Combining Ensembles of Regional Climate Model Output via a Multivariate Markov Random Field Model

Stephan R. Sain,¹ Reinhard Furrer,² and Noel Cressie³

July 28, 2008

SUMMARY: Climate models have become an important tool in the study of climate and climate change, and ensemble experiments consisting of multiple climate model runs are used in studying and quantifying the uncertainty in climate model output. However, there are often only a limited number of model runs available for a particular experiment, and the statistical challenge is to characterize the distribution of the model output. To that end, we have developed a multivariate hierarchical approach, at the heart of which is a new representation of a multivariate Markov random field model. This approach allows for flexible modeling of the multivariate spatial dependencies, including the cross-dependencies between variables. We demonstrate this statistical model on an ensemble arising from a regional climate model experiment over the western United States and focusing on the projected change in seasonal temperature and precipitation over the next 50 years.

KEY WORDS: Lattice Data, Conditional Autoregressive Model (CAR), Bayesian Hierarchical Model, Climate Change.

¹Geophysical Statistics Project, Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307, ssain@ucar.edu. Corresponding author.
²Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401, rfurrer@mines.edu
³Department of Statistics, The Ohio State University, Columbus, OH 43210, ncressie@stat.ohio-state.edu.
1 Introduction

Many processes in the Earth system cannot be directly observed and computer modeling has become a primary mode for studying these processes. These models often encapsulate entire fields of knowledge and provide a virtual laboratory for understanding physical relationships and serve as a basis for making predictions. The Earth’s climate, for example, is determined by the flows of energy, water, gasses, etc., within and between the different components of the climate system, including atmosphere, oceans, terrestrial and marine biospheres, sea ice, etc. Climate models attempt to represent this system, as well as incorporating anthropogenic forcings to assess the impact of human intervention.

Atmosphere-ocean general circulation models (GCMs) couple an atmospheric model with an ocean model and typically have grid boxes on the scale of 200 to 500 km. Most of the smaller-scale climate processes are simplified or approximated in such models. Recently, the focus of climate models has shifted from global perspectives to including impacts of climate and climate change on regional scales. Regional climate models (RCMs) are similar in concept to GCMs, but with grid boxes typically on the scale of 20 to 100 km. This allows for more explicit modeling of the smaller-scale processes that are often only approximated in GCMs.

While climate models are deterministic, the output generated by these models is complex and subject to a number of sources of uncertainty. Initial climate states, assumptions about future forcings, and, of course, our understanding (or lack thereof) of the physical processes and their representation in computer models are all issues that may lead to uncertainty in the model output. To gain a better understanding of this model uncertainty, there is an increasing use of ensembles consisting of multiple model runs. These experiments may vary initial conditions (simple ensembles), model physics (perturbed-physics ensembles), specific models (multi-model ensembles), or some combination of all three, in an attempt to capture the range of variation in the model output.
1.1 A Statistical Representation of Climate Model Output

We propose a statistical model for combining the output from simple ensembles of RCMs in order to characterize the distribution of the model output. This statistical model will be formulated through what has now become the standard three-level hierarchical formulation, namely data model, process model, and prior distributions. The data model links the RCM output to an unobserved spatial process, where this process model is formulated to capture the spatial variation in the RCM output. Both the data model and the process model depend on unknown parameters to which a prior distribution is assigned.

This basic hierarchical approach has been used in other settings for combining climate model output; see, for example, Tebaldi et al. (2005), Furrer et al. (2007), Berliner and Kim (2008), and Tebaldi and Sansó (2008). However, to our knowledge, this is the first approach of this kind for a spatial analysis of the multivariate output from RCMs. In addition, at the heart of this statistical model is a novel implementation of a multivariate Markov random field (MRF) for lattice data that offers improved flexibility in modeling the cross-dependencies.

MRF models are excellent tools for analyzing data laid out on regular spatial lattices, such as those associated with images, remote-sensing, climate models, etc., or on irregular spatial lattices, such as U.S. census divisions (counties, tracts, or block-groups) or other administrative units. In contrast to geostatistical methods that model spatial dependence through the specification of a covariance function that is typically based on distances between spatial locations, Gaussian MRF models represent observations at a spatial location as a linear combination of observations at neighboring locations, thus inducing spatial dependence through this autoregression and the neighborhood structure in the data. Using a MRF formulation will allow us to incorporate computational advantages due to the gridded nature of the climate model output and the sparseness that is characteristic of the spatial-precision matrices (inverse covariance matrices) that are specified in such models. Furthermore, the multivariate nature of the statistical model (more than one model output considered at each grid box) will allow for more complex inferences that are of use to those
studying impacts of climate and climate change.

