A functional regression model for heterogeneous BioGeo-Chemical Argo data in the Southern Ocean

Moritz Korte-Stapff, Stilian Stoev, Tailen Hsing

*University of Michigan, Ann Arbor, MI United States America.*

E-mail: kortest@umich.edu, sstoev@umich.edu, thsing@umich.edu

Drew Yarger

*Purdue University, 150 N. University Street, West Lafayette, IN, United States of America, 47907*

E-mail: anyarger@purdue.edu

**Summary.**

Leveraging available measurements of our environment can help us understand complex processes. One example is Argo Biogeochemical data, which aims to collect measurements of oxygen, nitrate, pH, and other variables at varying depths in the ocean. We focus on the oxygen data in the Southern Ocean, which has implications for ocean biology and the Earth’s carbon cycle. Systematic monitoring of such data has only recently begun to be established, and the data is sparse. In contrast, Argo measurements of temperature and salinity are much more abundant. In this work, we introduce and estimate a functional regression model describing dependence in oxygen, temperature, and salinity data at all depths covered by the Argo data simultaneously. Our model elucidates important aspects of the joint distribution of temperature, salinity, and oxygen. Due to fronts that establish distinct spatial zones in the Southern Ocean, we augment this functional regression model with a mixture component. By modelling spatial dependence in the mixture component and in the data itself, we provide predictions onto a grid and improve location estimates of fronts. Our approach is scalable to the size of the Argo data, and we demonstrate its success in cross-validation and a comprehensive interpretation of the model.

**Keywords:** clustering, functional regression, penalized splines, physical and biogeochemical oceanography, spatially-dependent data
1. Introduction

The rapid growth of high-resolution data motivates continued advances in flexible and efficient statistical methodology. One prime example is the Argo data, which is collected by autonomous devices drifting through the world oceans called floats that measure temperature and salinity. Argo floats collect their data in so-called profiles, collections of measurements at various depths for fixed points in time and space (see Figure 1 and Argo, 2020). The Argo data has greatly enhanced our understanding of the physical nature of the upper 2000 meters of the oceans with near-uniform spatial sampling of these variables (Johnson et al., 2022). Argo floats determine their own depth through pressure measurements as an increase in 1 decibar in pressure occurs for each additional meter of depth (approximately). From here on, we will thus often use pressure rather than depth to refer to a float’s vertical position in the ocean.

An important expansion of the Argo Project’s goals is Biogeochemical (BGC) Argo, which has added a limited number of floats that also measure oxygen and nitrate, among other variables. We focus on Argo data in the Southern Ocean because the majority of the BGC Argo data has been collected under the Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) project (Riser et al., 2023, soccom.princeton.edu). The increased availability of biogeochemical data is important to improve our understanding of vital biogeochemical processes such as the biological carbon pump and air-sea CO\textsubscript{2} exchanges, monitor changes such as ocean deoxygenation and acidification, and improve estimates of the carbon budget (Bittig et al., 2019; Claustre et al., 2020; Gray et al., 2018). However, BGC floats use additional sensors, leading to higher fixed and maintenance costs than Core Argo floats (that is, those that measure only temperature and salinity). Thus, it is not practical to introduce the same coverage as achieved by Core Argo floats. Indeed, in the Southern Ocean during 2020, Core Argo floats collected approximately 12000 profiles while BGC Argo floats collected just 3600 over the same time period (see Figure 1). When modelling biogeochemical data, it is thus desirable to utilize the more prevalently available Core Argo data \textit{in addition to} the available BGC data to improve prediction and estimates and reduce uncertainty. This paper is aimed towards this goal, and our main contributions are as
A functional regression model for heterogeneous BioGeoChemical Argo data

Fig. 1. (Left) Oxygen, temperature, and salinity data of ten profiles collected from a BGC Argo float (WMO #5904679) as a function of pressure (1 decibar of pressure is approximately 1 meter of depth). (Top Right) Locations of BGC profiles collected in the Southern Ocean in 2020 by the Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) Project and (Bottom Right) locations of nearby Core Argo profiles collected in 2020.

follows.

(a) We develop a descriptive statistical model on the joint distribution of oxygen, temperature, and salinity in the Southern Ocean. The model fully takes into account the functional nature and space-time information of the Argo data.

(b) Based on this model, and leveraging the more readily available temperature and salinity data, we address methodological and computational issues related to the functional prediction and interpolation of oxygen levels at arbitrary locations.

While both (a) and (b) are of significant scientific as well as statistical interests, we are unaware of any existing work that considers them comprehensively. Below, we explain our approaches in more detail.
The basic building block for our model is a spatio-temporal, function-on-function regression model that encodes the functional relationship between the temperature, salinity and oxygen profiles across all pressure levels ranging from 0 to 2000 decibars. This enables us to predict oxygen profiles using temperature and salinity profiles. It also facilitates the estimation and prediction of other quantities of scientific interests such as derivatives and integrals of the profiles (Park et al., 2023).

One major challenge of modelling Argo data in the Southern Ocean is that the transition from warmer subtropical waters in the north to colder Antarctic waters does not occur smoothly. Instead, ocean properties in that region are characterized by a series of sudden changes, occurring at so-called fronts — sharp transition zones generally aligned east–west (Chapman et al., 2020; Orsi et al., 1995). In particular, the contrasting behavior of oxygen in various regimes of the Southern Ocean has been well established in Bushinsky et al. (2017). Consequently, any stationarity assumption on the data across different fronts would be violated. However, it is established that ocean properties in any region between fronts tend to be homogeneous (Chapman et al., 2020). To address this challenge, we augment our functional regression model with a latent mixture component aimed at capturing regional information. Our model thus describes aspects of the joint distribution of oxygen, temperature, and salinity modulated by different regimes in the Southern Ocean. In our analyses, we demonstrate that the dependence structure in pressure of these three variables varies critically across the regimes estimated through the mixture model. As a byproduct we obtain estimates for the locations of ocean fronts and how they vary with time. These are of considerable scientific interest in their own right and have been the subject of research in the oceanographic literature (Chapman et al., 2020). Consequently, they constitute an additional key contribution of this work.

Identifying ocean fronts through the use of clustering and mixture models in oceanography has recently gained significant traction. For example, Jones et al. (2023, 2019) focus on the Southern Ocean region; Rosso et al. (2020), Boehme and Rosso (2021), and Fonvieille et al. (2023) consider subregions of the Southern Ocean, while Maze et al. (2017), Poropat et al. (2023), and Houghton and Wilson (2020) analyze other oceanographic areas. Most of these works use standard Gaussian mixture models. The
primary difference between this literature and our work here is that we account for the spatio-temporal dependence for the cluster memberships and the data within each cluster while these critical aspects are ignored in the above cited works. We accomplish this by encoding the spatio-temporal dependence of cluster memberships using a Markov random field model (Liang et al., 2020; Jiang and Serban, 2012; Besag, 1975). In particular, this enables predictions at any location whether temperature and salinity data are available or not. An important application of this is mapping, or the creation of gridded products, where data sampled at irregular locations in space and time are used to provide estimates at a standard grid (for example, Roemmich and Gilson, 2009; Gray and Riser, 2014).

When temperature and salinity measurements are available, predictions of oxygen may be straightforwardly made using off-the-shelf methodology. A well-known example in that regard is the random forest approach of Giglio et al. (2018), which produces highly accurate predictions of oxygen at fixed pressure levels using temperature and salinity data. For this scenario, we provide comparisons of our mode-based approach with the random forest approach of Giglio et al. (2018). However, when estimating the oxygen level at a location or a pressure level where temperature and/or salinity information is not available, the random forest approach requires an additional step of interpolating predicted oxygen data onto this location. Our model-based approach, on the other hand, readily provides direct predictions as well as estimated uncertainty of oxygen levels at arbitrary locations and pressure levels using nearby temperature and salinity data without interpolations.

Our approach also involves advancements in both statistical methodology and implementation. There is a rich literature on function-on-function regression, including the seminal paper Yao et al. (2005) and Zhou et al. (2008), while modelling spatio-temporal dependence in univariate functional data has been studied extensively by, e.g., Zhou et al. (2010); Jiang and Serban (2012); Liang et al. (2020). However, we are unaware of any development of function-on-function regression models that accounts for spatio-temporal correlation in the functional response and predictor variables. Furthermore, to the best of our knowledge, functional kriging and cokriging problems in the
context of functional mixture models have not been explored at all. The SOCCOM Argo data considered here alone consists of more than fifteen thousand locations and millions of observations. Consequently, reducing the runtime and memory usage of the model estimation and prediction procedure is essential. This work is accompanied by an R package implementing our methodology where crucial parts of the estimation and prediction procedure are written in C++.

We finally briefly describe the organization of the rest of the paper. In Section 2, we propose a novel mixture model for multivariate, functional data with spatial dependence. In Section 3, we discuss parameter estimation, prediction, and model selection. In Section 4, we describe our analysis on the Argo data, including the description of the estimated model, evaluation of oceanographic front estimates, and cross-validation experiments. Finally, we discuss limitations and potential improvements to our work in Section 5.

2. The Data and the Model

Argo data is publicly available from the Argo Global Data Assembly (GDAC, Argo, 2020) after it undergoes various data control measures (Maurer et al., 2021). A precise description of the data under consideration is given in the supplement Section S1.

Notationally, our data consists of profiles at \( n \) locations \( \{x_i\}_{i=1}^n \) in the Southern Ocean and corresponding points in time \( \{t_i\}_{i=1}^n \). We use \( D \) to denote the spatial domain and use the positive real axis \( \mathbb{R}_+ \) as time domain. To simplify notation, we will denote a space-time coordinate \((x_i,t_i) \in D \times \mathbb{R}_+\) by \( s_i \). At each point \( s_i \), we observe profiles of a biogeochemical quantity \( Y \) (oxygen) as well as temperature \( T \) and salinity \( S \) each at pressure levels \( \{p_{ij}\}_{j=1}^n \). Consequently, oxygen data for one profile for \( Y \) is then \( \{Y_{s_i}(p_{ij})\}_{j=1}^n \) and for temperature and salinity \( \{T_{s_i}(p_{ij})\}_{j=1}^n \) and \( \{S_{s_i}(p_{ij})\}_{j=1}^n \). Even though the number of observations as well as the pressure indexes may vary between temperature, salinity and oxygen, we suppress this in our notation for clarity of presentation. Motivated by this structure in the data, we model the observations as point evaluations from unknown random functions of pressure, observed with additive
measurement errors:
\[ Y_s(p_{ij}) = Y_s(p_{ij}) + \epsilon^Y_{ij}, \quad T_s(p_{ij}) = T_s(p_{ij}) + \epsilon^T_{ij}, \quad S_s(p_{ij}) = S_s(p_{ij}) + \epsilon^S_{ij}. \]

We use calligraphic letters to denote the underlying functions and regular letters for the observations. For each index \( s \), we model the functions \( Y_s, T_s \) and \( S_s \) as elements of a Hilbert space \( \mathbb{H} \) which we choose to be \( \mathbb{H} = L_2([0, 2000]) \), the Hilbert space of square integrable functions on the closed interval \([0, 2000]\) because most Argo profiles consist of measurements between 0 and 2000 decibars. In addition, we assume sufficient regularity of the functions to ensure pointwise evaluations, and we later impose a smoothness penalty on the 2nd derivative. The measurement errors \( \{ \epsilon^Y_{ij} \}, \{ \epsilon^T_{ij} \}, \) and \( \{ \epsilon^S_{ij} \} \) are assumed to be i.i.d. Gaussian random variables with zero mean and variances \( \sigma^2_Y, \sigma^2_T \) and \( \sigma^2_S \) to be estimated from the data. The main focus of our model is the biogeochemical quantity \( Y \), so we collect temperature and salinity in a covariate vector of functions \( \mathcal{X}_s(\cdot, \cdot) = (T_s(\cdot), S_s(\cdot)) \in \mathbb{H} \times \mathbb{H} \). Combined, this yields two space-time, function-valued random fields \( \{ Y_s : s \in D \times \mathbb{R}_+ \} \) and \( \{ \mathcal{X}_s : s \in D \times \mathbb{R}_+ \} \).

**Modelling Ocean Fronts via a Latent Mixture Component**

As stated in the introduction, one main challenge with the Argo data is its nonstationary. To account for this, much of the previous work on the Argo data uses local approaches for example via sliding windows; see for example, Kuusela and Stein (2018); Yarger et al. (2022); Park et al. (2023). A locally-stationary model is estimated for each window, and since parameters are estimated anew for each window, the model adapts to varying covariance structures in different locations. However, in the Southern Ocean, this approach may not be suitable due to the sharp and sudden changes in water properties at the ocean fronts. Moreover, due to the circumpolar nature of the ocean fronts, water properties may stay homogeneous across long distances which any local approach will fail to fully utilize. In order to explicitly model the ocean fronts, we partition the data into \( G \) clusters, with ocean fronts representing the boundaries between clusters. We use a latent field \( \{ Z(s) \} \) where each \( Z(s) \) is a random variable taking value in \( \{1, \ldots, G\} \) for some \( G \) to be specified. We then model

\[
\mathcal{Y}_s(p) = \sum_{g=1}^G \mathbb{I}(Z(s) = g)Y^g_s(p), \quad \mathcal{X}_s(p_1, p_2) = \sum_{g=1}^G \mathbb{I}(Z(s) = g)\mathcal{X}^g_s(p_1, p_2)
\]
where the random fields \( \{ Y^g \} \) and \( \{ X^g \} \) govern the spatio-temporal covariance structure in cluster \( g \). Here \( \mathbb{I} \) is the usual indicator function.

