Collaborative Project: The problem of bias in defining uncertainty in computationally enabled strategies for data-driven climate model development

Final Report

Gabriel Huerta (co-PI)
Department of Mathematics and Statistics
The University of New Mexico
311 Terrace NE
Albuquerque, NM 87131
Phone: (505) 277-4613
Fax: (505) 277-5505
e-mail: ghuerta@stat.unm.edu

In collaboration with: Charles S. Jackson (PI)
Institute for Geophysics
The University of Texas at Austin (lead institution)

Final Report

1 Project Synopsis

The objective of the project is to develop strategies for better representing scientific sensibilities within statistical measures of model skill that then can be used within a Bayesian statistical framework for data-driven climate model development and improved measures of model scientific uncertainty. One of the thorny issues in model evaluation is quantifying the effect of biases on climate projections. While any bias is not desirable, only those biases that affect feedbacks affect scatter in climate projections. The effort at the University of Texas is to analyze previously calculated ensembles of CAM3.1 with perturbed parameters to discover how biases affect projections of global warming. The hypothesis is that compensating errors in the control model can be identified by their effect on a combination of processes and that developing metrics that are sensitive to dependencies among state variables would provide a way to select version of climate models that may reduce scatter in climate projections. Gabriel Huerta at the University of New Mexico is responsible for developing statistical methods for evaluating these field dependencies. The UT effort will incorporate these developments into MECS, which is a set of python scripts being developed at the University of Texas for managing the workflow associated with data-driven climate model development over HPC resources. We also will consult with Peter Gleckler (LLNL), Mark Taylor (SNL), and Rich Neale (NCAR) to provide additional insight into issues concerning CAM5 model development and multi-model evaluation of the AR5 archive. This report reflects the main activities at the University of New Mexico where the PI (Huerta) and the Postdocs (Nosedal, Hattab and Karki) worked on the project.

2 Primary Research and Development Activities

2.1 A modified skill metric for climate models using Gaussian Markov Random Fields

2.1.1 Introduction

A new metric for climate model evaluation has been developed that potentially mitigates some of the limitations that exists for observing and representing field and space dependencies of climate phenomena. Traditionally such dependencies have been ignored when climate models have been evaluated against observational data, which makes it difficult to assess whether any given model is simulating observed climate for the right reasons. The new metric uses Gaussian Markov Random Fields for estimating field and space dependencies within a first order grid point neighborhood structure. We illustrate the ability of Gaussian Markov Random Fields to represent empirical estimates of field and space covariances using 'witch hat' graphs. We further use the new metric to evaluate the tropical response of a climate model (CAM3.1) to changes in two parameters important to its representation of

cloud and precipitation physics. Overall, the inclusion of dependency information did not alter significantly the recognition of those regions of parameter space that best approximated observations. However there were some qualitative differences in the shape of the response surface that suggest how such a measure could affect estimates of model uncertainty.

Within the climate assessment community, there is an interest to develop metrics of how well simulations reproduce observed climate for purposes of comparing models, driving model development, and evaluating model prediction uncertainties [4, 6, 7, 5, 11]. Nevertheless, a certain level of skepticism exists about whether a scalar metric can be sufficiently informative for these purposes. Climate phenomena involve interactions of multiple quantities on a wide range of time and space scales from minutes to decades (and longer) and from meters to planetary scales. Thus it can be challenging to summarize what is physically meaningful. The most common approach to climate model evaluation among climate scientists is to display maps of long-term means of well-known quantities (e.g. temperature, sea-level pressure, precipitation) whose distribution is familiar and well understood in order to identify the source of model error. The Taylor metric that is often generated as part of model evaluation is based on spatial means of squared grid point errors for individual quantities [8]. Such measures neglect field and space dependencies and thus may be insensitive to mechanisms giving rise to model errors. There is a need to develop metrics that can evaluate whether a model is capturing observed space and field relationships sufficiently well [2]. The hope is that by accounting for relationship information within climate model metrics, they will prove to be more useful for scientific evaluation.

Given that there is only a limited amount of observations available to quantify field and space relationships of climate phenomena, data assimilation is the most common approach to fill in gaps in the observational record of a climate model [9]. While assimilation data products help solve some aspects of the problem of how one compares point measurements to the scales resolved by climate models, these data products include the space and field dependencies of the model that was used to assimilate the data. Here we introduce a new kind of metric based on Gaussian Markov Random Fields that only needs limited data to decipher space and field dependencies of climate phenomena.

We define a new Z-test statistic, alternatively referred to as a log-likelihood or cost for assessing the significance of a discrepancy between model output and observations. The statistic makes use of Gaussian Markov Random Fields to estimate field and space dependencies that exist within gridded climate model output that can be assessed against space and field dependent observational data. The matrix form of the test statistic is given by:

$$\mathbf{v}^T\mathbf{S}^{-1} \otimes (\alpha\mathbf{I} + (1-\alpha)\mathbf{Q})\mathbf{v} \tag{1}$$

where $\mathbf{v}$ is the vector of differences between model output and observations with a length given by the product of the number of observational fields and number of grid points, $n_{obs} \times n_{pts}$, α is a scalar with a value close to zero, $\mathbf{I}$ stands for an identity matrix (a diagonal matrix of ones) of a dimension corresponding to $\mathbf{v}$ and $\mathbf{Q}$ is a precision matrix of dimension $n_{pts} \times n_{pts}$ from a Gaussian Markov Random Field (GMRF) induced by a first order neighborhood structure. This cost function captures field dependencies through $\mathbf{S}^{-1}$ which is a matrix

$n_{obs} \times n_{obs}$ where each of its elements represents a spatial-average of variances and covariances between fields. The spatial dependency between grids is approximated through $\mathbf{Q}$. The quantity α could be interpreted as a weight of the spatial relationship between grid cells. The Kronecker product $\otimes$ provides a means for associating the different matrix dimensions of the metric, essentially combining its field and space components.

The following sections explain, test, and provide examples of how various components of equation (1) work. Section 2.1.2 gives a brief introduction to GMRFs. This section will allow us to understand how $\mathbf{Q}$ is obtained and the information that it provides about spatial dependency between grid cells. In this section we also define and discuss Kronecker products, and how to use this concept to generalize GMRFs ideas to deal with more than one field. Section 2.1.3 introduces a graph for testing the extent to which equation (1) captures observed variance-covariances of tropical temperature, precipitation, sea level pressure, and upper level winds. Finally, in Section 2.1.5, we consider the field and space dependencies that are captured by the GMRF-based metric within the response of an atmospheric general circulation model CAM3.1 to two model parameters important to cloud and precipitation physics. What we learned in general is that including the space and field dependencies provides some qualitatively different perspectives about which model configurations are more similar to what is observed. For the example we consider, the effects of space dependencies turn out to be more critical than field dependencies.

2.1.2 Gaussian Markov Random Fields (GMRFs)

A Gaussian Markov Random Field (GMRF) is a special case of a multivariate normal distribution, one that satisfies additional properties related to conditional independence. The density of a normal random vector $\mathbf{x} = (x_1, x_2, ..., x_n)^T$ (where T denotes the operation of transposing a column to a row), with mean μ ($n \times 1$ vector) and covariance matrix $\mathbf{\Sigma}$ ($n \times n$ matrix), is

$$f(\mathbf{x}) = (2\pi)^{-n/2} |\mathbf{\Sigma}|^{-\frac{1}{2}} exp\left\{ -\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu) \right\} \tag{2}$$

Here, $\mu_i = E(x_i)$, $\Sigma_{ij} = Cov(x_i, x_j)$, and $\Sigma_{ii} = Var(x_i) > 0$. All eigenvalues of $\mathbf{\Sigma}$ must be greater than zero, otherwise $\mathbf{\Sigma}$ becomes a singular matrix and does define a valid multivariate normal distribution. It can also be shown that if all eigenvalues of $\mathbf{\Sigma}$ are positive then all eigenvalues of $\mathbf{\Sigma}^{-1}$ are also greater than zero. We define $\mathbf{Q} = \mathbf{\Sigma^{-1}}$ and refer to $\mathbf{Q}$ as the precision matrix, and denote $\mathbf{x} \sim \mathbf{N}(\mu, \mathbf{Q})$ to represent that $\mathbf{x}$ follows a multivariate normal distribution with vector mean μ and precision matrix $\mathbf{Q}$.

Precision matrix of a GMRF

The precision matrix $\mathbf{Q}$ is an operator for obtaining information about dependencies among neighboring grid cells. Although $\mathbf{Q}$ is sparse, its inverse, as a model for the covariance matrix $\mathbf{\Sigma}$, presumes all grid points are conditionally dependent. $\mathbf{Q}$ needs to be constructed such that it:

- Reflects the kind of spatial dependency we assume our data has.
- Yields a legitimate covariance matrix, $\mathbf{\Sigma}$, i.e. symmetric and positive definite, so that it can be used to compute a likelihood function.

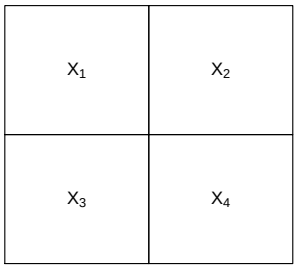


Figure 1: Graphical representation of 2×2 lattice and elements of $\mathbf{x}$.

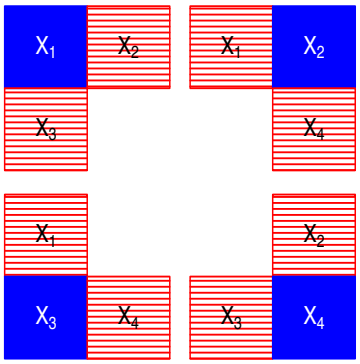


Figure 2: Neighbors of $x_1,\ x_2,\ x_3$ and x_4

Consider $\mathbf{x}$, a vector of measurements on a 2×2 lattice, as represented in Figure 1. Assume a neighborhood structure between the four elements of $\mathbf{x}$. In Figure 2, the neighbors for each element of $\mathbf{x}$ are defined graphically. Given the neighborhood structure shown in Figure 2, the precision matrix that works for this problem is

$$\mathbf{Q} = \begin{pmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & 0 & -1 \\ -1 & 0 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{pmatrix}$$

which follows these rules,

- $\mathbf{Q}_{ij} = -1$, if x_i and x_j are neighbors.
- $\mathbf{Q}_{ij} = 0$, if x_i and x_j are not neighbors.
- $\mathbf{Q}_{ii}$ gives the total number of neighbors of x_i.

While the implementation of GMRF is simple, the theory and mathematics are rather involved. It may also not be immediately clear to a physical scientist that such a simple specification, where only relationships among neighboring grid cells are taken into account, would be sufficient to quantify correlated quantities across large distances. The mathematics of working with precisions allows one to infer the net effect of long distance relationships through relationship information that exists among neighboring cells. While the GMRF approach does not include information about particular teleconnection structures such as ENSO, the approach is sensitive to how changes in large scale conditions induce local covariances across multiple fields within the entire domain. In this way teleconnections are represented through a conditional dependence.

A problem arises in that one of the eigenvalues of the $\mathbf{Q}$ matrix is 0, which implies that this definition of the precision matrix does not induce an invertible covariance matrix. This problem is solved by using $\alpha\mathbf{I} + (1-\alpha)\mathbf{Q}$, instead of $\mathbf{Q}$. If α is small, the neighborhood structure remains, basically, unchanged. A subsequent section describes our approach to specifying a value for α.

Generalizing concepts to deal with multiple fields

The generalization of $\mathbf{Q}$ to handle multiple fields will be illustrated by an example using $\mathbf{x}$ and $\mathbf{y}$ which represent observations for two different fields of interest. These observations are taken on a 2×2 lattice. First, $\mathbf{x}$ and $\mathbf{y}$ are combined to form one vector $\mathbf{v}$ as follows: $\mathbf{v}^T = (x_1, x_2, x_3, x_4, y_1, y_2, y_3, y_4)$. The average covariances among these observations can be represented by a 2×2 matrix between the first field, $\mathbf{x}$, and the second field, $\mathbf{y}$:

$$\mathbf{S} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

where $Var(\mathbf{x}) = \sigma_{11}$, $Var(\mathbf{y}) = \sigma_{22}$, and $Cov(\mathbf{x}, \mathbf{y}) = \sigma_{12}$. Recalling that the correlation between fields 1 and 2 is defined as: $\rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$, it is easy to show that the inverse of $\mathbf{S}$ is

$$\mathbf{S}^{-1} = \begin{pmatrix} \frac{1}{\sigma_{11}(1-\rho^2)} & \frac{-\rho}{(1-\rho^2)\sqrt{\sigma_{11}\sigma_{22}}} \\ \frac{-\rho}{(1-\rho^2)\sqrt{\sigma_{11}\sigma_{22}}} & \frac{1}{\sigma_{11}(1-\rho^2)} \end{pmatrix} = \begin{pmatrix} S_{11}^{-1} & S_{12}^{-1} \\ S_{12}^{-1} & S_{22}^{-1}. \end{pmatrix}$$

If we consider the Kronecker product in Equation 1 when $\alpha = 0$,

$$\mathbf{S}^{-1} \otimes \mathbf{Q} = \begin{pmatrix} S_{11}^{-1}\mathbf{Q} & S_{12}^{-1}\mathbf{Q} \\ S_{21}^{-1}\mathbf{Q} & S_{22}^{-1}\mathbf{Q} \end{pmatrix}$$

then

$$\mathbf{v}^T\mathbf{S}^{-1} \otimes \mathbf{Q}\mathbf{v} = S_{11}^{-1}\mathbf{x}^T\mathbf{Q}\mathbf{x} + S_{12}^{-1}\mathbf{y}^T\mathbf{Q}\mathbf{x} + S_{21}^{-1}\mathbf{x}^T\mathbf{Q}\mathbf{y} + S_{22}^{-1}\mathbf{y}^T\mathbf{Q}\mathbf{y}.$$

In this last expression, one can see that the inverse of $\mathbf{S}$ in combination with the Kronecker product with $\mathbf{Q}$ includes terms involving cross products between fields.

2.1.3 A test of GMRF estimates of variance

GMRF provide a way to approximate field and space dependencies contained in the inverse covariance matrix $\mathbf{\Sigma}^{-1}$ of equation (1) by its GMRF equivalent $\mathbf{S}^{-1} \otimes (\alpha\mathbf{I} + (1-\alpha)\mathbf{Q})$. In this section, we will test how well GMRF are able to reproduce observed space and field dependencies. This may be achieved by comparing field and spatial variance and covariance estimates obtained from the inverse of the GMRF equation with those obtained empirically from observational data. It turns out this comparison is sensitive to the value that is selected for α. Fortunately, the optimal choice of α depends only on geometric considerations of the neighborhood model that is used for GMRF and the number of grid points in the fields and not the properties of the field data. We introduce a 'witch hat' graph that provides a compact summary of variance-covariance information between these two methods in order to show that GMRF does a reasonable job approximating observed field and space relationships.