1.2 Outline

In the following section, an overview of MRF models is presented, followed by a description of the new formulation in Section 3. Section 4 contains the details of the hierarchical specification. An extensive study of an application using a simple ensemble of regional climate model output focusing on changes in seasonal temperature and precipitation will be presented in Section 5.

2 MRF and CAR Models

Besag (1974) laid out the basic framework for MRF models. For random variables \( y_1, \ldots, y_n \) observed at \( n \) locations on a spatial lattice structure, the collection of conditional distributions \( f(y_i|y_{-i}), \ i = 1, \ldots, n \) (where \( y_{-i} \) refers to all random variables except the \( i \)th one) can be combined under certain regularity conditions to form a joint distribution \( f(y_1, \ldots, y_n) \).

Rue and Held (2005) can be consulted for an excellent exposition of the theory of MRFs; see also the reviews in the texts by Cressie (1993), Banerjee et al. (2004), and Schabenberger and Gotway (2005). Conditional autoregressive (CAR) models are special cases of MRF models where the conditional distributions are assumed to be Gaussian. It should be noted that the conditional specification is not the same as the so-called simultaneously specified autoregressive (SAR) models and are actually a more general construction (see Haining, 1990, and Cressie, 1993, for discussion and comparison of these two formulations).

2.1 Univariate CAR Models

In the univariate setting and assuming Gaussian conditional distributions for \( f(y_i|y_{-i}) \), the conditional mean and conditional variance associated with \( f(y_i|y_{-i}) \) are specified as

\[
E[y_i|y_{-i}] = \mu_i + \sum_{j=1}^{n} b_{ij}(y_j - \mu_j) \quad \text{and} \quad \text{Var}[y_i|y_{-i}] = \tau^2_i,
\]
for \( i = 1, \ldots, n \). This collection of conditional distributions gives rise to a joint Gaussian distribution,

\[
\mathcal{N}(\mu, (I - B)^{-1}T),
\]

where \( \mu = (\mu_1, \ldots, \mu_n)' \), \( I \) is an \( n \times n \) identity matrix, \( B \) is the \( n \times n \) matrix with the \( i, j \)th element \( b_{ij} \), and \( T = \text{diag}(\tau_1^2, \ldots, \tau_n^2)' \). Of course, the spatial dependence parameters, \{\( b_{ij} \}\}, must be chosen to ensure that the resulting matrix, \((I - B)^{-1}T\), is a bona fide covariance matrix; that is, \((I - B)^{-1}T\) is symmetric and positive-definite.

The spatial dependence is induced by the autoregression defined by typically setting \( b_{ij} \neq 0 \) if \( j \in N_i \) (and 0 otherwise), where \( N_i \) is a collection of indices that define a neighborhood of the \( i \)th location in the spatial lattice. Note that \( b_{ii} = 0 \), as one cannot regress the value at a particular location on the lattice on itself. Often, for regular lattices or even irregular lattices such as census or other government or administrative units, neighborhoods are defined by lattice locations that share an edge, although more complex neighborhood structures are also possible. Symmetry of the covariance matrix \((I - B)^{-1}T\) is assured if \( b_{ij} \tau_j^2 = b_{ji} \tau_i^2 \).

One simple parameterization assumes that \( \tau_j^2 = \tau^2 \) (homogeneity) and \( b_{ij} = \phi I_{j \in N_i} \), which simplifies the covariance matrix to \( \tau^2(I - \phi C)^{-1} \), where \( C \) is an adjacency matrix such that \( c_{ii} = 0 \) and \( c_{ij} = 1 \) if \( j \in N_i \) (and 0 otherwise). There are two points to make concerning this formulation that also apply to more complicated parameterizations. First, the conditional mean simplifies to \( \mathbb{E}[y_i | y_{-i}] = \mu_i + \phi \sum_{j \in N_i} (y_j - \mu_j) \), and \( \phi \) controls the strength of the autocorrelation by controlling the strength of the conditional dependence between the \( i \)th location and the neighbors specified through \( N_i \). Of course, the range of possible values for \( \phi \) is restricted by the neighborhood structure implicit in the adjacency matrix, \( C \); that is, \( 1/c_1 < \phi < 1/c_n \), where \( c_1 < 0 \) and \( c_n > 0 \) are the smallest and largest eigenvalues of \( C \), respectively (see Cressie, 1993, Ch. 7). Secondly, the off-diagonal elements of the inverse covariance, \((I - \phi C)/\tau^2\), are either zero, implying conditional independence between the corresponding locations on the lattice, or \(-\phi/\tau^2\), implying that the locations are not conditionally independent.
2.2 Multivariate CAR Models

