Computer Model Calibration with Multivariate Spatial Output: A Case Study

K. Sham Bhat
Department of Statistics
The Pennsylvania State University
kgb130@psu.edu

Murali Haran
Department of Statistics
The Pennsylvania State University
mharan@stat.psu.edu

Marlos Goes
Department of Geosciences
The Pennsylvania State University
mpg14@psu.edu

January 2, 2010

Abstract

Computer model calibration involves combining information from simulations of a complex computer model with physical observations of the process being simulated by the model. Increasingly, computer model output is in the form of multiple spatial fields, particularly in climate science. We study a simple and effective approach for computer model calibration with multivariate spatial data. We demonstrate the application of this approach to the problem of inferring parameters in a climate model. We find that combining information from multiple spatial fields results in sharper posterior inference than obtained from a single spatial field. In addition, we investigate the effects of including a model discrepancy term and compare the use of a plug-in versus a fully Bayesian approach for accounting for emulator variances. We find that usually, although not always, inclusion of the model discrepancy term results in more accurate and sharper inference of the calibration parameter, and estimating emulator spatial variances in a fully Bayesian model results in wider posterior distributions.

1 Introduction

Complex computer models are widely used by scientists to understand and predict the behavior of complex physical processes. Examples of applications include climate science, weather forecasting, disease dynamics, and hydrology. Inference on these complex systems often combines information from simulations of the complex computer model with field data collected from experiments or observations on the real physical system. The computer model simulations are frequently very computationally expensive, with each simulation taking minutes, days or even weeks to complete, which makes Monte Carlo-based approaches to inference infeasible. Computer model emulation is a powerful approach pioneered by Sacks et al. (1989) to approximate the expensive computer model by a Gaussian process. Emulation allows approximate output at any parameter setting to be obtained from a computationally tractable Gaussian process fit to the output from the computer model at several parameter settings. This approach can then be used in a larger framework that includes a model for physical observations in order to do computer model calibration. Computer model calibration finds the value of the computer model parameters or 'inputs' most compatible with the observations of the process. Here we follow the general framework described in Kennedy and O'Hagan (2001) and further developed by many others (cf. Bayarri et al., 2007a; Sansò et al., 2008).

Increasingly, computer model output is multivariate (cf. Bayarri et al., 2007a; Higdon et al., 2008). Of particular interest are models where the output is in the form of multivariate spatial data. We consider as a case study the problem of inferring the value of a climate parameter based on climate model output and physical observations that are in the form of multivariate spatial data sets. This problem is motivated by the goal of assessing the risks of future climate change. Specifically, we focus on the problem of learning about the climate parameter 'background ocean vertical diffusivity' (K_v), which determines the strength of the heat and salt diffusion in the ocean component of the climate model, and is a key parameter in climate model predictions of the Atlantic Meridional Overturning Circulation (AMOC). The AMOC, part of the global ocean circulation system, plays an important role in global climate. A weakening or possible collapse of the AMOC can potentially result in major temperature and precipitation changes and a shift in terrestrial ecosystems. AMOC predictions may be obtained from climate models, which include several parameterizations in order to mimic real physical processes. Of the model parameters, K_v is particularly important for predictions of AMOC. Reducing the uncertainty about the value of K_v will also reduce the uncertainty of other key model parameters like climate sensitivity (Forest et al., 2002). While the value of the parameter K_v may not resemble the observed ocean diffusivity, because it is intended to represent several mechanisms that generate turbulent mixing in the ocean, ocean tracers

can provide information about large scale ocean patterns. Such information can be used to infer K_v in the model, since observed tracers are strongly affected by this parameter. For example, larger observed values of the tracer Δ^{14}C in the deep ocean suggest a higher intensity of vertical mixing. These tracer data are in the form of spatial fields. Hence, the computer model calibration problem here involves climate parameter inference based on multivariate spatial data.

In this book chapter, we consider inference based on three oceanic tracers, all in the form of relatively small one-dimensional spatial fields. We present a simple framework for combining information from multiple spatial fields from model simulations and physical observations in the context of inferring the climate parameter K_v. We study the impact of including a Gaussian process model for the discrepancy between the model and the true system. In addition, we study the impact of model assumptions by holding out model output at a particular parameter setting and treating noisy versions of this output as 'real data'. We consider two statistical models, one that combines observation error and model discrepancy into a single independent error term, the second where model discrepancy is modeled separately using a Gaussian process. We are particularly interested in studying the impact of the discrepancy term on climate parameter inference. We also then examine the effect of estimating emulator spatial variance in a Bayesian framework versus using a plug-in approach.

The rest of our this book chapter is organized as follows. In Section 2, we discuss our approach for calibration with spatial output. We build upon this framework to perform parameter inference with multiple spatial fields in Section 3, paying special attention to model discrepancy and emulator variances. In Section 4, we describe our case study, discussing both the data set and modeling and implementation details. We describe the results of our study in Section 5 and conclude with a summary and discussion in Section 6.

2 Computer model calibration with spatial output

In this section, we describe our model for inferring calibration parameters from the observations and model output of a single spatial field. We use the two-stage approach described below for model calibration. We will also discuss the importance of various modeling assumptions.

In the first stage of our approach, we emulate the computer model by fitting a Gaussian process to the spatial computer model output. In the second stage, we connect the calibration parameters to the observations using the emulator, while allowing for other sources of uncertainty, such as model discrepancy and observation error. This allows us to use a Bayesian

approach to obtain a posterior distribution for the parameters. Our approach of splitting inference into two stages has several advantages over fitting a single model in one inferential step including separating the parts of the statistical model that are known to be correct from the parts of the model that are questionable, improved diagnostics, and computational advantages (see Bayarri et al., 2007b; Liu et al., 2009; Rougier, 2008a).

We begin with some notation. Let $Z(\mathbf{s})$ be the observation of the spatial field at location $\mathbf{s}$, where $\mathbf{s} \in D$ with $D \in \mathbb{R}^d$. For simplicity, and given the case study in Section 4, we assume that $d = 1$, i.e. we have a one-dimensional spatial field. Let θ be the calibration or model parameter of interest; our framework may easily be expanded to allow for vectors of parameters. $Y(\mathbf{s}, \theta)$ denotes the computer model output at the location $\mathbf{s}$, and at the calibration parameter setting θ. In general, the spatial data from the computer model grid may or may not coincide with the locations of the observations. The objective here is to infer a posterior distribution of θ given the observed data and computer model output.

Let $\mathbf{Y} = (Y_{11}, \cdots, Y_{n1}, Y_{12}, \cdots, Y_{n2}, \cdots, Y_{1p}, \cdots, Y_{np})^T$, obtained by stacking computer model output at all calibration parameter settings, denote the computer model output for a single spatial field. Y_{ik} corresponds to the model output for location $\mathbf{s}_i$ and calibration parameter setting θ_k, and n is the number of model output locations and p is the number of calibration parameter settings. Similarly, $\mathbf{Z} = (Z_1, \cdots, Z_N)^T$ are the observations for the spatial field, where N is the total number of observations.