For \( \{ Z(s) \} \), we employ a commonly used Markov random field model (cf. Jiang and Serban, 2012; Liang et al., 2020), where the marginal distribution \( Z(s_i) \) given appropriately defined neighbors \( \partial s_i \) is given by

\[
P(Z(s_i) = g | \partial s_i) \propto \exp \left( \xi \sum_{s_j \in \partial s_i} \mathbb{I}(Z(s_j) = g) \frac{\text{dist}(s_i, s_j)}{\text{dist}(s_i, s_j)} \right).
\]

We use \( k \)-nearest neighbor graphs to define the neighborhood structure based on the distance \( \text{dist}(s_i, s_j) = \text{dist}_{\text{long}} + 3\text{dist}_{\text{lat}} + 12\text{dist}_{\text{dayofyear}} \) to account for the anisotropy in the Argo data. The additional weighting by the inverse distance follows the intuition that profiles that are close should provide better cluster information than profiles that are further away. This also reduces the importance of the choice of the number of nearest neighbors to construct the underlying graph which we chose to be 15. Consequently, we found that reasonably varying this number has negligible impact. The Markov random field model allows cluster memberships to be interpolated between the irregularly-sampled profile locations. Moreover, the Markov structure yields computational advantages by using a pseudo-likelihood approximation as in Besag (1975).

We found that incorporating a temporal component in the distance is important. Indeed, we implemented a Markov random field that used only space (without a seasonal component) and found that in many locations this gave conflicting information on cluster membership from winter versus summer profiles.

**Spatio-Temporal Function on Function Regression** Within each cluster, we use the following spatial regression model to specify the joint distribution of \( X^g \) and \( Y^g \):

\[
Y^g_s(\cdot) = \mu^g_{Y,t}(\cdot) + \mathcal{F}^g(X^g_s - \mu^g_{X,t})(\cdot) + \mathcal{E}^g_s(\cdot),
\]

where the \( \mathcal{F}^g : \mathbb{H} \times \mathbb{H} \to \mathbb{H} \) are linear operators and the \( \mathcal{E}^g_s \) are functional, spatio-temporally correlated residuals that are independent of \( \{ X^g \} \). Here, \( \mu^g_{Y,t}(\cdot) \) and \( \mu^g_{X,t}(\cdot) \) are mean functions of pressure for the response and predictors, respectively, in the \( g \)-th group and depend on time \( t \). In the mean function \( \mu^g_{Y,t}(\cdot) \), we include seasonality using...
a Fourier basis as in Roemmich and Gilson (2009):

$$\mu_{Y,t}(\cdot) = \gamma_{Y,0}(\cdot) + \sum_{l=1}^{R} \sin \left( 2\pi l \frac{t}{365.25} \right) \gamma_{Y,l}(\cdot) + \sum_{l=1}^{R} \cos \left( 2\pi l \frac{t}{365.25} \right) \gamma_{Y,R+1}(\cdot).$$

The evaluation $\gamma_{Y,0}(p)$ then represents the average (across the year) mean of oxygen for group $g$ at pressure $p$, while the $\gamma_{Y,l}(p)$ for $l = 1, \ldots, 2R$ describe seasonal variations in that mean based on sines and cosines. We take $R = 3$ in this case to capture the main seasonality component, which is half of the number basis functions used in Roemmich and Gilson (2009) because in Antarctica there are only two seasons (NASA, 2024). The mean functions of the predictors have identical form. Next, we employ a Karhunen-Loève type expansion along the pressure dimension,

$$X_{s}(\cdot, \cdot) = \mu_{X,t}(\cdot, \cdot) + \sum_{l=1}^{\infty} \alpha_{l}(s) \psi_{l}(\cdot, \cdot), \quad E_{s}(\cdot) = \sum_{l=1}^{\infty} \eta_{l}(s) \phi_{l}(\cdot)$$

in order to model the spatio-temporal correlation structure through the principal component random fields $\{\alpha_{l}(s) : s \in \mathbb{D} \times \mathbb{R}_{+}\}$ and $\{\eta_{l}(s) : s \in \mathbb{D} \times \mathbb{R}_{+}\}$ (Zhou et al., 2010). The functions $\{\psi_{l}\}_{l=1}^{\infty}$ and $\{\phi_{l}\}_{l=1}^{\infty}$ form an orthonormal basis for $H \times H$ and $H$ respectively for each $g$ and are called the principal component functions (PCFs). In practice, the infinite sums have to be truncated appropriately. Through the joint principal component decomposition of $X_{s}(\cdot, \cdot)$, dependence between predictors (temperature and salinity) is modeled. Quantifying this dependence in the model is critical, since these variables are known to be dependent. For example, temperature, salinity, and pressure jointly determine potential density which is known to be generally monotone as a function of pressure (Section 3.5 of Talley, 2011). The coefficient random fields $\{\alpha_{l}(s) : s \in \mathbb{D} \times \mathbb{R}_{+}\}$ and $\{\eta_{l}(s) : s \in \mathbb{D} \times \mathbb{R}_{+}\}$ are assumed to be independent across cluster $g$ and component $l$. Using (3) in (2), and afterwards truncating the infinite sums after $Q_{1}$ and $Q_{2}$ terms, we obtain

$$Y_{s}(\cdot) = \mu_{Y,t}(\cdot) + \sum_{l=1}^{Q_{1}} \alpha_{l}(s) \mathcal{F}_{l}^{y} \psi_{l}(\cdot) + \sum_{l=1}^{Q_{2}} \eta_{l}(s) \phi_{l}(\cdot).$$

We model the random fields constituting the principal component scores as Gaussian random fields as is common in spatial statistics. Because they have zero mean, it remains to specify their covariance structure. A popular choice is the Matérn covariance function
as advocated for by Stein (2013). However, for the Argo data an isotropic covariance function such as the standard Matérn covariance is likely not appropriate. For example, Kuusela and Stein (2018) found that, typically, the correlation range scales are different for longitude and latitude. We follow Guinness (2021) who, when modelling Argo data, used spatial deformation of the sphere via gradients of the spherical harmonics to model anisotropies. Concretely, we deform the sphere by mapping a point \( x \mapsto x + \Phi_{g,\alpha,l}(x) \) where \( \Phi_{g,\alpha,l}(x) \) is the weighted sum of gradients of the spherical harmonics of degree 2 at \( x \). The weights will be learned from the data. We then use the following covariance

\[
\text{Cov} \{ \alpha_{g,l}(s), \alpha_{g,l}(s') \} = \sigma_{g,\alpha,l}^2 \{ d_{g,\alpha,l}(s, s') \} \nu_{g,\alpha,l} \{ d_{g,\alpha,l}(s, s') \},
\]

(4)

where \( \nu \geq 0 \), \( K_\nu \) is the modified Bessel function of the second type and for \( s = (x, t) \),

\[
d_{g,\alpha,l}(s, s') = \frac{\| x + \Phi_{g,\alpha,l}(x) - \{ x' + \Phi_{g,\alpha,l}(x') \} \|}{\kappa_{g,\alpha,l,x}} + \frac{\| t - t' \|}{\kappa_{g,\alpha,l,t}},
\]

where \( \| \cdot \| \) is the standard Euclidean norm (chordal distance). An R implementation of this covariance function is readily available in the GpGp package (Guinness et al., 2021).

Generally, we found that allowing more flexibility in the covariance function improved prediction performance. The smoothness parameter \( \nu_{g,\alpha,l} \), variance parameter \( \sigma_{g,\alpha,l}^2 \), and range parameter \( \kappa_{g,\alpha,l,x} \) and \( \kappa_{g,\alpha,l,t} \) govern properties of the Matérn model for the \( l \)-th predictor principal component of the \( g \)-th group. We use the same covariance form for \( \{ \eta_{g,l}(s) \} \).

3. Model Estimation and Prediction

The mean functions, the PCFs, and the functions \( \{ \mathcal{F}_g \psi_g^l \} \) are unknown and have to be estimated from the data. To do so, we use a cubic B-spline basis, a common choice in functional data analysis with favourable computational and provable optimality properties (Wahba, 1990; Hsing and Eubank, 2015). Specifically, if \( \{ b_i \}_{i=1}^P \) are the elements of the cubic B-spline basis on \( [0, 2000] \) write \( B(p) = (b_1(p), \ldots, b_P(p)) \). Then we use

\[
\gamma_{y,l}^g(p) = B(p) Y_{y,l}^g, \quad \phi_l^g(p) = B(p) \Theta_{E,l}^g, \quad (\mathcal{F}_g \psi_l^g)(p) = B(p) \Lambda_l^g,
\]

\[
\gamma_{X,l}^g(p_1, p_2) = (B(p_1), B(p_2)) Y_{X,l}^g, \quad \psi_l^g(p_1, p_2) = (B(p_1), B(p_2)) \Theta_{X,l}^g,
\]

where ...
where \( \Upsilon^g_{Y,l} \) is the \( l \)-th column of the \( P \times 7 \) matrix of spline coefficients for the seasonal means of \( Y \) and \( \Upsilon^g_{X,l} \) is the corresponding matrix for the covariates, \( \Theta^g_{X,l} \) is the \( l \)-th column of the \( P \times Q_2 \) matrix containing the coefficients for the PCFs for \( E \), and the \( P \times Q_1 \) matrix \( \Lambda^g \) contains the coefficients for the \( \{ \mathcal{F}^g \psi^g_l \} \). Similarly, \( \Theta^g_{X} \) is a \( 2P \times Q_1 \) matrix. For the PCFs, we impose the following constraints to enforce orthonormality:

\[
(\Theta^g_{X})^\top \int B(t)^\top B(t) \, dt \Theta^g_{X} = I_{Q_1}, \quad (\Theta^g_{E})^\top \int B(t)^\top B(t) \, dt \Theta^g_{E} = I_{Q_2}.
\]

Here \( \otimes \) is the standard Kronecker product and \( I_q \) is the \( q \times q \) identity matrix. For identifiability reasons, we require the first entries in each column of \( \Theta^g_{X} \) and \( \Theta^g_{E} \) to be positive and order the PCFs by the marginal variances of the random fields constituting the principal component scores.

**Remark 1.** A common approach to enforce the orthogonality constraints in (5) is to use an orthogonalized spline basis \( \tilde{B}(t) \) and then require \( (\Theta^g_{X})^\top \Theta^g_{X} = I_{Q_1} \) and \( (\Theta^g_{E})^\top \Theta^g_{E} = I_{Q_2} \); see Zhou et al. (2010) and Liang et al. (2020), among others. However, for datasets as large as the Argo data, this becomes impractical because the evaluation matrices \( \tilde{B}(t) \) (at observed data points) for orthogonalized splines are less sparse compared to regular \( B \)-splines. We outline in the supplement Section S9 that regular \( B \)-splines can be used to obtain orthonormal principal components with a proper orthonormalization step.

### 3.1. Monte Carlo EM Algorithm

Notationally, we collect all observations \( \{ \{ Y_{s,ij}, p_{ij} \} \}_{j=1}^{n_i} \) in the vector \( Y \) and all temperature and salinity observations in the vector \( X \). Data from a single profile will be denoted by \( Y_i \) and \( X_i \) respectively. The corresponding latent principal component scores and the cluster memberships are collected in the vectors \( \alpha, \eta, \) and \( Z \) respectively. We collect the spline coefficient matrices, the covariance parameters, and the MRF parameter in the set \( \Omega \). To fit the model, we aim to maximize the penalized log-likelihood of the observed data

\[
\ell(\Omega, X, Y) - \mathcal{P}(\Omega) = \log \int L(\Omega, X, Y, \alpha, \eta, Z) \, d\alpha d\eta dZ - \mathcal{P}(\Omega),
\]

where \( L(\Omega, X, Y, \alpha, \eta, Z) \) is the likelihood function.
where $\mathcal{R}(\Omega)$ denotes a roughness penalty and $L$ denotes the likelihood function. Corresponding to our choice of cubic B-splines, we use integrated squared second derivatives of the respective functions as penalty (cf. Theorem 6.6.9; Hsing and Eubank, 2015).

Because the random vectors $\alpha, \eta$ and $Z$ are unobserved, direct maximization of (6) is not feasible. A useful tool for such scenarios is the EM algorithm (Dempster et al., 1977). It consists of initializing the parameters to $\Omega^{(1)}$ and alternating between the following two steps for $m = 1, 2, \ldots$ until convergence:

(a) **E-step**: Compute $Q(\Omega|\Omega^{(m)}) = \mathbb{E}\{\ell(\Omega, X, Y, \alpha, \eta, Z)|X, Y, \Omega^{(m)}\}$.

(b) **M-step**: Update the parameters via $\Omega^{(m+1)} = \arg\max\Omega Q(\Omega|\Omega^{(m)}) - P(\Omega)$.

Direct application of the EM algorithm in our case is not possible because $Q(\Omega|\Omega^{(m)})$ is not tractable. Instead, we use a Monte Carlo EM algorithm (Wei and Tanner, 1990).

