Geosci. Model Dev., 9, 2407–2414, 2016
www.geosci-model-dev.net/9/2407/2016/
doi:10.5194/gmd-9-2407-2016

Geoscientific
Model Development
Open Access

A new test statistic for climate models that includes field and spatial dependencies using Gaussian Markov random fields

Alvaro Nosedal-Sanchez[1,2], **Charles S. Jackson**[3], **and Gabriel Huerta**[1]

[1]Department of Mathematics and Statistics, The University of New Mexico, Albuquerque, USA
[2]Department of Mathematical and Computational Sciences, University of Toronto, Mississauga, USA
[3]Institute for Geophysics, The University of Texas at Austin, Austin, USA

Correspondence to: Charles Jackson (charles@ig.utexas.edu)

Received: 15 November 2015 – Published in Geosci. Model Dev. Discuss.: 15 January 2016
Revised: 10 June 2016 – Accepted: 12 June 2016 – Published: 20 July 2016

Abstract. A new test statistic for climate model evaluation has been developed that potentially mitigates some of the limitations that exist 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 statistic 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 statistic 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.

1 Introduction

Climate scientists are interested in developing new metrics for assessing how well climate simulations reproduce observed climate for purposes of comparing models, driving model development, and evaluating model prediction uncertainties (Gleckler et al., 2008; Reichler and Kim, 2008; Santer et al., 2009; Knutti et al., 2010; Weigel et al., 2010; Braverman et al., 2011). Formal methods for accomplishing these goals, such as Bayesian calibration, operate with a single test statistic[1] for determining likelihood measures of different model configurations. A level of skepticism exists within the climate assessment community concerning the sufficiency of any one metric to judge a climate model's scientific credibility. Climate phenomena involve interactions of multiple fields (observables) on a wide range of timescales and space scales from minutes to decades (and longer) and from meters to planetary scales. Thus there are plenty of challenges that exist for synthesizing the many ways that a climate model can be tested against observational data.

The most common approach to climate model evaluation among climate scientists is to display maps of long-term means of well-known fields (e.g., temperature, sea-level pressure, precipitation) whose distribution is familiar and well understood in order to identify sources of model error. Taylor metrics that are often generated as part of model evaluation are based on spatial means of squared grid point errors for individual fields (Taylor, 2001). Such measures neglect field and space dependencies that arise as a consequence of how the physics of the climate system correlate multiple quantities in space. Neglecting these dependencies therefore ignores additional information that can be used to test whether models are simulating observables for the right reasons.

Here we present a new test statistic based on Gaussian Markov random fields (GMRFs) that addresses some of the

[1] A test statistic is a metric that includes information about the significance of modeling errors.

challenges that currently exist for estimating the significance of modeling errors across multiple fields that takes into account field and space dependencies that exist within observations. Perhaps one of the under-recognized challenges in this regard is the limited number of observations available to quantify dependencies. Data assimilation is commonly used to fill in gaps in the observational record (Trenberth et al., 2008). While assimilation products help address some aspects of the problem of how one compares point measurements to the scales resolved by climate models, these products include the space and field dependencies of the model that was used to assimilate observations. The imprint of the reanalysis model is readily seen when comparing two or more assimilation products, particularly quantities that are directly related to parameterized physics such as precipitation and radiation. One of the advantages of GMRFs is that they only need a limited amount of data to decipher space and field dependencies of climate phenomena. This is because GMRFs summarize relationship information as it is expressed across fields of gridded data.

The present application of GMRFs operates on long-term means. While it may be possible to extend GMRFs to capture time dependencies (Cressie and Wikle, 2011), the present application represents an advance over more traditional metrics.

The sections of this paper explain, test, and provide examples of how various components of GMRFs work. Section 2 gives a brief introduction to GMRFs and the use of a neighborhood structure for estimating dependency information using a precision operator **Q**. In this section we also define and discuss the Kronecker product and how it is used to generalize GMRFs to deal with more than one field. Section 3 introduces a graph for testing the extent to which GMRFs represent observed variance–covariances of tropical temperature, precipitation, sea level pressure, and upper level winds. Finally, in Sect. 4, 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 Gaussian Markov random fields (GMRFs)

A Gaussian Markov random field (GMRF) is a special case of a multivariate normal distribution. The density of a normal random vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ (where T denotes the operation of transposing a column to a row), with mean $\boldsymbol{\mu}$ ($n \times 1$ vector) and covariance matrix $\boldsymbol{\Sigma}$ ($n \times n$ matrix), is