2.1 Computer model emulation

We model the computer model output $\mathbf{Y}$ using a Gaussian process:

$$\mathbf{Y} \mid \boldsymbol{\beta}, \theta, \boldsymbol{\xi}_m \sim N(\mu_{\boldsymbol{\beta}}(\theta), \Sigma_M(\boldsymbol{\xi}_m))$$

where we assume a linear mean function, $\mu_{\boldsymbol{\beta}}(\theta) = X\boldsymbol{\beta}$, with X a covariate matrix of dimension $np \times b$, where there are $(b-1)$ covariates. The covariates we use are location and the calibration parameter. $\boldsymbol{\xi}_m$ is a vector of covariance parameters that specify the covariance matrix $\Sigma_M(\boldsymbol{\xi}_m)$ and $\boldsymbol{\beta}$ is a vector of regression coefficients. We use a Gaussian covariance function as described below:

$$(\Sigma_M)_{ij}(\boldsymbol{\phi}, \kappa) = \zeta I(i = j) + \kappa \exp\left(-\frac{\|\mathbf{s}_i - \mathbf{s}_i\|^2}{\phi_s^2} - \frac{|\theta_i - \theta_j|^2}{\phi_c^2}\right) \tag{1}$$

where $\boldsymbol{\phi} = (\phi_s, \phi_c)$, $\kappa, \zeta, \phi_s, \phi_c > 0$. The covariance function is separable over space and calibration parameters, although a nonseparable covariance could be chosen if appropriate (see Gneiting, 2002). Note that this function can be easily adapted to models for multiple calibration parameters, as well as multiple spatial dimensions.

Let the maximum likelihood estimate of $(\boldsymbol{\xi}_m, \boldsymbol{\beta})$ be $(\hat{\boldsymbol{\xi}}_m, \hat{\boldsymbol{\beta}})$. Let $\mathbf{S}$ be the set of locations where the observations were collected. Following the standard kriging framework (Cressie, 1993; Stein, 1999), the multinormal predictive distribution for the computer model output at a new θ at $\mathbf{S}$ is obtained by substituting $(\hat{\boldsymbol{\xi}}_m, \hat{\boldsymbol{\beta}})$ in place of $(\boldsymbol{\xi}_m, \boldsymbol{\beta})$ and conditioning on $\mathbf{Y}$. We denote the random variable with this predictive distribution by $\boldsymbol{\eta}(\mathbf{Y}, \theta)$ in the second stage of our inference below.

2.2 Computer model parameter inference

In order to infer θ based on the observations $\mathbf{Z}$, we need a probability model connecting θ and $\mathbf{Z}$. The predictive distribution from Section 2.1 provides a model for computer model output at any θ and any set of new locations. We now model the observations $\mathbf{Z}$ as realizations from a stochastic process obtained by accounting for additional error to the computer model emulator from Section 2.1. Our model for the observations $\mathbf{Z}$ is therefore

$$\mathbf{Z} = \boldsymbol{\eta}(\mathbf{Y}, \theta) + \boldsymbol{\delta}(\mathbf{S}) + \boldsymbol{\epsilon}$$

where $\boldsymbol{\eta}(\mathbf{Y}, \theta)$ is as described in Section 2.1, $\boldsymbol{\epsilon} \sim N(0, \psi I)$, where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_N)^T$ is the observation error with $\psi > 0$ as the observation error variance. The model discrepancy, $\boldsymbol{\delta}(\mathbf{S})$, is modeled as a zero-mean Gaussian process. Hence, $\boldsymbol{\delta}(\mathbf{S}) \sim N(\mathbf{0}, \Sigma_d(\boldsymbol{\xi}_d))$, where $\boldsymbol{\xi}_d$ is a vector of covariance parameters that specify the covariance matrix $\Sigma_d(\boldsymbol{\xi}_d)$. We have in essence 'inferred a likelihood' for use in our Bayesian framework, since for any fixed $\mathbf{Z}$, we can obtain a value of the likelihood for any value of θ. We will discuss the merits of including a model discrepancy term in Section 3.3.

We may allow the emulator spatial variance scale parameter from the first stage, κ, to vary, rather than plugging in the MLE $\hat{\kappa}$. We can now perform inference on θ, ψ, κ, and $\boldsymbol{\xi}_d$ by specifying a prior for these parameters. Using Markov Chain Monte Carlo (MCMC), we can estimate a posterior distribution for θ. It should be noted that the computational complexity of the matrix operations involved in the second stage of our approach is solely dependent on N, the size of $\mathbf{Z}$, and not $M = np$, where M is the size of the ensemble of model output $\mathbf{Y}$. We will discuss prior selection for θ, ψ, κ, and $\boldsymbol{\xi}_d$ in Section 4.2.

3 Calibration with multivariate spatial output

In this section, we discuss how our approach can be used to combine information from multiple spatial fields. We use a separable covariance model (see for instance, Banerjee et al., 2004) to

model the relationship of the computer model output from the three spatial fields. The similar shape of the empirical variograms of the model output from three spatial fields in our case study in Section 4 justify the use of a separable covariance model. We extend our notation to allow for multiple spatial fields. Let $\mathbf{Y}_1 = (Y_{11} \cdots Y_{1np})^T$, $\mathbf{Y}_2 = (Y_{21} \cdots Y_{2np})^T$, and $\mathbf{Y}_3 = (Y_{31} \cdots Y_{3np})^T$. denote the computer model output for three spatial fields. Similarly, $\mathbf{Z}_1 = (Z_{11} \cdots Z_{1N})^T$, $\mathbf{Z}_2 = (Z_{21} \cdots Y_{2N})^T$ and $\mathbf{Z}_3 = (Z_{31} \cdots Y_{3N})^T$ are the observations for the same three spatial fields. For convenience, we write $\mathbf{Y} = (Y_{11}, Y_{21}, Y_{31}, Y_{12} \cdots Y_{1np}, Y_{2np}, Y_{3np})^T$, and $\mathbf{Z} = (\mathbf{Z}_1\ \mathbf{Z}_2\ \mathbf{Z}_3)$.

3.1 Stage 1: Emulation for multivariate data

The computer model output for the spatial fields are modeled using using a separable cross-covariance function as described below in equation (2):

$$\begin{aligned} \mathbf{Y} \mid \boldsymbol{\beta}, \theta, \boldsymbol{\xi}_m &\sim N(\mu_{\boldsymbol{\beta}}(\theta), \Sigma_M(\boldsymbol{\xi}_m)) \\ \mu_{\boldsymbol{\beta}} &= (\mu_{\boldsymbol{\beta}_1}, \mu_{\boldsymbol{\beta}_2}, \mu_{\boldsymbol{\beta}_3})^T \\ \Sigma_M(\boldsymbol{\xi}_m) &= P(\boldsymbol{\zeta}) + H(\boldsymbol{\phi}) \otimes T(\boldsymbol{\kappa}, \boldsymbol{\rho}) \end{aligned} \tag{2}$$

where $\mu_{\boldsymbol{\beta}_i}$ is a function of the calibration parameters, and $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ are the coefficient vectors for $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ respectively. We note that we are implicitly assuming a linear relationship among the spatial fields. For an approach based on a flexible hierarchical model allowing for non-linear relationships among spatial fields, see Bhat et al. (2009). A computationally inexpensive approach that avoids computer model emulation and hence utilizes several simplifying assumptions is described in Goes et al. (2009).