**Monte Carlo E-Step.** We use importance sampling to estimate $Q(\Omega|\Omega^{(m)})$. That is, we fix $T \in \mathbb{N}$ and create samples $\alpha^{(\tau)}, \eta^{(\tau)}, z^{(\tau)}$ for $\tau = 1, \ldots, T$ from a proposal distribution $h$ and approximate $Q(\Omega|\Omega^{(m)})$ via

$$
\hat{Q}(\Omega|\Omega^{(m)}) = \left(\sum_{\tau=1}^{T} w^{(\tau)}\right)^{-1} \sum_{\tau=1}^{T} w^{(\tau)} \ell(\Omega, X, Y, \alpha^{(\tau)}, \eta^{(\tau)}, z^{(\tau)})
$$

(7)

where the quantities $w^{(\tau)} = f(\alpha^{(\tau)}, \eta^{(\tau)}, z^{(\tau)}|X, Y, \Omega^{(m)})/h(\alpha^{(\tau)}, \eta^{(\tau)}, z^{(\tau)}|\Omega^{(m)})$ are called the importance sampling weights and $\bar{w}^{(\tau)}$ are the self-normalized weights. For details on importance sampling, see Robert and Casella (2004). In a similar situation, Liang et al. (2020) use a Markov Chain Monte Carlo EM algorithm by employing a Gibbs sampler to generate samples of the latent random variables. When we applied their sampler to the Argo data, we found that the resulting Markov Chain mixes slowly and thus obtained suboptimal results. Both Monte Carlo approaches are motivated by the fact that, due to the dependence in the data and the discrete nature of the cluster memberships, the conditional distribution $Z|X, Y$ is not tractable in closed form. A brief discussion and comparison as well as details on our proposal $h$ can be found in the supplement Section S4.
M-step. To update the parameters set \( \Omega^{(m+1)} = \arg \max_{\Omega} Q(\Omega|\Omega^{(m)}) - \mathcal{D}(\Omega) \). The log-likelihood of the complete data factors into separate terms

\[
Q(\Omega|\Omega^{(m)}) = \sum_{\tau=1}^{T} \tilde{w}^{(\tau)} \left\{ \ell(Y|z^{(\tau)}, \alpha^{(\tau)}, \eta^{(\tau)}) + \ell(X|z^{(\tau)}, \alpha) + \ell(\alpha^{(\tau)}|\alpha) + \ell(\eta^{(\tau)}|\eta) + \ell(z^{(\tau)}) \right\},
\]

allowing for most parameters to be updated separately. A closed form solution exists for most of the parameters, except for the spatial parameters and the MRF parameter for which numerical maximization is necessary. The updating formulas are given in the supplement Section S8. For EM algorithms, initialization can play an important role. We provide a detailed initialization procedure that we found to work well in the supplement Section S10.

3.2. Model Selection

We choose the number of clusters \( G = 5 \) by domain knowledge (see e.g. Chapman et al., 2020) as indicated in the introduction and have already specified our MRF structure. It remains to choose the number of principal components and the smoothing penalties. The number of knots for the spline basis plays a small role in model fit as long as there are sufficiently many to approximate relevant functions because we impose a smoothing penalty (cf. Zhou et al., 2008). For the remaining tuning parameters, we use an AIC criterion where we use again importance sampling to approximate the intractable observed data likelihood. Let \( \mathcal{D}(\Omega) \) denote the model degrees of freedom, \( z^{(1)}, \ldots, z^{(T)} \) samples of the latent field from a proposal \( \tilde{h} \) and \( \tilde{w}^{(\tau)} \) the corresponding weights. Then our AIC criterion is

\[
-2 \log \left( \hat{f}(X, Y) \right) + 2 \mathcal{D}(\Omega) = -2 \log \left( \sum_{\tau=1}^{T} \tilde{w}^{(\tau)} \hat{f}(X, Y|Z = z^{(\tau)}) \right) + 2 \mathcal{D}(\Omega). \tag{8}
\]

The proposal \( \tilde{h} \) is slightly different from the one used during the EM-algorithm; details are found in the supplement Section S4. To determine \( \mathcal{D}(\Omega) \), we take into account the smoothing penalty imposed on functions as in Wei and Zhou (2010). We choose different smoothing parameters for \( \{\mu_{g,Y,t}^0\}, \{\mu_{g,X,t}^0\}, \{\phi_{g,l}^0\}, \{\psi_{g,l}^0\}, \{\mathcal{F}^{p}\psi_{g,l}^0\}; \) for simplicity, we assume that these parameters do not vary over the cluster or principal component index.
A frequently used alternative choice is to use the output of the Monte-Carlo E-step $\hat{Q}$ from the last EM-iteration as an approximation to the true likelihood (Huang et al., 2014; Liang et al., 2020). This did not work well for the Argo data as anticipated by Ibrahim et al. (2008), and we found that the cost of one additional E-step of (8) was offset by a more reliable model selection.

### 3.3. Prediction

To provide functional predictions at new locations $\hat{s} = \{\hat{s}_i\}$ using the observed data $Y$ and $X$ we use the conditional expectation $\hat{Y}_{\hat{s}}(\cdot) = E\{Y_{\hat{s}}(\cdot)|X, Y\}$. In this setting, we may include additional data (for example, temperature and salinity data at or nearby $\hat{s}$) in $X$ not included in model training. Furthermore, we stress that one does not need temperature and salinity at $\hat{s}$ to construct predictions at these locations. Similar to the E-step, closed-form expressions of the conditional expectation $E\{Y_{\hat{s}}(\cdot)|X, Y\}$ are not available due to the latent field $Z$. Therefore, we again use importance sampling. Concretely, we generate samples of the latent field (which now includes the new locations $\hat{s}$) $z^{(1)}, \ldots, z^{(T)}$ from the Potts model (1) and then set

$$\hat{Y}_{\hat{s}}(\cdot) = \sum_{i=1}^{T} \bar{w}^{(\tau)} E\left\{Y_{\hat{s}}(\cdot)|X, Y, z^{(\tau)}\right\},$$

where $\bar{w}^{(\tau)}$ are the importance weights. The conditional (co)variances for oxygen at varying pressures or locations are formed using the importance samples and the law of total (co)variance. For example, for one location $\hat{s}_1$ and pressure $p$, we compute:

$$\text{Var} \left\{ \hat{Y}_{\hat{s}_1}(p) \Big| X, Y \right\} = \sum_{\tau=1}^{T} \bar{w}^{(\tau)} \left[ \text{Var} \left\{ \hat{Y}_{\hat{s}_1}(p) \Big| X, Y, z^{(\tau)} \right\} \right.$$  
$$+ \left\{ E \left( Y_{\hat{s}_1}(p) \Big| X, Y, z^{(\tau)} \right) - \hat{Y}_{\hat{s}_1}(p) \right\}^2 \right].$$

Due to the Gaussianity of the latent fields $\{\alpha_{g,l}(s)\}$ and $\{\eta_{g,l}(s)\}$, closed from expressions for the quantities appearing in (9) and (10) are available, see the supplement Section S5. Inference for the latent field $Z(\hat{s})$ at new locations is based on the samples $z^{(1)}, \ldots, z^{(T)}$.

Modeling functional dependence enables us to construct confidence bands for $Y_{\hat{s}_i}(\cdot)$ for a location $\hat{s}_i$ such that we expect $Y_{\hat{s}_i}(p)$ to lie entirely (for all $p$) within the band.
We construct bands based on the estimates of $\text{Var}(\alpha_{g,l}(s)|X, Y)$ and $\text{Var}(\eta_{g,l}(s)|X, Y)$ using the approach of Choi and Reimherr (2018). When we predict that $Z(\hat{s}_i)$ belongs to a certain cluster with probability very close to 1, these bands may be directly applied; otherwise, bands from different groups should be considered individually. Our functional approach also yields principled predictions for linear functionals such as integrals and derivatives of $Y_s$ that are frequently of interest for Argo data.

### 3.4. Computational Details

Due to the size of the Argo data considered here – alone the SOCCOM data consists of more than fifteen thousand locations and millions of observations – naive implementation of the above EM-algorithm is computationally expensive. The most computationally demanding steps are the covariance parameter estimation of the latent random fields $\{\alpha_g(s)\}$ and $\{\eta_g(s)\}$ during the M-step as well as computing their conditional distribution for the E-step and for predictions. The main tool we employ for these steps is Vecchia’s approximation Katzfuss et al. (2020) but additional steps are necessary. We outline them in the supplement Section S7.

### 4. Data Analysis Results

Here, we present results from our analysis of the Argo data, beginning with the estimated model in Section 4.1. We then provide a cross-validation study to assess prediction performance in Section 4.2 and present our data product of comprehensive functional predictions in Section 4.3.

#### 4.1. Description of estimated model

Our final model uses 18 principal components for the response and 18 for the predictors even though our AIC criterion (provided in the supplement Section S2) indicates that even more principal components could be used. However, we found that additional principal components did not improve prediction performance and increased computation time.
Fig. 2. (Top) Cluster predictions for the 15th day of four months. We plot in black lines traditional estimates of fronts using the same criteria for fronts as Bushinsky et al. (2017) based on the 2004-2018 Roemmich and Gilson (2009) estimate. (Bottom) Cluster membership of BGC Argo profiles, in a histogram summarized by the day of the year.

In Figure 2, we plot the predicted clusters for four months of the year. Our clusters are ordered by their mean temperature function’s value at 1000 decibars. We also plot in black lines previous estimates of fronts using the same criteria as Bushinsky et al. (2017) based on the 2004-2018 Roemmich and Gilson (2009) estimate, see also Orsi et al. (1995). These can be described as follows. South of the southern-most line is the Seasonal Ice Zone, whose northern boundary is defined by the maximum extent of sea ice. The next zone is defined by the maximum extent of sea ice in the south and the Subantarctic front in the north, which contains the Antarctic Circumpolar Current (ACC), a strong ocean current that encircles Antarctica (Bushinsky et al., 2016). The final plotted front is the Subtropical front. These traditional estimates of fronts are generally represented well by our clustering plotted in Figure 2. For example, our
first cluster matches the Seasonal Ice Zone for most months. Clusters 2 and 3 make up the main component of the ACC, with their northern boundary generally aligning with the Subantarctic front. Clusters 4 and 5 define waters above the Subantarctic front, with cluster 4 more concentrated in the Pacific and Atlantic. There are also substantial seasonal variations, underscoring the importance of modeling the contrast between winter and summer in the Southern Ocean. For example, in the austral winter (July) clusters 2 and 4 are disproportionately represented while cluster 5 is nearly absent. Between cluster boundaries, the predicted cluster has relatively high probability, while there is some uncertainty on their boundaries. In the supplement Section S3, we also compare with Jones et al. (2019)’s analysis of 6 clusters from temperature data.

In general, the clusters are influenced by a variety of processes occurring in the Southern Ocean. Near the ocean surface, interactions with the atmosphere result in higher levels of oxygen dissolved near oxygen’s saturation point in water. In this level, the primary determinant of dissolved oxygen is the water’s capacity to hold oxygen, which increases as temperature or salinity decreases (see Section 4.5 of Talley, 2011). Oxygen is generally decreasing for a water mass as a function of “age,” that is, the time since it has interacted with the atmosphere (Section 3.6 of Talley, 2011), due to the utilization of oxygen by ocean life. In the Southern Ocean, upwelling (water at lower depths rising to the surface) of older, low-oxygen water occurs near around the Southern ACC front (approximately, our cluster 2), after which it splits both poleward and toward the equator at the surface (Gray et al., 2018). During the winter, ice cover blocks the potential uptake of oxygen to the upper layer of the ocean in the Seasonal Ice Zone; freezing water also ejects its salinity as it freezes, increasing the salinity of the surrounding water (Chamberlain et al., 2018; Talley, 2011). Finally, photosynthesis from plants increases oxygen, while remineralization of organic matter decreases the oxygen concentration (Stramma and Schmidtko, 2021).

We find evidence of such processes in the mean functions for each of the clusters in Figure 3. For example, since temperature and salinity is low for cluster 1, the water’s capacity to hold oxygen is very high. Furthermore, there is no ice blocking the flux of oxygen into the ocean during the summer. As a result, dissolved oxygen for cluster 1
Fig. 3. (Top Left) Functional time-averaged mean functions $\gamma_{Y,0}$ (for oxygen) and $\gamma_{X,0}$ (for temperature and salinity); (Bottom Left) these mean functions are summarized in a temperature-salinity diagram. Contours of potential density are plotted in the background. (Right) Seasonal variation in mean functions for four months of the year and 0 to 400 decibars.

at the surface has the highest oxygen levels across all pressures and clusters in Figure 3. The sharp gradient in cluster 1’s mean between low and high oxygen represents the contrast and mixing of the old oxygen-sparse water with water that is highly-saturated in oxygen at the surface. In general, we associate cluster 2 with upwelling water. During the winter, the ice cover blocks the upwelling water flowing in the Seasonal Ice Zone from interacting with the atmosphere, and thus its oxygen levels are lower compared to the summer. We attribute the extent of cluster 2 in Figure 2 during the winter in the Seasonal Ice Zone to this phenomenon. Clusters 4 and 5 generally have similar mean functions for depths of more than 800 meters, representing shared deep bottom water across the Southern Ocean. In addition, the seasonal variations in the right panel of Figure 3 follow the expected changes: a decrease in oxygen and increase in salinity for winter months in cluster 1 at the surface, and increases in temperature during summer months at the surface for all clusters. We also provide a temperature and salinity plot, a visualization tool commonly used in oceanography, of the mean functions in
A functional regression model for heterogeneous BioGeoChemical Argo data

Figure 3. Comparing this to Figure 13.7 of Talley (2011), clusters 4 and 5 correspond to the Subantarctic zone, cluster 3 corresponds to the Polar Frontal Zone, clusters 2 and 1 correspond to the Antarctic Zone north and south of the Southern ACC front, respectively.

We plot the first functional principal components that describe the variance of the data in Figure 4. We found that the leading principal components are relatively smooth compared to higher-order principal components. In general, the principal components’ higher values (in absolute terms) near the surface indicate more variability there. The first principal component of the predictors measures a general increase or decrease in temperature across nearly all pressures. Based on the first transformation functions, a higher value for the first predictors PCFs leads to different responses in oxygen by group and pressure. At the surface, clusters 2 through 5 have decreasing oxygen for higher temperature, the expected direction based on a decrease in oxygen solubility. For cluster 1, this relationship is likely confounded by higher temperatures that are associated with less ice cover, which would increase oxygen levels. For greater depths, oxygen’s response to changes in the first predictor component is varied across groups, emphasizing the importance of accounting for heterogeneity in the Southern Ocean.