2.1.4 Finding an appropriate value of α

In the effort to compare space and field dependences approximated by GMRF with empirical estimates we need to determine an optimal value for α. In order to carry out this comparison, we need to find the inverse of $\mathbf{S}^{-1} \otimes (\alpha\mathbf{I} + (1-\alpha)\mathbf{Q})$, our proposed precision matrix based on GMRF. Using results of Kronecker products, we have that $[\mathbf{S}^{-1} \otimes (\alpha\mathbf{I} + (1-\alpha)\mathbf{Q})]^{-1} = \mathbf{S} \otimes (\alpha\mathbf{I} + (1-\alpha)\mathbf{Q})^{-1}$. Let $\mathbf{Q}^* = (\alpha\mathbf{I} + (1-\alpha)\mathbf{Q})^{-1}$, then $\mathbf{S} \otimes \mathbf{Q}^*$ for two fields can be written as

$$\begin{pmatrix} S_{11}\mathbf{Q}^* & S_{12}\mathbf{Q}^* \\ S_{12}\mathbf{Q}^* & S_{22}\mathbf{Q}^* \end{pmatrix}.$$

If $n = n_{pts}$ is the total number of grid points of the lattice, $\mathbf{S} \otimes \mathbf{Q}^*$ is a $(2 \times n) \times (2 \times n)$ covariance matrix. Note that each element of $diag(S_{ij}\mathbf{Q}^*)$ contains the estimated variance or covariance at each grid point for fields i and j using a GMRF where i can be equal to j. If we average these estimates across the whole lattice, we obtain G_{ij}, the GMRF estimate of the variance or covariance. Therefore,

$$G_{ij} = \frac{S_{ij} \sum_{k=1}^{n} Q^*_{kk}}{n} = \frac{S_{ij} tr(\mathbf{Q}^*)}{n} \tag{3}$$

where $tr(\mathbf{Q}^*)$ denotes the trace of $\mathbf{Q}^*$ and Q^*_{kk} are its diagonal elements. We will now select a value for α that allows the GMRF estimate for field variances and covariances to be equal, on average, to what has been calculated for $\mathbf{S}$. In order to achieve this, G_{ij} needs to equal S_{ij}. Satisfying this condition is equivalent to finding the solution for

$$\frac{tr(\mathbf{Q}^*)}{n} = 1. \tag{4}$$

It may not be so obvious what the diagonal elements of $\mathbf{Q}^*$ are, however, one can use the fact that for any matrix $\mathbf{A}$ that admits a Singular Value Decomposition, $tr(\mathbf{A})$ is equal to sum of its eigenvalues. In our case, if the eigenvalues of $\mathbf{Q}$ are $\lambda_1, \lambda_2, ..., \lambda_n$, the eigenvalues of $\alpha\mathbf{I} + (1-\alpha)\mathbf{Q}$ are $\alpha + (1-\alpha)\lambda_1, \alpha + (1-\alpha)\lambda_2, ..., \alpha + (1-\alpha)\lambda_n$. The eigenvalues of $\mathbf{Q}^* = (\alpha\mathbf{I} + (1-\alpha)\mathbf{Q})^{-1}$ are $(\alpha + (1-\alpha)\lambda_1)^{-1}, (\alpha + (1-\alpha)\lambda_2)^{-1}, ..., (\alpha + (1-\alpha)\lambda_n)^{-1}$. This implies that in order to satisfy Equation 4, we need to find α from

$$f(\alpha) = \sum_{i=1}^{n} \frac{1}{n(\alpha + (1-\alpha)\lambda_i)} = 1. \tag{5}$$

Figure 3 shows the relationship between various values of α and $f(\alpha)$. The eigenvalues used to obtain this figure correspond to a precision matrix, $\mathbf{Q}$, for a GMRF induced by a first order neighborhood structure and considering a 128×22 lattice (which is the dimension of our data). From the figure we can see that the curve crosses the value of 1 when α is close to 0. By using linear interpolation, we determine that α is approximately 0.0024.

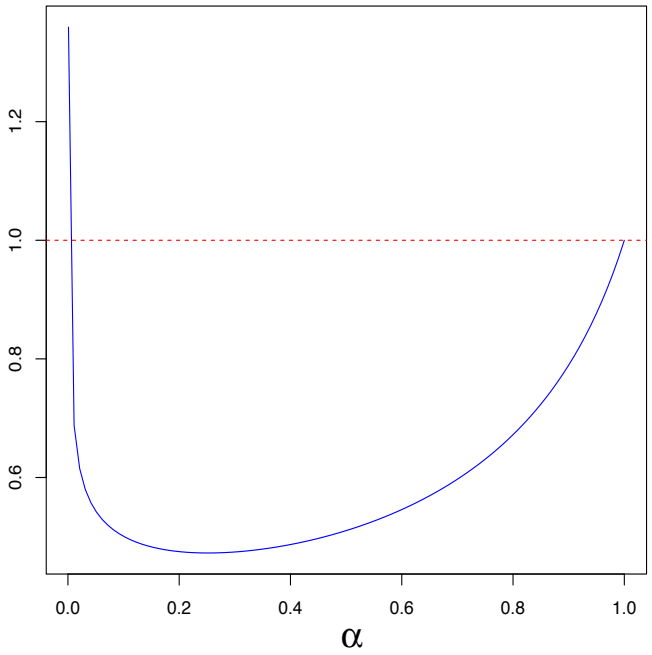


Figure 3: α vs $f(\alpha)$.

Witch hat comparison test

To illustrate any differences that may exist between empirical estimates of the covariance matrix $\mathbf{\Sigma}$ and its GMRF equivalent $\mathbf{S} \otimes (\alpha\mathbf{I} + (1-\alpha)\mathbf{Q})^{-1}$, we rely on a graph that shows

the spatial average grid point variance and covariances as a function of distance for cells and their neighbors. We compute the average entries of the covariance matrix corresponding to each grid cell and the corresponding element to the north (for the positive distances) or to the south (for the negative distances) relative to the main diagonal of the matrix. The zero distance case is the average of variances of the main diagonal. Alternatively, we can produce a graph that considers the east and west directions. On average, covariances decrease with distance making the graph have the shape of a witch's hat. This graph is symmetric because covariance matrices are symmetric.

Figure 4 shows a witch hat test of estimated variances for air temperatures simulated by the Community Atmosphere Model version 3.1 (CAM3.1). The variances are estimated from 15 samples of two year mean summertime temperatures. Setting $\alpha = 1$ provides a solution to equation (5), however, this will shut down the effect of $\mathbf{Q}$ and only the variances at the reference point (lag 0) will be well estimated. On the other hand, when $\alpha = 0.024$, we allow $\mathbf{Q}$ to play more of a role which results in a better representation of covariances at neighboring points (lags different of zero).

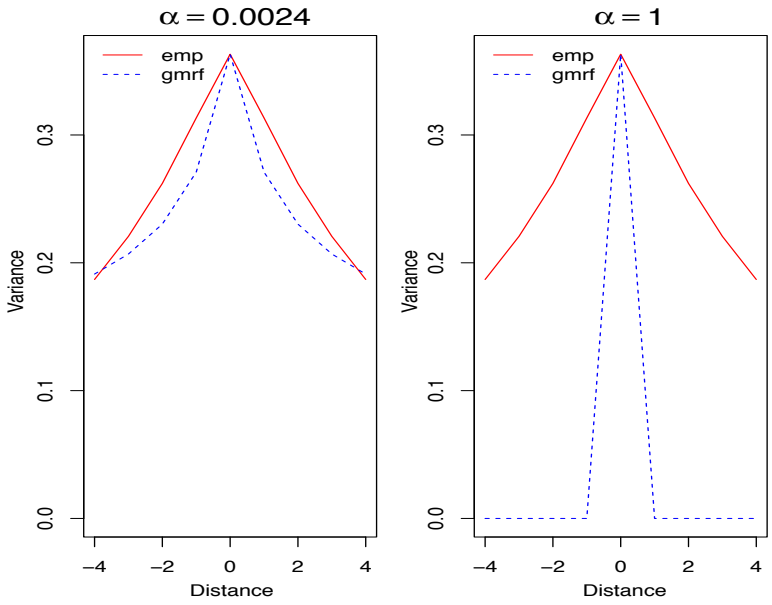


Figure 4: 'Witch hat' graphs for air temperature on a 128×22 lattice of the tropics from 30°S to 30°N.

2.1.5 Climate response to uncertain parameters

In this section we show how inclusion of field and space dependencies using GMRF affect comparisons of the Community Atmosphere Model (CAM3.1) [3] with observations. We con-

sider CAM3.1's response to to changes in parameter ke, which controls rain drop evaporation rates, and parameter $c0$, which controls precipitation efficiency through conversion of cloud water to rain water. For this comparison we only consider the response for the June, July, and August (JJA) seasonal mean between 30°S to 30°N on four variables including 2 meter air temperature (TREFHT), 200-millibar zonal winds (U), sea level pressure (PSL), and precipitation (PRECT). Experiments with CAM3.1 use observed climatological sea surface temperatures and sea ice extents. Each experiment with CAM3.1 is 32-years in duration.

The observational data that is used to evaluate the model comes from a reanalysis product ECMWF-ERA interim [10] for 2 m air temperature, 200-millibar zonal winds, and sea level pressure and GPCP [1] for precipitation. We make use of approximately 30 years of JJA mean fields between 1979 and 2009. For constructing S, we calculate variances from 2-year means (i.e. 15 samples).

A total of 64 experiments were completed, varying each of the two parameters within an 8×8 lattice. For each experiment we calculate three versions of GMRF-based cost (equation 1); The first version is the traditional cost based on the assumption of space and field independence set here by setting the off diagonal components of $\mathbf{S}$ to zero and setting $\alpha = 1$. This approach is similar to what has been done previously for **(author?)** [8]. The second version of evaluating the cost takes field dependencies into account by including all components of $\mathbf{S}$ and setting $\alpha = 1$. The third version for the cost takes field and space dependencies into account by including all components of $\mathbf{S}$ and setting $\alpha = 0.0024$.

The correlation matrix, $\mathbf{R}$, corresponding to the $\mathbf{S}$ matrix of 2-year JJA seasonal mean variances and covariances, as estimated from 30 years of observations, is:

	PRECT	PSL	TREFHT	U
PRECT	1	-0.219	-0.047	0.015
PSL	-0.219	1	-0.313	-0.112
TREFHT	-0.047	-0.313	1	-0.145
U	0.015	-0.112	-0.145	1

The primary field correlations are the values of (-0.313) and (-0.219) occurring between sea level pressure (PSL) and 2 m air temperature (TREFHT), and precipitation (PRECT) and sea level pressure (PSL), respectively. These correlations make physical sense in that precipitation mainly occurs within low pressure storm systems which tends to cool the underlying surface. The other correlations are minimal and there is not a good physical argument supporting their relationship. Figure 5 shows a comparison of the three versions of the GMRF-based cost for the 64 experiments within an 8×8 lattice. All versions of cost result in qualitatively similar results with high and low cost values roughly in the same portions of parameter space. The main difference among the versions of cost comes from taking space dependencies into account within the field-space version. In this case, extremely low values of ke result in higher metric values. Figure 6 examines the reasons for this by graphing the different field contributions to the GMRF-based costs for a slice where $c0 = 0.0035$ which corresponds to one of the rows of the lattice. By plotting everything differenced from metric values at $k\text{e} = 3 \times 10^{-6}$, one can learn that the biggest qualitative difference comes from cost

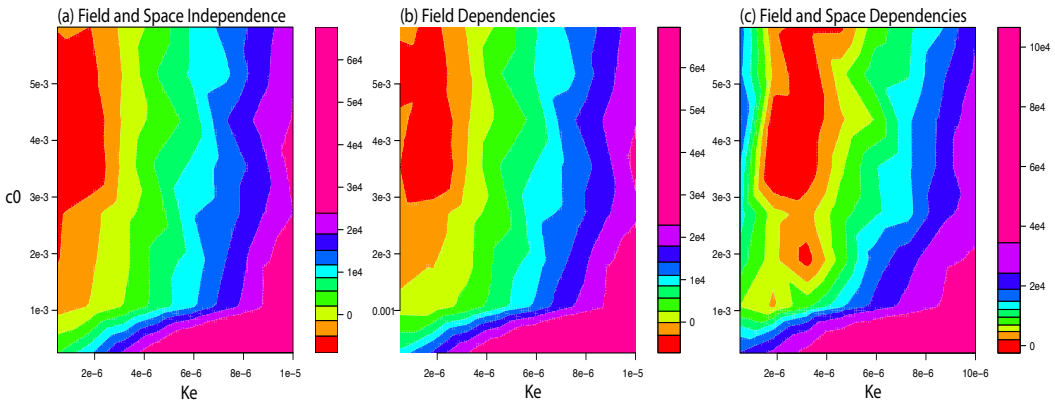


Figure 5: Three versions of the GMRF-based cost as a function of two CAM3.1 parameters ke and $c0$ that assumes the data has (a) field and space independence (Traditional), (b) field dependencies (Field), and (c) field and space dependencies (Field-Space). Each color represents ten percentiles of the cost distribution. The cost is shown relative to the value of the default model configuration.

values associated with 2 m air temperature. Closer inspection of differences between model output and observations of 2 m air temperature (not shown) indicates that the traditional cost is likely reflecting large-scale differences over the southern hemisphere oceans. Inclusion of space dependencies places much greater significance on smaller-scale anomalies occurring over the continents, particularly over the Andes Mountains. This finding is a result of the mathematics of GMRF. It does not imply that the large-scale errors are of lesser scientific importance. It only means that GMRF is less sensitive to large-scale anomalies, perhaps because they are associated with fewer degrees of freedom than highly structured errors. Understanding whether and how these distinctions aid model assessment needs further study. We do find it reassuring that GMRF-based metrics of distance to observations are similar, at least in the example provided, to a traditional metric.

2.1.6 Summary of work on GMRF's measure of model skill

We have developed a new test statistic as a scalar measure of model skill or cost for evaluating the extent to which climate model output captures observed field and space relationships using Gaussian Markov Random Fields (GMRFs). The challenge has been that few observations exit for establishing a meaningful observational basis for quantifying field and space relationships of climate phenomena. Much of the data that is typically used for model evaluation is suspected of having its own relationship biases introduced by the numerical model that is used to synthesize measurements into gridded products. The GMRF-based metric overcomes some of these limitations by considering field and space variations within a neighborhood structure thereby lowering the metric's data requirements. The form of the metric separates space and field dependencies using a Kronecker product that, when multiplied out,

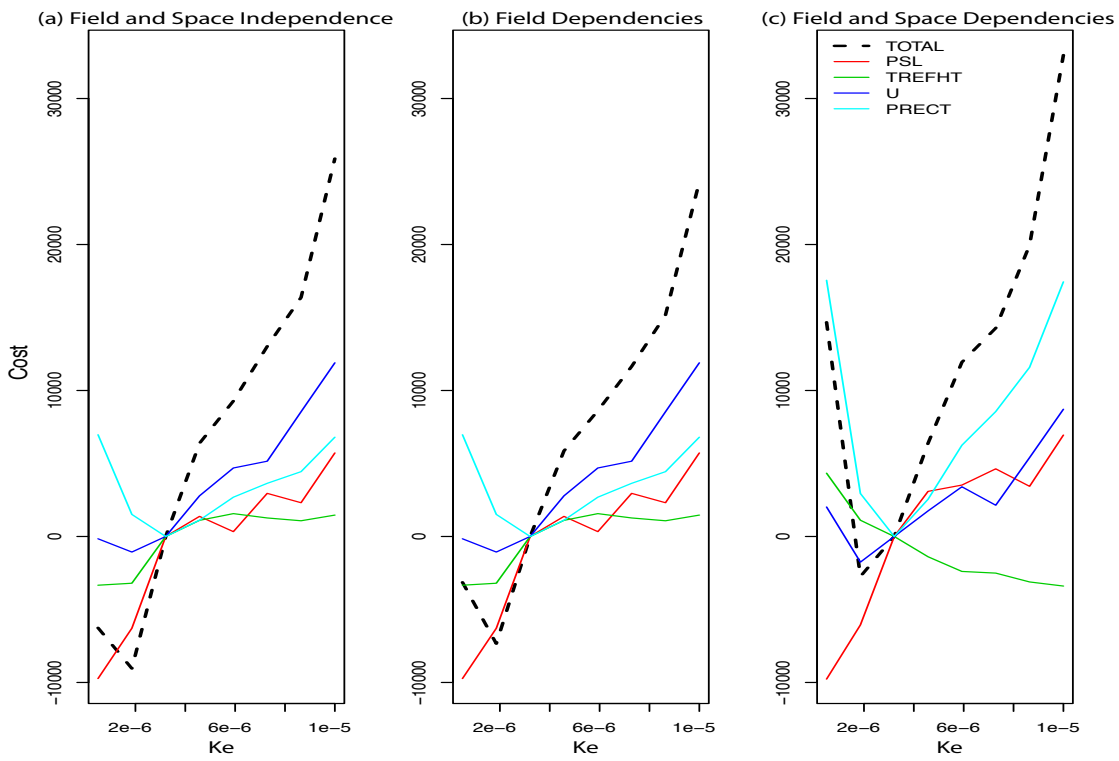


Figure 6: Different field contributions to the GMRF-based costs for a slice of Figure 5 where $c0 = 0.0035$. Cost values are relative to the default parameter setting for ke. Note that total cost (black dashed line) is a weighted sum of field contributions as given by $\mathbf{S}^{-1}$ with contributions from sea level pressure (PSL, red line), 2-m air temperature (TREFHT, green line), 200-millibar zonal winds (U, blue line), and total precipitation (PRECT, cyan line).

has all the terms necessary to represent how different points in space are tied together across multiple field. We also include a scalar α that weights the importance of spatial relationships between grid cells. Its optimal value turns out to be independent of the data type which aids the use of GMRF for comparing model output to data across multiple fields. Using 'witch hat' graphs, we show a first order (nearest neighborhood) structure does an excellent job of capturing empirical estimates of field and space relationships. We have applied three versions of cost that selectively turn on or off field and space dependencies in a climate model (CAM3.1) output against observational products for tropical JJA climatologies for 2 m air temperature, sea level pressure, precipitation, and 200-millibar zonal winds. The results show subtle, but potentially important differences among these versions of the cost which may prove beneficial for selecting models that capture observed climate phenomena for the right reasons.