We now assume $\mu_{\boldsymbol{\beta}_i}(\theta) = X\boldsymbol{\beta}_i$ (for i=1,2) where X is the covariate matrix of dimension $M \times b$, with covariates (depth and calibration parameters) as specified in Section 2.1. $\boldsymbol{\xi}_m$ is a vector of covariance parameters that specify the cross-covariance matrix $\Sigma_M(\boldsymbol{\xi}_m)$. $H(\boldsymbol{\phi})$ explains spatial dependence, while $T(\boldsymbol{\kappa}, \boldsymbol{\rho})$ is interpreted as the cross-covariance between spatial fields. $P(\boldsymbol{\zeta})$ is a matrix that describes microscale variance of the process. The covariance matrices are defined as follows:

$$H_{ij}(\boldsymbol{\phi}) = \exp\left(-\frac{\|\mathbf{s}_i - \mathbf{s}_i\|^2}{\phi_s} - \frac{|\theta_i - \theta_j|^2}{\phi_c}\right) \tag{3}$$

$$T = \begin{bmatrix} \kappa_1 & \rho_{12}\sqrt{\kappa_1\kappa_2} & \rho_{13}\sqrt{\kappa_1\kappa_3} \\ \rho_{12}\sqrt{\kappa_1\kappa_2} & \kappa_2 & \rho_{23}\sqrt{\kappa_2\kappa_3} \\ \rho_{13}\sqrt{\kappa_1\kappa_3} & \rho_{23}\sqrt{\kappa_2\kappa_3} & \kappa_3 \end{bmatrix} \quad P = \begin{bmatrix} \zeta_1\mathbf{I}_N & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \zeta_2\mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \zeta_3\mathbf{I}_N \end{bmatrix} \tag{4}$$

where $\boldsymbol{\phi} = (\phi_s, \phi_c)$, and $\kappa_i, \zeta_i, \phi_s, \phi_c > 0$, $-1 \leq \rho_{ij} \leq 1$. We reduce parameters and ensure that $\Sigma_{\mathbf{Y}}$ is positive definite and symmetric by letting $\rho_{ii} = 1$ and $\rho_{ij} = \rho_{ji}$. We estimate MLEs for the following parameters using the computer model output: $\mathbf{Y}$: $\zeta_1, \zeta_2, \zeta_3, \kappa_1, \kappa_2, \kappa_3, \phi_s, \phi_c$. In principle, $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ and $\boldsymbol{\rho}$ may be estimated using maximum likelihood, but in the case study in Section 4, we estimate $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ using least squares regression and $\boldsymbol{\rho}$ using empirical sample correlations. As in Section 2.1, we obtain a multinormal predictive distribution $\boldsymbol{\eta}(\mathbf{Y}, \theta)$ each θ at $\mathbf{S}$ by plugging in the MLEs and conditioning on $\mathbf{Y}$. For ease of computation, we order the model output by depth and calibration parameter, and we write $\mathbf{Y} = (Y_{11}, Y_{21}, Y_{31}, Y_{12} \cdots Y_{1np}, Y_{2np}, Y_{3np})^T$.

3.2 Stage 2: Inference for multiple spatial fields

We write the model for the observed data as follows:

$$\mathbf{Z} = \boldsymbol{\eta}(\mathbf{Y}, \theta) + \boldsymbol{\delta}(\mathbf{S}) + \boldsymbol{\epsilon}$$

where $\boldsymbol{\eta}(\mathbf{Y}, \theta)$ is as described earlier in Section 3.1, and $\boldsymbol{\epsilon} = (\epsilon_{11}, \cdots, \epsilon_{N1}, \epsilon_{12}, \cdots, \epsilon_{N2}, \epsilon_{13}, \cdots, \epsilon_{N3})^T$ is the observation error. We assume that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma_{\boldsymbol{\epsilon}})$ with

$$\Sigma_{\boldsymbol{\epsilon}} = \begin{bmatrix} \psi_1 \mathbf{I}_N & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \psi_2 \mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \psi_3 \mathbf{I}_N \end{bmatrix}$$

where $\psi_1, \psi_2, \psi_3 > 0$ are the observation error variances for the three spatial fields. The model error or discrepancy $\boldsymbol{\delta}(\mathbf{S})$ is modeled as a vector of three independent zero mean Gaussian processes. The model for discrepancy is given in Section 3.3, and includes covariance parameters $\boldsymbol{\xi}_{d1}$, $\boldsymbol{\xi}_{d2}$, and $\boldsymbol{\xi}_{d3}$.

Using the observations $\mathbf{Z}$, we obtain the posterior distribution of θ, ψ_1, ψ_2, ψ_3, $\boldsymbol{\xi}_{d1}$, $\boldsymbol{\xi}_{d2}$, and $\boldsymbol{\xi}_{d3}$ using MCMC as discussed in Section 2.2. We may also allow the emulator spatial variance parameters from the first stage, κ_1, κ_2, and κ_3 to be reestimated, rather than using plug-in MLEs. This can be done using by estimating the matrix T using an inverse Wishart prior. More details about prior selection are discussed in Section 4.2.

3.3 Model discrepancy

An important concern in the process of calibration is whether the model is an adequate representation of the true phenomena in the system. When this is not the case, there is a

need to consider ways to incorporate the difference between the computer model and reality. The latter is usually referred to as model discrepancy.

A framework to account for model discrepancy is introduced in Kennedy and O'Hagan (2001), and strong arguments in favor of inclusion of a model discrepancy term in any calibration approach is made in Bayarri et al. (2007b). Specifically, the argument is made that neglecting to account for the model discrepancy results in overfitting, resulting in potentially biased and incorrect inference of the calibration parameters. A test is introduced for whether model discrepancy is needed in Bayarri et al. (2009), which almost always results in rejecting the hypothesis that the model represents the truth. However, O'Hagan (2009) suggests that the inclusion of a model discrepancy does not always result in a less biased estimates of calibration parameters, rather more biased estimates are possible. A further difficulty in including a model discrepancy term in our statistical model is the high dependence between the calibration parameter and model discrepancy term (Liu et al., 2009). Previous work has shown that attempting to separate observation error and model error can impose nontrivial computation and conceptual problems (Kennedy and O'Hagan, 2001; Sansò et al., 2008). An approach that combines observation error and model error into a single term, rather than estimate them separately is described in Sansò et al. (2008). Even this approach requires substantial compromises in computing techniques in order to fit the model. In our case study, we consider the two different approaches to incorporate the error into our statistical model as follows:

Approach 1: **No model discrepancy term (model discrepancy and observation error combined)**:

$$\mathbf{Z} = \boldsymbol{\eta}(\mathbf{Y}, \theta) + \boldsymbol{\epsilon}$$

where $\boldsymbol{\eta}(\mathbf{Y}, \theta)$ is the predictive distribution as described earlier in Section 3.1, and $\boldsymbol{\epsilon} = (\epsilon_{11}, \cdots, \epsilon_{N1}, \epsilon_{12}, \cdots, \epsilon_{N2}, \epsilon_{13}, \cdots, \epsilon_{N3})^T$ is the observation error, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma_{\boldsymbol{\epsilon}})$, with

$$\Sigma_{\boldsymbol{\epsilon}} = \begin{bmatrix} \psi_1 \mathbf{I}_N & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \psi_2 \mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \psi_3 \mathbf{I}_N \end{bmatrix}$$

where $\psi_1, \psi_2, \psi_3 > 0$ are the observation error variances for the three spatial fields.