In Figure 5, we plot the functional cross-correlation estimated between each of the
processes, again demonstrating cluster heterogeneity. At the surface, the dependence between oxygen and temperature as well as oxygen and salinity again follow the expected (negative) direction in terms of solubility (Bushinsky et al., 2017). The thickness of this layer in the oxygen and temperature correlations matches the general increase in the depth of the mixed layer, the area at the ocean surface with relatively constant ocean properties, as one moves north in the Southern Ocean (Holte et al., 2017). This also generally matches well with water mass analysis of the Southern Ocean (for example, Figure 13.4 of Talley, 2011). Finally, between temperature and salinity, the overall signs of the relationship match with exploratory analysis in Salvana and Jun (2022), where there is a band of positive correlations for measurements at the same pressure.

Our model thus comprehensively estimates the complex nonstationarities and dependence between temperature, salinity, and oxygen at different pressures through the clusters and estimated model properties for each cluster. Our results are in accordance with known processes and characteristics of these variables in the Southern Ocean. We further elucidate seasonal variations in the clusters and the variables across the entire pressure dimension. These results motivate more detailed analysis of the seasonal variations in clusters which is beyond the scope of the current paper.

4.2. Prediction Performance

We now turn to the prediction of oxygen based on temperature and salinity. We assess the prediction performance of our model in two separate cross-validation scenarios: leave-profile-out (LPO) and leave-float-out (LFO). For LPO, we hold out oxygen data from 1/5 of BGC profiles using a random partition, train the model on the remaining data, predict the left-out oxygen data, and repeat this process for of the five sets in the partition. For LFO, we follow a similar process, but we instead partition floats into 10 sets, so that we hold out all oxygen data from all profiles from selected floats at a time. In the LPO setting, nearby oxygen data from the same float is available while in the LFO setting only data from other floats can be used. We find that as a consequence, prediction errors are generally smaller in the LPO setting. Since Argo floats either have or do not have BGC sensors, LFO is closer to the practical prediction setting.
First, we consider predictions of oxygen at locations where temperature and salinity data is available mimicking the scenario in which one is interested in reconstructing the Biogeochemical data at Core Argo floats. In this setting, we can compare to the random forest approach of Giglio et al. (2018). Afterwards, we consider the more challenging scenario of predicting oxygen at locations where no data is available and only nearby data can be used.

4.2.1. Oxygen Reconstruction with Temperature and Salinity

The results of the cross-validation exercise can be found in Tables 1 and 2 and plots of example predictions can be found in Figure 6. The shaded areas in Figure 6 correspond to pointwise prediction intervals (dark grey) and confidence bands for the entire function (light grey). We found that for held-out locations with temperature and salinity data available, there is virtually no variation in cluster assignment across Monte Carlo samples used for prediction. Consequently, each profile is assigned entirely into a particular cluster, and it is straightforward to construct confidence intervals and bands using the resulting Gaussian distribution using solely that cluster for that profile. A description
Fig. 6. Example profile prediction in the leave-profile-out (Left two panels, float 5905072 cycle 1, looking at different pressure ranges) and leave-float-out (Right two panels, float 5906249 cycle 13) settings. The dark gray shaded area is based on 95% pointwise intervals, while the light gray shared area is a 95% confidence band. MAD for measurements between 145 and 155 dbar are 3.99 and 6.6 for these profiles, respectively.

on the validity of our uncertainty quantification can be found in Table 3.

To compare prediction performance with the random forest employed in Giglio et al. (2018), we replicate their setup and consider Argo measurements between 145 and 155 decibars and compare the point predictions for each held-out data experiment in Table 1. The hyperparameters of the random forest are as in Giglio et al. (2018). In the LPO setting, the resulting median prediction error is close to the instrument measurement error around 3 $\mu$mol/kg (Maurer et al., 2021; Mignot et al., 2019) for our functional prediction error as well as the random forest model. The random forest obtains slightly better pointwise prediction results, though we believe they are qualitatively similar: relative to the range over which oxygen ranges for pressures between 145 and 155 dbar, the median errors are 1.2% (random forest) vs. 1.5% based our functional model and both these errors are very near the instrument error. More importantly, our model produces predictions across the entire pressure dimension which the random forest cannot do. It needs all covariates (including temperature and salinity) available in order to produce
A functional regression model for heterogeneous BioGeoChemical Argo data

Fig. 7. Median absolute error of oxygen (in $\mu$mol/kg) for leave-floats-out (Left) and leave-profiles-out (Right), summarized in 100 decibar bins. We compare predictions based on the cluster means only ("Mean"), the cluster means combined with temperature and salinity data ("Temperature and salinity"), and the full prediction ("Full prediction").

Table 1. Comparison with random forest (RF) and our functional cokriging (FC) approach at 145-155 decibars. The median absolute deviation (MAD) and root-mean-squared-error (RMSE) are the metrics used for each of the methods and leave-out settings.

| Approach | LPO MAD ($\mu$mol/kg) | LPO RMSE ($\mu$mol/kg) | LFO MAD ($\mu$mol/kg) | LFO RMSE ($\mu$mol/kg) |
|----------|------------------------|-------------------------|------------------------|-------------------------|
| RF       | 3.01                   | 6.98                    | 6.29                   | 13.22                   |
| FC       | 3.73                   | 8.68                    | 7.02                   | 13.99                   |

4.2.2. Prediction at Locations without Temperature and Salinity

We next assess the prediction performance of our model when no temperature and salinity data is available at the location of prediction. We use the same set-up and data splits as in the previous cross-validation experiment, however, this time we predict without the available temperature and salinity data at the prediction location. On the other hand, we use approximately 15000 additional Core profiles to illustrate how our model can utilize additional Core data to improve prediction performance. The result are displayed in Table 2. Table 2 also provides a comparison of prediction errors when a Core Argo profile is nearby vs. when no Core Argo data is nearby. Here, nearby means
Table 2. Prediction performance when no temperature or salinity data is available at the location of prediction.

| Pressure in [145,155] | LPO MAD (µmol/kg) | LPO RMSE (µmol/kg) | LFO MAD (µmol/kg) | LFO RMSE (µmol/kg) |
|-----------------------|-------------------|--------------------|-------------------|-------------------|
| Overall               | 6.83              | 16.96              | 12.06             | 27.32             |
| Core Data Nearby      | 6.80              | 16.70              | 10.55             | 23.27             |
| No Core Data Nearby   | 6.88              | 17.40              | 15.77             | 33.04             |
| All Pressure Levels   |                   |                    |                   |                   |
| Overall               | 3.10              | 10.07              | 6.80              | 16.38             |
| Core Data Nearby      | 3.23              | 9.83               | 6.23              | 14.13             |
| No Core Data Nearby   | 2.88              | 10.48              | 8.02              | 19.75             |

Table 3. Leave-out uncertainty quantification performance. The values represent average coverage of 95% confidence intervals.

| Leave-out type | Pointwise prediction intervals, by pressure | Functional Band |
|----------------|--------------------------------------------|-----------------|
|                | All            | 0–500          | 500–1000        | 1000–1500        | 1500–2000        |
| LFO            | 0.91           | 0.91           | 0.88            | 0.88             | 0.91             |
| LPO            | 0.95           | 0.94           | 0.96            | 0.97             | 0.97             | 0.92             |

that there is at least one Core Argo profile collected within 30 days and no further than 2 degrees in longitude or latitude away. This distance yields \(\approx 65\%\) of the observations having nearby Core data. Unsurprisingly, the difference is particularly pronounced in the leave-float-out setting where there may be few BGC profiles nearby. However, even in the leave-profile-out setting there is marginal improvement.

4.3. Data Product: Gridded predictions

We now present overall prediction results using the entire BGC data onto a grid in the Southern Ocean. As an example in Figure 8, we plot oxygen predictions at 300 decibars for four different months encompassing the seasons. The clustering approach results in relatively hard boundaries in the oxygen predictions. By plotting four months of the year, we demonstrate that the Markov random field and predicted functions model some seasonality in the data, though this is less pronounced at 300 decibars compared to the
Fig. 8. (Left) Oxygen prediction $\hat{Y}_s(p = 300$ decibars) and various $t$. (Right) Their standard deviations. We also plot in white lines traditional estimates of fronts using the same criteria as Bushinsky et al. (2017) based on the 2004-2018 Roemmich and Gilson (2009) estimate.

surface compared to the surface. We additionally plot the oxygen standard deviations for the same level and months. At 300 decibars, the prediction standard deviations are generally higher at and North of the ACC. There are also smaller areas where the prediction standard deviation is lower, representing areas where a float collected data nearby in space and time.

In Figure 9, we decompose the variance into the two terms of (10) for one month. In general, the second term is smaller, and captures variability due to uncertainty in the cluster membership with elevated levels around fronts, while the first term captures variability to the uncertainty in the principal component scores that is decreased when floats are nearby. By comparing the total variance before and after prediction on the right panel of Figure 9, we provide an estimate of an R-squared statistic per location, which is high in most areas.

A principal advantage of our approach is the ability to provide functional predictions. We plot an example in Figure 10 for locations with the same longitude and varying
Fig. 9. (Left) Model-based uncertainty of oxygen based on the scores in term (10) for April 15th, 2020 at 300 decibars. (Middle) Model-uncertainty based on the clustering in term (10). The combination of this uncertainty represents the overall uncertainty in the oxygen estimates (that is, the right panel of Figure 8. (Right) Proportion of oxygen variance explained by the covariates and spatial correlation in prediction, April 15th, 2020.

latitude. As suggested by the previous plots, we can identify which cluster contributes the most to each of the locations for the most part, with few functions lying in between the clustered groups. We also provide functional standard deviation estimates, which have different shape depending on latitude. Oxygen standard deviation is highest in the Southern-most sections of the Southern Ocean around 125 decibars.

5. Discussion

We have developed and implemented a comprehensive space-time functional regression mixture model for temperature, salinity, and oxygen. The model yields a functional cokriging framework, where the oxygen concentration as a function of depth and the season can be estimated over the entire Southern Ocean along with (functional) uncertainties. The model has been validated against the most accurate existing non-functional approaches, and the mixture component of the model identifies contrasting distributions of the measured variables.

While our analysis introduces a comprehensive study of the dependence structure and prediction of oxygen in the Southern Ocean, there are a number of areas where
A functional regression model for heterogeneous BioGeoChemical Argo data

Fig. 10. Estimated oxygen (left) and oxygen standard deviation (right) for longitude -90° for varying latitude and pressure. As one moves from south (dark) to north (light), the predicted oxygen and their uncertainties take different shapes.

the richness and complexity of the Argo data calls for improvements. As the principal components are entirely based on the pressure domain, predictions may be better if the components considered spatial information as well (cf., Kuenzer et al., 2020). A non-functional prediction approach, as done in Salvana and Jun (2022), might also fare well, but requires more computing than readily available at the scale of the entire Southern Ocean. Our model also does not take into account the monotonicity of density or the soft constraint on the saturation of oxygen near the ocean surface, which could improve predictions; however, applying standard functional data analysis approaches were not successful due to the limited domain in the density dimension for some profiles. Addressing this problem with the recent research of (for example) Lin and Wang (2022) may be possible. The current model is also developed with independence assumed between different clusters; incorporating dependence between clusters may better represent the available data. This, as well as evaluating whether independence should be assumed for different principal components the principal component scores may be assumed independent (as in Liang et al., 2023) would more comprehensively justify the data analysis. Including scalar covariates like satellite data in the model is straightforward and may be of practical interest. Finally, using more sophisticated model selection for the Markov random field might improve the cluster predictions.
Statistical approaches like the one developed here will play an important role for the analysis of data from the planned global deployment of BGC Argo floats (see NSF Award 1946578).

**Data Availability Statement**

All data used in this work is publicly available. The Argo data is available at Argo (2020), and we use the SOCCOM-specific data files from Riser et al. (2023). For plotting of traditional estimates of fronts, we use the sea ice data Fetterer et al. (2017) and the Roemmich-Gilson mean Roemmich and Gilson (2009). Finally, we use the grid for prediction from the Southern Ocean State Estimate (Verdy and Mazloff, 2017). Code to produce results presented in the paper are available at [https://github.com/dyarger/argo_fstmr_oxygen/](https://github.com/dyarger/argo_fstmr_oxygen/).

**Acknowledgements**

Data were collected and made freely available by the Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) Project funded by the National Science Foundation, Division of Polar Programs (NSF PLR-1425989 and OPP-1936222), supplemented by NASA, and by the International Argo Program and the NOAA programs that contribute to it. ([https://argo.ucsd.edu](https://argo.ucsd.edu), [https://www.ocean-ops.org](https://www.ocean-ops.org)). The Argo Program is part of the Global Ocean Observing System. Computational resources for the SOSE (Verdy and Mazloff, 2017) were provided by NSF XSEDE resource grant OCE130007 and SOCCOM NSF award PLR-1425989. This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor. Authors Korte-Stapff, Stoev, and Hsing are supported for this work by NSF Grant DMS-1916226. Author Yarger is supported for this work by NSF Grant DGE-1841052.
References

Argo (2020) Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). URL: https://doi.org/10.17882/42182#85023.

Besag, J. (1975) Statistical analysis of non-lattice data. *J. R. Statist. Soc. Series D (The Statistician)*, 24, 179–195.

Bittig, H. C., Maurer, T. L., Plant, J. N., Schmechtig, C., Wong, A. P. S., Claustre, H., ... and Xing, X. (2019) A BGC-Argo guide: planning, deployment, data handling and usage. *Frontiers in Marine Science*, 6.

Boehme, L. and Rosso, I. (2021) Classifying oceanographic structures in the Amundsen Sea, Antarctica. *Geophysical Research Letters*, 48, e2020GL089412.

Bushinsky, S. M., Emerson, S. R., Riser, S. C. and Swift, D. D. (2016) Accurate oxygen measurements on modified Argo floats using in situ air calibrations: In situ Argo float air O$_2$ calibration. *Limnology and Oceanography: Methods*, 14, 491–505.

