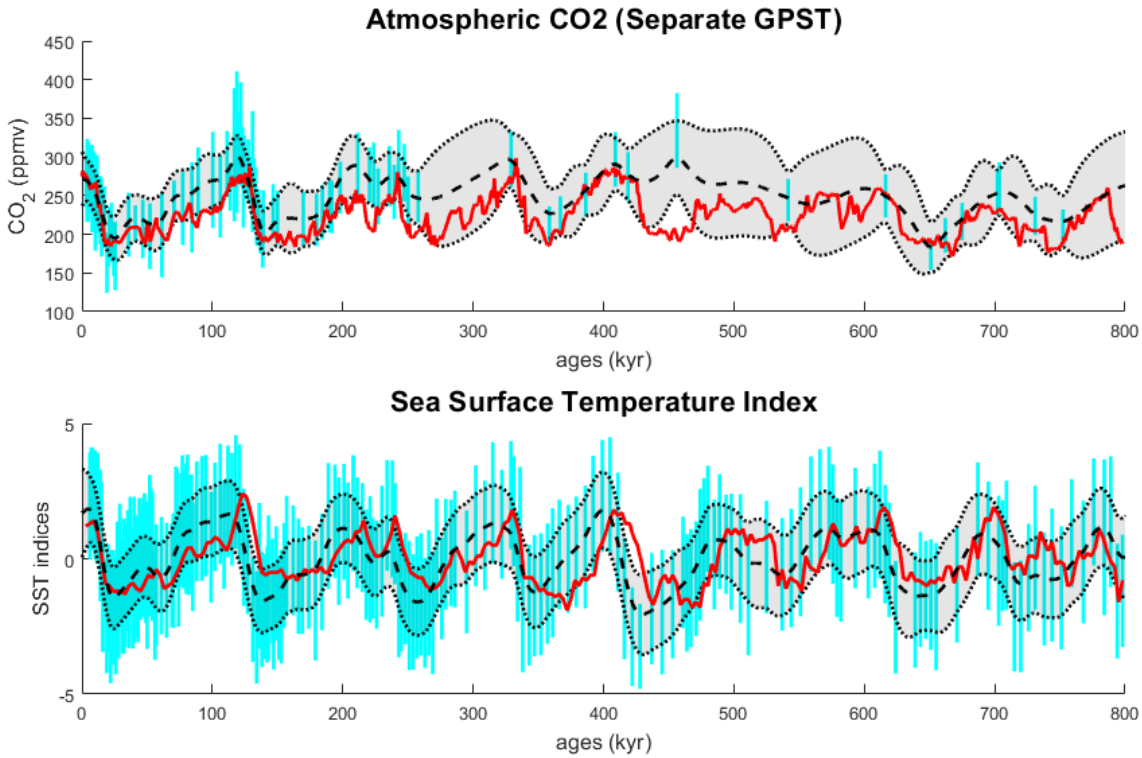

Figure 6: The reconstruction results of atmospheric CO_2 (above) and SST index (below) by GPST separately. In each case, the shaded region indicates the 95% confidence bands of the inferred events, the black dashed line is the mean function, blue bars represent the benchmark from the individual inference, and the red curve shows the "true" events.