In the univariate case, $\boldsymbol{\epsilon} \sim N(0, \psi I)$, where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_N)^T$ is the observation error with $\psi > 0$ as the observation error variance.

Approach 2: **Model discrepancy modeled as a zero-mean Gaussian process.**

$$\mathbf{Z} = \boldsymbol{\eta}(\mathbf{Y}, \theta) + \boldsymbol{\delta}(\mathbf{S}) + \boldsymbol{\epsilon}$$

where $\boldsymbol{\eta}(\mathbf{Y}, \theta)$ and $\boldsymbol{\epsilon}$ are the same as in Approach 1 above. The model error or discrepancy $\boldsymbol{\delta}(\mathbf{S})$ is modeled as a vector of three independent zero mean Gaussian processes below:

$$\boldsymbol{\delta}(\mathbf{S}) \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\Sigma_{d1}(\boldsymbol{\xi}_{d1}) & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \Sigma_{d2}(\boldsymbol{\xi}_{d2}) & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \Sigma_{d3}(\boldsymbol{\xi}_{d3})\end{bmatrix}\right)$$

where $\boldsymbol{\xi}_{dk}$=$((\phi_{dk})_i, \kappa_{dk})$ covariance matrix $\Sigma_{dk}(\boldsymbol{\xi}_{dk})$ is as follows:

$$(\Sigma_{dk})_{ij}(\phi_{dk}, \kappa_{dk}) = \kappa_{dk} \exp\left(-\frac{\|\mathbf{s}_i - \mathbf{s}_j\|^2}{\phi_{dk}^2}\right), \quad \kappa_{dk}, (\phi_{dk}) > 0 \tag{5}$$

In the univariate case, $\boldsymbol{\delta}(\mathbf{S}) \sim N(\mathbf{0}, \Sigma_D(\boldsymbol{\xi}_d))$, where the covariance matrix $\Sigma_D(\boldsymbol{\xi}_d)$ is the same form as equation (5). While it can be argued that the model discrepancy should not be assumed to have zero mean, in practice it may be too hard to identify a non-zero mean. The Gaussian process is flexible enough to correct for an incorrect mean structure. Further, additional parameters to model the mean of the model discrepancy would be confounded with the climate calibration parameter.

3.4 Estimation of emulator spatial variance parameters

Bayarri et al. (2007b) discusses the issue of estimating emulator spatial parameters in a modularization framework; specifically the question of whether to estimate these parameters in a full Bayesian approach as opposed to a plug-in MLE approach. Bayarri et al. (2007b) argues that while a full Bayesian approach would be more informative because uncertainty in the emulator parameters is taken into account, such uncertainties are often small compared to the uncertainties due to the model discrepancy resulting in little difference in the final results. Further, using a full Bayesian approach often leads to a significant increase in computation time. The full Bayesian approach also results in identifiability issues. For example, attempting to estimate the microscale variation (emulator nugget) in the second stage is difficult because it is clearly confounded with the observation error variance. We have therefore used a plug-in approach so far.

We now study the estimation of the emulator spatial variance in a Bayesian framework in the second stage for Model 2, where a model discrepancy term is included. For the univariate case, this consists of estimating κ, in the multivariate case, we need to estimate the cross-covariance matrix $T(\boldsymbol{\kappa}, \boldsymbol{\rho})$ in the second stage.

4 Application to climate parameter inference

4.1 Ocean tracer data

In this study, we focus on three tracers that have previously been shown to be informative about K_v in ocean models: Δ^{14}C, trichlorofluoromethane (CFC11), and ocean temperature (T) (cf. Schmittner et al., 2009). ^{14}C (radiocarbon) is a radioactive isotope of carbon, which may be produced naturally and by detonation of thermonuclear devices. ^{14}C and CFC11 enter the oceans from the atmosphere by air-sea gas exchange and is transported from the ocean by advection, diffusion, and to a lesser degree by biological processes (Key et al., 2004; McCarthy et al., 1977).

Δ^{14}C, CFC11, and ocean temperature (T) measurements were collected for all oceanic basins in the 1990s, with locations denoted by a latitude, longitude, and depth. The data have been controlled for quality and gridded by Key et al. (2004). We use the observations from the data synthesis project by Key et al. (2004), which are then aggregated globally (i.e. aggregated over latitude and longitudes), resulting in a data set with $N = 13$ depths. In addition, model output at $p = 10$ different values of K_v, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, and 0.5 cm^2/s, on a grid of locations of latitude, longitude, and ocean depth were evaluated from University of Victoria (UVic) Earth System Climate Model as described in Schmittner et al. (2009). The model output were also aggregated globally, providing a 'depth profile' representing an average between 1990-2000 (cf. Goes et al., 2009). The total number of depths in the model output is $n = 13$, resulting in $M{=}np{=}130$ model output values per tracer. Depths below 3000 m are excluded to minimize problems due to sparse sampling (Key et al., 2004) and model artifacts (Schmittner et al., 2009).

To perform statistical inference on climate parameters, we need to establish a relationship between the observations and the climate parameters. We accomplish this by using an earth system model, which simulate the complex phenomenon of the atmosphere and the oceans under specific input parameter settings to obtain output. The climate models are complex computer codes representing the solution to a large set of differential equations that approximate physical, chemical, and biological processes (Weaver et al., 2001). These climate models often take weeks to months to execute for any given calibration parameter setting, making it very computationally expensive to obtain output at a large number of parameter settings. This provides a compelling argument for using emulation.

4.2 Implementation details

In this section, we discuss some of the details of the application of our approach to the ocean tracer data described in Section 1. We verify the emulator using a leave one out cross-validation approach, where we leave out the model output for one calibration parameter (Rougier, 2008b) and predict at all locations for that calibration parameter setting. Plots for the model output and predictions using both cross-validation approaches for K_v=0.2 are shown in Figure 4. The cross-validation appear to result in predictions for the removed locations visually similar to the original model output (Figure 4).