Bushinsky, S. M., Gray, A. R., Johnson, K. S. and Sarmiento, J. L. (2017) Oxygen in the Southern Ocean from Argo floats: determination of processes driving air-sea fluxes. *Journal of Geophysical Research: Oceans*, 122, 8661–8682.

Chamberlain, P. M., Talley, L. D., Mazloff, M. R., Riser, S. C., Speer, K., Gray, A. R. and Schwartzman, A. (2018) Observing the ice-covered Weddell Gyre with profiling floats: Position uncertainties and correlation statistics. *Journal of Geophysical Research: Oceans*, 123, 8383–8410.

Chapman, C. C., Lea, M.-A., Meyer, A., Sallée, J.-B. and Hindell, M. (2020) Defining Southern Ocean fronts and their influence on biological and physical processes in a changing climate. *Nature Climate Change*, 10, 209–219.

Choi, H. and Reimherr, M. (2018) A geometric approach to confidence regions and bands for functional parameters. *J. R. Statist. Soc. B*, 80, 239–260.

Claustre, H., Johnson, K. S. and Takeshita, Y. (2020) Observing the global ocean with Biogeochemical-Argo. *Annual Review of Marine Science*, 12, 23–48.
Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39, 1–22.

Fetterer, F., Knowles, K., Meier, W., Savoie, M. and Windnagel, A. (2017) Sea Ice Index, Version 3. URL: https://nsidc.org/data/G02135/versions/3.

Fonvieille, N., Guinet, C., Saraceno, M., Picard, B., Tournier, M., Goulet, P., Campagna, C., Campagna, J. and Nerini, D. (2023) Swimming in an ocean of curves: A functional approach to understanding elephant seal habitat use in the Argentine Basin. *Progress in Oceanography*, 218, 103120.

Giglio, D., Lyubchich, V. and Mazloff, M. R. (2018) Estimating oxygen in the Southern Ocean using Argo temperature and salinity. *Journal of Geophysical Research: Oceans*, 123, 4280–4297.

Gray, A. R., Johnson, K. S., Bushinsky, S. M., Riser, S. C., Russell, J. L., Talley, L. D., Wanninkhof, R., Williams, N. L. and Sarmiento, J. L. (2018) Autonomous biogeochemical floats detect significant carbon dioxide outgassing in the high-latitude Southern Ocean. *Geophysical Research Letters*, 45, 9049–9057.

Gray, A. R. and Riser, S. C. (2014) A global analysis of Sverdrup balance using absolute geostrophic velocities from Argo. *Journal of Physical Oceanography*, 44, 1213–1229.

Guinness, J. (2021) Gaussian process learning via Fisher scoring of Vecchia’s approximation. *Statistics and Computing*, 31, 25.

Guinness, J., Katzfuss, M. and Fahmy, Y. (2021) *Fast Gaussian Process Computation Using Vecchia’s Approximation*. URL: https://cran.r-project.org/web/packages/GpGp/index.html. R package version 0.3.2.

Holte, J., Talley, L. D., Gilson, J. and Roemmich, D. (2017) An Argo mixed layer climatology and database. *Geophysical Research Letters*, 44, 5618–5626.

Houghton, I. A. and Wilson, J. D. (2020) El Niño detection via unsupervised clustering of Argo temperature profiles. *Journal of Geophysical Research: Oceans*, 125.
Hsing, T. and Eubank, R. (2015) *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. John Wiley and Sons, Inc.

Huang, H., Li, Y. and Guan, Y. (2014) Joint modeling and clustering paired generalized longitudinal trajectories with application to cocaine abuse treatment data. *J. Am. Statist. Ass.*, 109, 1412–1424.

Ibrahim, J. G., Zhu, H. and Tang, N. (2008) Model selection criteria for missing-data problems using the EM algorithm. *J. Am. Statist. Ass.*, 103, 1648–1658.

Jiang, H. and Serban, N. (2012) Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics*, 54, 108–119.

Johnson, G. C., Hosoda, S., Jayne, S. R., Oke, P. R., Riser, S. C., Roemmich, D., Suga, T., Thierry, V., Wijffels, S. E. and Xu, J. (2022) Argo—two decades: global oceanography, revolutionized. *Annual Review of Marine Science*, 14, 379–403.

Jones, D. C., Holt, H. J., Meijers, A. J. S. and Shuckburgh, E. (2019) Unsupervised clustering of Southern Ocean Argo float temperature profiles. *Journal of Geophysical Research: Oceans*, 124, 390–402.

Jones, D. C., Sonnewald, M., Zhou, S., Hausmann, U., Meijers, A. J., Rosso, I., Boehme, L., Meredith, M. P. and Naveira Garabato, A. C. (2023) Unsupervised classification identifies coherent thermohaline structures in the Weddell Gyre region. *Ocean Science*, 19, 857–885.

Katzfuss, M., Guinness, J., Gong, W. and Zilber, D. (2020) Vecchia approximations of Gaussian-Process predictions. *Journal of Agricultural, Biological and Environmental Statistics*, 25, 383–414.

Kuenzer, T., Hörmann, S. and Kokoszka, P. (2020) Principal component analysis of spatially indexed functions. *J. Am. Statist. Ass.*, 1–13.

Kuusela, M. and Stein, M. L. (2018) Locally stationary spatio-temporal interpolation of Argo profiling float data. *Proc. R. Soc. A: Mathematical, Physical and Engineering Sciences*, 474.
Liang, D., Huang, H., Guan, Y. and Yao, F. (2023) Test of weak separability for spatially stationary functional field. *J. Am. Statist. Ass.*, **118**, 1606–1619.

Liang, D., Zhang, H., Chang, X. and Huang, H. (2020) Modeling and regionalization of China’s PM$_{2.5}$ using spatial-functional mixture models. *J. Am. Statist. Ass.*, 1–17.

Lin, Z. and Wang, J.-L. (2022) Mean and covariance estimation for functional snippets. *J. Am. Statist. Ass.*, **117**, 348–360.

Maurer, T. L., Plant, J. N. and Johnson, K. S. (2021) Delayed-mode quality control of oxygen, nitrate, and pH data on SOCCOM biogeochemical profiling floats. *Frontiers in Marine Science*, **8**, 1118.

Maze, G., Mercier, H., Fablet, R., Tandeo, P., Lopez Racencio, M., Lenca, P., Feucher, C. and Le Goff, C. (2017) Coherent heat patterns revealed by unsupervised classification of Argo temperature profiles in the North Atlantic Ocean. *Progress in Oceanography*, **151**, 275–292.

Mignot, A., D’Ortenzio, F., Taillandier, V., Cossarini, G. and Salon, S. (2019) Quantifying observational errors in Biogeochemical-Argo oxygen, nitrate, and chlorophyll a concentrations. *Geophysical Research Letters*, **46**, 4330–4337.

NASA (2024) Nasa spaceplace. https://spaceplace.nasa.gov/antarctica/en/. Accessed: 2024-01-26.

Orsi, A. H., Whitworth, T. and Nowlin, W. D. (1995) On the meridional extent and fronts of the Antarctic Circumpolar Current. *Deep Sea Research Part I: Oceanographic Research Papers*, **42**, 641–673.

Park, B., Kuusela, M., Giglio, D. and Gray, A. (2023) Spatiotemporal local interpolation of global ocean heat transport using Argo floats: a debiased latent Gaussian process approach. *Ann. Appl. Statist.*, **17**, 1491–1520.

Poropat, L., Jones, D., Thomas, S. D. and Heuzé, C. (2023) Unsupervised classification of the Northwestern European seas based on satellite altimetry data. *EGUsphere*, **2023**, 1–20.
A functional regression model for heterogeneous BioGeoChemical Argo data

Riser, S. C., Talley, L. D., Wijffels, S. E., Nicholson, D., Purkey, S., Takeshita, Y., Fassbender, A., Gray, A., Robbins, P., Gilson, J., Plant, J. N., Clark, E., Swift, D. D., Rupan, R. A., Maurer, T. L. and Johnson, K. S. (2023) SOCCOM and GO-BGC float data - Snapshot 2023-08-28. In Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) and Global Ocean Biogeochemistry (GO-BGC) Biogeochemical-Argo Float Data Archive. URL: https://library.ucsd.edu/dc/object/bb9362707g.

Robert, C. P. and Casella, G. (2004) Monte Carlo Statistical Methods. Springer Texts in Statistics. New York, NY: Springer.

Roemmich, D. and Gilson, J. (2009) The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo Program. Progress in Oceanography, 82, 81–100.

Rosso, I., Mazloff, M. R., Talley, L. D., Purkey, S. G., Freeman, N. M. and Maze, G. (2020) Water mass and biogeochemical variability in the Kerguelen Sector of the Southern Ocean. Journal of Geophysical Research: Oceans, 125.

Salvana, M. L. and Jun, M. (2022) 3D bivariate spatial modelling of Argo ocean temperature and salinity profiles. arXiv preprint arXiv:2210.11611.

Stein, M. L. (2013) Interpolation of Spatial Data: Some Theory for Kriging. Springer.

Stramma, L. and Schmidtko, S. (2021) Spatial and temporal variability of oceanic oxygen changes and underlying trends. Atmosphere-Ocean, 59, 122–132.

Talley, L. D. (2011) Descriptive Physical Oceanography: An Introduction. Academic press.

Verdy, A. and Mazloff, M. R. (2017) A data assimilating model for estimating Southern Ocean biogeochemistry. Journal of Geophysical Research: Oceans, 122, 6968–6988.

Wahba, G. (1990) Spline Models for Observational Data. Society for Industrial and Applied Mathematics.
Wei, G. C. G. and Tanner, M. A. (1990) A Monte Carlo implementation of the EM algorithm and the Poor Man’s data augmentation algorithms. *J. Am. Statist. Ass.*, **85**, 699–704.

Wei, J. and Zhou, L. (2010) Model selection using modified AIC and BIC in joint modeling of paired functional data. *Statistics & Probability Letters*, **80**, 1918–1924.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005) Functional linear regression analysis for longitudinal data. *Ann. Statist.*, **33**, 2873–2903.

Yarger, D., Stoev, S. and Hsing, T. (2022) A functional-data approach to the Argo data. *Ann. Appl. Statist.*, **16**, 216–246.

Zhou, L., Huang, J. and Carroll, R. (2008) Joint modeling of paired sparse functional data using principal components. *Biometrika*, **95**, 601–619.

Zhou, L., Huang, J. Z., Martinez, J. G., Maity, A., Baladandayuthapani, V. and Carroll, R. J. (2010) Reduced rank mixed effects models for spatially correlated hierarchical functional data. *J. Am. Statist. Ass.*, **105**, 390–400.
Supplemental Material to “Spatial Prediction and Clustering of Oceanographic Oxygen in the Southern Ocean using Functional Regression”

Moritz Korte-Stapff, Drew Yarger, Stilian Stoev, Tailen Hsing

September 2, 2024

Contents

S1 Data Description ......................................................... 1
S2 Additional Data Analysis Plots ....................................... 2
S3 Comparison with Jones et al. (2019) .................................. 2
S4 Details on Importance Sampling ....................................... 5
S5 Conditional Distributions ............................................... 6
S6 Description of Prediction Formulas .................................... 8
S7 Computation of Conditional Distributions ......................... 9
S8 Maximization Step Updates ............................................ 11
S9 Avoiding the Use of Orthogonal B-splines ......................... 13
S10 Details for the Initialization .......................................... 13
S11 Clustering Accuracy Simulation Study ............................. 14

S1 Data Description

We describe the data used in the analysis here. We use:

• the August 28, 2023 high-resolution MLR snapshot of the SOCCOM data:
  https://doi.org/10.6075/J0542NS9 (Riser et al., 2023),

• the June 6, 2021 snapshot of the Argo data for nearby core data:
  http://doi.org/10.17882/42182#85023 (Argo, 2020).

We remove observations with pressure less than 0 decibar or greater than 2,000 decibars. For the SOCCOM data, we remove measurements with: a) latitude greater than −25 degrees, b) oxygen quality control denotes “bad” (8) c) temperature, salinity, or pressure quality control is not good (0) d) missing pressure or location data (this keeps linearly interpolated locations while under ice part of the data).

Following these reductions, we remove profiles with:

• any measurements of salinity less than 33 PSU or greater than 37.2 PSU

• a pressure gap of 200 decibars or more between consecutive measurements

• a minimum pressure greater than 100 decibars or a maximum pressure less than 1,450 decibars
• fewer than 15 measurements
• had a potential density estimated derivative less than $-0.005 \text{ kg/L/decibar}$ for either of its core or BGC measurements
• there are fewer than five oxygen measurements with good (0) or probably good (4) quality control
• an oxygen measurement greater than $400 \mu\text{mol/kg}$
• any pressure, temperature, and salinity measurements without good (0) quality control flag.

For Core Argo data with temperature and salinity only, we make similar, but not identical, exclusions for profiles (note that the Argo and SOCCOM data have different quality flag meanings):
• any measurements of salinity less than 33 PSU or greater than 37.2 PSU
• a pressure gap of 200 decibars or more between consecutive measurements
• a minimum pressure greater than 100 decibars or a maximum pressure less than 1,450 decibars
• temperature, salinity, or pressure quality control is not good (3, 4)
• missing pressure or location data (this keeps linearly interpolated locations while under ice part of the data)
• the pressure adjusted error was greater than 16 decibars
• the closest SOCCOM profile was more than 500 kilometers away

S2 Additional Data Analysis Plots

We show additional data analysis plots of interest in a variety of figures. In Figure S1, we demonstrate the EM algorithm reduces the measurement error variances for the three variables, suggesting that the algorithm improves the fit to the data. We also run the model with the number of principal component functions for the response and predictors each set to the same levels: 12, 14, 16, and 20. The resulting AIC values are plotted in Figure S2. In Figures S3 and S4, we plot predictions and predictive variances for temperature and salinity, respectively, which are obtained naturally from our analysis of temperature, salinity, and oxygen.