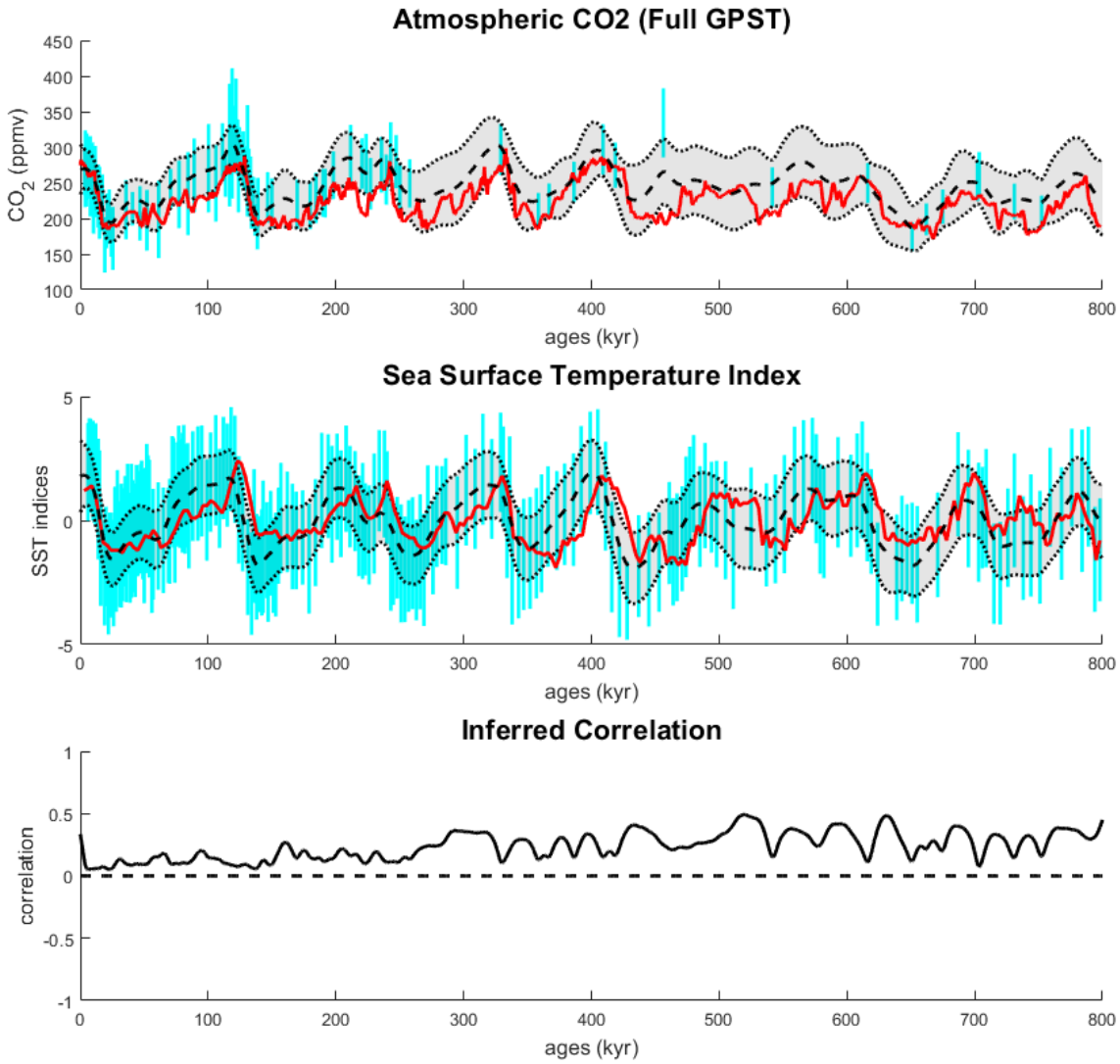

Figure 7: The reconstruction results of atmospheric CO_2 (top), SST index (middle) and their correlation (bottom) by the full GPST.

of δ^{18}O is converted into a generalized Student's t-distribution as (2). Figure 1 gives the spatial information of those three cores.

Ages are rescaled by $x \rightarrow (x - 263.1929)/229.2451$. Unlike the other kernel hyperparameters, periods r_1 and r_2 in (4) are both set to $100 \rightarrow 100/229.2451$, as the ages are standardized, which implies that the periods of atmospheric CO_2 and SST events are 100 kiloyears. The values are restored from their standardized forms in the final step.

4 RESULTS

To establish benchmarks, we first ran the Metropolis-Hastings algorithm [10, 18, 20] on the planktonic δ^{18}O proxy of MD97-2140 to get the 95% confidence intervals of the associated SST indices individually by using (2) as the emission model only. For atmospheric CO_2, the published confidence intervals in [5] are used. We also have "true" atmospheric CO_2 and SST indices from the Antarctic ice core records [12, 16, 21, 22, 24, 26] and from the Shakun's stack, respectively.