In the second stage, we use MCMC to obtain the posterior distributions of θ. We use a Lognormal (-1.55, 0.59) on θ which reflects the geoscientists' prior uncertainty about K_v based on previous research (Bhat et al., 2009). We use a wide inverse gamma prior for the observation error and model discrepancy variances, specifically $\psi_1 \sim IG(2, 10)$ and $\kappa_{d1} \sim IG(2, 1000)$ for Δ^{14}C, $\psi_2 \sim IG(2, 0.1)$ and $\kappa_{d2} \sim IG(2, 0.6)$ for CFC11, and $\psi_3 \sim IG(2, 0.1)$ and $\kappa_{d3} \sim IG(2, 15)$ for T. We use wide uniform priors for the model discrepancy range parameter. For the emulator spatial variances, we use $\kappa_1 \sim IG(5, 24000)$, $\kappa_2 \sim IG(5, 2.4)$, and $\kappa_3 \sim IG(5, 60)$. When combining multiple spatial fields we can instead place an inverse Wishart prior on the cross covariance matrix T, specifically, $T \sim IW(10, 8T_{MLE})$. Here T_{MLE} is matrix for T obtained by plugging in the MLE from the first stage. The other parameters for the Inverse Wishart were determined using formulae from Anderson (2003) to ensure that the distribution is centered around T_{MLE} and variances of individual matrix elements are relatively small. Specifically, the variances of individual matrix elements decrease as the first parameter of the Inverse Wishart distribution is increased. These priors were obtained after an exploratory analysis of the data suggested the approximate scale of these parameters. While we understand that one needs to be careful about using the data in any way to determine priors, our priors are fairly wide with infinite variance (except for the emulator spatial variance terms), and are not strongly informative.

To ensure convergence of our MCMC based estimates in the second stage, we obtained Monte Carlo standard errors for the posterior mean estimates of θ and other parameters computed by consistent batch means (Flegal et al., 2008; Jones et al., 2006). The posterior mean estimates of θ had MCMC standard errors below 10^{-4} for both the univariate and bivariate approaches. The MCMC standard errors for the other parameters were less than 10^{-3} for both the univariate and bivariate approaches.

5 Results

5.1 Ocean tracer data

In this section we present the results from our analyses using the tracers Δ^{14}C, CFC11, and T. While there is substantial overlap among the posterior distributions of K_v (with model discrepancy is included) obtained by using Δ^{14}C, CFC11, and T separately and then jointly, there are also clear differences (Figure 3). We calculated credible regions using the Highest Posterior Density (HPD) method (Chen et al., 2000). The 90% credible region for K_v using the single tracer Δ^{14}C is between 0.057 and 0.352 cm^2/s, the 90% credible region for K_v using the single tracer CFC11 is between 0.170 and 0.407 cm^2/s, the 90% credible region for K_v using the single tracer T is between 0.156 and 0.420 cm^2/s, and the 90% credible region for K_v using the tracers jointly is between 0.164 and 0.313 cm^2/s. Combining the information from all three tracers results in a sharper posterior distribution of K_v when model discrepancy is included (Figure 3).

Inclusion of the model discrepancy term appears to shift the posterior probability distribution to the left when we combine the three tracers (Figure 2) and when we use the CFC and T tracers individually (Figure 1), suggesting that an approach without taking model discrepancy into account results in a bias. The 90% credible region for K_v using the tracers jointly is between 0.194 and 0.390 cm^2/s when model discrepancy *is not* included, between 0.164 and 0.313 cm^2/s when model discrepancy *is* included, and between 0.130 and 0.395 cm^2/s, when emulator spatial variance is also estimated in a fully Bayesian approach. There is little difference when the tracer Δ^{14}C is used, however, between the approach including model discrepancy and the approach that does not do so. Further, it appears that estimating emulator spatial variance in a fully Bayesian approach results in wider posterior probability distributions, likely due to the additional uncertainty contributed by the emulator. It appears that the posterior distribution for K_v for the tracers jointly is clearly sharper than for the tracers individually for all of the approaches (Figures 1 and 2).

5.2 Simulation study

We investigated the impact of including a model discrepancy term as described in Section 3.3. Our goal in this study is to determine whether the inclusion of model discrepancy and estimation of emulator spatial variance actually results in better inference of the calibration parameter under different error situations. We hold out a calibration parameter setting, say K_v=0.35, and treat the model output (for all the tracers) for that parameter setting as the

observations. We then apply our two stage approach as described earlier to the remainder of the model output and the synthetic observations for all three modeling approaches; exclusion of the model discrepancy term, inclusion of the model discrepancy term, and inclusion of the model discrepancy term plus estimation of the emulator spatial variance. This procedure is executed for all three scenarios below:

Scenario 1: No error. In this scenario, we simply define the observations as the model output at the held out calibration parameter setting. That is, $\mathbf{Z}_k^*(\mathbf{s}_i) = \mathbf{Y}_k(\mathbf{s}_i, \theta^*)$ for $i = 1, \cdots, N$, where θ^* is the held out calibration parameter setting and $k = 1, 2, 3$ denotes the tracer of interest.

Scenario 2: Independent and identically distributed (i.i.d.) error. In this scenario, we add $N(0, \sigma_k^2)$ to the model output at the calibration parameter setting for each location for tracer i. Specifically, $\mathbf{Z}_k^*(\mathbf{s}_i) = \mathbf{Y}_k(\mathbf{s}_i, \theta^*) + \epsilon_k^*$, where $\epsilon_{ki}^* \sim N(0, \sigma_k^2)$. Since the scale of the three tracers are different, we must select different values for σ_k^2.

Scenario 3: Model discrepancy plus i.i.d. error. In this scenario, we add a $GP(\mu_k, \Sigma_k)$ to the observations in Scenario 2. Specifically, $\mathbf{Z}_k^*(\mathbf{s}_i) = \mathbf{Y}_k(\mathbf{s}_i, \theta^*) + \boldsymbol{\delta}_k^* + \boldsymbol{\epsilon}_k^*$, where $\boldsymbol{\delta}_k^* \sim GP(\mu_k, \Sigma_k)$. Again since the scale of the three tracers are different, so are the parameters of the Gaussian processes.

The results of this experiment suggest that adding a model discrepancy term results in more accurate inference and less overfitting for all three scenarios (Figure 5(a)-(c)). Estimation of the emulator spatial variance term results in much wider posterior probability distributions for all three scenarios (Figure 5(a)-(c)). In Scenario 1, both approaches of excluding and including the model discrepancy term result in having a posterior distribution centered near the held out parameter K_v=0.35. However the posterior distribution is sharper and slightly more accurate when the model parameter is included (Figure 5(a)). Estimation of the emulator spatial variance term results in a wider posterior probability distribution, but correctly centered around the held out parameter K_v=0.35 (Figure 5(a)). In Scenario 2, excluding the model discrepancy term results in a bias to the left (smaller values of K_v), while including the model discrepancy results in more accurate inference and a sharper posterior for the calibration parameter (Figure 5(b)). Estimating the emulator spatial variance results in a slightly biased and wider posterior for the calibration parameter (Figure 5(b)). In Scenario 3, excluding the model discrepancy term results in a clearly biased distribution that has a wide bimodal posterior, while including the model discrepancy term results in a posterior distribution that has much less bias, is unimodal, and is sharper (Figure 5(c)). Estimating the emulator spatial variance results in a wider posterior distribution that is biased slightly to the left (Figure 5(c)). It is important to stress that obtaining such results required much

experimentation in determining instructive parameters for σ_k and Σ_k. Simulating observations with too much noise would clearly result in the signal being too weak, and thus the inability to obtain reasonable inference about the calibration parameter, while adding too little error would result in virtually the same inference as in the case with no added error.