Climate models can be used to estimate changes in climate as a difference from the modern. While it is acknowledged that climate models can contain significant biases or

errors in simulating what is observed, by using models to estimate climate changes there is an implicit assumption that biases that exist for modern climate remain fixed even as system evolves. In a linear system such differencing will eliminate the biases in predictions. Since important facets of climate can respond non-linearly, the preferred solution to this problem is to minimize model biases during model development by capturing observed phenomena for the right physical reasons. However selecting which of the existing climate model biases are most problematic for predictions is not well established. Current approaches to highlighting the most problematic biases emphasize scientific intuition, which has its limitations.

2.2 A regression between bias and climate sensitivity within a perturbed physics ensemble of CAM3.1

2.2.1 Introduction

The objective here is to identify which model errors are most relevant to a models predictions of global warming response to CO2. This calculation involves creating a statistical summary of a perturbed physics ensemble that through multivariate regression relates model bias to what occurs in a climate change experiment. The results we seek are latitude-longitude maps indicating how various quantities and locations need to be weighted to predict a models response to CO2 forcing. The solution to this problem is not unique insofar as one is potentially relating 10^4 or more predictors (errors at each grid point for any number of fields) to estimate only 10^2 response variables such as the level of global warming for a limited set of models. Thus part of the present objective is to describe a test that can differentiate between robust and non-robust predictors. The calculation is formulated with both frequentist and Bayesian approaches, which result in slightly different interpretations and outcomes.

This calculation is also related to a growing literature concerning the identification of "emergent constraints" which are observables that are predictive of the level of warming that occurs in response to a doubling of atmospheric CO_2 concentrations within a multi-model ensemble (MME) (Piani et al., 2005; Harris et al., 2006; Piani et al., 2007; Volodin, 2007; Huber et al., 2011; Abe et al., 2011; Klocke et al., 2011; Yip et al., 2011; Fasullo and Trendberth, 2012; Ingram 2013, Masson and Knutti, 2013; Sanderson 2013; Caldwell et al. 2014; Sherwood et al. 2014) Identification of emergent constraints is an essential ingredient within probabilistic frameworks for climate predictions because of its role in weighting model output by each model's skill in representing what is important to predictions. The identification of emergent constraints has proven challenging (Caldwell et al., 2014), perhaps because the causes for differences in sensitivity are dependent on structures that are model specific. By focusing on a single model, the Community Atmosphere Model version 3.1 (CAM3.1), our objective is use tools of statistical regression to diagnose those structures that are important determinants of its climate sensitivity. We do not attempt to address here how this calculation can be applied to other models.

2.2.2 Description of ensemble

A 165 member ensemble was generated representing parametric uncertainties resulting from a calibration of a set of 16 parameters important to clouds, convection, and radiation within the Community Atmosphere Model version 3.1 (CAM3.1) at an approximate resolution of 2.8∘ longitude by 2.8∘ latitude with 26 vertical levels. The ensemble design follows Jackson et al. (2008) with differences relating to shorter model experiments (4 years instead of 11 years), an expanded list of uncertain model parameters, and a revised test statistic or "cost function". The updated test statistic is based on quantities, observations, and regions that are used within the development of CAM through a set of Taylor diagnostics within the Atmosphere Model Diagnostic Package (http://www.cgd.ucar.edu/cms/diagnostics/). These metrics emphasize fields between 30S and 30N including 2 m air temperature (Willmott and Matsuura 2000), vertically averaged air temperature (ERA40, Uppala et al. 2005), latent heat fluxes of the ocean (Yu et al. 2008), zonal winds at 300 mb (ERA40, Uppala et al. 2005), longwave and shortwave cloud forcing (CERES2, Loeb et al. 2009), precipitation over land and ocean (GPCP, Adler et al. 2003), sea level pressure (ERA40, Uppala et al. 2005), vertically averaged relative humidity (ERA40, Uppala et al. 2005). Other quantities include Pacific Ocean wind stress between 5S and 5N (ERS-2, Bentamy et al. 1999) and the global mean annual mean radiative balance.

In total 3336 experiments were conducted with fixed sea surface temperatures. 1800 of these experiments were identified by the sampling algorithm to represent the parametric uncertainties. All model configurations selected attained a test statistic skill score equal to or better than the default value and were within 1 Wm^{-2} of radiative balance at the top of the atmosphere. These 1800 were ordered by their test statistic values and every 10th member for a total 180 model configurations were selected for representing parametric uncertainties in global warming experiments. Global warming experiments consist of coupling the model to a slab ocean, integrating the model 50 years under modern (year 2000) CO2 values and again for another 50 years under doubled CO_2 values. Unlike Jackson et al. (2008) heat flux adjustments were only estimated from the default model, which were subsequently used for all global warming experiments. 15 model configurations resulted in runway cooling along the equator which is a result of increased cloud cover over regions where there exists negative heat flux adjustments. Because the slab ocean lacks the correct physics to account for changes in ocean heat transports, such a circumstance is unphysical and is one of the limitations of using AGCM-slab ocean models as a computationally inexpensive way to estimate the equilibrium response of a coupled climate system model to changes in atmospheric CO_2 concentrations. The 165 member ensemble has a range of equilibrium climate sensitivities (the amount of warming in response to a doubling of atmospheric CO_2 concentrations) ranging between 2∘ C and 4∘ C with the maximum likelihood centered at 2.7∘ C. An evaluation of this ensemble relative to other models has been presented in Yokohata et al. (2011) and Yokohata et al. (2013).

2.2.3 Regression Modeling

Consider the linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}_n, \quad E(\boldsymbol{e}_n) = \boldsymbol{0}, \qquad Cov(\boldsymbol{e}_n) = \sigma^2 \boldsymbol{I}_n, \tag{6}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of observable response random values, $\boldsymbol{X}$ is an $n \times p$ known model matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, $\boldsymbol{I}_n$ is an identity matrix of size n, σ^2 is an unknown parameter, $\boldsymbol{e}_n$ is an $n \times 1$ vector of independent and unobservable errors, $E(.)$ and $Cov(.)$ are the expectation and the covariance operators respectively. Depending on the context, the latter operator will be also used to indicate the sample covariance. $\boldsymbol{D}(z_i)$ defines a diagonal matrix that has $z_1, \ldots, z_k$ as diagonal entries. The size of the matrix should be clear from the context. It is usually assumed that $\boldsymbol{e}_n$ follows a multivariate normal distribution. This assumption is not necessary for estimation or prediction but needed in order to compute confidence intervals for parameters and prediction intervals for future values of the response variable. We assume throughout that the data follow a normal distribution and, thus, model (1) can be re-expressed as $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$.

In our problem, $\boldsymbol{Y}$ represents the climate sensitivity which is the average change in global mean temperature when CO2 is doubled. The design matrix $\boldsymbol{X}$ forms a set of predictors that takes on values on 11 fields measured over grid points (a set of spatial locations) that form part of the climate model output. There are 8192 locations represented by their latitudes and longitudes for the first seven fields which are denoted by TREFHT, SWCF, PRECT, LWCF, FSNT, FLNT and TAUX. For the rest of the fields RELHUM, U, T and CLOUD, the locations are represented by their latitudes and heights. Currently, we have 165 climate sensitivity experiments. Hence, the sample size and number of rows of $\boldsymbol{X}$, $n = 165$. CLOUD could not be measured over a small set of locations producing missing values for $\boldsymbol{X}$. After omitting these missing values, p, the number of regression parameters (predictors) or equivalently the number of columns of $\boldsymbol{X}$, is 63873 including the intercept. $\boldsymbol{\beta}$ represents predictors effects.

One of the main purposes of this study is to present sensible estimate of $\boldsymbol{\beta}$. Another goal is to produce predictions of climate sensitivity, y_0, when a set of predictor values, $\boldsymbol{x}_0$, is provided. $\boldsymbol{x}_0$ is a vector column of size p representing a new set of values of the 11 fields over all locations. A prediction of y_0 is computed as $\hat{y}_0 = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{\beta}$. Note that p is much larger than n, thus the least squares method to estimate $\boldsymbol{\beta}$ will interpolate the data points without giving a reasonable approximation of the underlying process that generated them. Therefore, we will rely on Principal Component Regression (PCR) to estimate $\boldsymbol{\beta}$ and perform predictions.

2.2.4 PCA

Principal Component Analysis (PCA) is primarily used as a data reduction technique by constructing a small set of new variables that carry and summarize most of the variability of the data matrix $\boldsymbol{X}$. These new variables (Principal Components), which are linear combinations of the original variables, are chosen so that they are uncorrelated and represent

the maximum variability contained in $\boldsymbol{X}$. PCA can accomplish a huge reduction especially when p is much greater than n which is our case. For the following discussion, p is assumed to be greater than n. For more details on PCA and PCR, see Johnson and Wichern (2007), Jolliffe (2002), and Massy (1965).

Since the sample covariance of $\boldsymbol{X}$, $\boldsymbol{S} = Cov(\boldsymbol{X})$, is a symmetric non-negative definite matrix, it can be expressed through the spectral decomposition as $\boldsymbol{S} = \tilde{\boldsymbol{V}}\boldsymbol{D}(\lambda_i)\tilde{\boldsymbol{V}}^T$ where $\tilde{\boldsymbol{V}}$ is an orthogonal matrix of size p, i.e. $\tilde{\boldsymbol{V}}\tilde{\boldsymbol{V}}^T = \tilde{\boldsymbol{V}}^T\tilde{\boldsymbol{V}} = \boldsymbol{I}_p$, with columns representing eigenvectors of $\boldsymbol{S}$ and $\boldsymbol{D}(\lambda_i)$ is a diagonal matrix carrying the eigenvalues of $\boldsymbol{S}$. The vectors $\boldsymbol{v}_i$'s are usually termed in the geophysical sciences as Empirical Orthogonal Functions (EOFs).

Define $\tilde{\boldsymbol{W}} = \begin{bmatrix}\boldsymbol{w}_1, & \boldsymbol{w}_2, & \ldots, & \boldsymbol{w}_p\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}\boldsymbol{v}_1, & \boldsymbol{X}\boldsymbol{v}_2, & \ldots, & \boldsymbol{X}\boldsymbol{v}_p\end{bmatrix} = \boldsymbol{X}\tilde{\boldsymbol{V}}$. $\tilde{\boldsymbol{W}}$ has the same dimension as $\boldsymbol{X}$, $n \times p$.

2.2.5 PCR

2.3 Estimation

PCR is concerned in performing regression analysis using the principal components as predictors rather than the original variables. Model (1) can be represented as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}_n = \boldsymbol{W}\boldsymbol{\alpha} + \boldsymbol{e}_n \tag{7}$$

where $\boldsymbol{\alpha} = \boldsymbol{V}^T\boldsymbol{\beta}$ which is a vector of length r. Equivalently, $\boldsymbol{Y} \sim N(\boldsymbol{W}\boldsymbol{\alpha}, \sigma^2\boldsymbol{I}_n)$ under normality of $\boldsymbol{e}_n$. The estimation problem has been reduced from estimating p coefficients to r. For the NCAR CAM 3.1 data described previously, $p = 63872$ and $r = 164$, so a large reduction has been accomplished.

This system does not have a unique solution. In fact, it has infinitely many solutions and all are given by $\boldsymbol{V}\hat{\boldsymbol{\alpha}} + \boldsymbol{u}$ or $\hat{\boldsymbol{\beta}} + \boldsymbol{u}$ where $\boldsymbol{u} = (\boldsymbol{I}_p - \boldsymbol{V}\boldsymbol{V}^T)\boldsymbol{z}$ for an arbitrary vector $\boldsymbol{z}$.

2.3.1 Prediction

A prediction for a future response value, y_0, at a new point $\boldsymbol{x}_0$, which is a vector of length p, can be obtained by finding $\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$ or $\boldsymbol{x}_0^T\boldsymbol{V}\hat{\boldsymbol{\alpha}}$. The latter quantity is an estimation of $\boldsymbol{x}_0^T\boldsymbol{\beta}$ which is $E(y_0)$. The prediction variance is $\sigma^2 + \boldsymbol{x}_0^T\boldsymbol{V}Cov(\hat{\boldsymbol{\alpha}})\boldsymbol{V}^T\boldsymbol{x}_0$. More precisely, the prediction and prediction variance are $\overline{\boldsymbol{Y}} + \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$ and $\sigma^2 + \sigma^2/n + \boldsymbol{x}_0^T\boldsymbol{V}Cov(\hat{\boldsymbol{\alpha}})\boldsymbol{V}^T\boldsymbol{x}_0$ respectively. For now we will ignore $\overline{\boldsymbol{Y}}$ and concentrate on $\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$.

2.3.2 Bayesian Paradigm

The paper by West (2003) imposes a generalized shrinkage prior on $\boldsymbol{\alpha}^T = \begin{bmatrix}\alpha_1, & \alpha_2, & \ldots, & \alpha_r\end{bmatrix}$ for Model (2). In particular, he assumes that $\alpha_i \sim N\left(0, c_i/\phi_i\right)$ for $i = 1, \ldots, r$, where the ϕ_i's are independent and identically distributed as $Gamma(j/2, j/2)$ where j is a known positive constant. In one of the examples in West (2003), he chooses j that would minimize

the mean square prediction error. It follows that $\boldsymbol{\alpha} \sim N_r(\mathbf{0}, \boldsymbol{D}(c_i/\phi_i))$. As mentioned in Section 4.2, there are multiple ways to extract $\boldsymbol{\beta}$ from $\boldsymbol{\alpha}$ and West (2003) properly justifies the use of the Moore-Penrose generalized inverse to make inferences on $\boldsymbol{\beta}$. This choice is irrelevant for predictions. Thus $\boldsymbol{\beta} = \boldsymbol{V}\alpha \sim N_p\left(\mathbf{0}, \boldsymbol{V}\boldsymbol{D}(c_i/\phi_i)\boldsymbol{V}^T\right)$ is the implied prior (singular distribution) for $\boldsymbol{\beta}$. Samples from the posterior distribution of $\boldsymbol{\beta}$ are easily obtained once samples for $\boldsymbol{\alpha}$ have been drawn. The components of $\boldsymbol{\beta}$ are usually estimated by either their posterior means or posterior medians. Percentiles of the posterior distribution provide credible intervals. All these quantities are approximated from the posterior samples.

The scalar c_i can be interpreted as weights to indicate the importance of the ith PC, $\boldsymbol{w}_i$. Usually, the PCs with small variances are deleted from the model. This can be directly done by setting $c_i = 0$. Instead of this arbitrary practice, West (2003) suggests to assume that $c_i = gi^{-2}$ where g is a scale factor that has a flat prior. This is a shrinkage prior and requires a stronger evidence from the data to see the influential role for higher ordered PCs. In accordance with Jolliffe (1982), we prefer to use $c_i = g$ to give all PCs the same chance to appear in the model. Assuming an Inverse-Gamma prior for σ^2 or a uniform distribution on $\log\sigma$ completes the prior specification.

Posterior sampling for this model is easily implemented using MCMC and simply executed by standard software tools such as OpenBugs or rjags. The predictions over $\boldsymbol{X}_0$ are naturally found by treating the future responses, $\boldsymbol{Y}_0$, as missing values and to be imputed during the MCMC sampling iterations. We reiterate that $\boldsymbol{X}_0$ must be stacked to $\boldsymbol{X}$ before performing PCA. Thus the analysis has to be repeated if $\boldsymbol{X}_0$ changes. The prediction is usually given by the median or the mean of the posterior predictive distribution, $E(\boldsymbol{Y}_0|\boldsymbol{Y})$. These quantities along with 95% credible predictive intervals are easily approximated from the posterior predictive samples.