Figure 6 visualizes the GPST results that were obtained separately, i.e., not assuming the correlation between CO_2 and SST index, whereas figure 7 shows those by the full model in section 2. The reconstructed SST indices of two models are similar to each other and give tighter and more accurate inference than the individual ones. The advantage of our GPST model in section 2 appears in the reconstruction of atmospheric CO_2 in figure 7: though δ^{11}B

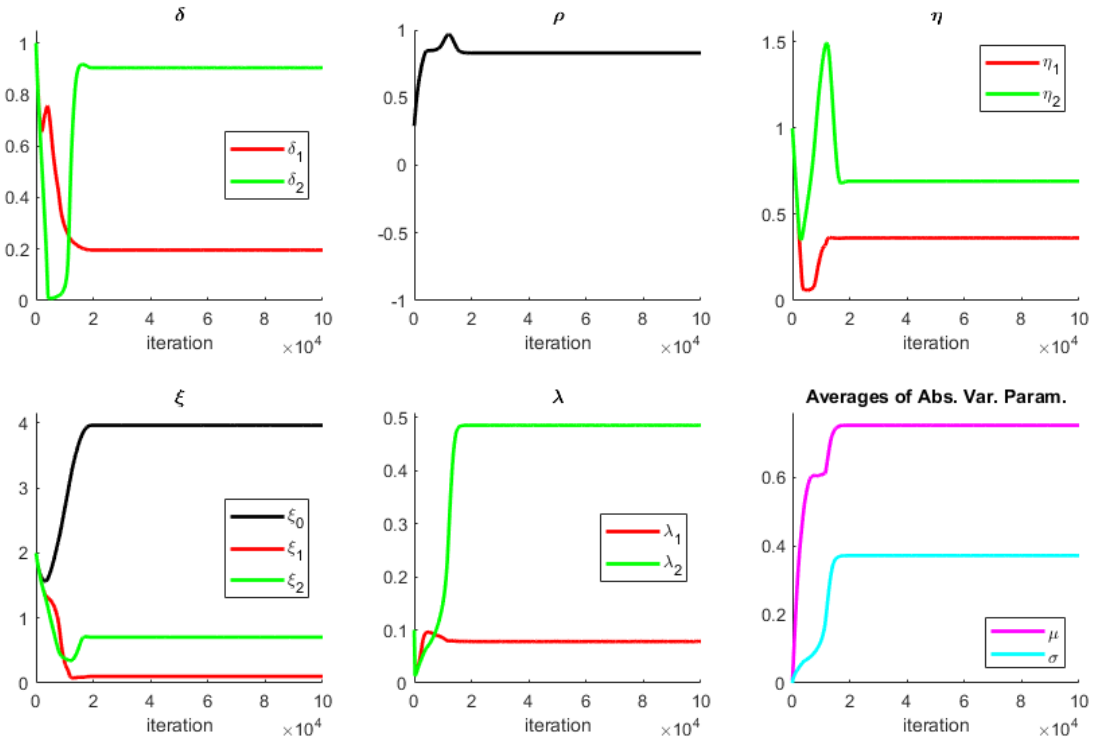

Figure 8: The tuned kernel hyperparameters over iteration. The last panel shows the average values of the absolute variational parameters as indicators of the convergence of variational parameters over iterations

after 250 kiloyears are sparser than those before that moment, the inferences are tighter and more accurate than those of figure 6. The assumption that atmospheric CO_2 and SST index are correlated brings such an advantage by "borrowing" the information from the planktonic δ^{18}O proxy indirectly to the reconstruction of atmospheric CO_2. How much information is brought from one to another is roughly measured by the inferred correlation over ages in the bottom panel of figure 7. The results are improved further than the individual inference at 456.3 kiloyears that stems from the apparent outlier of δ^{11}B. Figure 8 shows that the kernel hyperparameters and variational parameters are converged after 20000 iterations of the gradient descent steps.