Inspired by the suggestion from O'Hagan (2009) that the inclusion of a model discrepancy term may actually result in more biased estimates of the calibration parameter, we attempted to find a situation where the inclusion of the model discrepancy term 'makes the situation worse'. To do so we added a function proportional to 1/depth or 1/depth^2 as 'error' to the model output at K_v=0.35, and we obtained a situation where the inference was more biased and less accurate by including model discrepancy than without model discrepancy (see Figure 5(d)). Hence, in some situations, adding model discrepancy may make inference about calibration parameters worse.

6 Summary

We develop and apply an approach for inferring calibration parameters by combining information from observations and climate model output for multiple tracers while taking into account multiple sources of uncertainty. We find that, as one would expect, combining information from multiple spatial fields results in tighter posterior distributions for the climate model parameter. We studied the impact of modeling the model discrepancy and observation error. Based on our study, we find that it is important to include a model discrepancy term, and modeling the discrepancy via a zero mean Gaussian process seems to be the safest approach to guard against bias and overfitting. These results corroborate, in the spatial output setting, the conclusions of Bayarri et al. (2007b). We note, however, that when the computer model is a poor representation of reality, the resulting inference may be more biased when model discrepancy is included. Our study suggests that estimating the emulator spatial variance in a fully Bayesian framework appears to simply reflect the uncertainty from the prior distribution of the emulator spatial variance to the posterior distribution of the calibration parameter. Hence, we recommend using a plug-in estimate of the emulator spatial variance unless there is clear prior information for these parameters.

A possible issue with calibration in general is the known confounding between the calibration parameters and the model discrepancy parameters. We also note that the climate parameter inference obtained here is based on heavily aggregated data, which neglects local spatial effects and small-scale behavior across the ocean, and uses a simple covariance function. Hence, computationally tractable approaches, for large datasets, such as those explored

in Bhat et al. (2009), may provide more scientifically rigorous conclusions than those reported here.

Acknowledgements

We thank Andreas Schmittner for providing us the output of the published runs in Schmittner et al. (2009). The authors are grateful to Susie Bayarri, Jim Berger, and others in the 'Interaction of Deterministic And Stochastic Models' working group in the Statistical and Applied Mathematical Sciences (SAMSI) research program on Space-time Analysis in Environmental Mapping, Epidemiology and Climate Change for helpful discussions. We acknowledge support from the National Science Foundation. Opinions, findings, and conclusions expressed in this work are those of the authors alone, and do not necessarily reflect the views of the NSF.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis.* Wiley Series in Probability and Statistics.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Models and Analysis for Spatial Data*. Chapman & Hall CRC.

Bayarri, M., Berger, J., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R., Paulo, R., Sacks, J., and Walsh, D. (2007a). Computer model validation with functional output. *The Annals of Statistics*, 35(5):1874–1906.

Bayarri, M., Berger, J., Higdon, D., Kennedy, M., Kottas, A., Paulo, R., Sacks, J., Cafeo, J., Cavendish, J., Lin, C., et al. (2007b). A Framework for Validation of Computer Models. *Technometrics*, 49(2):138–154.

Bayarri, M., Berger, J., Kennedy, M., Kottas, A., Paulo, R., Sacks, J., Cafeo, J., Lin, C., and Tu, J. (2009). Predicting Vehicle Crashworthiness: Validation of Computer Models for Functional and Hierarchical Data. *Journal of the American Statistical Association*, 104(487):929–943.

Bhat, K., Haran, M., Tonkonojenkov, R., and Keller, K. (2009). Inferring likelihoods and climate system characteristics from climate models and multiple tracers. Technical report, Pennsylvania State University, Department of Statistics.

Chen, M., Shao, Q., and Ibrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer.

Cressie, N. A. (1993). *Statistics for Spatial Data*. John Wiley & Sons, New York, 2nd. edition.

Flegal, J., Haran, M., and Jones, G. (2008). Markov Chain Monte Carlo: Can We Trust the Third Significant Figure? *Statist. Sci*, 23(2):250–260.

Forest, C., Stone, P., Sokolov, A., Allen, M., and Webster, M. (2002). Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science*, 295(5552):113–117.

Gneiting, T. (2002). Nonseparable, Stationary Covariance Functions for Space-Time Data. *Journal of the American Statistical Association*, 97(458):590–601.

Goes, M., Urban, N., Tonkonojenkov, R., Haran, M., and Keller, K. (2009). The skill of different ocean tracers in reducing uncertainties about projections of the Atlantic Meridional Overturning Circulation. Technical report, Pennsylvania State University, Department of Geosciences.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). Computer Model Calibration Using High-Dimensional Output. *Journal of the American Statistical Association*, 103(482):570–583.

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101:1537–1547.

Kennedy, M. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(3):425–464.

Key, R., Kozyr, A., Sabine, C., Lee, K., Wanninkhof, R., Bullister, J., Feely, R., Millero, F., Mordy, C., and Peng, T. (2004). A global ocean carbon climatology: Results from Global Data Analysis Project (GLODAP). *Global Biogeochem. Cycles*, 18(4).

Liu, F., Bayarri, M., and Berger, J. (2009). Modularization in Bayesian Analysis, with Emphasis on Analysis of Computer Models. *Bayesian Analysis*, 4(1):119–150.

McCarthy, R., Bower, F., and Jesson, J. (1977). The fluorocarbon-ozone theoryI. Production and releaseworld production and release of CCl3F and CCl2F2 (fluorocarbons 11 and 12) through 1975. *Atmospheric Environment (1967)*, 11(6):491–497.

O'Hagan, A. (2009). Reification and true parameters: Discussion of Goldstein and Rougier. *Journal of Statistical Planning and Inference*, 139(3):1240–1242.

Rougier, J. (2008a). Comment on article by Sanso et al. *Bayesian Analysis*, 3(1):45–56.

Rougier, J. (2008b). Efficient emulators for multivariate deterministic functions. *Journal of Computational and Graphical Statistics*, 17(4):827–843.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments (C/R: P423-435). *Statistical Science*, 4:409–423.

Sansò, B., Forest, C., and Zantedeschi, D. (2008). Inferring climate system properties using a computer model. *Bayesian Analysis*, 3(1):1–38.

Schmittner, A., Urban, N., Keller, K., and Matthews, D. (2009). Using tracer observations to reduce the uncertainty of ocean diapycnal mixing and climate carbon-cycle projections. *Global Biogeochemical Cycles*.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag Inc.

Weaver, A., Eby, M., Wiebe, E., Bitz, C., Duffy, P., Ewen, T., Fanning, A., Holland, M., MacFadyen, A., Matthews, H., et al. (2001). The UVic Earth System Climate Model: Model description, climatology, and applications to past, present and future climates. *Atmosphere-Ocean*, 39(4):361–428.