2.3.3 Analysis

In this section we present a frequentist analysis of the data. It is assumed that PCA has been already performed on $\boldsymbol{X}$. The fitting procedure for Model (2) is introduced and described under the general framework. Recall that Model (2) is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}_n = \boldsymbol{W}\boldsymbol{\alpha} + \boldsymbol{e}_n$$

For our data, $n = 165$ and $r = 164$. One degree of freedom has been lost to estimate the intercept (the response mean which is 2.789) leaving us with a saturated model, i.e. 0 degrees of freedom and an estimate of the error cannot be obtained. Formal tests to assess predictors significance are not feasible. In such cases, predictors screening process such as normal probability plot is pretty useful and presented next. One might suggest to consider the PCs that are associated with small eigenvalues as non-significant and to be removed from the model since they explain a little of the variability of the original variables. However, Jolliffe (1982) demonstrated that these components can be equivalently important as much as those with large eigenvalues. He claims that a selection process depending solely on the magnitude of the eigenvalues should not be considered. An alternative solution is presented next.

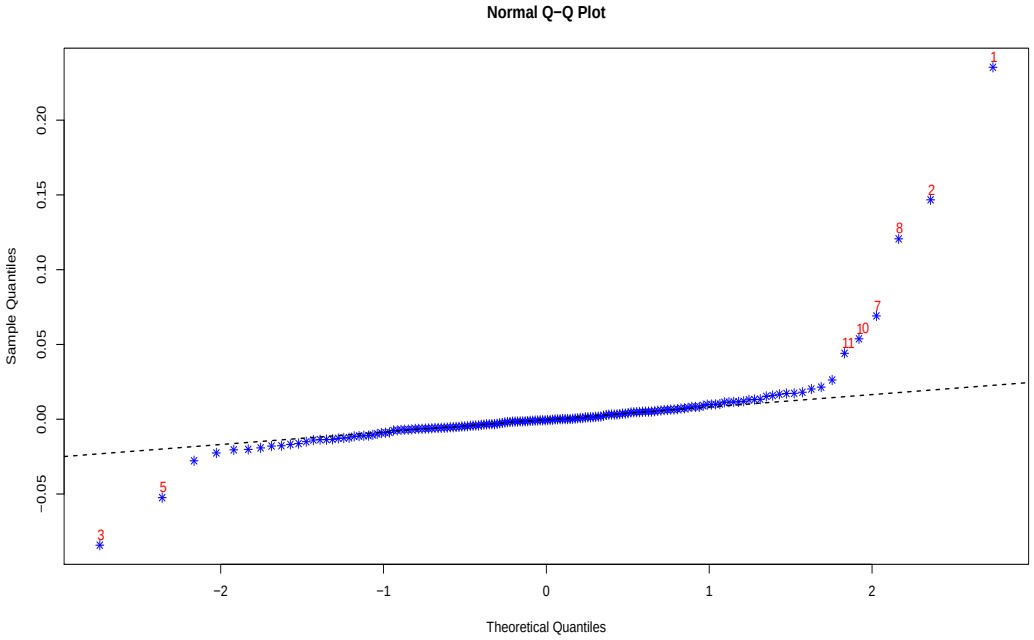
Table 1: PC Estimates and their Standard Errors multiplied by 10^3

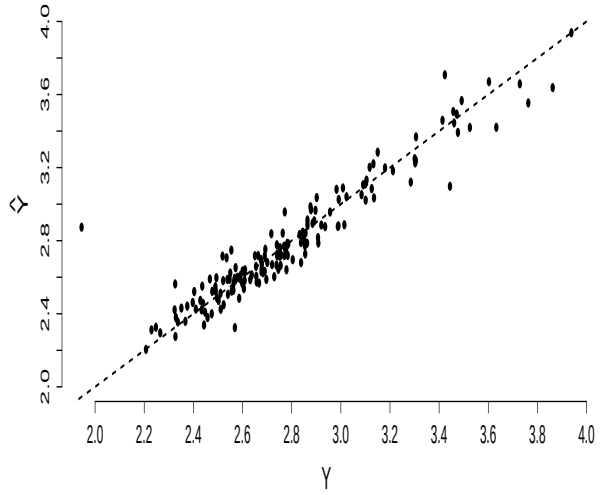
PC #	1	2	3	5	7	8	10	11
$\hat{\alpha}_i$	1.298	1.844	-1.187	-1.004	1.690	3.341	1.938	1.708
SE	0.050	0.115	0.129	0.175	0.224	0.253	0.329	0.354

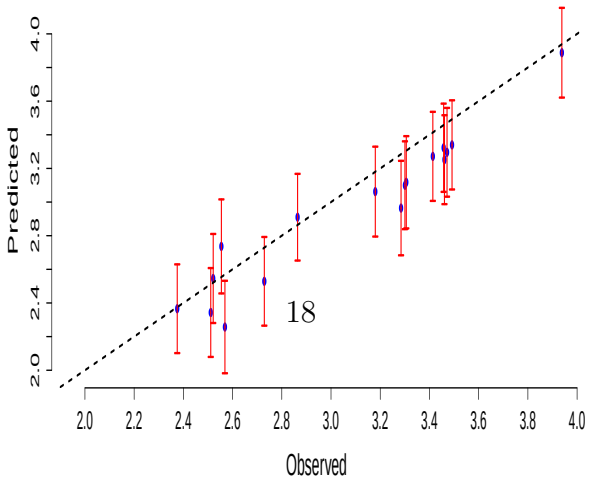
Model (2) constitutes an orthogonal design because $\boldsymbol{W}^T\boldsymbol{W}$ is a diagonal matrix. The least squares estimate of $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{W}^T\boldsymbol{W}\right)^{-1}\boldsymbol{W}^T\boldsymbol{Y} = \boldsymbol{D}\left(\frac{1}{\lambda_i(n-1)}\right)\boldsymbol{W}^T\boldsymbol{Y}$ and $Cov(\hat{\boldsymbol{\alpha}}) = \sigma^2\left(\boldsymbol{W}^T\boldsymbol{W}\right)^{-1} = \sigma^2\boldsymbol{D}\left(\frac{1}{\lambda_i(n-1)}\right)$. In particular, the ith component of the least squares estimate $\hat{\alpha}_i = \dfrac{\boldsymbol{w}_i^T\boldsymbol{Y}}{\lambda_i(n-1)}$ and has $var(\hat{\alpha}_i) = \dfrac{\sigma^2}{\lambda_i(n-1)}$. Clearly, $\hat{\alpha}_i$ depends solely on the ith PC in isolation of the other PCs. Orthogonality also implies that these components are uncorrelated (independent under normality). Note also that the significance of a PC depends directly on the magnitude of λ_i. Low ordered PC's have smaller standard errors. As a result of that, higher ordered PC's might not be declared significant even though they may be highly correlated with $\boldsymbol{Y}$. If the $\boldsymbol{w}_i$'s are all scaled to have the same length (divide by $\sqrt{\lambda_i}$) , then $Cov(\hat{\boldsymbol{\alpha}}) = \dfrac{\sigma^2\boldsymbol{I}_r}{n-1}$. It follows that $\hat{\boldsymbol{\alpha}} \sim N\left(\boldsymbol{\alpha}, \dfrac{\sigma^2\boldsymbol{I}_r}{n-1}\right)$ and PC selection based on normal probability plot is now feasible. As we will see shortly, this method is very simple and yet effective. Originally, it was introduced by Daniel (1959) for unreplicated 2^k factorial design when the number of factors is large and obtaining another replication of the experiment is not possible. Daniel (1959) suggests to investigate a normal probability plot of the estimates of the predictors. Negligible or inactive predictors, i.e. have a zero mean, should fall on a straight line. Critical or active predictors, have nonzero means, will tend to deviate from the line. Only active predictors will be represented in the model. For more details and examples see Montgomery (1996).

A normal probability plot for the scaled PCs is shown in Figure 1-a. According to Daniel (1959), PC#: 1, 2, 3, 5 , 7, 8, 10 and 11 are deemed significant and exclusively considered in the preliminary model. The estimates for the rest of the PCs are set to zero. Notice that higher ordered PC's such as 10 and 11 are more important than the relatively low ordered PCs 4 and 6. Table 1 presents the coefficient estimates $\hat{\alpha}_i$ along with the standard errors for the significant PCs according to Daniel's plot. Additionally, Figure 1-b displays the actual values of $\boldsymbol{Y}$ versus the fitted values $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{W}\hat{\boldsymbol{\alpha}}$. The correlation between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ exceeds 94 %. Residual analysis and diagnostic plots (not shown here) support the appropriateness of the model apart of an apparent outlier with an observed climate sensitivity of about 1.95 °C. Refitting the model with this point removed yielded somewhat similar results in terms of parameter estimates, standard errors and predictions.

As mentioned earlier an estimate of $\boldsymbol{\beta}$, the standardized regression coefficient vector, can be obtained by $\hat{\boldsymbol{\beta}} = \boldsymbol{V}\hat{\boldsymbol{\alpha}}$ where $\hat{\boldsymbol{\alpha}}^T = \left[\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, 0, \hat{\alpha}_5, 0, \hat{\alpha}_7, \hat{\alpha}_8, 0, \hat{\alpha}_{10}, \hat{\alpha}_{11}, \boldsymbol{0}_{1\times(153)}\right]$. It


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−0.05
0.00
0.05
0.10
0.15
0.20
−2
−1
0
1
2
1
2
3
5
7
8
10
11


ŷ
y
2.0
2.4
2.8
3.2
3.6
4.0
2.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0


Predicted
Observed
18
2.0
2.4
2.8
3.2
3.6
4.0
2.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0

follows that $\hat{\boldsymbol{\beta}} = \sum_{i \in d} \hat{\alpha}_i \boldsymbol{v_i}$ and $Cov(\hat{\boldsymbol{\beta}}) = \frac{\sigma^2}{n-1} \sum_{i \in d} \frac{\boldsymbol{v}_i \boldsymbol{v}_i^T}{\lambda_i}$ where $d = \{1, 2, 3, 5, 7, 8, 10, 11\}$. Note that each term $\boldsymbol{v}_i \boldsymbol{v}_i^T$ is a $p \times p$ matrix, therefore finding $Cov(\hat{\boldsymbol{\beta}})$ may not be computationally feasible if p is excessively large. We concentrate, instead, on its diagonal elements which represent $var(\hat{\beta}_j), j = 1, \ldots, p$. The latter quantity is exactly $\frac{\sigma^2}{n-1} \sum_{i \in d} \frac{v_{ji}^2}{\lambda_i}$ where v_{ji} is the jth component of $\boldsymbol{v}_i$. The standard error of $\hat{\beta}_j$ is found by taking the square root and replacing σ^2 by its estimate $\hat{\sigma}^2$.

Figures 2 and 3 display the components of $\hat{\boldsymbol{\beta}}$ distributed over their corresponding geographic locations within each underlying field generated by the CAM. Figures 4 and 5 present their standard errors. The quantities in these maps are multiplied by 10^5. In particular, Figures 2 and 3 provide a visual display of how fields are impacting climate sensitivity across the geographical regions.

To assess the model predictive ability, 17 observations selected randomly (around 10% of the data) have been held out and removed for prediction validation. Figure 1-c displays the actual values of these observations versus predicted values along with 95% prediction intervals. This procedure has been repeated several times with a random selection of points chosen to be removed and then predicted. The results for these repetitions are similar to what is observed in Figure 1-C. The prediction at $\boldsymbol{x}_0$ is $\overline{\boldsymbol{Y}} + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} = \overline{\boldsymbol{Y}} + \sum_{i \in d} \hat{\alpha}_i w_{00i}$. The prediction variance is $\sigma^2 + \sigma^2/n + \boldsymbol{w}_{00}^T Cov(\hat{\boldsymbol{\alpha}}) \boldsymbol{w}_{00} = \sigma^2 + \sigma^2/n + \frac{\sigma^2}{n-1} \sum_{i \in d} \frac{w_{00i}^2}{\lambda_i}$, where $\boldsymbol{w}_{00} = \boldsymbol{V}^T \boldsymbol{x}_0$ and w_{00i} is its ith component.

2.3.4 Bayesian Analysis

Now we reanalyze the climate model output from CAM 3.1 using the Bayesian model presented in Subsection 4.3. Two priors are introduced for $\boldsymbol{\alpha}$. The first one, denoted by M_1, assumes $\alpha_i \sim N(0, c_i/\phi_i)$, where $c_i = gi^{-2}$. The other one, M_2,assumes $\alpha_i \sim N(0, c_i/\phi_i)$, where $c_i = g$. In both cases, we adopt adopt flat priors for β_0 (the intercept) and g and also assume that $\phi_i \sim Gamma(0.8, 0.8)$ and $1/\sigma^2 \sim Gamma(1, 1)$. These choices of priors are not critical. In fact, the posterior analysis was minimally affected when we assumed different Gamma priors. Posterior analysis, posterior medians and 95% credible intervals, for $\boldsymbol{\alpha}$ for both M_1 and M_2 are presented in Figure 6. The Red bars represent the PCs that are significantly different from 0 from Bayesian perspective.

Both M_1 and M_2 agree with the previous analysis presented in Section 5 in terms of significant PCs. M_1 intervals narrow down to 0 as the index, i, increases as a result of the prior specification. Also, M_1 yielded slightly shorter intervals for the first PCs. Next we removed the non-significant PCs by setting $c_i = 0$ for $i \notin d$. We also set $c_i = g$ for $i \in d$ where d is defined in Section 5. This model is denoted by M_3.