Mardia (1988) extended the Markov random field model of Besag (1974) to the multivariate setting where there is more than one measurement at each lattice point. In particular, let $y_i$ be a $p$-dimensional random vector, and, for $i = 1, \ldots, n$, let $f(y_i|y_{-i})$ be a Gaussian conditional distribution with

$$E[y_i|y_{-i}] = \mu_i + \sum_{j=1}^{n} B_{ij}(y_j - \mu_j)$$

and

$$\text{Var}[y_i|y_{-i}] = T_i,$$

where $\mu_i$ is a $p$-vector, $B_{ij}$ is a general $p \times p$ matrix, and $T_i$ is a $p \times p$ covariance matrix. Assume that $B_{ij} T_j = T_i B_{ji}'$ for all $i, j = 1, \ldots, n$ (to ensure symmetry), $B_{ii} = -I$, and $B_{ij} \neq 0$ for $j \in N_i$ and $i = 1, \ldots, n$. Let $\text{Block}(A_{ij})$ denote the generic block matrix,

$$\begin{bmatrix}
A_{11} & \cdots & A_{1n} \\
\vdots & \ddots & \vdots \\
A_{m1} & \cdots & A_{mn}
\end{bmatrix}.$$ 

Under the assumption that $\text{Block}(-B_{ij})$ is positive-definite, Mardia (1988) establishes that $y = (y_1', \ldots, y_n')'$ follows a $\mathcal{N}(\mu, \Sigma)$ distribution where

$$\mu = (\mu_1', \ldots, \mu_n')' \quad \text{and} \quad \Sigma = \left(\text{Block}(-T_i^{-1}B_{ij})\right)^{-1}.$$

As written, this formulation is overparameterized and there have been a number of efforts in the literature that focus on ways of specifying the parameters in the basic model. See, for example, Billheimer et al. (1997), Kim et al. (2001), Sain and Cressie (2002), Pettitt et al. (2002), Carlin and Banerjee (2003), Gelfand and Vounatsou (2003), Jin et al. (2005), Daniels et al. (2006), Jin et al. (2007), Sain and Cressie (2007), among others.

3 A New Multivariate MRF Formulation

The multivariate extension of the framework laid out by Besag (1974) and explored by Mardia (1988) is based on the assumption of a multivariate observation at each point on a standard two-dimensional spatial lattice. We expand upon a different way of representing multivariate lattice data that was proposed by Sain and Cressie (2002, 2007). Fundamental
to this new approach is the basic idea of thinking of multivariate lattice data as univariate data on a more complex lattice structure.

In particular, this more complex lattice structure is conceptualized as a collection or a “stacking” of the lattices associated with each variable. Neighborhoods are defined by connections between locations for each variable within a lattice and again for locations across each lattice structure. A two-dimensional example, with representations of these different types of neighborhoods, is shown in Figure 1. A within-variable dependence structure is induced by connecting locations within a lattice associated with a particular variable (left frame). Cross-dependencies, both within a location (middle frame) and across locations (right frame) are induced through connections between the lattices for different variables.

The key feature of this approach is that it still falls within the original univariate framework of Besag (1974) outlined in Section 2.1. Let $y_{ij}$ denote the $j$th variable observed at the $i$th location on the lattice. Then, for each Gaussian conditional distribution, the mean and variance need to be specified. With sums on the right-hand side corresponding
to specific types of neighborhoods in Figure 1, the conditional mean is given by

\[
E[y_{ij}|y_{-\{ij\}}] = \mu_{ij} + \sum_{k} b_{ijkj}(y_{kj} - \mu_{kj}) \quad \text{(left)}
\]

\[
+ \sum_{\ell} b_{ij\ell\ell}(y_{\ell\ell} - \mu_{\ell\ell}) \quad \text{(middle)}
\]

\[
+ \sum_{k,\ell} b_{ijk\ell}(y_{k\ell} - \mu_{k\ell}) \quad \text{(right)}
\]

(2)

with the conditional variance given by

\[
\text{Var}[y_{ij}|y_{-\{ij\}}] = \tau_{ij}^2,
\]

for all lattice points \(i = 1, \ldots, n\), variables \(j = 1, \ldots, p\), and with \(y_{-\{ij\}}\) denoting all components of \(y\) except for the \(i, j\)th.