S3 Comparison with Jones et al. (2019)

We also compare with Jones et al. (2019)’s analysis of 6 clusters, who specifically focus on temperature data in the Southern Ocean. Our January results for the first three clusters match well spatially with the first three clusters of Jones et al. (2019). For our fourth and fifth cluster, Jones et al. (2019) has three clusters that are more spatially divided into the Atlantic, Indian, and Pacific sectors. Our clusters are also determined by salinity and oxygen, which may result in less regional separation compared to when using temperature only. In looking at our seasonality plot, there are areas near Antarctica that are classified in cluster 2 instead of cluster 1 for some months; Jones et al. (2019) also identifies a number of profiles in the same areas with an outlying cluster. Our results also demonstrate substantial space-time variability in the clusters, which Jones et al. (2019) does not analyze extensively.
Figure S1: Estimated measurement error variables by Monte Carlo iteration for oxygen, salinity, and temperature.

Figure S2: Akaike information criterion (AIC) for different number or principal component functions.
Figure S3: Temperature (left) and salinity (right) predictions for various $t$ at 300 decibar. We plot in white lines traditional estimates of fronts using the same criteria for fronts as Bushinsky et al. (2017) based on the 2004-2018 Roemmich & Gilson (2009) estimate.

Figure S4: Temperature (left) and salinity (right) predicted standard deviations for various $t$ at 300 decibar. We plot in white lines traditional estimates of fronts using the same criteria for fronts as Bushinsky et al. (2017) based on the 2004-2018 Roemmich & Gilson (2009) estimate.
S4 Details on Importance Sampling

Monte-Carlo E-step. Our goal is to construct a proposal \( h \) that is easy to sample from and that is close to the true conditional distribution \( f(\alpha, \eta, Z|X, Y) \). To do so, the key ingredient is that sampling from \( f(\alpha, \eta|Z = z^{(r)}, X, Y) \) is tractable; see Proposition S5.1. It then only remains to specify a proposal distribution \( P_I \) for sampling the latent process of cluster memberships \( Z|X, Y \) and employ the following procedure:

1. Sample \( z^{(r)} \) from \( P_I \) and compute \( w^{(r)} \).
2. Sample \( \alpha^{(r)}, \eta^{(r)} \) from \( f(\alpha, \eta|Z = z^{(r)}, X, Y) \).

We now construct \( P_I \). By Bayes’ Theorem \( P(Z = z|X, Y) \propto f(X, Y|Z = z)P(Z = z) \). Now replace \( f(X, Y|Z = z) \) with \( f_I(X, Y|Z = z) := \prod_{i=1}^{n} f(X_i, Y_i|Z_i = z_i) \), and set

\[
P_I(Z = z|X, Y) \propto f_I(X, Y|Z = z)P(Z = z)
\]

\[
\propto \exp \left( \sum_{i=1}^{n} \log(f(X_i, Y_i|Z_i = z_i)) + \xi \sum_{j_1 \sim j_2} \mathbb{I}(z_{j_1} = z_{j_2}) \right), \tag{S4.1}
\]

where \( z = \{z_i\}_{i=1}^{n} \) and \( i \sim j \) indicates that \( i \) and \( j \) are neighbors so that the second sum is over all possible neighboring pairs \( j_1 \) and \( j_2 \). Sampling from (S4.1), known as the Potts model, has been studied extensively, see for example Feng et al. (2012) and references therein. We use a Gibbs sampler based on a single site updating scheme similar to Feng et al. (2012) to sample from \( P_I(Z|X, Y) \). With this, the final proposal distribution is \( h(\alpha, \eta, Z) = P_I(Z = z|X, Y)f(\alpha, \eta|Z = z, X, Y) \).

Utilizing the fact that we can sample from \( f(\alpha, \eta|Z = z, X, Y) \), the importance sampling weights
can be computed via Bayes’ Theorem

\[ w^{(\tau)} = \frac{P(Z = z^{(\tau)}|X, Y)}{P_I(Z = z^{(\tau)}|X, Y)} = C \frac{f(X, Y|Z = z^{(\tau)})}{\prod_{i=1}^n f(X_i, Y_i|Z_i = z_i^{(\tau)})}. \]

The unknown constant \( C \) cancels when self-normalizing the weights.

**Model Selection.** For model selection, we need a different proposal because the one used in the E-step can be sampled from but cannot easily be evaluated. Again we only need a proposal for the latent field \( Z \) given the observed data. Set

\[ \mathbb{P}_{I_2}(Z = z|X, Y) = \prod_{i=1}^n \frac{f(X_i, Y_i|Z_i = z_i)}{f(X_i, Y_i)} \] (S4.2)

which corresponds to a Gaussian mixture model with independent observations. For any fixed \( z \), (S4.2) can easily be evaluated exactly. Using Bayes’ Theorem, we propose to approximate the observed data likelihood based on the samples \( z^{(1)}, \ldots, z^{(T)} \) from \( \mathbb{P}_{I_2} \) via

\[ f(X, Y) = \sum_{z \in \mathcal{Z}} f(X, Y|Z = z) \mathbb{P}(Z = z) \approx \frac{1}{T} \sum_{\tau=1}^T f(X, Y|Z = z^{(\tau)}) \mathbb{P}(Z = z^{(\tau)}) \mathbb{P}_{I_2}(Z = z^{(\tau)}|X, Y) \]

Every term on the right hand side can be evaluated analytically with the exception of \( \mathbb{P}(Z = z^{(\tau)}) \) which can only be evaluated up to a constant. However, this constant is constant across Monte-Carlo samples and hyper parameters to be selected and thus cancels when comparing the AIC criterion for different hyperparameters.

**S5 Conditional Distributions**

Here, we provide the conditional distributions required for our MCEM algorithm. The first proposition gives the conditional distribution of the observed data given the latent field \( \{Z(s_i)\}_{i=1}^n \) in various scenarios. In the following, for arbitrary matrices \( A_1, \ldots, A_M \), let \( \text{bdiag}(A_1, \ldots, A_M) \) denote the block diagonal matrix with the matrices \( A_1, \ldots, A_M \) on the diagonal.
Proposition S5.1. Let $\alpha$ be a normally distributed random vector of size $m_1$ with mean 0 and covariance $\Sigma_\alpha$. Let $\eta$ be another normally distributed random vector of size $m_2$ with mean 0 and covariance $\Sigma_\eta$. Furthermore, let $\epsilon_X$ and $\epsilon_Y$ be two random vectors with

$$
\epsilon_Y \sim N(0, \sigma_Y^2 I_{N_2}) \quad \text{and} \quad \epsilon_X \sim N\left(0, \text{bdiag}(\sigma_X^2 I_{n_1}, \ldots, \sigma_X^2 I_{n_K})\right) =: N(0, D_X),
$$

where the parameters $\sigma_Y^2 > 0$ and $\sigma_X^2 > 0$ for $k = 1, \ldots, K$ are measurement error variances. Assume that $\alpha, \eta, \epsilon_X$ and $\epsilon_Y$ are mutually independent of each other. Let $N_1 = \sum_{k=1}^K n_k$ and assume $\min(N_1, N_2) > \max(m_1, m_2)$. Furthermore, let $A_1, A_2, A_3$ be three matrices of dimensions $N_1 \times m_1$, $N_2 \times m_1$ and $N_2 \times m_2$ respectively. Define

$$
X = A_1 \alpha + \epsilon_X, \quad Y = A_2 \alpha + A_3 \eta + \epsilon_Y.
$$

Then

$$
\begin{pmatrix}
X \\
Y
\end{pmatrix} \sim N\left(0, \begin{pmatrix}
D_X + A_1 \Sigma_\alpha A_1^\top & A_1 \Sigma_\alpha A_2^\top \\
A_2 \Sigma_\alpha A_1^\top & A_2 \Sigma_\alpha A_2^\top + A_3 \Sigma_\eta A_3^\top
\end{pmatrix}
\right).
$$

Proof. This follows directly from standard properties of the multivariate normal distribution. \qed

Next, we derive the conditional density $f(\alpha, \eta|X, Y, Z)$. As observations from different clusters are assumed to be independent, it suffices to derive the distribution for specific cluster. Thus to ease notation, we drop cluster dependence.

Proposition S5.2. Let $B_i$ be the $n_i \times P$ matrix of basis evaluations for profile $i$ for the response. Similarly, let $B_i$ be the respective $K n_i \times K P$ be the respective matrix for the covariates, which is the block-diagonal matrix of evaluations for each predictor. Write $\tilde{B} = \text{bdiag}(B_1, \ldots, B_n)$ and $B = \text{bdiag}(B_1, \ldots, B_n)$. For any matrix or vector $\Theta$ with appropriate dimension, write

$$
B \Theta = \text{bdiag}(B_1 \Theta, \ldots, B_n \Theta) \quad \text{and} \quad \tilde{B} \Theta = \text{bdiag}(B_1 \Theta, \ldots, B_n \Theta)
$$

as well as $D_X = \text{bdiag}(D_{X_1}, \ldots, D_{X_n})$ where $D_{X_i} = \text{Var} \left( [\epsilon_{1i}, \ldots, \epsilon_{m_i}, \ldots, \epsilon_{1i}, \ldots, \epsilon_{m_i}] \right)$. For the response, let $\sigma_Y^2 = \text{Var}(\epsilon_{11})$. In addition, let $\Sigma_\alpha = \text{Var}(\alpha)$ and $\Sigma_\eta = \text{Var}(\eta)$ be defined by the spatial covariance. Let

$$
A = \begin{pmatrix} \Sigma_\alpha & 0 \\ 0 & \Sigma_\eta \end{pmatrix} \quad \text{and} \quad U = \begin{pmatrix} D_X^{-1/2} B \Theta X & 0 \\ \sigma_Y^{-1} \tilde{B} \Lambda & \sigma_Y^{-1} \tilde{B} \Theta \varepsilon \end{pmatrix}
$$

Within the same cluster, we have

$$
\alpha, \eta|X, Y \sim N\left( (A^{-1} + U^\top U)^{-1} \begin{pmatrix} X - B \Upsilon_X \\ Y - B \Upsilon_Y \end{pmatrix} , (A^{-1} + U^\top U)^{-1} \right)
$$

$$
= N\left( (A^{-1} + U^\top U)^{-1} \begin{pmatrix} X_0 \\ Y_0 \end{pmatrix} , (A^{-1} + U^\top U)^{-1} \right)
$$

Here, $\Theta_\varepsilon$, $\Lambda$, and $\Theta_X$ are the respective matrices for the principal component/random effect function coefficients, $\Upsilon_X = [\Upsilon_{X_{r1}}]_{r=1}^{2R+1}$ and $\Upsilon_Y = [\Upsilon_{Y_{r1}}]_{r=1}^{2R+1}$ are the respective vectors for the mean coefficients, ignoring notation on the cluster, and $B_t = \text{bdiag}(B_1 \otimes \delta(t_1), \ldots, B_n \otimes \delta(t_n))$, with $\otimes$ denoting the Kronecker product.
\[ \tilde{B}_t = \text{bdiag}(B_1 \otimes \delta(t_1), \ldots, B_n \otimes \delta(t_n)), \text{ where} \]
\[ \delta(t) = (1, \sin(2\pi t/365.25), \ldots, \sin(2\pi R_t/365.25), \cos(2\pi t/365.25), \ldots, \cos(2\pi R_t/365.25)) \quad (S5.1) \]

is the row vector of covariates for the mean functions.

### S6 Description of Prediction Formulas

As in other parts of the supplement, we ignore the dependence on clusters in the following developments. In Proposition S5.1, we found that, in each cluster,

\[ \alpha, \eta | X, Y \sim \mathcal{N} \left( (A^{-1} + U^\top U)^{-1} U^\top \left( \begin{array}{c} X_0 \\ Y_0 \end{array} \right), (A^{-1} + U^\top U)^{-1} \right) \quad (S6.1) \]

\[ = \mathcal{N} \left( \Sigma U^\top \left( \begin{array}{c} X_0 \\ Y_0 \end{array} \right), \Sigma \right) \quad (S6.2) \]

where \( X_0 \) and \( Y_0 \) are centered versions of the data, \( A \) and \( U \) are defined as in Proposition S5.2.