Three MCMC chains were initiated using different starting values. Each chain has a

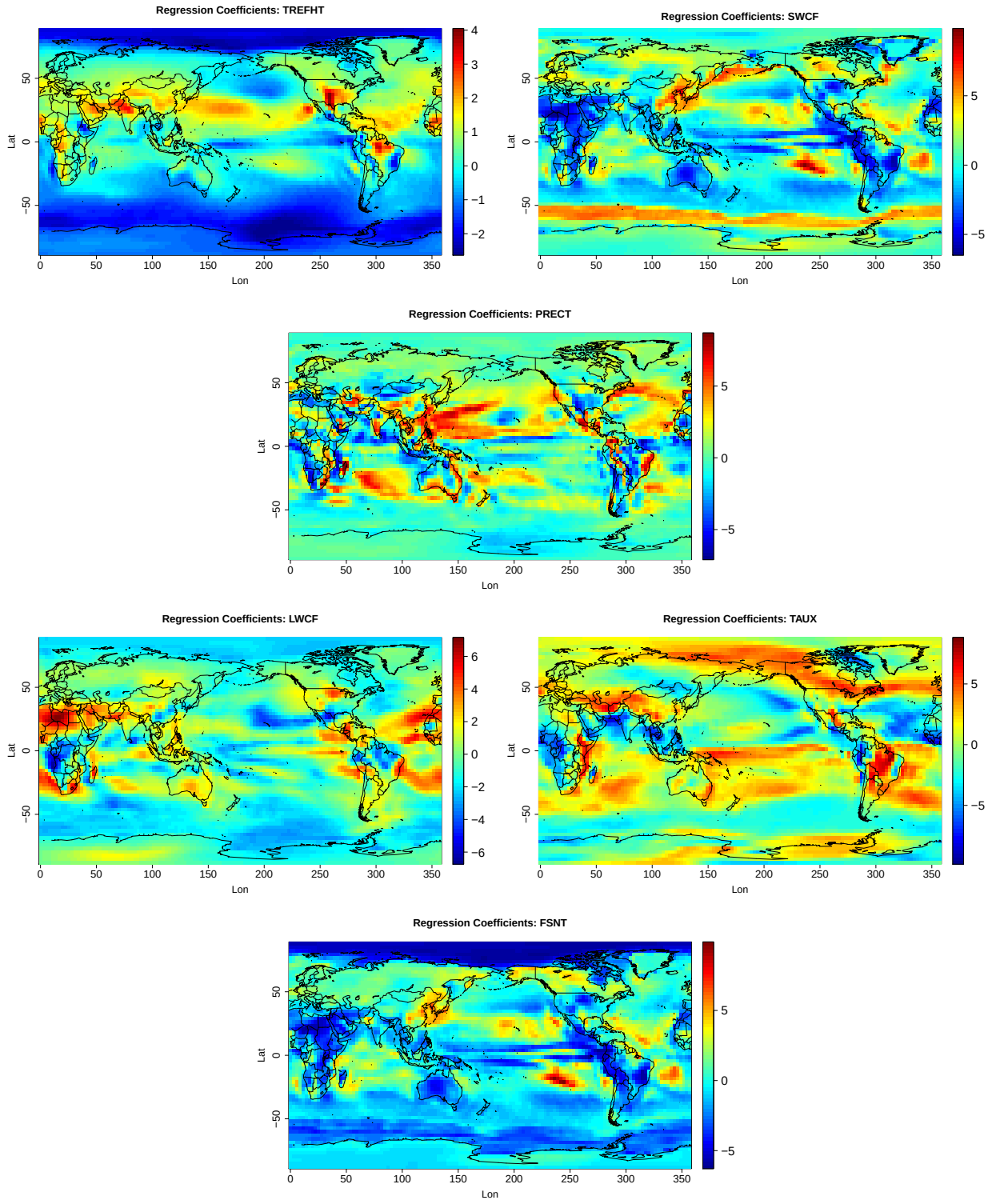


Figure 8: Estimate of $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}} \times 10^5$ (a) TREFHT. (b) SWCF. (c) PRECT. (d) LWCF. (e) TAUX. (f) FSNT

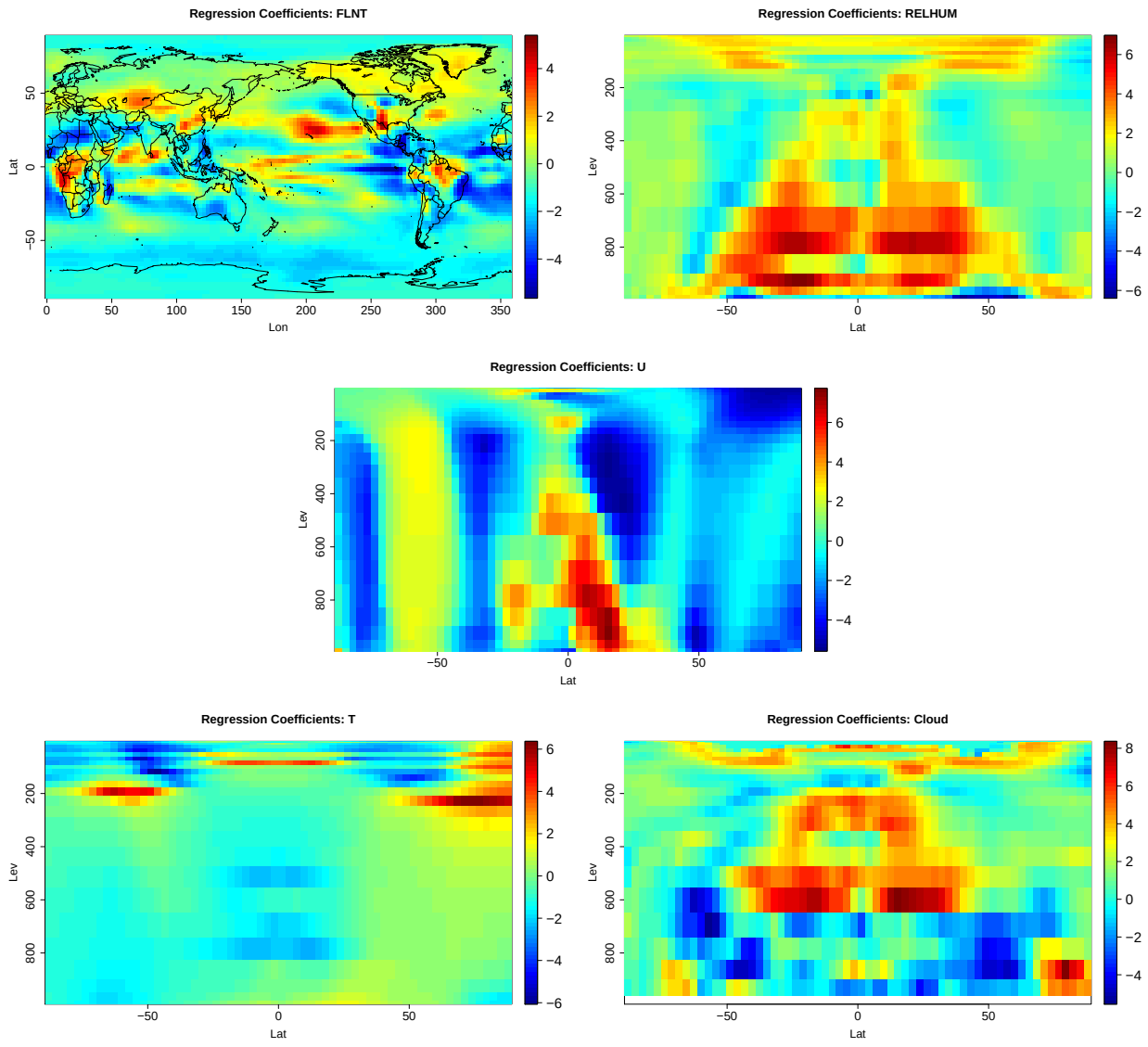


Figure 9: Estimate of $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}} \times 10^5$ (a) FLNT. (b) RELHUM. (c) U. (d)T.

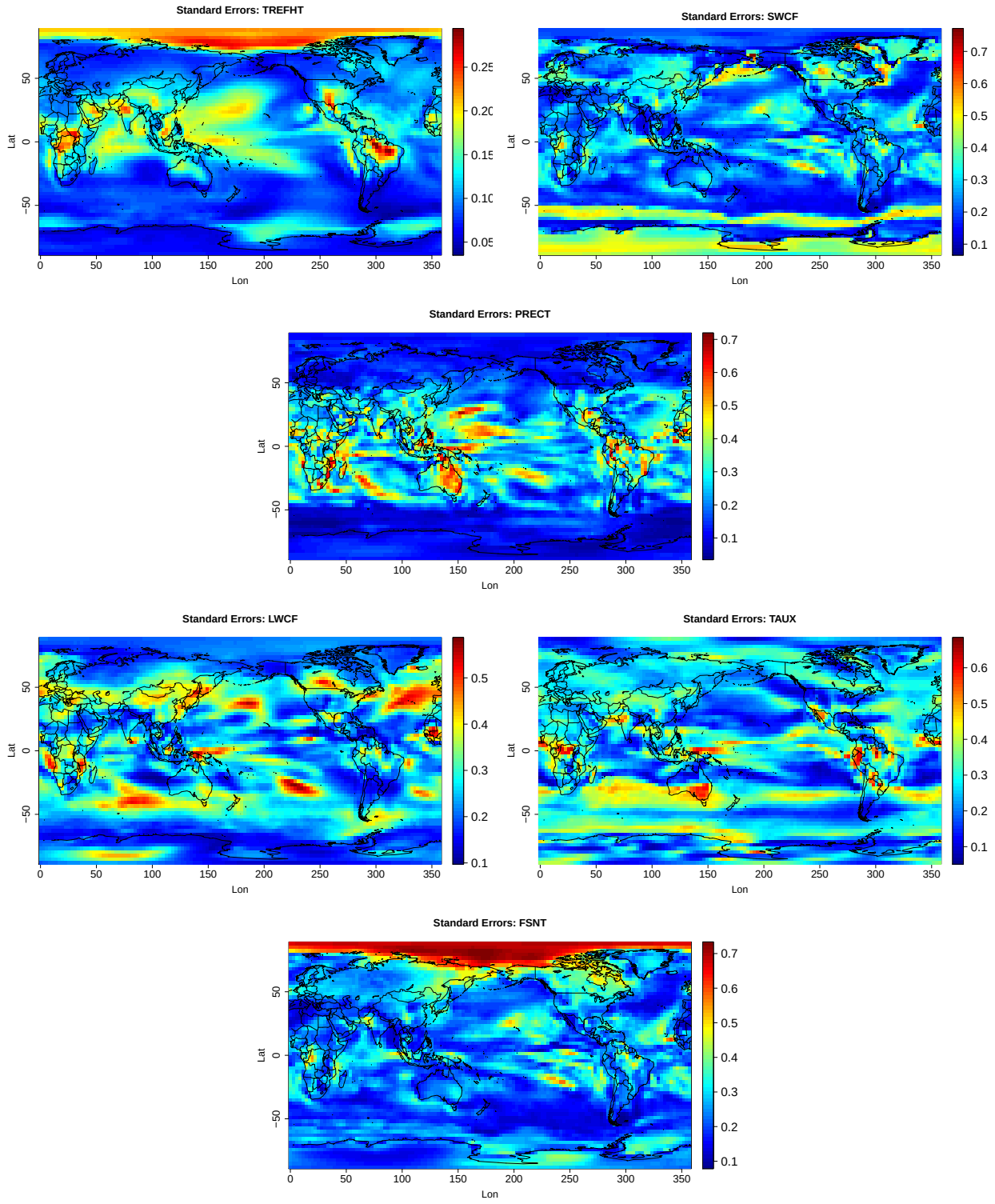


Figure 10: Standard Error of $\hat{\boldsymbol{\beta}} \times 10^5$. (a) SE:TREFHT. (b) SE:SWCF. (c) SE:LWCF. (d) SE: TAUX. (d) SE:FSNT.

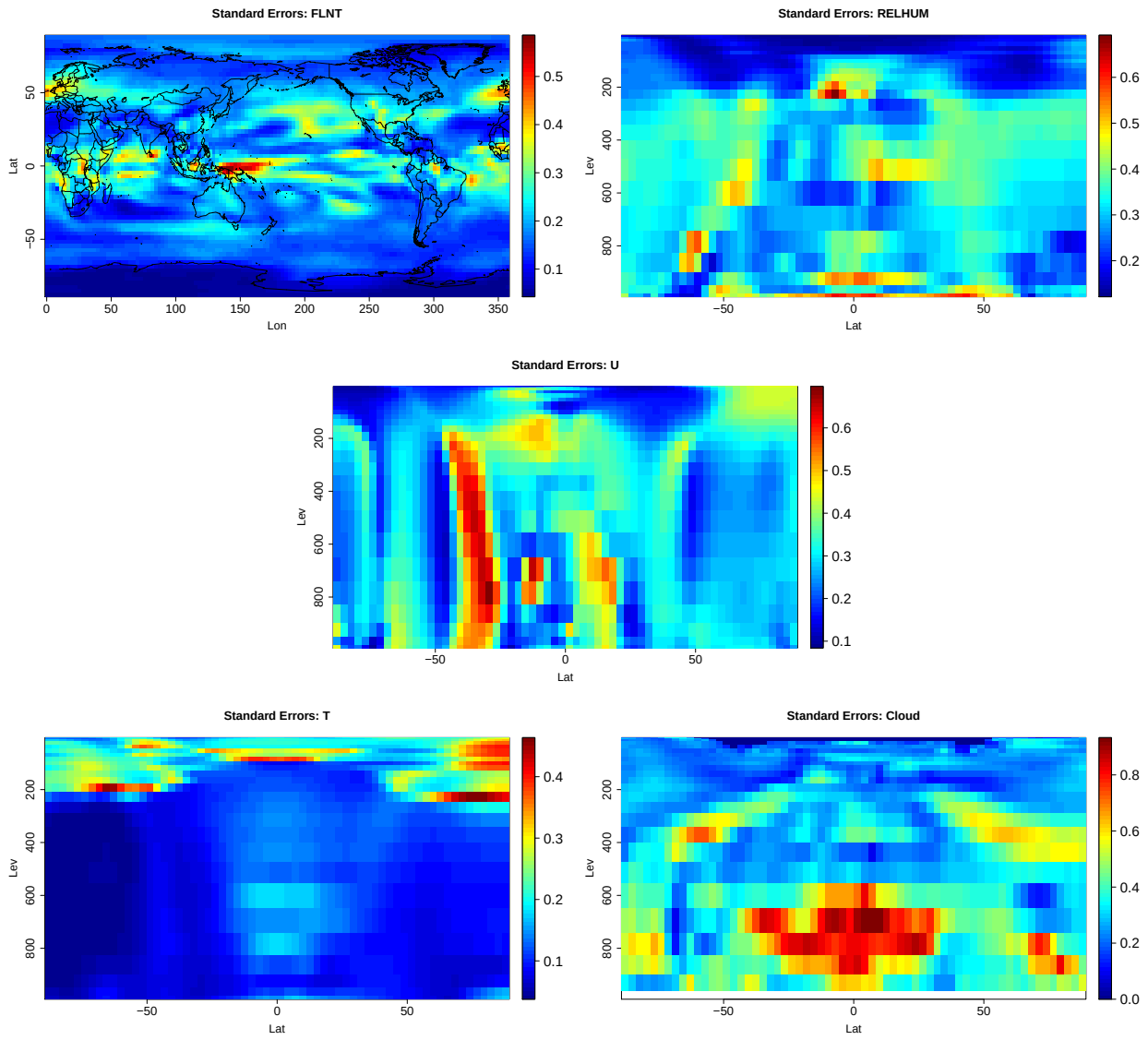


Figure 11: Standard Error of $\hat{\boldsymbol{\beta}} \times 10^5$. (a) SE:FLNT. (b) SE:RELHUM. (c) SE:U. (d) SE:T

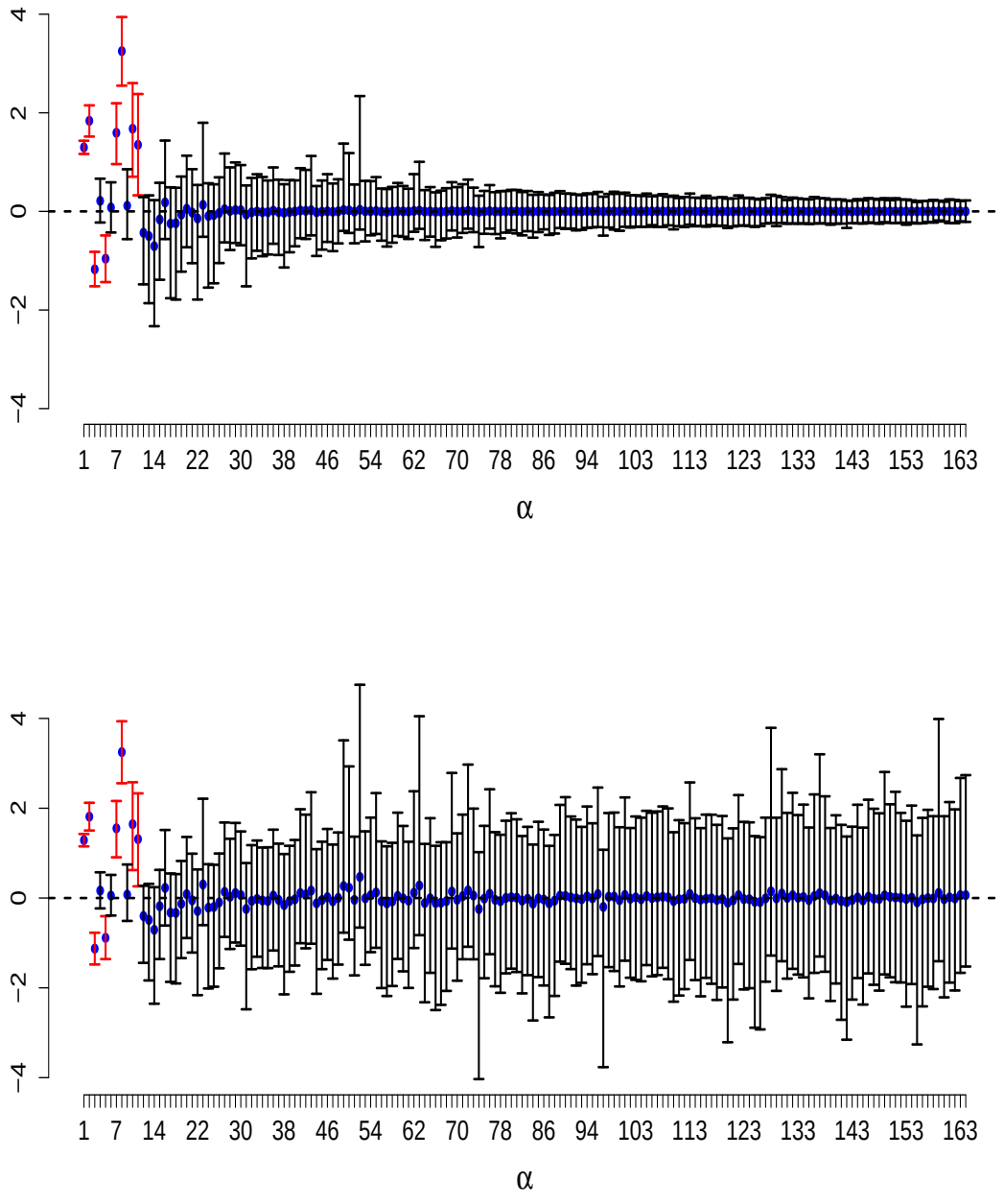


Figure 12: Posterior Analysis of $\boldsymbol{\alpha}$ multiplied by 10^3. Top (M_1), Bottom (M_2)