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

Here, $\mu_i = E(x_i)$, $\Sigma_{ij} = \mathrm{Cov}(x_i, x_j)$, $\Sigma_{ii} = \mathrm{Var}(x_i) > 0$, and $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. Estimating $\boldsymbol{\Sigma}$ can be quite challenging in many contexts, especially for climate models where there are only limited data. All eigenvalues of $\boldsymbol{\Sigma}$ must be greater than zero, otherwise $\boldsymbol{\Sigma}^{-1}$ becomes a singular matrix and it does not define a valid multivariate normal distribution. It can also be shown that if all eigenvalues of $\boldsymbol{\Sigma}$ are positive, then all eigenvalues of $\boldsymbol{\Sigma}^{-1}$ are also greater than zero. Rather than estimating $\boldsymbol{\Sigma}$ and ensuring all eigenvalues of $\boldsymbol{\Sigma}^{-1}$ are positive, GMRFs make use of the precision matrix $\mathbf{P} = \boldsymbol{\Sigma}^{-1}$. We denote $\boldsymbol{x} \sim \mathbf{N}(\boldsymbol{\mu}, \mathbf{P})$ to represent $\boldsymbol{x}$ as a multivariate normal distribution with vector mean $\boldsymbol{\mu}$ and precision matrix $\mathbf{P}$. GMRFs approximate $f(\boldsymbol{x})$ using a sparse representation for $\mathbf{P}$ by setting all precisions outside a neighborhood structure to zero. Thus GMRFs make the assumption that points outside a neighborhood structure are conditionally independent. As we shall show below, this limitation does not prevent GMRFs from capturing covariances outside the neighborhood structure used to define precisions.

The GMRF-based expression that we have developed for quantifying the significance of differences between model output and observations is

$$\boldsymbol{v}^T \mathbf{S}^{-1} \otimes (\alpha \mathbf{I} + (1-\alpha)\mathbf{Q}) \boldsymbol{v}, \tag{2}$$

where $\boldsymbol{v}$ 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, $n_{\mathrm{obs}} n_{\mathrm{pts}}$, α is a scalar with a value close to zero, $\mathbf{I}$ stands for an identity matrix (a diagonal matrix of ones) of dimension n_{pts} corresponding to $\boldsymbol{v}$, and $\mathbf{Q}$ is a precision operator of dimension $n_{\mathrm{pts}} \times n_{\mathrm{pts}}$ from a GMRF induced by a first-order neighborhood structure. This cost function captures field dependencies through $\mathbf{S}^{-1}$, which is a matrix of dimension $n_{\mathrm{obs}} \times n_{\mathrm{obs}}$ where each of its elements represents a spatial average of grid point 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 of associating the different matrix dimensions of the metric, essentially combining its field and space components. Each of the following subsections provides additional information about the derivation and application of Eq. (2).

2.1 Precision operator of a GMRF

The precision operator of a GMRF **Q** provides a way to estimate dependencies among neighboring grid cells. **Q** needs to be constructed such that it

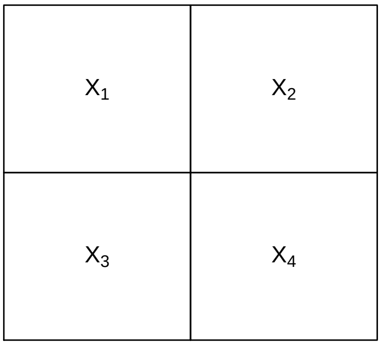

Figure 1. Graphical representation of a 2 × 2 lattice and elements of $\boldsymbol{x}$.

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

Consider $\boldsymbol{x}$, a vector of measurements on a 2×2 lattice, as represented in Fig. 1. Assume a neighborhood structure between the four elements of $\boldsymbol{x}$. In Fig. 2, the neighbors for each element of $\boldsymbol{x}$ are defined graphically. Given the neighborhood structure shown in Fig. 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 GMRFs is simple, the theory and mathematics are rather involved. A more full description of the mathematics of this example is provided in the Supplement. 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.