We now consider new scores \( \tilde{\alpha} \) and \( \tilde{\eta} \) where we want to predict, potentially with additional predictor data \( \tilde{X} \). For simplicity, suppose that \( \tilde{X} \) is available; however, we discuss later how to form predictions when \( \tilde{X} \) is not available. Consider the matrix

\[
\tilde{A} = \begin{pmatrix}
\Sigma_{\alpha} & \Sigma_{\alpha,\tilde{\alpha}} & 0 & 0 \\
\Sigma_{\alpha,\tilde{\alpha}}^\top & \Sigma_{\tilde{\alpha}} & 0 & 0 \\
0 & 0 & \Sigma_{\eta} & \Sigma_{\eta,\tilde{\eta}} \\
0 & 0 & \Sigma_{\eta,\tilde{\eta}}^\top & \Sigma_{\tilde{\eta}}
\end{pmatrix}
\]

where \( \Sigma_{\alpha,\tilde{\alpha}} = \text{Cov}(\alpha, \tilde{\alpha}), \Sigma_{\tilde{\alpha}} = \text{Var}(\tilde{\alpha}) \), and likewise for the other matrices involving \( \eta \). Also, note that \( \tilde{A}^{-1} \) can be formed using a Vecchia approximation for the entire matrix

\[
\begin{pmatrix}
\Sigma_{\alpha} & \Sigma_{\alpha,\tilde{\alpha}} \\
\Sigma_{\alpha,\tilde{\alpha}}^\top & \Sigma_{\tilde{\alpha}}
\end{pmatrix}^{-1}
\]

and likewise for \( \eta \) and \( \tilde{\eta} \). Recall the definition of \( U \) in Proposition S5.2, and similarly define the matrices

\[ U_{X,\alpha} = D_X^{-1/2} B \Theta_X, \quad U_{Y,\alpha} = \sigma_Y^{-1} \tilde{B} \Lambda, \quad U_{Y,\eta} = \sigma_Y^{-1} \tilde{B} \Theta_Y \]

Furthermore, let

\[ U_{\text{all}} = \begin{pmatrix}
U_{X,\alpha} & 0 & 0 & 0 \\
U_{Y,\alpha} & 0 & U_{Y,\eta} & 0 \\
0 & U_{X,\tilde{\alpha}} & 0 & 0
\end{pmatrix}
\]

where \( U_{X,\tilde{\alpha}} = D_X^{-1/2} B \tilde{X} \Theta_X \) is the respective matrix for the new data. The three groups of rows correspond to \( X \) and \( Y \), and \( \tilde{X} \), respectively, while the four groups of columns correspond to \( \alpha \), \( \tilde{\alpha} \), \( \eta \), and \( \tilde{\eta} \).
Proposition S6.1. In this case, the conditional distribution of the scores can be written:

$$\alpha, \hat{\alpha}, \eta, \hat{\eta} \mid X, Y, \bar{X} \sim \mathcal{N}\left(\left(\bar{A}^{-1} + U_{\text{all}}^T U_{\text{all}}\right)^{-1} U_{\text{all}}^T\begin{pmatrix} X_0 \\ Y_0 \end{pmatrix}, \left(\bar{A}^{-1} + U_{\text{all}}^T U_{\text{all}}\right)^{-1}\right)$$

where $X_0, Y_0,$ and $\bar{X}_0$ are centered versions of the data.

The proof is similar to that of Proposition S5.2. Likewise, when $\bar{X}$ is not available, we form $U_{\text{all}}$ with one group of rows deleted:

$$U_{\text{all}} = \begin{pmatrix} U_{X,\alpha} & 0 & 0 & 0 \\ U_{Y,\alpha} & 0 & U_{Y,\eta} & 0 \end{pmatrix}$$

and compute the same formulas with only $X$ and $Y$ as the available data.

S7 Computation of Conditional Distributions

Consider the notation in Section S5, and note that our goal is to sample from $f(\alpha, \eta \mid X, Y)$ efficiently. All distributions should be considered conditional on $Z$, but this notation is omitted as in Section S5. Write again

$$\Sigma^{-1} = A^{-1} + U^T U.$$

In practice, $\Sigma^{-1}$ will be considerably sparse due to the independence assumption on the scores and because we apply Vecchia’s approximation (Guinness, 2021) resulting in a sparse matrix $A^{-1}$. Note that $U^T U$ is (essentially) a block diagonal matrix. However, the actual covariance matrix $\Sigma$ is not sparse. We next illustrate how to use sparseness of $\Sigma^{-1}$ instead.

1. Instead of computing $\Sigma$ explicitly, we compute the conditional expectation in (S6.2) using the conjugate Gradient solver as implemented in the ConjugateGradient class of the Eigen C++ library (Guennebaud et al., 2010). This takes explicit advantage of the sparsity of $\Sigma^{-1}$. If $(\hat{\mu}_\alpha^t, \hat{\mu}_\eta^t)^T$ is output of running the conjugate Gradient solver after $t$ iterations, then we stop when

$$\left\| \Sigma^{-1} \begin{pmatrix} \hat{\mu}_\alpha^t \\ \hat{\mu}_\eta^t \end{pmatrix} - U^T \begin{pmatrix} X_0 \\ Y_0 \end{pmatrix} \right\| \leq 10^{-18} \left\| U^T \begin{pmatrix} X_0 \\ Y_0 \end{pmatrix} \right\| \approx 10^{-12},$$

where the final approximation varies from EM step to EM step but only very slightly.

2. To generate a mean-zero random normal vector with covariance $\Sigma$, we use Lanczos quadrature (cf. Chapter 10; Golub & Van Loan, 2013). That is, we instantiate a standard normal random vector $\tilde{W}$ and then normalize it to $W = \tilde{W} / \|\tilde{W}\|$. Then, we compute a tridiagonal decomposition

$$\Sigma^{-1} = VTV^\top$$

where $V = \text{nrow}(\Sigma^{-1}) \times K$ is orthogonal with the first column being equal to $W$ and $T$ is tridiagonal $K \times K$ matrix. We then compute

$$\Sigma^{1/2} W \approx VT^{-1/2} V^\top W = VT^{1/2} \epsilon_1.$$
where $e_1$ is a vector with 1 for the first entry and zeros everywhere else. Here $K$ needs to be chosen sufficiently large, (note that if $K = \text{nrow}(\Sigma^{-1})$, then the procedure is exact). We stop if

$$\|W^T - W^{T-10}\|_\infty \leq 10^{-6}.$$ 

3. We need to compute the marginal likelihood of $X$ and $Y$ (once again conditional on cluster memberships) for the importance sampling weights and for model selection. Write

$$D_{X,Y} = \text{diag}\left(D_{X}^{1/2}, \sigma_Y I_{\sum_{i=1}^{n_i}}\right),$$

$$\Sigma_{\text{data}} = \text{Var}\left(\begin{bmatrix} X \\ Y \end{bmatrix} | \Omega \right) = D_{X,Y} U A U^\top D_{X,Y} + D_{X,Y}^2.$$ 

Finally, write $n_{\text{data}} = (K + 1) \sum_{i=1}^{n_i}$ as the number of total observed measurements. Now, let $X_0$ and $Y_0$ be mean-centered versions of the data, so that

$$\ell(\Omega, X, Y) = -\frac{1}{2} \left( \log(2\pi)n_{\text{data}} + \log(|\Sigma_{\text{data}}|) + \left( X_0^\top Y_0^\top \right) \Sigma_{\text{data}}^{-1} \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix} \right)$$

where $|\Sigma_{\text{data}}|$ denotes the determinant of $\Sigma_{\text{data}}$. Using the matrix determinant lemma,

$$|\Sigma_{\text{data}}| = |A^{-1} + U^\top U| |A^{-1}|^{-1} |D_{X,Y}^2|.$$ 

The determinant of $A^{-1}$ can be computed efficiently due to Vecchiia’s approximation and final determinant (sum of a diagonal matrix) can be computed quickly as well. For the quadratic from, using the Woodbury matrix identity,

$$\begin{bmatrix} X_0 \\ Y_0 \end{bmatrix}^\top \Sigma_{\text{data}}^{-1} \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix} = \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix}^\top \left( D_{X,Y}^{-2} - D_{X,Y}^{-1} U \left( A^{-1} + U^\top U \right)^{-1} U^\top D_{X,Y}^{-1} \right) \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix}$$

$$= \left\| D_{X,Y}^{-1} \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix} \right\|^2 - V^\top \left( A^{-1} + U^\top U \right)^{-1} V,$$

where

$$V = U^\top D_{X,Y}^{-1} \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix}.$$ 

The first term can be quickly computed. For the second term, we once again use Lanczos quadrature using the algorithm described in Chapter 9.3 in Golub & Van Loan (2013) for quadratic forms. We also use this approach to estimate the determinant of $A^{-1} + U^\top U$ combined with Hutchinson’s trick (Hutchinson, 1990).

After tailoring this likelihood computation to a particular clustering $Z$, this solves the evaluation of $f(X, Y | Z)$ in the importance sampling weights. Evaluation of $f(X_i, Y_i | Z_i)$ can be done similarly.

These three computational derivations allow working with matrices the size of the total number of principal component scores, rather than the total number of data points.
In this section, we outline the parameter updates during the M-step. Throughout, we will define \( \zeta_{i,g}^{(\tau)} = \tilde{w}^{(\tau)} \mathbb{I}(Z_i^{(\tau)} = g) \), and \( \mathcal{B} \) denote the relevant penalty matrices. Since we use self-normalizing weights, we have \( \sum_{\tau=1}^{T} \tilde{w}^{(\tau)} = T \), which are occasionally used interchangeably in the ensuing formulas. We denote the sampled predictor scores for the \( i \)-th profile from the \( \tau \)-th Monte Carlo iteration as \( \alpha_i^{(\tau)} = (\alpha_{i,1}^{(\tau)}, \ldots, \alpha_{i,1}^{(\tau)})^\top \), which correspond to principal components from the cluster \( Z_i^{(\tau)} \), and likewise for \( \eta_{i,l}^{(\tau)} \).

- Use the notation for \( \delta(t) \) defined in (S5.1), and let \( \Upsilon_{\mathcal{Y}}^{g} = [\Upsilon_{\mathcal{Y},r}^{g}]_{r=1}^{2R+1} \) and \( \Upsilon_{\mathcal{X}}^{g} = [\Upsilon_{\mathcal{X},r}^{g}]_{r=1}^{2R+1} \) be vectors. The means for the response are then updated via

\[
\Upsilon_{\mathcal{Y}}^{g} = \left[ \sum_{\tau=1}^{T} \sum_{i=1}^{n} \zeta_{i,g}^{(\tau)} (B_i \otimes \delta(t_i))^\top (B_i \otimes \delta(t_i)) + \lambda_{\theta_y} (\mathcal{B} \otimes I_{2R+1}) \right]^{-1} \left[ \sum_{\tau=1}^{T} \sum_{i=1}^{n} \zeta_{i,g}^{(\tau)} (B_i \otimes \delta(t_i))^\top (X_i - B_i \lambda^g \alpha_i^{(\tau)} - B_i \Theta_i \eta_i^{(\tau)}) \right].
\]

The updates for the means of the covariates are similar:

\[
\Upsilon_{\mathcal{X}}^{g} = \left[ \sum_{\tau=1}^{T} \sum_{i=1}^{n} \zeta_{i,g}^{(\tau)} (B_i \otimes \delta(t_i))^\top (B_i \otimes \delta(t_i)) + \lambda_{\theta_x} (\mathcal{B} \otimes I_{2R+1}) \right]^{-1} \left[ \sum_{\tau=1}^{T} \sum_{i=1}^{n} \zeta_{i,g}^{(\tau)} (B_i \otimes \delta(t_i))^\top (X_i - B_i \Theta_i \alpha_i^{(\tau)}) \right].
\]

- We update the parameters in the matrices defining the principal components sequentially by column. The update for the \( l \)-th column of \( \Lambda^g \) is

\[
\Lambda_i^g = \left[ \sum_{\tau=1}^{T} \sum_{i=1}^{n} \zeta_{i,g}^{(\tau)} (\alpha_{i,l}^{(\tau)})^2 B_i^\top B_i + \lambda_{\delta} \mathcal{B} \right]^{-1} \left[ \sum_{\tau=1}^{T} \sum_{i=1}^{n} \zeta_{i,g}^{(\tau)} \alpha_{i,l}^{(\tau)} B_i \left( Y_i - (B_i \otimes \delta(t_i)) \Upsilon_{\mathcal{Y}}^{g} - B_i \Theta_i \eta_{i,l}^{(\tau)} - \sum_{l' \neq l} \alpha_{i,l'}^g B_i \Lambda_{i,l'}^g \right) \right].
\]

The update for the \( l \)-th column of \( \Theta_{\mathcal{X},g} \) is

\[
\Theta_{\mathcal{X},l}^g = \left[ \sum_{\tau=1}^{T} \sum_{i=1}^{n} \zeta_{i,g}^{(\tau)} (\alpha_{i,l}^{(\tau)})^2 B_i^\top D_{\mathcal{X}}^{-1} B_i + \lambda_{\theta_x} \mathcal{B} \right]^{-1} \left[ \sum_{\tau=1}^{T} \sum_{i=1}^{n} \zeta_{i,g}^{(\tau)} \alpha_{i,l}^{(\tau)} B_i^\top D_{\mathcal{X}}^{-1} \left( X_i - (B_i \otimes \delta(t_i)) \Upsilon_{\mathcal{X}}^{g} - \sum_{l' \neq l} \alpha_{i,l'}^g B_i (\Theta_{\mathcal{X},l'}^g) \right) \right].
\]

The update for the \( j \)-th column of \( \Theta_{\mathcal{E},g} \) is

\[
\Theta_{\mathcal{E},l}^g = \left[ \sum_{\tau=1}^{T} \sum_{i=1}^{n} \zeta_{i,g}^{(\tau)} (\eta_{i,l}^{(\tau)})^2 B_i^\top B_i + \lambda_{\delta} \mathcal{B} \right]^{-1}.
\]
\[
\sum_{\tau=1}^{T} \sum_{i=1}^{n} c_{i,g}^{(\tau)} \eta_{i,l}^{(\tau)} B_i^T \left( Y_i - (B_i \otimes \delta(t_i)) Y_{i,l}^g - B_i \Lambda g^{(\tau)} - \sum_{l \neq l}^{G} \eta_{i,l}^{(\tau)} B_i (\Theta_{g,l}^\alpha) \right).
\]

- The update for the measurement error of the response is

\[
\sigma_{\gamma}^2 = \frac{1}{T \sum_{i=1}^{T} \sum_{l=1}^{T} \sum_{g=1}^{G} c_{i,g}^{(\tau)}} \left\| Y_i - (B_i \otimes \delta(t_i)) Y_{i,l}^g - B_i \Lambda g^{(\tau)} - B_i \Theta_{g,l}^\alpha \right\|^2.
\]

Furthermore, for \( \Theta_{X}^{\theta} \) and \( \theta_{X}^{\theta} \), let \( \Theta_{X,,-T}^{\theta} \) and \( \Theta_{X,-T}^{\theta} \) denote the rows corresponding to temperature. Then, defining \( \sigma_{\epsilon}^2 = \text{Var}(\epsilon_{11}) \), we obtain an update via

\[
\sigma_{\epsilon}^2 = \frac{1}{T \sum_{i=1}^{T} \sum_{l=1}^{T} \sum_{g=1}^{G} c_{i,g}^{(\tau)}} \left\| T_i - (B_i \otimes \delta(t_i)) T_{i,l} - B_i \Theta_{g,l}^\alpha \right\|^2
\]

where \( T_i \) are the measurements for the \( i \)th temperature profile. The corresponding update for salinity is similar.