Table 2: $E(\alpha_i|Y)$ and $SD(\alpha_i|Y)$ multiplied by 10^3

PC #	1	2	3	5	7	8	10	11
$E(\alpha_i\|Y)$	1.295	1.832	-1.173	-0.984	1.647	3.284	1.841	1.603
$SD(\alpha_i\|Y)$	0.070	0.160	0.179	0.243	0.310	0.352	0.457	0.493

length of 30000 and the first 5000 iterations had been used as burn-in. The computation time did not exceed 15 seconds. Several diagnostics procedures have shown that convergence has been reached quickly. In particular trace plots for g, σ^2 and the intercept versus the iteration index are shown in Figure 7. Posterior analysis for the parameters α_i and ϕ_i is displayed in Figure 8. Table 2 shows posteriors means and standard deviations for α_i. Note that $E(\alpha_i|\boldsymbol{Y}) \approx \hat{\alpha}_i$, the least squares estimate of α_i, and $SD(\alpha_i|Y) > SE(\hat{\alpha}_i)$. The actual values for the climate sensitivity versus the fitted values, $E(\boldsymbol{X}\boldsymbol{\beta}|\boldsymbol{Y})$, are shown in Figure 7-d which almost coincides with Figure 1-b. Assuming other priors on g, σ^2 and ϕ_i yielded very similar results. However, assuming $\phi_i = \phi$ for all i produced poorer fit and less stable results. Similar issues were observed when g was fixed at 1.

The posterior analysis for $\boldsymbol{\beta}$, displayed in Figures 9–12 produced similar results to the frequentist approach presented in Section 4 in terms of point estimates, i.e. $E(\beta_j|\boldsymbol{Y}) \approx \hat{\beta}_j$ but with higher variability, $SD(\beta_j|\boldsymbol{Y}) > SE(\hat{\beta}_j)$. It can be seen again that there are some locations having larger effects on climate sensitivity than other locations depending on each field. The posterior median and 95% probability intervals for the coefficients of one location that has longitude of 261.5625 and latitude of 29.30136 are shown in Figure 13. Notice that none of these intervals include zero. For example, we expect that for every one standard deviation increase in TREFHT on this location corresponds to, on average, 3.559223×10^{-5} increase in climate sensitivity keeping all other predictors fixed. We have a 95% probability that this increase falls between 2.846493×10^{-5} and 4.233791×10^{-5}. The other estimates and intervals on Figure 13 can be interpreted similarly. Notice that climate sensitivity in this location is positively associated with some fields and negatively over others. These relationships can change across locations.

For illustration purposes, the data matrix, $\boldsymbol{X}$ has been divided into two parts, $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_R \\ \boldsymbol{X}_0 \end{bmatrix}$. $\boldsymbol{X}_0$ consists of the 17 points in Figure 1-c. Climate sensitivity is considered missing and to be predicted over these points. Figure 14 shows the resulting Bayesian predictions which are slightly better in terms of the mean squared error compared to the Frequentist predictions. It is evident that the Bayesian credible intervals are wider than the Frequentist intervals. In fact they are on average 46% wider. We also found $\boldsymbol{V}_R\boldsymbol{V}_R^T\boldsymbol{X}_0^T \approx \boldsymbol{X}_0^T$ where $\boldsymbol{V}_R$ is the $\boldsymbol{V}$ matrix that resulted from performing PCA on $\boldsymbol{X}_R$. This supports the suitability of $\boldsymbol{X}_0\hat{\boldsymbol{\beta}}$ for predictions. The two approaches can produce results that might be drastically different if $\boldsymbol{V}_R\boldsymbol{V}_R^T\boldsymbol{X}_0^T$ does not give a reasonable approximation for $\boldsymbol{X}_0^T$.

Climate sensitivity is a quantity of great interest in climatological research which is defined as the average change in global mean temperature when CO2 is doubled. To study this quantity, we considered output/data generated by the NCAR Community Atmospheric

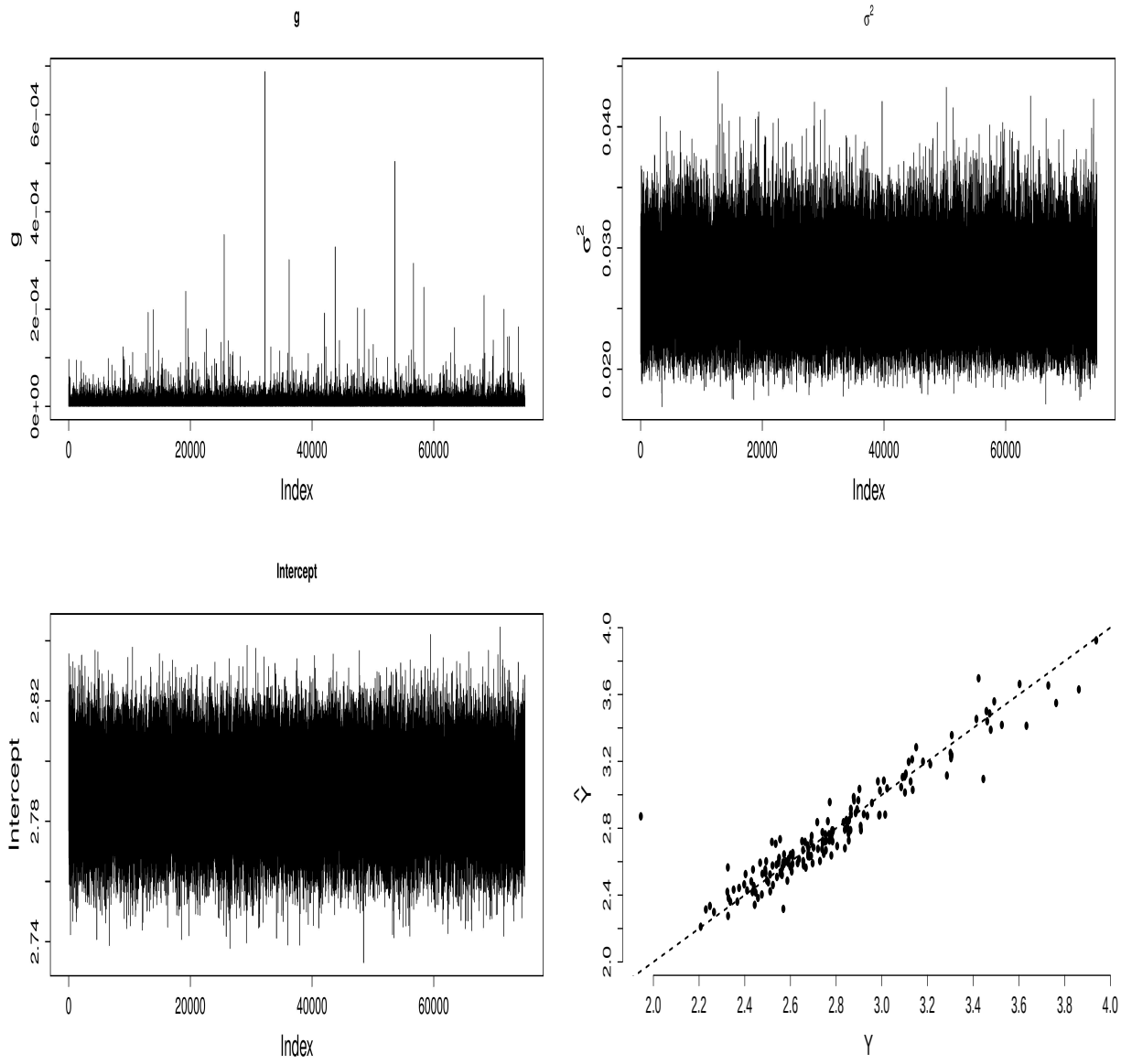


Figure 13: M_3 Posterior Analysis. (a) Trace Plot: g. (b) Trace Plot: σ^2. (c) Trace Plot: β_0. (d) Actual Vs. Fitted

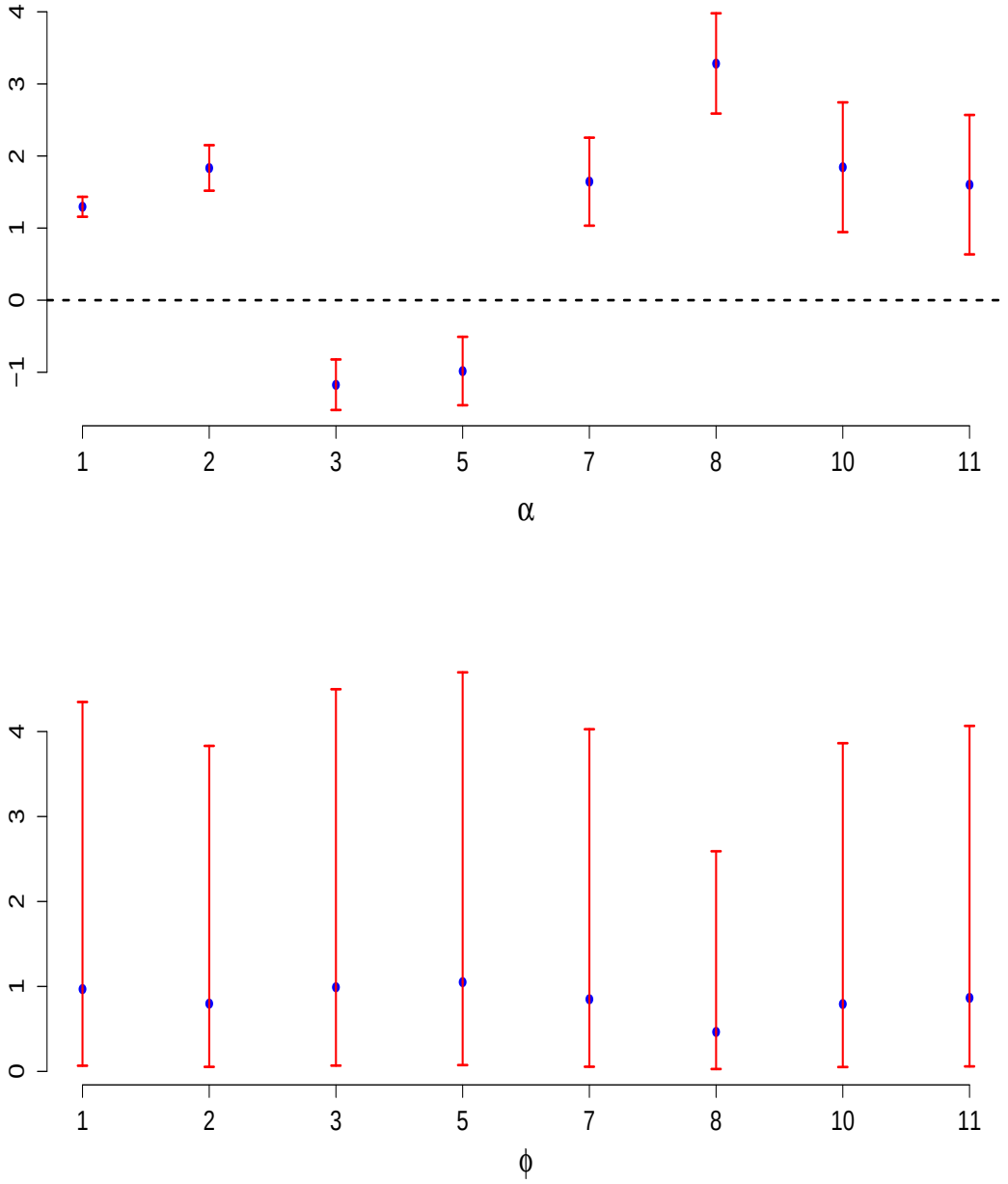


Figure 14: M_3 Posterior Analysis. (a) $\alpha_i \times 10^3$ (b) ϕ_i

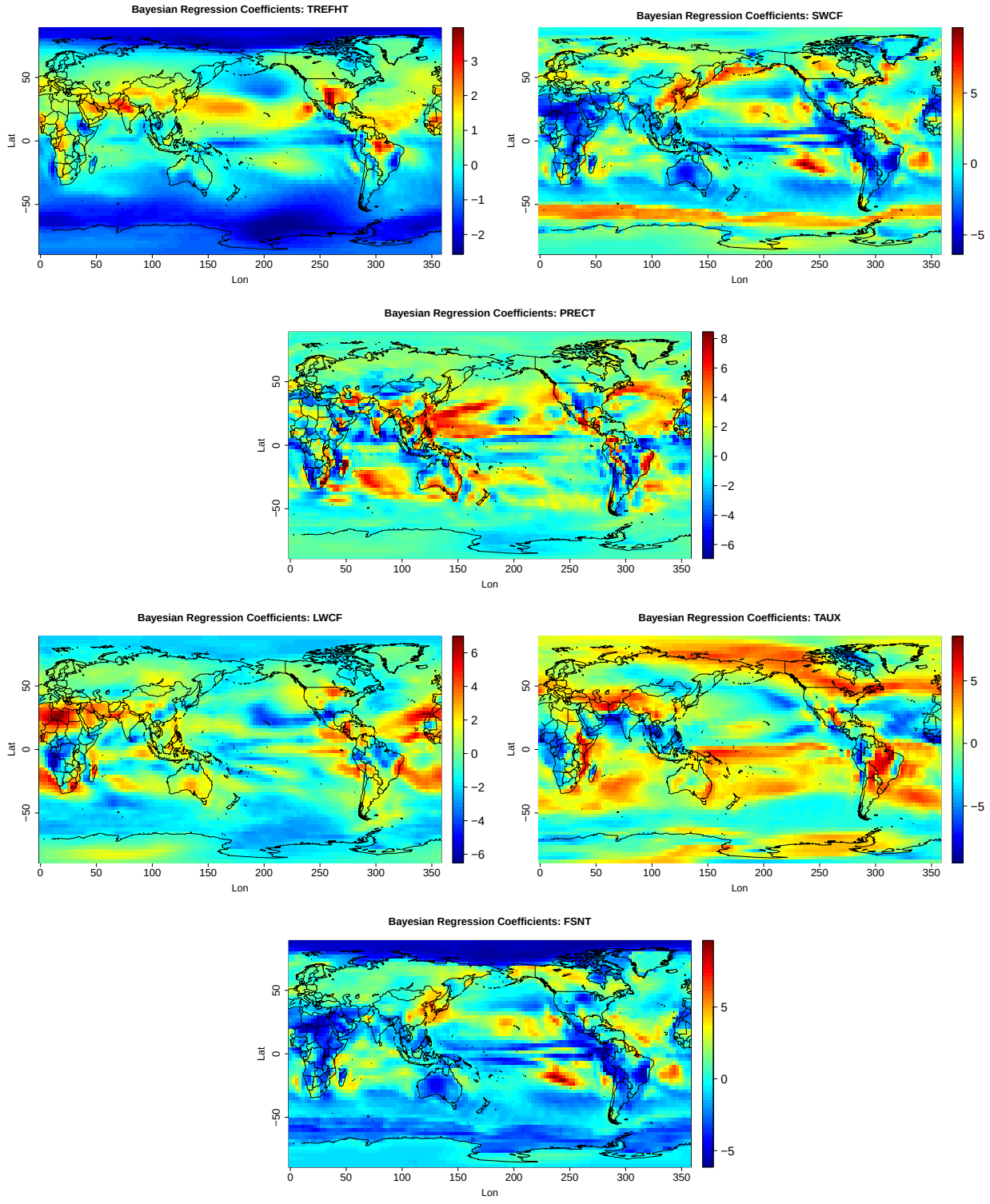


Figure 15: $E(\boldsymbol{\beta}|\boldsymbol{Y}) \times 10^5$. (a) TREFHT. (b) SWCF. (c) PRECT. (d) LWCF. (e) TAUX. (f) FSNT.