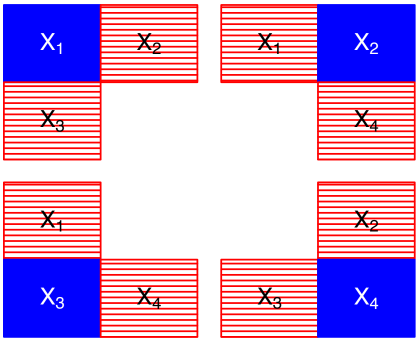

Figure 2. Neighbors of x_1, x_2, x_3 and x_4

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. Although $\mathbf{Q}$ may be inverted using the Moore–Penrose pseudoinverse, we have solved this problem by using $\alpha\mathbf{I}+(1-\alpha)\mathbf{Q}$, instead of $\mathbf{Q}$. If α is small, the neighborhood structure remains essentially unchanged. Section 3 describes our approach to specifying a value for α.

2.2 Generalizing concepts to deal with multiple fields

The generalization of $\mathbf{Q}$ to handle multiple fields involves a Kronecker product (⊗) between $\mathbf{S}^{-1}$ and $\mathbf{Q}$. For reference, a Kronecker product of $A\otimes B$ where

$$\mathbf{A}=\begin{pmatrix} 1 & 4\\ 2 & 5 \end{pmatrix} \text{ and } \mathbf{B}=\begin{pmatrix} 1 & 3\\ 0 & 4 \end{pmatrix}$$

is given by

$$\mathbf{A}\otimes\mathbf{B}=\begin{pmatrix} 1(\mathbf{B}) & 4(\mathbf{B})\\ 2(\mathbf{B}) & 5(\mathbf{B}) \end{pmatrix}=\begin{pmatrix} 1 & 3 & 4 & 12\\ 0 & 4 & 0 & 16\\ 2 & 6 & 5 & 15\\ 0 & 8 & 0 & 20 \end{pmatrix}.$$

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

$$\mathbf{S}=\begin{pmatrix} \sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22} \end{pmatrix},$$

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

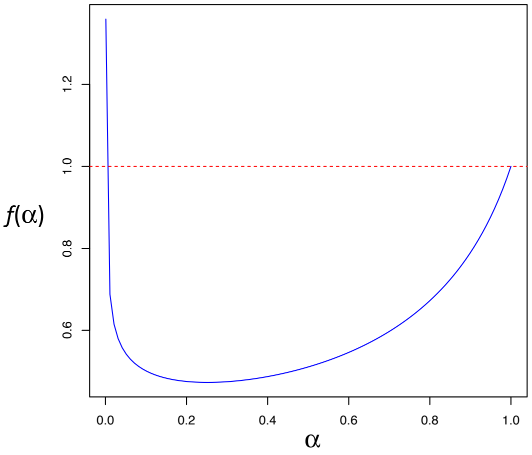

Figure 3. α vs. $f(\alpha)$.

$$\mathbf{S}^{-1} = \begin{pmatrix} \dfrac{1}{\sigma_{11}(1-\rho^2)} & \dfrac{-\rho}{(1-\rho^2)\sqrt{\sigma_{11}\sigma_{22}}} \\ \dfrac{-\rho}{(1-\rho^2)\sqrt{\sigma_{11}\sigma_{22}}} & \dfrac{1}{\sigma_{22}(1-\rho^2)} \end{pmatrix} = \begin{pmatrix} S_{11}^{-1} & S_{12}^{-1} \\ S_{21}^{-1} & S_{22}^{-1} \end{pmatrix}.$$

If we consider the Kronecker product in Eq. (2) 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

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

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. The Supplement carries this expression one step further by estimating the conditional mean for the first element of $\boldsymbol{v}$ to illustrate how this element is related to itself and its neighbors across multiple fields.