5 CONCLUSION

Our GPST model reconstructs both atmospheric CO_2 and SST index by considering not only their proxies, δ^{11}B and planktonic δ^{18}O, but also their mutual dependency over ages which a Gaussian process specifies. Emission models are defined by Generalized Student's t-distributions that reflect the uncertainty of published inference. A variational inference approximates the intractable posterior distribution with a Gaussian distribution to make the inference at arbitrary query ages explicitly. Both kernel hyperparameters and variational parameters are optimized with the ELBO by a gradient descent. Our model that deals with both climate events shows an advantage over one that treats each event separately. This is particularly the case for CO_2 where the associated δ^{11}B is of low resolution after 250 kiloyears. This advantage stems from the information that is borrowed from the dense planktonic δ^{18}O proxy.

As discussed in [15], Gaussian process models themselves have an innate disadvantage: they become intractable as the size of data increases because matrix inversions are required in both learning and inference. [28] deals with this drawback in the framework of variational inference that considers pseudo-inputs and sufficient statistics. Extending our model by adding that step is easy but not applied here because we have only 285 observations. The extension is, however, required after all to exploit more hidden climate events and relevant proxy observations. Another problem rises as the number of hidden climate events increases: the number of kernel hyperparameters in our setting is quadratic to it. This would not be problematic in practice because only events that are regarded as correlated to one another are worth being coupled. Nevertheless, our GPST model provides an effective and general way of taking data that are spaced irregularly and treating transition models nonparametrically. The MATLAB codes that we have run are in https://github.com/eilion/GPST_CI2020.

ACKNOWLEDGMENTS

This paper is based on the works supported by the Division of Applied Mathematics in Brown University and by the National Science Foundation (NSF) under a grant number OCE-1760838.

REFERENCES

[1] Zexun Chen, Bo Wang, and Alexander Gorban. 2020. Multivariate Gaussian and Student−t Process Regression for Multi-output Prediction. *Neural Computing and Applications* (04 2020), 3005–3028.

[2] J. Christen and E. Sergio. 2009. A New Robust Statistical Model for Radiocarbon Data. *Radiocarbon* 51, 3 (2009), 1047–1059.

[3] Thibault de Garidel-Thoron, Yair Rosenthal, Franck Bassinot, and Luc Beaufort. 2005. Stable sea surface temperatures in the Western Pacific warm pool over the past 1.75 million years. *Nature* 433 (2005), 294–298.

[4] A. Doucet, N. de Freitas, and N. (Eds.) Gordon. 2001. *Sequential Monte Carlo methods in practice.* Springer.

[5] Kelsey A. Dyez, Bärbel Hönisch, and Gavin A. Schmidt. 2018. Early Pleistocene Obliquity-Scale pCO2 Variability at 1.5 Million Years Ago. *Paleoceanography and Paleoclimatology* 33, 11 (2018), 1270–1291.

[6] Stefanos Eleftheriadis, Thomas F.W. Nicholson, Marc P. Deisenroth, and James Hensman. 2017. Identification of Gaussian Process State Space Models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 5315–5325.

[7] Gavin L. Foster and James W.B. Rae. 2016. Reconstructing Ocean pH with Boron Isotopes in Foraminifera. *Annual Review of Earth and Planetary Sciences* 44, 1 (2016), 207–237.

[8] Roger Frigola, Yutian Chen, and Carl E. Rasmussen. 2014. Variational Gaussian Process State-Space Models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Vol. 2. MIT Press, 3680–3688.

[9] Marc G. Genton. 2002. Classes of Kernels for Machine Learning: A Statistics Perspective. *Journal of Machine Learning Research* 2 (2002), 299–312.

[10] W. K. Hastings. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 1 (1970), 97–109.