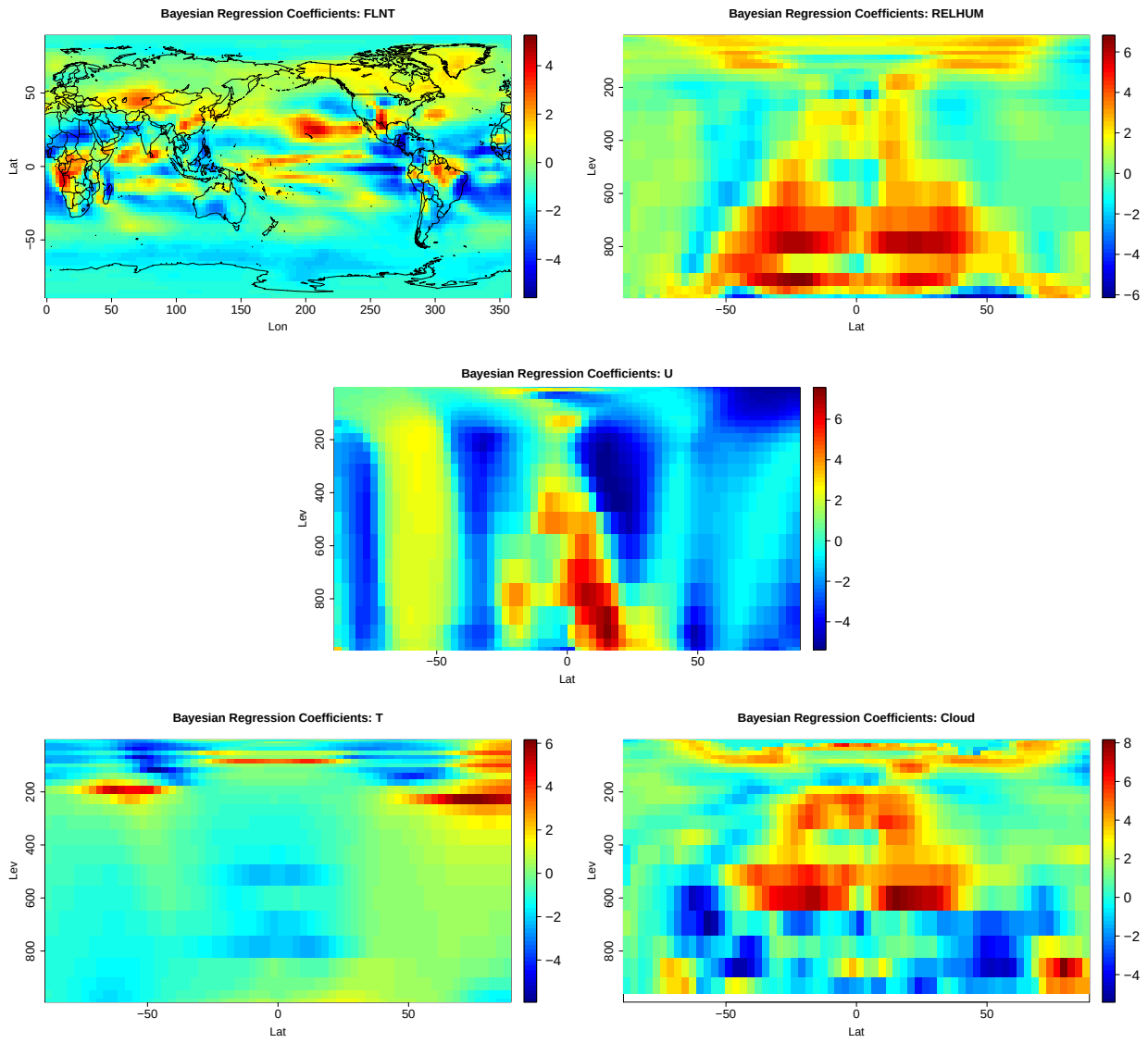


Figure 16: $E(\boldsymbol{\beta}|\boldsymbol{Y}) \times 10^5$. (a) FLNT. (b) RELHUM. (c) U. (d) CLOUD.

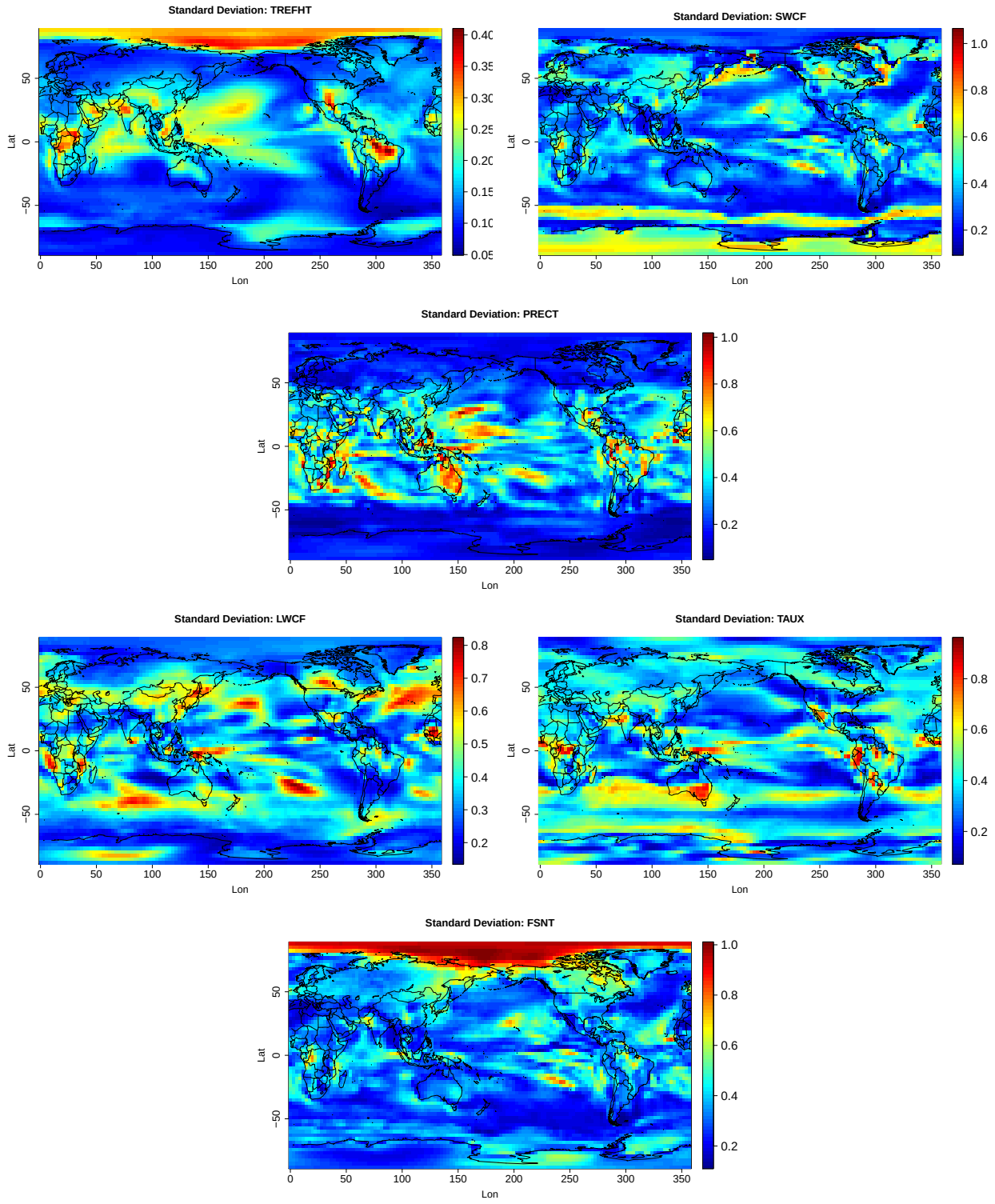


Figure 17: $SD(\boldsymbol{\beta}|\boldsymbol{Y}) \times 10^5$ (a) TREFHT. (b) SWCF. (c) PRECT. (d) LWCF. (e) TAUX. (f) FSNT.

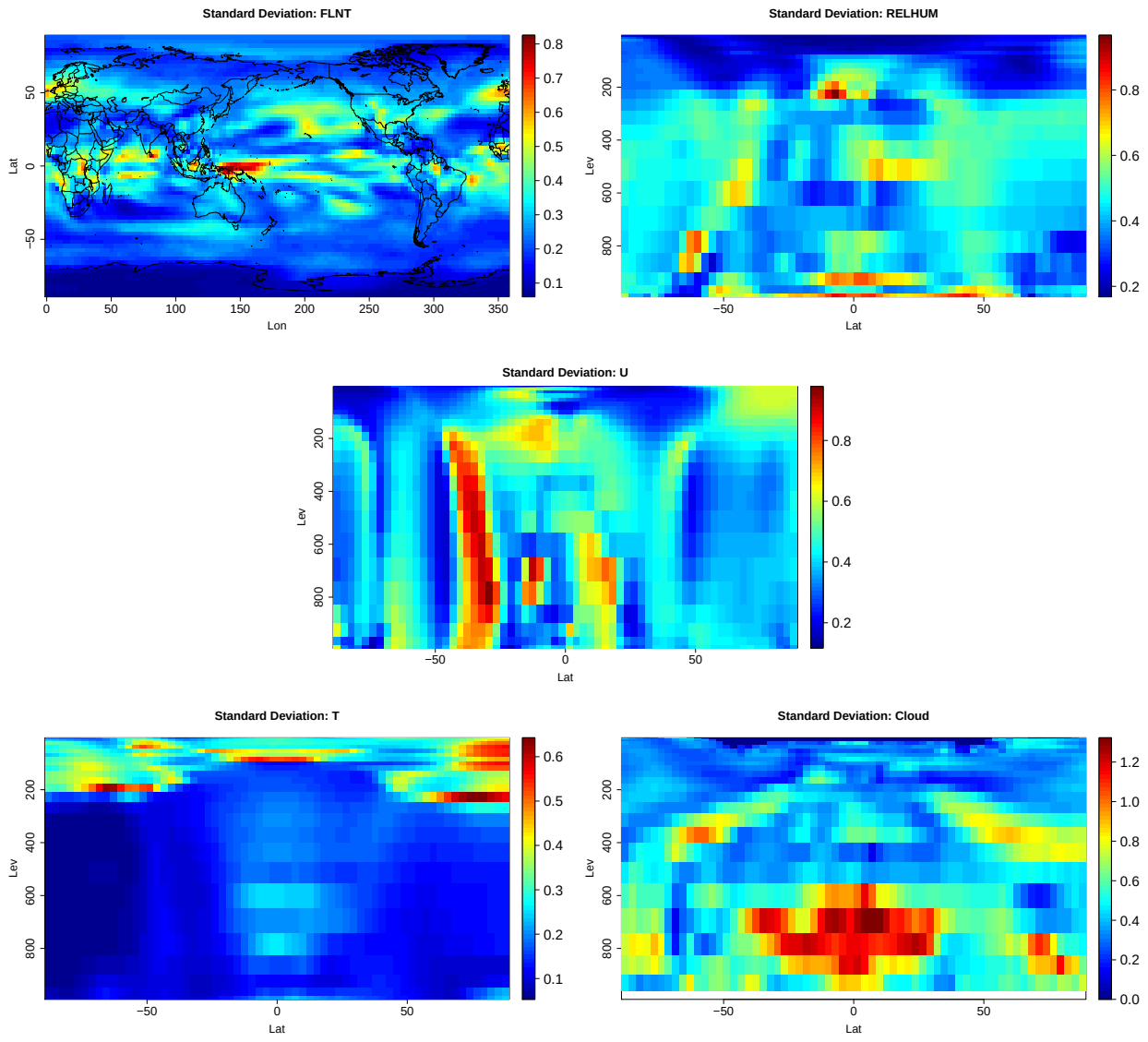


Figure 18: $SD(\boldsymbol{\beta}|\boldsymbol{Y}) \times 10^5$ (a) FLNT. (b) RELHUM. (c) U. (d) T. (e) CLOUD.

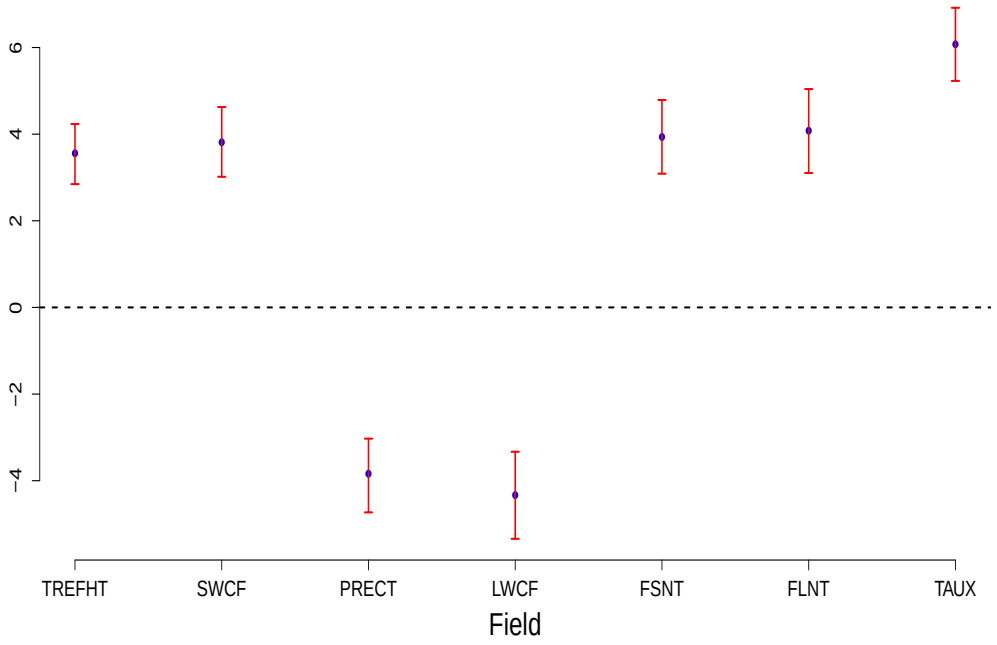


Figure 19: 95% Posterior Probability Intervals for location $(29.30, 261.56)$

Model (CAM 3.1). In addition to climate sensitivity, CAM 3.1 produces model runs of climate simulation generating output related to 11 different fields or variables, such as humidity, and measured over a spatial grid consisting of 8192 locations. There are only 165 of such model runs which induces a problem where the number of predictors is overwhelmingly greater than the number of observations. The main goal of this study was to introduce sensible and flexible methods to predict climate sensitivity based on the current CAM 3.1 models runs and for those available for other models. The problem was tackled via Principal Component Regression (PCR) analysis based on Bayesian and Frequentist perspectives. Specifically, the Bayesian approach is needed to solve identifiability issues that are needed to differentiate between robust and non-robust predictors, an issue that is usually ignored in the literature of PCR. Our framework provides maps to visualize the relationship between climate sensitivity and the 11 fields considered

3 Project Participants

1. Gabriel Huerta (PI) is Professor, Department of Mathematics and Statistics. The University of New Mexico.

2. Alvaro Nosedal-Sanchez. Postdoctoral fellow. Now Assistant Professor at University of Toronto-Mississauga.

3. Mohammad Hattab. Postdoctoral fellow. Now Postdoctoral fellow at Virginia Commonwealth University.

4. Menuka Karki. Postdoctoral Fellow. Now Visiting Assistant Professor at Louisiana Tech University.

5. Yonghua Wei. Graduate Student. Now Data Scientist at Walmart Headquarters in Arkansas.

4 Publications and Software

Jackson, C. and Huerta, G. (2016). Empirical Bayes Approach to Climate Model Calibration. *Geosci. Model Dev. Discuss.*, doi:10.5194/gmd-2016-20, in review.