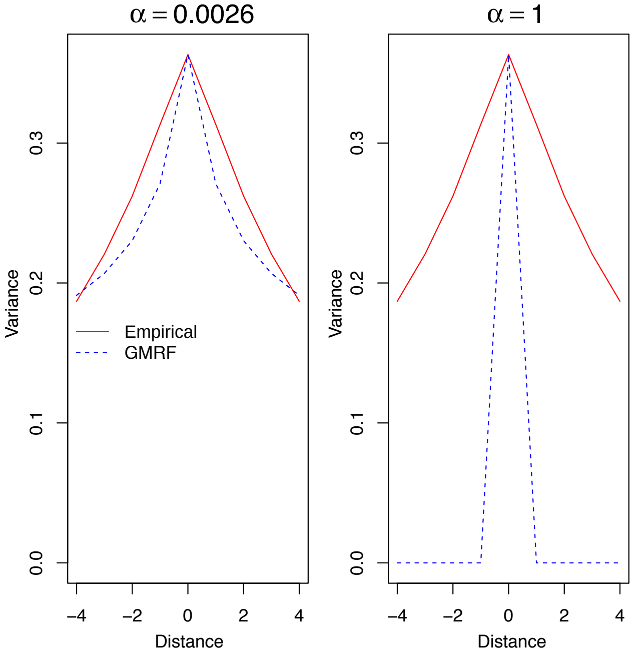

Figure 4. "Witch hat" graphs for air temperature on a 128 × 22 lattice of the tropics from 30° S to 30° N. The empirical estimates are given by the solid red line. The GMRF estimate is given by the dashed blue line.

Table 1. Correlation matrix between four fields from CAM 3.1.

	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

3 A test of GMRF estimates of variance

GMRFs provide a way to approximate field and space dependencies contained in the inverse covariance matrix $\mathbf{\Sigma}^{-1}$ of Eq. (1) by its GMRF equivalent $\mathbf{S}^{-1} \otimes (\alpha\mathbf{I} + (1-\alpha)\mathbf{Q})$. In this section, we will test how well GMRFs 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 estimate of the precision matrix with those obtained empirically from observational data. It turns out this comparison is sensitive to the value that is selected for α. By construction, 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 GMRFs do a reasonable job approximating observed field and space relationships.

3.1 Finding an appropriate value of α

In the effort to compare space and field dependencies 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

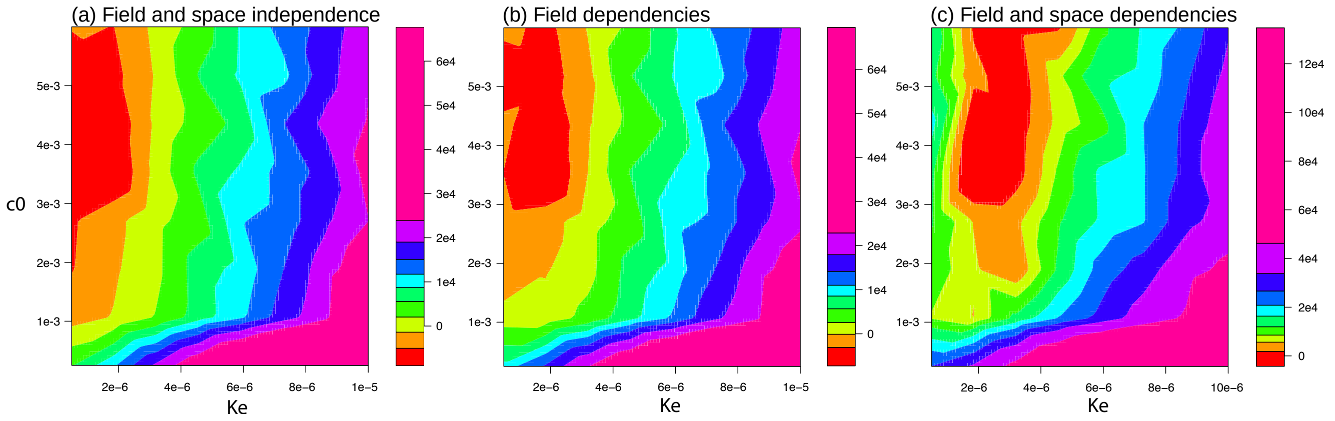

Figure 5. Three versions of the GMRF-based cost as a function of two CAM3.1 parameters *ke* and *c0* that assumes the data have **(a)** field and space independence, **(b)** field dependencies, and **(c)** field and space dependencies. Each color represents ten percentiles of the cost distribution. The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

$\left[\mathbf{S}^{-1}\otimes(\alpha\mathbf{I}+(1-\alpha)\mathbf{Q})\right]^{-1}=\mathbf{S}\otimes(\alpha\mathbf{I}+(1-\alpha)\mathbf{Q})^{-1}$. Letting $\mathbf{Q}^*=(\alpha\mathbf{I}+(1-\alpha)\mathbf{Q})^{-1}$, then $\mathbf{S}\otimes\mathbf{Q}^*$ for two fields can be written as

$$\begin{pmatrix} S_{11}\mathbf{Q}^* & S_{12}\mathbf{Q}^* \\ S_{21}\mathbf{Q}^* & S_{22}\mathbf{Q}^* \end{pmatrix}.$$