In the conditional mean, the coefficients in the first summation, \(\{b_{ijkj}; i = 1, \ldots, n, k \in N_i, j = 1, \ldots, p\}\), represent connections within a particular layer and control conditional dependence between the \(i\)th lattice point and neighboring points for the \(j\)th variable (left frame in Figure 1). The coefficients in the second summation, \(\{b_{ij\ell\ell}; i = 1, \ldots, n, j \neq \ell = 1, \ldots, p\}\), represent connections across layers at the same lattice point and control conditional dependence between variables \(j\) and \(\ell\) at the \(i\)th lattice point (middle frame in Figure 1). Finally, the coefficients in the third summation, \(\{b_{ijk\ell}; i = 1, \ldots, n, k \in N_i, j \neq \ell = 1, \ldots, p\}\), represent connections between locations across layers for different variables and control conditional cross-spatial dependence (right frame in Figure 1). Of course, all of these conditional dependence parameters and the variances \(\{\tau_{ij}^2\}\) must be chosen to yield a covariance matrix for the joint distribution that is symmetric and positive-definite.

Some simplification of this basic structure is necessary. Ordering the data as \(y = [y'_1, \ldots, y'_n]'\) where \(y_i = [y_{i1}, \ldots, y_{ip}]'\) represents variables at the \(i\)th lattice point, we see from (1) that the joint distribution is Gaussian with mean given by \(\mu = [\mu'_1, \ldots, \mu'_n]'\) and
\( \mu_i = [\mu_{i1}, \ldots, \mu_{ip}]', \) and with \((np \times np)\)-covariance matrix given by

\[
\begin{bmatrix}
A_1 & B_{i2}I_{12} & \cdots & B_{1n}I_{1n} \\
B_{21}I_{21} & A_2 & \vdots & \\
\vdots & \ddots & \ddots & \\
B_{n1}I_{n1} & \cdots & \cdots & A_n
\end{bmatrix}
\begin{bmatrix}
\tau_{i1}^{-2} & 0 & \cdots & 0 \\
0 & \tau_{i2}^{-2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
-b_{i\ell ij}/\tau_{i1} & -b_{i\ell ij}/\tau_{i2} & \cdots & 1
\end{bmatrix}
^{-1}

T,
\] (3)

where \(I_{ik} = 1\) if \(k \in N_i\) and 0 otherwise. Each \((p \times p)\) block is given by

\[
A_i = \begin{bmatrix} 1 & -b_{iji\ell} \\ -b_{i\ell ij} & 1 \end{bmatrix} \quad \text{or} \quad B_{ik} = \begin{bmatrix} -b_{i\ell kj} & -b_{ipkp} \\ -b_{i\ell kj} & 1 \end{bmatrix},
\]

where \(-b_{iji\ell}\) and \(-b_{i\ell ij}\) are arbitrary off-diagonal elements of \(A_i\), and \(-b_{i\ell kj}\) and \(-b_{ipkp}\) are arbitrary off-diagonal elements of \(B_{ik}\). Finally, \(T = \text{diag}(\tau_1^2, \ldots, \tau_p^2)\).

In general, we will assume that the neighborhood structure is symmetric; that is, if the \(k\)th lattice point is a neighbor of the \(i\)th lattice point, then the \(i\)th lattice point is a neighbor of the \(k\)th \((I_{ik} = I_{ki})\). We will also assume that \(\tau_{ij}^2 = \tau_j^2\) for all \(j\), implying a separate variance for each variable and generalizing the homogeneity assumption in the univariate setting. Hence, \(T = I_n \otimes \text{diag}(\tau)\), where \(\tau = [\tau_1^2, \ldots, \tau_p^2]'\).