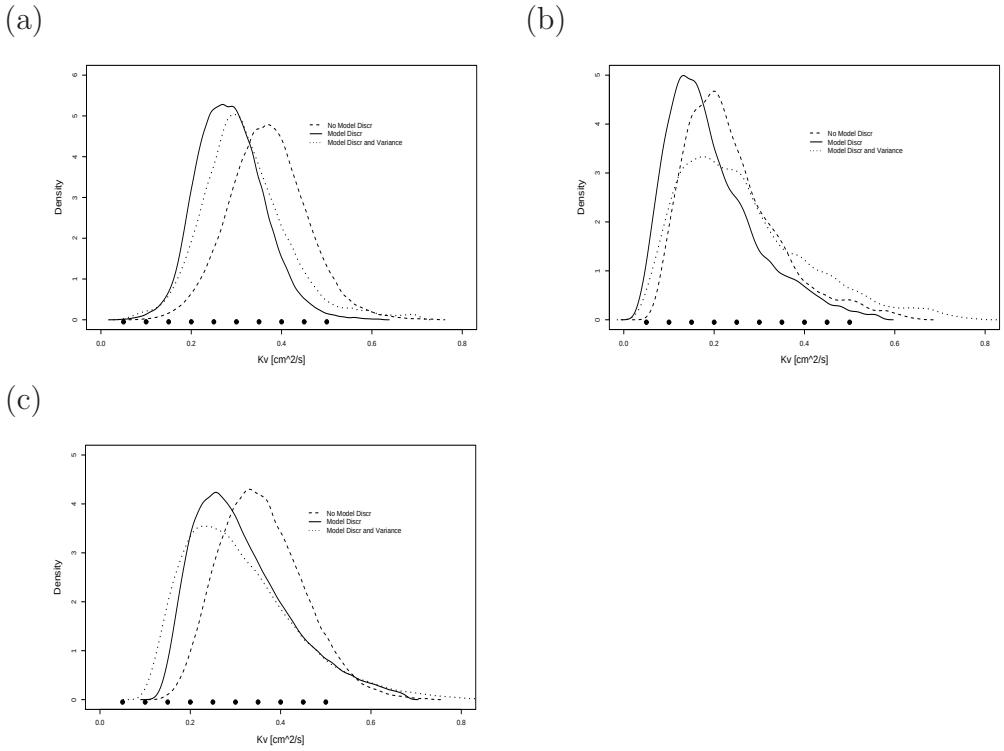

Figure 1: Posterior for K_v for the three tracers: (i) excluding model error (dotted black line), (ii) including model error (solid black line), (iii) including model error and estimation of emulator spatial variances (dotted-dashed black line). Top left: Δ^{14}C, Top right: CFC11, Bottom left: T.

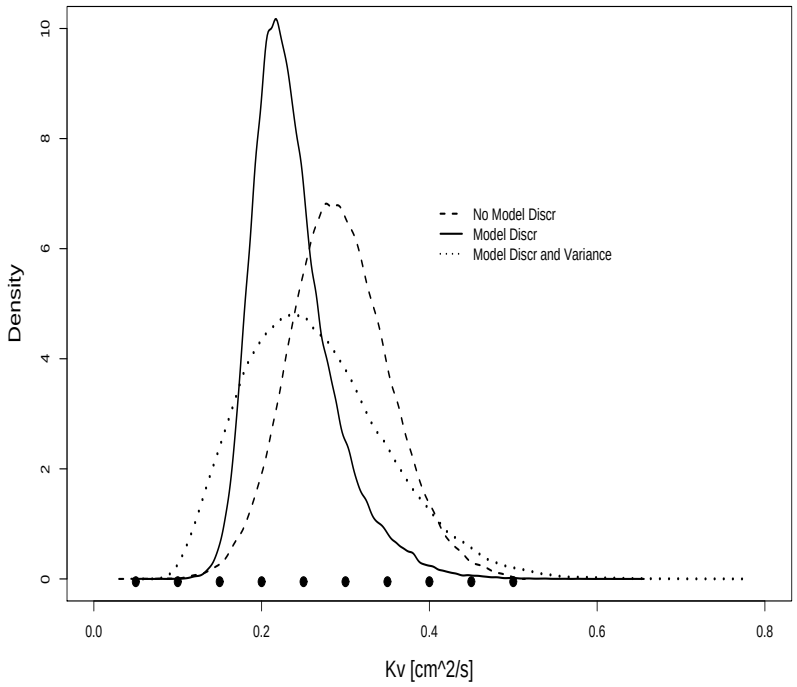

Figure 2: Posterior probability distribution for vertical diffusivity (K_v) for all three tracers jointly: (i) excluding model error (dotted black line), (ii) including model error (solid black line), (iii) including model error and estimation of emulator spatial variances in second stage (dotted-dashed black line).

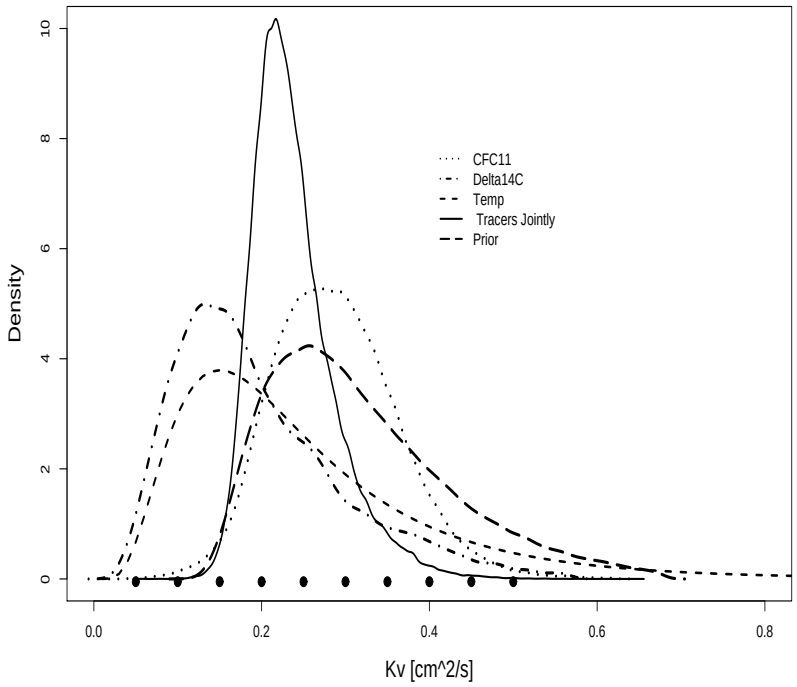

Figure 3: Log Normal Prior (dotted black line) and posterior probability distributions of vertical diffusivity (K_v) with model discrepancy term included using (i) CFC11 tracer (dotted black line), (ii) Δ^{14}C tracer (dotted-dashed black line), (iii) T tracer (dotted-dashed black line), (iv) all three tracers jointly (solid black line).