Rodriguez, S., Huerta, G. and Reyes, H. (2016) A study of trends for Mexico City ozone extremes:2001-2014. Accepted for *Revista Atmosfera-UNAM* , v. 29, n.2, p. 107-120, mar. 2016. ISSN 2395-8812

Nosedal, A., Jackson, C. and Huerta, G. (2016) A new metric for climate models that includes field and spatial dependencies using Gaussian Markov Random Fields, in review. *Geoscientific Model Development.* doi:10.5194/gmd-2015-250. (with assistance from M. Karki).

Qian, Y., Jackson C. Giorgi, F., Booth, B., Duan, Q. Forest, C., Higdon, D., Hou, J.Z. and Huerta, G. (2016) Uncertainty Quantification in Climate Modeling and Projection.

Accepted in *the Bulletin of the American Meteorological Society, BAMS* doi:10.1175/BAMS-D-15-00297.1, in press.

Khider, D., Huerta, G., Jackson, C., Stott, L.D. and Emile-Geay, J. (2015) A Bayesian, multivariate calibration for *Globigerinoides ruber* Mg/Ca *Geochem. Geophys. Geosyst.*, 16, doi:10.1002/2015GC005844.

Huerta, G. and Stark, G.A. (2013) Dynamic and Spatial Modeling of Block Maxima Extremes. *In Bayesian Inference and Markov Chain Monte Carlo: In Honor of Adrian Smith*. Oxford University Press, Chapter 10, 183-199.

Villagran, A., Huerta, G. , Vannucci, M., Jackson, C. and Nosedal, A. (2014). Non-Parametric Sampling Approximation via Voronoi Tessellations *Communications in Statistics - Simulation and Computation* DOI No. 10.1080/03610918.2013.870798

Wei, Y. and Huerta, G. (2016). Dynamic Generalized Extreme Value Modeling via Particle Filters (re-submitted).

Khider, D., Stott, L.D., Jackson, C. and Huerta, G. (2014) How unusual is the twentieth century within the Indo-Pacific Warm Pool? (in review).

Code for paper Khider, D.,Huerta, G., Jackson, C., Stott, L.D. and Emile-Geay, J. (2015) A Bayesian, multivariate calibration for *Globigerinoides ruber* Mg/Ca. Written in Matjags and Rjags.
Available from GitHub. https://github.com/khider/ *JAGSMgCA* repository.

Jackson, C. and Huerta, G. (2015). Code for "Empirical Bayes approach to climate model calibration". Written in Matlab and R. Available from Zenodo.
http://dx.doi.org/10.5281/zenodo.33545

Nosedal, A., Jackson, C. and Huerta, G. (2015). Code for "A new metric for climate models that includes field and spatial dependencies using Gaussian Markov Random Fields". Written in R. Available from Zenodo
http://dx.doi.org/10.5281/zenodo.33765

5 Presentations or Meetings Attended

1. Invited Speaker to the *Semanario Aleatorio* of the Instituto Tecnológico Autónomo de México (ITAM) "Inference and analyses of climate models via Bayesian Approaches". December 3, 2015, Mexico City, Mexico. (visit funded by ITAM).

2. Seminar at the Mathematics Department of New Mexico Tech. "Inference and analyses of climate models via Bayesian Approaches". November 20, 2015, Socorro, NM.

3. Introduction to a Bayesian Probability Framework. Invited Lecturer for Uncertainty Quantification in Climate Modeling and Projection. ICTP Campus, July 2015, Trieste, Italy. (1 hr. lecture + 1.5hr computer lab session)

4. Inference and Analysis of Climate Sensitivity via Data Reduction Techniques: A Detailed Approach. Joint Statistical Meetings, August 2015, Seattle, WA (with M. Hattab and C. Jackson).

5. (non) Emergent Constraints. Talk for the American Geophysical Union 2014 Fall Meeting, December 2014, San Francisco, CA (with M. Hattab and C. Jackson).

6. What Forced Holocene Millennial-Scale Variability? A Tale from the Western Tropical Pacific. Talk for the American Geophysical Union 2014 Fall Meeting, December 2014, San Francisco, CA. (with D. Khider and C. Jackson)

7. The Trouble in Quantifying Climate Uncertainty. Presentation at the Pacific Northwest National Laboratories (PNNL), November 2014, Richland, WA.

8. Statistical approaches for calibration of climate models. Seminar speaker at the Mind Research Network, August 2014, Albuquerque, NM.

9. Statistical approaches for calibration of climate models. Invited talk for the ICSA-KISS Joint Applied Statistics Symposium in Portland, Oregon, June 15-18, 2014

10. Toward a Statistical Framework for Quantifying Uncertainties in Climate using CAM. Invited talk for DOE Principal Investigators Meeting. Climate Sciences division, Washington, DC, May 2014. (joint with C. Jackson).

11. A new metric for climate models that includes spatial and field dependencies using GMRF. Invited talk at SIAM Conference in Uncertainty Quantification, Savannah, GA, April 2014. (joint with A. Nosedal-Sanchez and C. Jackson).

12. A new metric for climate models that includes spatial and field dependencies. Department of Mathematics. University of Nevada-Reno, June 2014, Reno, NV (with C. Jackson and A. Nosdeal-Sanchez).

13. A metric of CAM performance that includes field dependencies. National Center for Atmospheric Research, February 2014, Boulder, CO (with C. Jackson and A. Nosdeal-Sanchez)

14. Selecting, weeding, and weighting biased climate model ensembles. American Geophysical Union (AGU) meeting, December 2012, San Francisco, CA (with C. Jackson and A. Nosedal-Sanchez).

15. SIAM Conference on Uncertainty Quantification, Huerta, G, C. Jackson, A. Nosedal, G. Stark, J. Gattiker, D. Higdon (Spring 2012) "Measures of model skill and parametric uncertainty estimation in climate models"

16. International Society of Bayesian Analysis World Meeting, Huerta, G, C. Jackson, A. Nosedal, G. Stark, J. Gattiker, D. Higdon (Summer 2012) "Measures of model skill and parametric uncertainty estimation in climate models"

17. Joint Statistical Meetings of the ASA, Huerta, G., Stark G. (Summer 2012) "Dynamic and spatial modeling of precipitation extremes".

18. Using Stochastic Sampling of Parametric Uncertainties to Quantify Relationships Between CAM3.1 Bias and Climate Sensitivity. Department of Energy 2011 Climate Modeling Meeting, September 2011, Washington, DC (with C. Jackson).

19. DOE PI meeting, Jackson, C. S., Huerta, G. (Fall 2011) Assessing which climate model biases affect predictions

20. SAMSI, Program on Uncertainty Quantification, Pleasanton CA. Jackson, C. S., Huerta, G. (Fall 2011) Assessing which climate model biases affect predictions

21. AGU, Fall Meeting 2011, Jackson, C. S., Huerta G., Tobis, M. (2011) Using stochastic sampling of parametric uncertainties to quantify relationships between CAM3.1 bias and climate sensitivity (Invited)"

References Cited

[1] Robert F Adler, George J Huffman, Alfred Chang, Ralph Ferraro, Ping-Ping Xie, John Janowiak, Bruno Rudolf, Udo Schneider, Scott Curtis, David Bolvin, Arnold Gruber, Joel Susskind, Philip Arkin, and Eric Nelkin. The Version-2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979–Present). *http://dx.doi.org/10.1175/1525-7541(2003)004¡1147:TVGPCP¿2.0.CO;2*, 4(6):1147–1167, July 2009.

[2] Amy Braverman, Noel Cressie, and Joao Teixeira. A likelihood-based comparison of temporal models for physical processes. *Statistical Analysis and Data Mining*, 4(3):247–258, April 2011.

[3] W D Collins, P J Rasch, B A Boville, J J Hack, J R McCaa, D L Williamson, and B P Briegleb. The formulation and atmospheric simulation of the Community Atmosphere Model version 3 (CAM3). *Journal of Climate*, 19(11):2144–2161, 2006.

[4] P J Gleckler, K E Taylor, and C Doutriaux. Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 113(D6):1–20, March 2008.

[5] Reto Knutti, R Furrer, C Tebaldi, J Cermak, and Gerald A Meehl. Challenges in combining projections from multiple climate models. *Journal Of Climate*, 23(10):2739–2758, 2010.

[6] Thomas Reichler and Junsu Kim. How Well Do Coupled Models Simulate Today's Climate? *Bulletin of the American Meteorological Society*, 89(3):303–311, March 2008.

[7] B D Santer, K E Taylor, P J Gleckler, C Bonfils, T P Barnett, D W Pierce, T M L Wigley, C Mears, F J Wentz, W Brüggemann, N P Gillett, S A Klein, S Solomon, P A Stott, and M F Wehner. Incorporating model quality information in climate change detection and attribution studies. *Proceedings of the National Academy of Sciences*, 106(35):14778–14783, September 2009.

[8] KE Taylor. Summarizing multiple aspects of model performance in a single diagram. *Journal Of Geophysical Research-Atmospheres*, 106(D7):7183–7192, 2001.

[9] Kevin E Trenberth, Toshio Koike, and Kazutoshi Onogi. Progress and Prospects for Reanalysis for Weather and Climate. *Eos, Transactions American Geophysical Union*, 89(26):234–235, June 2008.

[10] S M Uppala, P W Kållberg, A J Simmons, U Andrae, V Da Costa Bechtold, M Fiorino, J K Gibson, J Haseler, A Hernandez, G A Kelly, X Li, K Onogi, S Saarinen, N Sokka, R P Allan, E Andersson, K Arpe, M A Balmaseda, A C M Beljaars, L Van De Berg, J Bidlot, N Bormann, S Caires, F Chevallier, A Dethof, M Dragosavac, M Fisher, M Fuentes, S Hagemann, E Hólm, B J Hoskins, L Isaksen, P A E M Janssen, R Jenne,

A P Mcnally, J F Mahfouf, J J Morcrette, N A Rayner, R W Saunders, P Simon, A Sterl, K E Trenberth, A Untch, D Vasiljevic, P Viterbo, and J Woollen. The ERA-40 re-analysis. *Quarterly Journal Of The Royal Meteorological Society*, 131(612):2961–3012, October 2005.

[11] Andreas P Weigel, Reto Knutti, Mark A Liniger, and Christof Appenzeller. Risks of Model Weighting in Multimodel Climate Projections. *Journal Of Climate*, 23(15):4175–4191, August 2010.

References Cited

[1] Daniel, C. 1959. Use of Half Normal Probability plots in Interpreting Factorial Two Level Experiments. *Technometrics***1**, 311–342

[2] Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., et al. 2009. The Version-2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979-Present). *J. Hydrometeor.***4**, 1147–1167.

[3] Bentamy, A., Queffeulou, P., Quilfen Y., and Katsaros, K. 1999. Ocean surface wind fields estimated from satellite active and passive microwave instruments. *IEEE Trans. Geosci. Remote Sensing* **37**,2469–2486.

[4] Caldwell, P. M., Bretherton, C. S., Zelinka, M. D., Santer, B. D., Klein, S. A., and Sanderson, B. M. (2014). Statistical significance of climate sensitivity predictors obtained by data mining. Geophysical . http://doi.org/10.1002/2014GL059205

[5] Fasullo, J. T., and Trenberth, K. E. (2012). A Less Cloudy Future: The Role of Subtropical Subsidence in Climate Sensitivity. Science, 338(6108), 792-794. http://doi.org/10.1126/science.1227465

[6] Harris, G. R., Sexton, D. M. H., Booth, B. B. B., Collins, M., Murphy, J. M., and Webb, M. J. (2006). Frequency distributions of transient regional climate change from perturbed physics ensembles of general circulation model simulations. Climate Dynamics, 27(4), 357-375. http://doi.org/10.1007/s00382-006-0142-8

[7] Huber, M., Mahlstein, I., Wild, M., Fasullo, J., and Knutti, R. (2011). Constraints on Climate Sensitivity from Radiation Patterns in Climate Models. Journal of Climate, 24(4), 1034-1052. http://doi.org/10.1175/2010JCLI3403.1

[8] Ingram, W. (2013). Some implications of a new approach to the water vapour feedback. Climate Dynamics, 40(3-4), 925-933. http://doi.org/10.1007/s00382-012-1456-3

[9] Klocke, D., Pincus, R., and Quaas, J. (2011). On Constraining Estimates of Climate Sensitivity with Present-Day Observations through Model Weighting. Journal of Climate, 24(23), 6092-6099. http://doi.org/10.1175/2011JCLI4193.1

[10] Loeb, N. G., Wielicki, B. A., Doelling, D. R., Smith, G. L., Keyes, D. F., Kato, S., et al. (2010). Toward Optimal Closure of the Earth's Top-of-Atmosphere Radiation Budget. Dx.Doi.org, 22(3), 748-766. http://doi.org/10.1175/2008JCLI2637.1

[11] Masson, D., and Knutti, R. (2013). Predictor Screening, Calibration, and Observational Constraints in Climate Model Ensembles: An Illustration Using Climate Sensitivity. Journal of Climate, 26(3), 887-898. http://doi.org/10.1175/JCLI-D-11-00540.1

[12] Piani, C., Frame, D. J., Stainforth, D. A., and Allen, M. R. (2005). Constraints on climate change from a multi-thousand member ensemble of simulations. Geophysical Research Letters, 32(23), L23825. http://doi.org/10.1029/2005GL024452

[13] Piani, C., Sanderson, B., Giorgi, F., Frame, D. J., Christensen, C., and Allen, M. R. (2007). Regional probabilistic climate forecasts from a multithousand, multi-model ensemble of simulations. Journal of Geophysical Research, 112(D24), D24108. http://doi.org/10.1029/2007JD008712

[14] Sanderson, B. M. (2013). On the estimation of systematic error in regression-based predictions of climate sensitivity. Climatic Change, 118(3-4), 757-770. http://doi.org/10.1007/s10584-012-0671-6

[15] Sherwood, S. C., Bony, S., and Dufresne, J.-L. (2014). Spread in model climate sensitivity traced to atmospheric convective mixing. Nature, 505(7481), 37-42. http://doi.org/10.1038/nature12829

[16] Uppala, S. M., Kallberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., et al. (2005). The ERA-40 re-analysis. Quarterly Journal of the Royal Meteorological Society, 131(612), 2961-3012. http://doi.org/10.1256/qj.04.176

[17] Volodin, E. M. (2008). Relation between temperature sensitivity to doubled carbon dioxide and the distribution of clouds in current climate models. Izvestiya, Atmospheric and Oceanic Physics, 44(3), 288-299. http://doi.org/10.1134/S0001433808030043

[18] Willmott, C. J. and K. Matsuura (2001) Global Air Temperature: Regridded Monthly and Annual Climatologies (V. 2.02) http://climate.geog.udel.edu/c̃limate/html_pages/README.lw2.html

[19] Yokohata, T., Annan, J. D., Collins, M., Jackson, C. S., Shiogama, H., Watanabe, M., et al. (2013). Reliability and importance of structural diversity of climate model ensembles. Climate Dynamics. http://doi.org/10.1007/s00382-013-1733-9

[20] Yokohata, T., Annan, J. D., Collins, M., Jackson, C. S., Tobis, M., Webb, M. J., and Hargreaves, J. C. (2011). Reliability of multi-model and structurally different single-model ensembles. Climate Dynamics, 39(3-4), 599-616. http://doi.org/10.1007/s00382-011-1203-1

[21] Multidecade Global Flux Datasets from the Objectively Analyzed Air-sea Fluxes (OAFlux) Project: Latent and sensible heat fluxes, ocean evaporation, and related surface meteorological variables. OAFlux Project Technical Report. OA-2008-01, 64pp. (2008) by L. Yu, X. Jin, R. A. Weller