Our specification of the conditional dependence structure must ensure that the resulting covariance is symmetric and positive-definite. Some assumptions must also be made to reduce the dimensionality of the parameters. It will be seen below that the conditional dependence parameters are assumed to constant across all locations; that is, \(A_i \equiv A\) and \(B_{ik} \equiv B\).

To address symmetry and the simplification across location, it suffices to examine the components of specific blocks in the inverse of the covariance matrix. First, the diagonal blocks are given by

\[
\text{diag}(\tau)^{-1}A_i = \begin{bmatrix}
\tau_{i1}^{-2} & 0 & \cdots & 0 \\
0 & \tau_{i2}^{-2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
1/\tau_{i1}^2 & -b_{i\ell ij}/\tau_{i2}^2 & \cdots & 1/\tau_{ip}^2
\end{bmatrix}
\]
By symmetry, the corresponding off-diagonal elements should be equal; that is,

\[ \frac{b_{ij\ell}}{\tau_j^2} = \frac{b_{\ell ij}}{\tau_\ell^2}. \]

Setting \( b_{ij\ell} = \rho_{j\ell} \tau_j / \tau_\ell \) with \( \rho_{ij} = \rho_{j\ell} \) is one way of achieving the desired result. Then,

\[
\text{diag}(\tau)^{-1} A_i = \begin{bmatrix}
1/\tau_1^2 & -\rho_{j\ell}/(\tau_j \tau_\ell) \\
-\rho_{j\ell}/(\tau_j \tau_\ell) & 1/\tau_p^2
\end{bmatrix}
= \text{diag}(\tau)^{-1/2} A \text{diag}(\tau)^{-1/2},
\]

where

\[ A = \begin{bmatrix}
1 & -\rho_{j\ell} \\
\cdots & \cdots \\
-\rho_{j\ell} & 1
\end{bmatrix}. \]

For the off-diagonal blocks, symmetry demands that

\[
\text{diag}(\tau)^{-1} B_{ik} = [\text{diag}(\tau)^{-1} B_{ki}]'.
\] (4)

Assuming \( i > k \), the left-hand side of (4) is given by

\[
\text{diag}(\tau)^{-1} B_{ik} = \begin{bmatrix}
\tau_1^{-2} & 0 & \cdots & -b_{i1k1} & -b_{ijkl} \\
0 & \tau_2^{-2} & \cdots & -b_{i\ell kj} & -b_{i\ell kp} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
-b_{i1k1}/\tau_1^2 & -b_{ijkl}/\tau_j^2 & \cdots & -b_{i\ell kj}/\tau_\ell^2 & -b_{ipkp}/\tau_p^2
\end{bmatrix}
= \begin{bmatrix}
\tau_1^{-2} & 0 & \cdots & -b_{i1k1} & -b_{ijkl} \\
0 & \tau_2^{-2} & \cdots & -b_{i\ell kj} & -b_{i\ell kp} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
-b_{i1k1}/\tau_1^2 & -b_{ijkl}/\tau_j^2 & \cdots & -b_{i\ell kj}/\tau_\ell^2 & -b_{ipkp}/\tau_p^2
\end{bmatrix}.
\]

Likewise, setting \( b_{ijkl} = \phi_{j\ell} \tau_j / \tau_\ell \) gives

\[
\text{diag}(\tau)^{-1} B_{ik} = \begin{bmatrix}
-\phi_{11}/\tau_1^2 & -\phi_{j\ell}/(\tau_j \tau_\ell) \\
-\phi_{\ell j}/(\tau_\ell \tau_j) & -\phi_{pp}/\tau_p^2
\end{bmatrix}
= \text{diag}(\tau)^{-1/2} B \text{diag}(\tau)^{-1/2},
\]
where

\[ B = \begin{bmatrix}
-\phi_{11} & -\phi_{j\ell} \\
\vdots & \ddots \\
-\phi_{\ell j} & -\phi_{pp}
\end{bmatrix}. \]

Then from the symmetry condition in (4), for \( i > k \),

\[ \text{diag}(\tau)^{-1}B_{ki} = \text{diag}(\tau)^{-1/2}B' \text{diag}(\tau)^{-1/2}. \]

The covariance in (3) then simplifies to

\[ \left[ I_n \otimes \tau^{1/2} \right] \left[ I_n \otimes A - \begin{bmatrix} 0 & BI_{ij} \\ B'_{ij} & 0 \end{bmatrix} \right]^{-1} \left[ I_n \otimes \tau^{1/2} \right], \quad (5) \]

where \( \tau^{1/2} = [\tau_1, \ldots, \tau_p]' \).