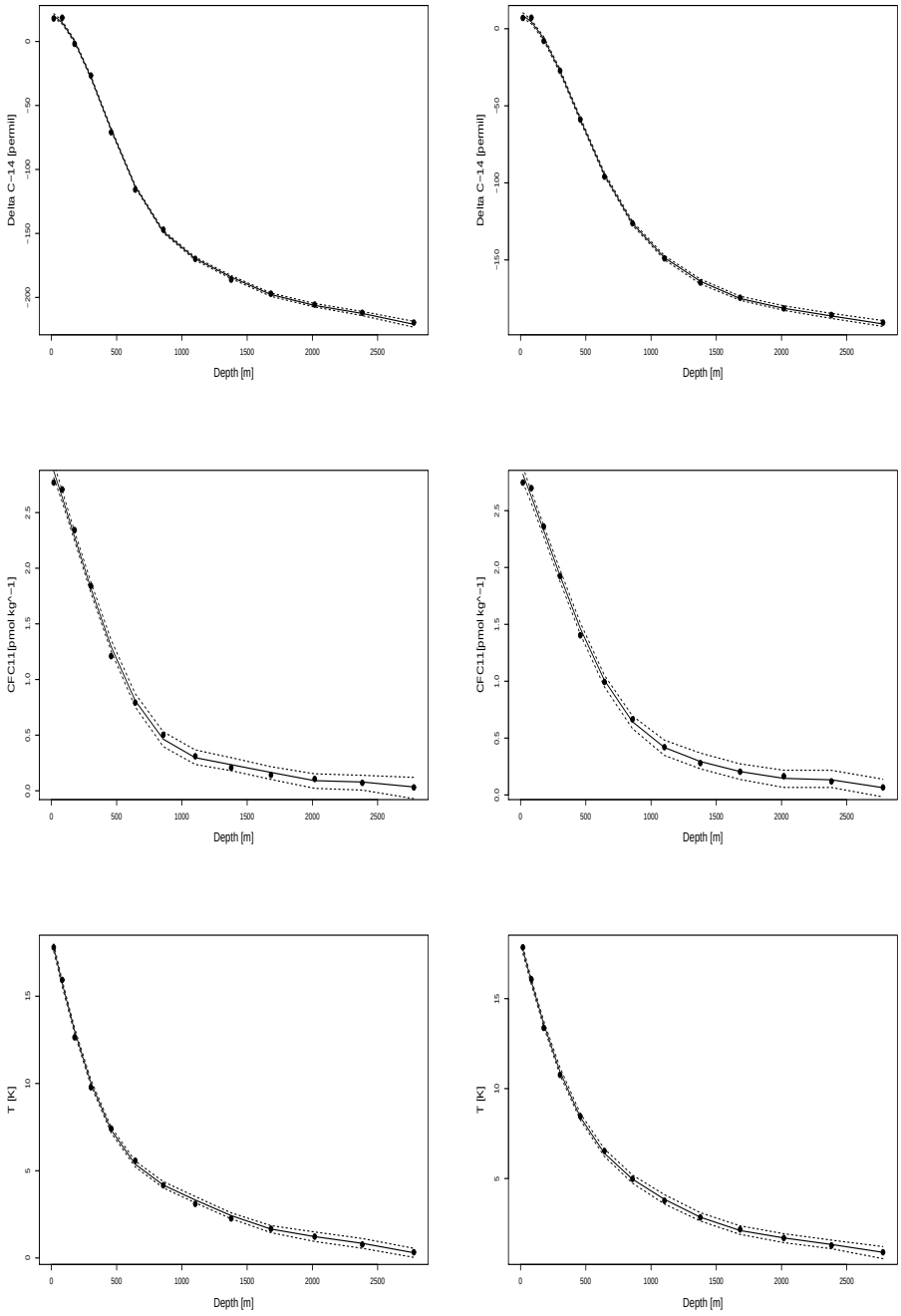

Figure 4: Cross-validation plots of predictions at K_v=0.2 and 0.4 with model output at K_v value held out. Black dots: model output, solid black lines: predictions, dotted black lines: 95 % confidence regions. Left: K_v=0.2, Right: K_v=0.4). Top row: Δ^{14}C, Middle row: CFC11, Bottom row: T.

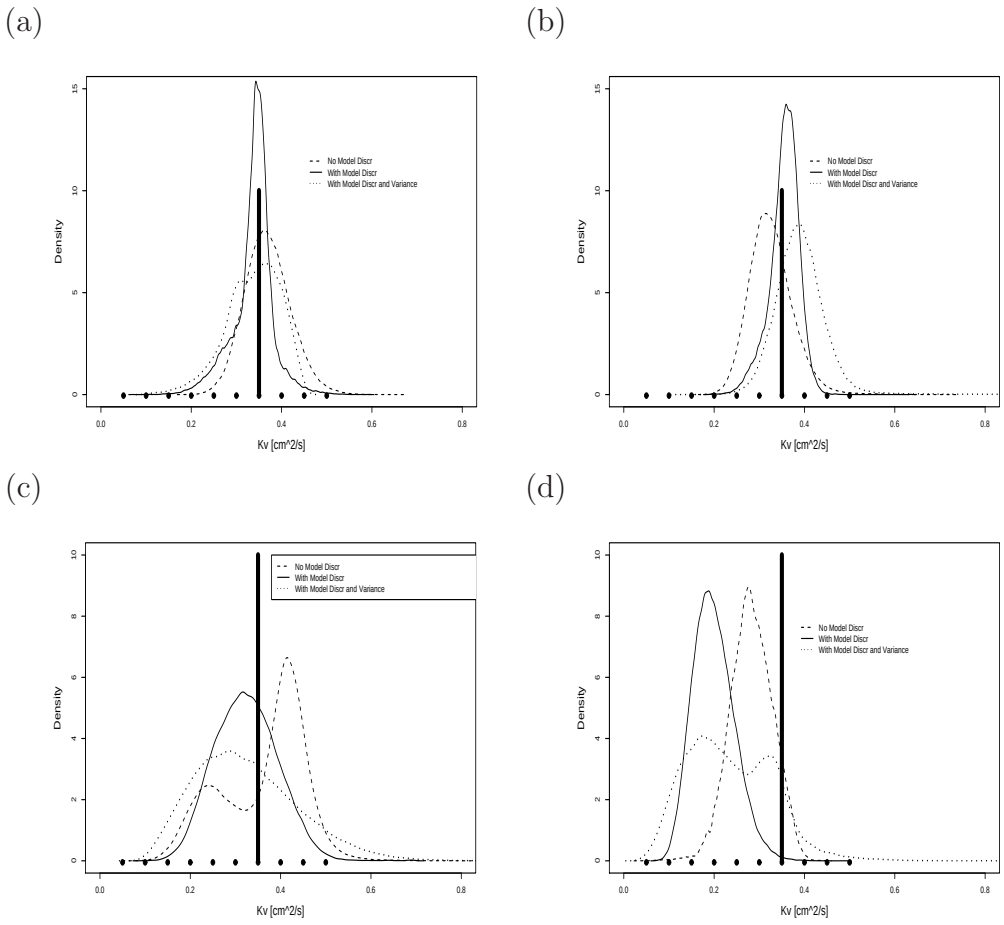

Figure 5: Posterior probability distribution for vertical diffusivity (K_v) for the three tracers jointly: (i) excluding model error (dotted black line), (ii) including model error (solid black line), (iii) including model error and estimation of emulator spatial variances in second stage (dotted-dashed black line) for simulation experiments. Top left: 'Observations' as model output at K_v=0.35 with no error, Top right: simulated iid error added, Bottom left: simulated model discrepancy. Bottom right: error function proportional to 1/depth^2 added. True parameter value of K_v=0.35 denoted by thick black line.