[11] Bärbel Hönisch, N. Gary Hemming, David Archer, Mark Siddall, and Jerry F. McManus. 2009. Atmospheric Carbon Dioxide Concentration Across the Mid-Pleistocene Transition. *Science* 324, 5934 (2009), 1551–1554.

[12] Andreas Indermühle, Eric Monnin, Bernhard Stauffer, Thomas F. Stocker, and Martin Wahlen. 2000. Atmospheric CO2 concentration from 60 to 20 kyr BP from the Taylor Dome Ice Core, Antarctica. *Geophysical Research Letters* 27, 5 (2000), 735–738.

[13] Diederik P Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat.ML] arXiv:1312.6114v10 [stat.ML].

[14] M. Klaas, M. Briers, N. de Freitas, A. Doucet, S. Maskell, and D. Lang. 2006. Fast particle smoothing: if I had a million particles. In *ICML*.

[15] T. Lee. 2020. *State-space Models and Gaussian Processes in Paleoceanography and Paleoclimatology.* Ph.D. Dissertation. Brown University.

[16] Dieter Lüthi, Martine Floch, Bernhard Bereiter, Thomas Blunier, Jean-Marc Barnola, Urs Siegenthaler, Dominique Raynaud, Jean Jouzel, Hubertus Fischer, Kenji Kawamura, and Thomas Stocker. 2008. High-resolution carbon dioxide concentration record 650,000-800,000 years before present. *Nature* 453 (2008), 379–382.

[17] D. J. MacKay. 1998. Introduction to Gaussian processes. *NATO ASI Series F Computer and Systems Sciences* 168 (1998), 133–166.

[18] L. Martino, J. Read, and D. Luengo. 2015. Independent Doubly Adaptive Rejection Metropolis Sampling Within Gibbs Sampling. *IEEE Transactions on Signal Processing* 63, 12 (2015), 3123–3138.

[19] B. Matérn. 1986. *Spatial Variation.* Springer-Verlag.

[20] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* 21 (1953), 1087–1092.

[21] Eric Monnin, Andreas Indermühle, André Dällenbach, Jacqueline Flückiger, Bernhard Stauffer, Thomas F. Stocker, Dominique Raynaud, and Jean-Marc Barnola. 2001. Atmospheric CO2 Concentrations over the Last Glacial Termination. *Science* 291, 5501 (2001), 112–114.

[22] J. R. Petit, Jean Jouzel, D. Raynaud, N. I. Barkov, J.-M. Barnola, Isabelle BASILE-DOELSCH, M Bender, J Chappellaz, M Davis, G Delaygue, M Delmotte, V. M. Kotlyakov, Michel Legrand, V. Y. Lipenkov, C Lorius, L Pepin, C Ritz, E Saltzman, and M Stievenard. 1999. Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* 399, 6735 (1999), 429–436. https://hal.archives-ouvertes.fr/hal-00756651

[23] C. E. Rasmussen and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning).* The MIT Press.

[24] Dominique Raynaud, Jean-Marc Barnola, Roland Souchez, Reginald Lorrain, Jean-Robert Petit, Paul Duval, and Vladimir Y. Lipenkov. 2005. The record for marine isotopic stage 11. *Nature* 436 (2005), 39–40.

[25] J. D. Shakun, D. W. Lea, L. E. Lisiecki, and M. E. Raymo. 2015. An 800-kyr record of global surface ocean δ^{18}O and implications for ice volume-temperature coupling. *Earth and Planetary Science Letters* 426 (2015), 58–68.

[26] Urs Siegenthaler, Thomas F. Stocker, Eric Monnin, Dieter Lüthi, Jakob Schwander, Bernhard Stauffer, Dominique Raynaud, Jean-Marc Barnola, Hubertus Fischer, Valérie Masson-Delmotte, and Jean Jouzel. 2005. Stable Carbon Cycle - Climate Relationship During the Late Pleistocene. *Science* 310, 5752 (2005), 1313–1317.

[27] M. L. Stein. 1999. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer-Verlag New York.

[28] Michalis Titsias. 2009. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 5).* PMLR, 567–574.