- Let \( \alpha_{g,l}^{(\tau)} \) be the random vector of observations of the random field of the \( l \)-th principal component score for the covariates in group \( g \) from the \( \tau \)-th sample from the importance sampling procedure. Let \( R_{g,l,\alpha}^{(\tau)} \) denote the corresponding correlation matrix. Furthermore, let \( n_{g}^{(\tau)} = \sum_{i=1}^{n} I \left( Z_i^{(\tau)} = g \right) \) denote the number of sites belonging to cluster \( g \) for sample \( Z^{(\tau)} \).

Correspondingly, let \( \sigma_{\alpha}^2_{g,l,\alpha} \) denote the marginal variance of the \( l \)-th predictor score for the \( g \)-th group. Then the update for \( \sigma_{g,l,\alpha}^2 \) is

\[
\sigma_{g,l,\alpha}^2 = \frac{1}{\sum_{\tau=1}^{T} w^{(\tau)}} \sum_{\tau=1}^{T} w^{(\tau)} \frac{1}{n_{g}^{(\tau)}} \left( \alpha_{g,l}^{(\tau)} \right) \left( R_{g,l,\alpha}^{(\tau)} \right)^{-1} \alpha_{g,l}^{(\tau)}.
\]

The marginal variances of \( \eta \) for each group and principal component are updated in the same way.

- The number and kind of spatial (spatio-temporal) correlation parameters depends on the choice of correlation function. For the updates, we use numerical optimization via Fisher scoring as in Guinness (2021).

- We update \( \xi \) by numerically finding the root of the gradient of the pseudo-likelihood with respect to \( \xi \); that is, if \( d_{ij} = \text{dist}(s_i, s_j) \)

\[
\frac{1}{T} \sum_{\tau=1}^{T} \sum_{i=1}^{G} \sum_{j=1}^{G} \frac{c_{i,j}^{(\tau)}}{Z_{ij}^{(\tau)}/d_{ij}} = \frac{1}{\sum_{\tau=1}^{T} \sum_{j=1}^{G} \sum_{j \neq i}^{G} Z_{ij}^{(\tau)}/d_{ij}} \exp \left( \xi \sum_{j \in \partial i} Z_{ij}^{(\tau)}/d_{ij} \right)
\]

where \( Z_{ij}^{(\tau)} = \mathbb{1} \left( Z_i^{(\tau)} = g \right) \).

Finally, an orthogonalization step is necessary to comply with the orthogonality constraints on the principal component functions. The details are outlined in Section S9 next.
S9  Avoiding the Use of Orthogonal B-splines

In previous functional data work (for example, Zhou et al., 2008; Liang et al., 2020), orthogonal B-splines are proposed for estimation of the principal components. First, for a B-spline basis $B(p)$ (in a row vector), a corresponding orthogonal B-spline basis can be written $\tilde{B}(p) = B(p)R^{-T}$, where $R^{-T} = (R^{-1})^{T}$ is a full rank matrix, with $RR^{T}$ is defined by the matrix of inner products $\int_{0}^{2000} B(p)^{T}B(p) dp$. Therefore, creating orthogonalized B-splines corresponds to an invertible change of basis. As a result, the estimated fitted functions from any penalized least squares estimate should be identical. However, they may be represented by different bases.

For both our work and work that uses orthogonalized B-splines, an orthonormalization step is used to ensure that the resulting functions are orthonormal. Consider the univariate setting, and following Zhou et al. (2008), let $\Sigma$ be the empirical covariance matrix of the scores. Also, let $\Theta$ be an estimated matrix of principal component functions for one group. Then the orthonormalization step is based on the eigendecomposition $R\Theta\Sigma\Theta^{T}R^{T} = ODO^{T}$ and $\Theta_{new} = R^{-1}O$ where $D$ is diagonal and $O$ is orthogonal. We can see that

$$
\int \Theta_{new}^{T}B(t)^{T}B(t)\Theta_{new} dt = O^{T}R^{-T}R^{T}RR^{-1}O = I_{Q_{2}}
$$

ensuring that the resulting functions are orthonormal. When using orthogonal B-splines, the same step is taken, but in this case $\int_{0}^{2000} \tilde{B}(p)^{T}\tilde{B}(p) dp$ and therefore $R$ are both the identity matrix.

Regular B-splines reduces the number of non-zero entries required for one variable from approximately $P\sum_{i=1}^{n} n_{i}/2$ to $4\sum_{i=1}^{n} n_{i}$, so we use them to obtain the same estimates. The complete orthonormalization step then consists of orthonormalizing separately $\Theta_{g,\alpha}$ and $\Theta_{g,\beta}$ for each $g$ as described above and as in Zhou et al. (2008), while also adjusting $\Lambda_{g}$ appropriately.

S10  Details for the Initialization

Here we provide more details on the initialization procedure. First, we run $k$-means for interpolated values of the data to obtain an initial cluster assignment. After this initialization, we use the model introduced in Zhou et al. (2008) extended to a Gaussian mixture model. We now assume that the pairs $(X_{s_{i}}, Y_{s_{i}})$ are $n$ i.i.d. functional random variables. Furthermore, we model the unobserved latent variables $\{Z_{i}\}$ as i.i.d. uniformly distributed on $\{1, \ldots, G\}$. Write

$$
Y_{i} = \mu_{i,Y} + \sum_{l=1}^{Q_{i}} \beta_{il}\phi_{Z_{i,l}}, \quad X_{i} = \mu_{i,X} + \sum_{l=1}^{Q_{i}} \alpha_{il}\psi_{Z_{i,l}}.
$$

Let $\alpha_{i} = (\alpha_{i1}, \ldots, \alpha_{iQ_{i}})$ and $\beta_{i} = (\beta_{i1}, \ldots, \beta_{iQ_{i}})$. Furthermore, let $D_{Z_{i,\alpha}} = \text{Var}(\alpha_{i}|Z_{i})$ and $D_{Z_{i,\beta}} = \text{Var}(\beta_{i}|Z_{i})$ be the diagonal matrices of marginal variances of the principal component scores. To model the correlation between $X_{i}$ and $Y_{i}$, we use a linear regression model for the principal component scores, that is

$$
\begin{pmatrix}
\beta_{i} \\
\alpha_{i}
\end{pmatrix}
\sim \mathcal{N}
\left(0,
\begin{pmatrix}
D_{Z_{i,\beta}} & D_{Z_{i,\alpha}}
C_{Z_{i}}^{T} & D_{Z_{i,\alpha}}
\end{pmatrix}
\right).
$$

Using the similar notation as in the main document, we approximate

$$
(B(p_{1}), B(p_{2}))\Theta_{X,l}^{g} \approx \psi_{g,l}(p_{1}, p_{2}) \quad \gamma_{X,r}^{g}(p_{1}, p_{2}) = (B(p), B(p))\Upsilon_{X,r}^{g}
$$

13
\[ B(p)\Theta_{g,1}^p \approx \phi_{g,1}(p) \quad \gamma_{g,1,r}^p(p) = B(p)\Upsilon_{g,1,r}. \]

The parameters are then estimated similarly to the procedure in the main document, with the only change being that the MCEM procedure is simpler as all spatial correlation is ignored so that obtaining the samples \{\alpha_i^{(\tau)}, \beta_i^{(\tau)}, Z_i^{(\tau)}\}_{i=1}^n for \tau = 1, \ldots, n is straightforward. Most M-step updates are analogous to the ones outlined in Section S8, and the M-step updates for \{D_{g,\alpha}\}, \{D_{g,\beta}\} and \{C_g\}_{g=1}^G can be found in Zhou et al. (2008).

After estimating the above model by running the MCEM algorithm for a few iterations, we use the final estimates for the covariates as well as the mean estimate for the response of the independent model as initial values for the spatio-temporal model. Additionally, we set \( \Lambda_g = \hat{\Theta}_{g,Y} C_g \) and
\[
\Theta_{g,\mathcal{E}} = \hat{\Theta}_{g,Y} P.
\]

where \( P \) is the matrix consisting the eigenvectors of the matrix \( D_{g,\beta} - C_g D_{g,\alpha} C_g^\top \).

**S11 Clustering Accuracy Simulation Study**

We test the performance of the Monte Carlo estimation strategy in a simulation study. The domains are \( s_i \in [0,1] \times [0,1] \) and \( p_{ij} \in [0,1] \), respectively. Locations \( s_i \) are drawn uniformly on \( S \), and the \( p_{ij} \)'s are observed on a uniform grid. For the scores, we use an isotropic exponential covariance function \( \text{Cov}(\alpha_{g,l}(s), \alpha_{g,l}(s')) = \sigma_{g,\alpha,l}^2 \exp(-\|s - s'/\kappa_{g,\alpha,l}) with \sigma_{g,\alpha,l}^2 and \kappa_{g,\alpha,l} > 0 \). The study is based on 100 independently-simulated datasets.

To assess our clustering strategy, we consider a setting comparable to Liang et al. (2020), who provide an extensive comparison of clustering methodologies for univariate functional data. We choose \( G = 2 \) groups, 3 principal components per group, and \( n = 200 \) locations. The mean was taken as the same for both groups as \( \mu(p) = e^{2-2p}\cos(5(p - 0.1)) \) (ignoring seasonality), while the \( l \)-th PC of the \( g \)-th group is taken as the \((g + 2l)\)-th function of an orthogonal spline basis of dimension 11 with uniformly-spaced knots. We set the measurement error variance to be 1.0. We use a five-nearest-neighbors graph to describe the dependence structure of the Markov random field and set \( \xi = 0.5 \). The score variances were taken to be the same for both groups, with values of 1/3, 1/6, and 1/12, respectively. The spatial range parameters \( \kappa_{g,l} \) are defined by the matrix:
\[
\kappa = \begin{pmatrix} 0.10 & 0.05 & 0.07 \\ 0.07 & 0.10 & 0.05 \end{pmatrix} \in \mathbb{R}^{2 \times 3}. \]

We consider two settings for the number of measurements per location (\( n_i = 20 \) and \( n_i = 100 \)).

We consider the Gibbs approach of Liang et al. (2020) with our importance sampling (IS) approach. We take \( G \) and the number of principal components (PCFs) as known. As Liang et al. (2020) compare their approach with others, we focus on theirs as a contrast.

In Figure S7, we plot the percent of locations whose cluster memberships are correctly identified averaged over the simulations. In the dense setting (\( n_i = 100 \)), the Gibbs sampler from Liang et al. (2020) only slightly improves upon its initialization. This is because, across Gibbs samples and EM iterations, the sampled clusters do not change enough, indicating inadequate exploration of the set of possible cluster assignments. In contrast, importance sampling appears to perform comparatively well in both scenarios with significant improvements over the initialization. Thus, ignoring spatial correlation of the functional data in our proposal distribution appears to be appropriate. We have considered a variety of simulation settings (varying \( G \), the type and number of PCFs, and \( \kappa \)), and often Gibbs sampling performs well when \( m_i < 50 \), while it essentially does not work when \( m_i \geq 75 \).
Figure S7: Percent of profiles correctly identified, averaged over simulations. The first 5 iterations are under the independent model for all models. IS refers to importance sampling.

References

Argo (2020). Argo float data and metadata from Global Data Assembly Centre (Argo GDAC).

Bushinsky, S. M., Gray, A. R., Johnson, K. S., & Sarmiento, J. L. (2017). Oxygen in the Southern Ocean from Argo floats: determination of processes driving air-sea fluxes. *Journal of Geophysical Research: Oceans*, 122(11), 8661–8682.

Feng, D., Tierney, L., & Magnotta, V. (2012). MRI tissue classification using high-resolution Bayesian hidden Markov normal mixture models. *J. Am. Statist. Ass.*, 107(497), 102–119.

Golub, G. H. & Van Loan, C. F. (2013). *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition.

Guennebaud, G., Jacob, B., et al. (2010). Eigen v3. http://eigen.tuxfamily.org.

Guinness, J. (2021). Gaussian process learning via Fisher scoring of Vecchia’s approximation. *Statistics and Computing*, 31(3), 25.

Hutchinson, M. (1990). A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 19(2), 433–450.

Jones, D. C., Holt, H. J., Meijers, A. J. S., & Shuckburgh, E. (2019). Unsupervised clustering of Southern Ocean Argo float temperature profiles. *Journal of Geophysical Research: Oceans*, 124(1), 390–402.

Liang, D., Zhang, H., Chang, X., & Huang, H. (2020). Modeling and regionalization of China’s PM$_{2.5}$ using spatial-functional mixture models. *J. Am. Statist. Ass.*, (pp. 1–17).

Riser, S. C., Talley, L. D., Wijffels, S. E., Nicholson, D., Purkey, S., Takeshita, Y., Fassbender, A., Gray, A., Robbins, P., Gilson, J., Plant, J. N., Clark, E., Swift, D. D., Rupan, R. A., Maurer, T. L., & Johnson, K. S. (2023). SOCCOM and GO-BGC float data - Snapshot 2023-08-28. In Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) and Global Ocean Biogeochemistry (GO-BGC) Biogeochemical-Argo Float Data Archive.
Roemmich, D. & Gilson, J. (2009). The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo Program. *Progress in Oceanography, 82*(2), 81–100.

Zhou, L., Huang, J., & Carroll, R. (2008). Joint modeling of paired sparse functional data using principal components. *Biometrika, 95*, 601–619.