The specifications above simply ensure symmetry and reduce the number of parameters that must be estimated. Of course, the collection of spatial dependence parameters, \( \{\rho_{j\ell}\} \) and \( \{\phi_{j\ell}\} \), must be chosen to ensure that (5) is a positive-definite covariance matrix. The final model has \( p(p-1)/2 \) within-location dependence parameters, \( \{\rho_{j\ell}\} \), and \( p^2 \) between-location spatial dependence parameters, \( \{\phi_{j\ell}\} \), in addition to the \( p \) scale parameters, \( \{\tau_j^2\} \), and any parameters that are used to define the means.

We note that a more general formulation can be considered by setting \( \tau_{ij}^2 = \tau_j^2/M_{ij} \) where \( M_{ij} \) are known constants. Such a formulation would allow for known sources of heterogeneity, such as when the \( y_{ij} \) represent counts or where one wishes to link the conditional variance to the number of neighbors on an irregular lattice. See, for example, the discussion in Sain and Cressie (2007). Of course, assuming the conditional variances have such a form will necessarily lead to a different structure in the simplification of the spatial dependence parameters to ensure that (5) is symmetric. However, our experience in this setting suggests that this more complex structure does not improve the analysis, in particular since the regional climate model output is on a regular lattice.
4 A Hierarchical Model

Let the \( n \)-vector \( y_{rj} \) denote the output of an RCM, in particular the \( r \)th ensemble member for the \( j \)th variable. In this work, we will focus solely on simple ensembles; that is, each member of the ensemble represents a perturbation of initial conditions for a single model. Potential extensions of this basic framework for perturbed physics or multi-model ensembles is discussed in Section 6.

At the first level of the hierarchy, the data model assumes that the vectors \( y_{rj}, r = 1, \ldots, m, j = 1, \ldots, p, \) are independent with

\[
y_{rj} \sim \mathcal{N} \left( X_1 \alpha_j + X_2 \beta_{rj} + h_{rj}, \Sigma_j \right),
\]

where \( m \) indicates the number of ensembles. In the mean structure, we allow for effects common to all ensemble members within a particular variable (\( X_1 \alpha_j \)) and effects specific to each ensemble member within a particular variable (\( X_2 \beta_{rj} \)). Spatial random effects are included through the \( \{ h_{rj} \} \), and \( \Sigma_j \) represents an \( n \times n \), variable-specific covariance matrix.

The process model has two parts. First, the vectors \( [\beta'_{r1}, \ldots, \beta'_{rp}]' \), \( r = 1, \ldots, m, \) are assumed to be independent with

\[
\begin{pmatrix} \beta_{r1} \\ \vdots \\ \beta_{rp} \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \Sigma_b \right),
\]

where \( \Sigma_b \) is \( pq \times pq \) covariance matrix with \( q \) indicating number of columns of \( X_2 \). Second, the vectors \( [h'_{r1}, \ldots, h'_{rp}]' \), \( r = 1, \ldots, m, \) are assumed to be independent with

\[
\begin{pmatrix} h_{r1} \\ \vdots \\ h_{rp} \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} h_1 \\ \vdots \\ h_p \end{pmatrix}, V \left( \{ \tau_j^2 \}, \{ \rho_{j\ell} \}, \{ \phi_{j\ell} \} \right) \right).
\]

The first part, (7), focuses on linking the ensemble-specific random regression coefficients, while the second part, (8), imposes a multivariate structure on the spatial random effects. The covariance matrix \( V \) takes its form from the multivariate Markov random field in (5).

The final level of the hierarchy assumes prior distributions on \( \{ \Sigma_j \}, \{ \alpha_j \}, \{ \beta_j \}, \{ h_j \}, \Sigma_b \), and the parameters of the spatial covariance, namely \( \{ \tau_j^2 \}, \{ \rho_{j\ell} \}, \) and \( \{ \phi_{j\ell} \} \). Typically
these priors will be vague or non-informative as well as independent. In addition, the prior distribution on $\{\rho_{j\ell}\}$ and $\{\phi_{j\ell}\}$ must ensure that the resulting covariance matrix is positive-definite.