If n is the total number of grid points of the lattice, $\mathbf{S}\otimes\mathbf{Q}^*$ is a $2n\times 2n$ covariance matrix. Note that each element of $\mathrm{diag}(S_{ij}\mathbf{Q}^*)$ 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 for fields i and j. Therefore,

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

where $\mathrm{tr}(\mathbf{Q}^*)$ 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

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

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

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

Figure 3 shows the relationship between various values of α and $f(\alpha)$. The eigenvalues used to obtain this figure correspond to the precision operator, $\mathbf{Q}$, 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 0.0026. Note that this value is independent of fields since Eq. (5) does not contain any field-specific information.

3.2 "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 or east (for the positive distances) or to the south or west (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. The cells corresponding to one or more grid cells away are mostly on entries in parallel with the main diagonal. 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 2-year mean summertime temperatures. Setting $\alpha=1$ provides a solution to Eq. (5); however, this will shut down the effect of $\mathbf{Q}$ and only the variances at the reference point (lag 0) will be well represented. On the

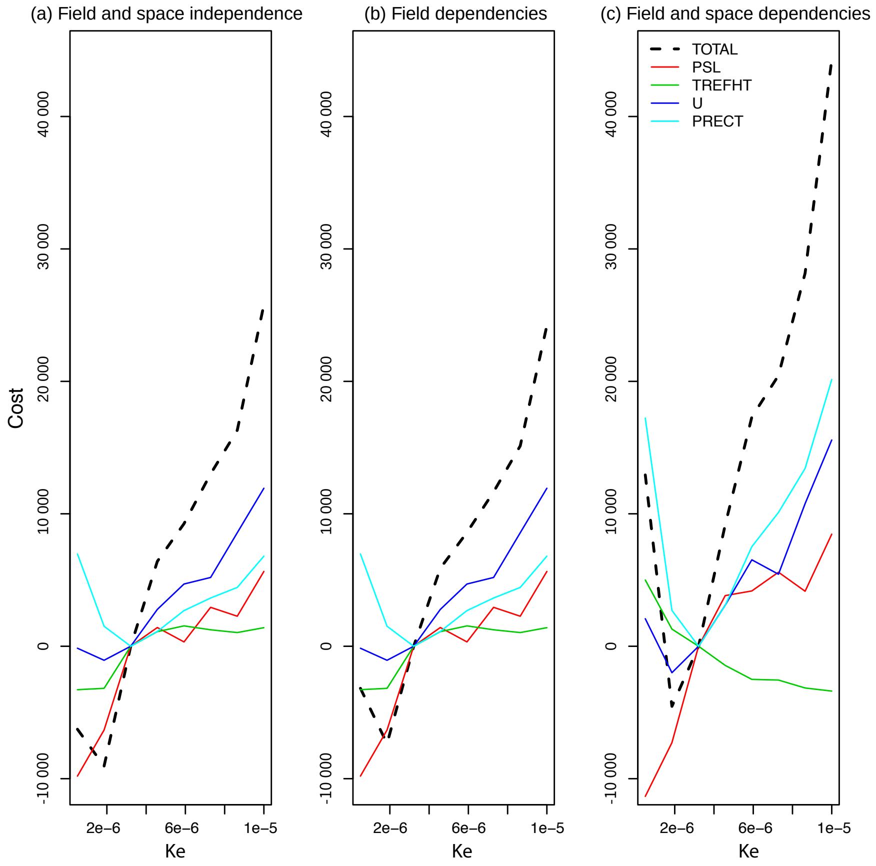

Figure 6. Different field contributions to the GMRF-based costs for a slice of Fig. 5 where $c0 = 0.0035$. Cost values are relative to the default parameter setting for *ke*. Note that total cost (black dashed line) is a weighted sum of field contributions as given by $\mathbf{S}^{-1}$ with contributions 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).

other hand, when $\alpha = 0.0026$, 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).

4 Climate response to uncertain parameters

In this section we show how inclusion of field and space dependencies using GMRF affects comparisons of the Community Atmosphere Model (CAM3.1) (Collins et al., 2006) with observations. We consider CAM3.1's response to changes in parameter *ke*, which controls raindrop 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 and 30° N on four variables including 2 m 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 are used to evaluate the model come from a ECMWF-ERA interim reanalysis product (Uppala et al., 2005) for 2 m air temperature, 200-millibar zonal winds, and sea level pressure and GPCP (Adler et al., 2009) for precipitation. We make use of approximately 30 years of JJA mean fields between 1979 and 2009. To construct 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 the GMRF test statistic which we refer to as a "cost" (Eq. 2). The first version is the traditional cost based on the assumption of space and field independence where the off diagonal components of $\mathbf{S}$ are set to zero and setting $\alpha = 1$. This approach is similar to what has been done previously for Taylor (2001). 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 $\alpha = 0.0026$.

The correlation matrix, **R**, corresponding to the **S** matrix of 2-year JJA seasonal mean variances and covariances, as estimated from 30 years of observations, is shown in Table 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. Maps of the grid point correlations between these fields show a lot of structure with regions of both positive and negative correlations. Therefore, providing a mechanistic explanation of the spatially averaged correlation is not particularly meaningful. Despite losing regional information in the **S** matrix summary of field covariances, GMRF estimated field covariances as seen within "witch hat" graphs are reasonable as compared to empirical estimates (see Supplement).

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 $ke = 3 \times 10^{-6}$, one can learn that the biggest qualitative difference comes from cost 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 GMRFs are 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.

5 Summary

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 exist for establishing a meaningful observational basis for quantifying field and space relationships of climate phenomena. Much of the data that are typically used for model evaluation are suspected of having their 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, has all the terms necessary to represent how different points in space are tied together across multiple fields. 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 GMRFs 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 for various lag windows or distances. 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.

6 Code and data availability

R code and data for generating Figs. 5 and 6 can be obtained through https://zenodo.org/record/33765 (Nosedal-Sanchez et al., 2015).

The Supplement related to this article is available online at doi:10.5194/gmd-9-2407-2016-supplement.

Acknowledgements. This material is based upon work supported by the US Department of Energy Office of Science, Biological and Environmental Research Regional & Global Climate Modeling Program under award numbers DE-SC0006985 and DE-SC0010843. Alvaro Nosedal-Sanchez was partially supported by the National Council of Science and Technology of Mexico (CONACYT).

Edited by: P. Ullrich
Reviewed by: two anonymous referees

References

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., Gruber, A., Susskind, J., Arkin, P., and Nelkin, E.: The Version-2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979–Present), J. Hydrometeorol., 1147–1167, doi:10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2, 2009.

Braverman, A., Cressie, N., and Teixeira, J.: A likelihood-based comparison of temporal models for physical processes, Statistical Analysis and Data Mining, 4, 247–258, 2011.

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

Cressie, N. and Wikle, C. K.: Statistics for Spatio-Temporal Data, Wiley, Hoboken, NJ, 2011.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res.-Atmos., 113, 1–20, 2008.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in combining projections from multiple climate models, J. Climate, 23, 2739–2758, 2010.

Nosedal-Sanchez, A., Jackson, C. S., and Huerta, G.: Code for "A new metric for climate models that includes field and spatial dependencies using Gaussian Markov Random Fields", Zenodo, doi:10.5281/zenodo.33765, 2015.

Reichler, T. and Kim, J.: How Well Do Coupled Models Simulate Today's Climate?, B. Am. Meteorol. Soc., 89, 303–311, 2008.

Santer, B. D., Taylor, K. E., Gleckler, P. J., Bonfils, C., Barnett, T. P., Pierce, D. W., Wigley, T. M. L., Mears, C., Wentz, F. J., Brüggemann, W., Gillett, N. P., Klein, S. A., Solomon, S., Stott, P. A., and Wehner, M. F.: Incorporating model quality information in climate change detection and attribution studies, P. Natl. Acad. Sci. USA, 106, 14778–14783, 2009.

Taylor, K.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res.-Atmos., 106, 7183–7192, 2001.

Trenberth, K. E., Koike, T., and Onogi, K.: Progress and Prospects for Reanalysis for Weather and Climate, Eos T. Am. Geophys. Un., 89, 234–235, 2008.

Uppala, S. M., Kållberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Berg, L. V. D., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J. F., Morcrette, J. J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J.: The ERA-40 re-analysis, Q. J. Roy. Meteor. Soc., 131, 2961–3012, 2005.

Weigel, A. P., Knutti, R., Liniger, M. A., and Appenzeller, C.: Risks of Model Weighting in Multimodel Climate Projections, J. Climate, 23, 4175–4191, 2010.