From Bayes’ Theorem, the posterior distribution for the three-level hierarchical model is given by

$$P(\{\beta_{rj}\}, \{h_{rj}\}, \{\alpha_j\}, \{\Sigma_j\}, \{\beta_j\}, \Sigma_b, \{h_j\}, \{\tau_{j}^2\}, \{\rho_{j\ell}\}, \{\phi_{j\ell}\}|Y) \propto P(Y|\{\alpha_j\}, \{\beta_{rj}\}, \{h_{rj}\}, \{\Sigma_j\})$$

$$\times P(\{\beta_{rj}\}|\{\beta_j\}, \Sigma_b)P(\{h_{rj}\}|\{h_j\}, \{\tau_{j}^2\}, \{\rho_{j\ell}\}, \{\phi_{j\ell}\})$$

$$\times P(\{\alpha_j\})P(\{\Sigma_j\})P(\{\beta_j\})P(\{\Sigma_b\})P(\{h_j\})P(\{\tau_{j}^2\})P(\{\rho_{j\ell}\}, \{\phi_{j\ell}\})$$.

It is clear that there is no closed-form solution for the posterior and Markov chain Monte Carlo (MCMC) (e.g. Gilks et al., 1996) is used to simulate realizations from the posterior distribution. In particular, we implement a Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990; Gelfand et al., 1990), incorporating Metropolis-Hastings steps (Metropolis et al., 1953; Hastings, 1970) where necessary.

One benefit of a MRF is that the specification involves the precision or inverse covariance matrix, and this matrix is typically sparse; that is, many of the elements of the matrix are zero. Methods for storing and manipulating such matrices have been widely established (e.g. Davis, 2006), and there is great potential for computational efficiency associated with sparse-matrix methods. There are now several sparse-matrix packages in the R statistical computing environment (R Development Core Team, 2007). However, the spam package (Furrer, 2008) has functionality that is well suited for implementing MCMC with a MRF model. For example, the sparse Cholesky decomposition is one of the most important computational devices used when implementing the Gibbs sampler for MRF models such as those developed in this work. A typical sparse Cholesky decomposition involves three steps: 1) reorganizing the matrix by permuting the rows/columns to achieve a pattern of sparsity that is more efficient for the sparse Cholesky algorithm, 2) a symbolic step that identifies the pattern of sparsity in the matrix, and 3) the numerical computation. The first
two steps do not change when manipulating matrices repeatedly with the same patterns of sparsity (as is the case here). The \texttt{spam} package allows us to achieve even greater computational efficiency by not repeating these steps during the course of the MCMC. For more on this and other computational benefits gained from incorporating sparse-matrix methods in such applications, see Furrer and Sain (2008).

5 Case Study

Leung \textit{et al.} (2004) describe an RCM experiment using the NCAR/DOE Parallel Climate Model to drive the NCAR/Penn State Mesoscale Model (MM5) as an RCM. The experiment produced a control run from 1995–2015 and three future runs (ensembles) from 2040–2060. The domain consisted of the western United States and part of western Canada, and the model used a “business as usual” climate scenario incorporating a 1% annual increase in the amount of greenhouse gases.
The \( n = 44 \times 56 = 2464 \) grid boxes form a regular lattice, and, for this analysis, twenty-year winter (December, January, and February) average temperature and total precipitation were computed for each grid box and for each of the control and the three future runs. Differences between the future and the control were calculated. Hence, there are \( p = 2 \) variables (change in average temperature and change in total precipitation) and \( m = 3 \) ensembles, giving six fields to be analyzed. These spatial fields for the winter season are shown in Figure 2. A second, separate analysis with the same structure was also conducted for the twenty-year summer (June, July, and August) change in average temperature and change in total precipitation.

5.1 Model Specification

We now outline some specifics about the statistical-model specification. After some exploratory analysis, (scaled) latitude, longitude, and elevation were used as covariates in the common regression component \( (X_1 \alpha_j) \), and a random intercept across ensembles was included \( (X_2 \beta_{rj}) \). The variable-specific covariance matrix was simplified to \( \Sigma_j = \sigma_j^2 I_n \), for \( j = 1, 2 \). The prior covariance matrix for the random intercept was also simplified to \( \sigma_b^2 I_p \).

The prior distribution for the variance parameters, \( \{\sigma_j^2\} \) and \( \sigma_b^2 \), were taken to be non-informative; that is \( P(\sigma^2) \propto 1/\sigma^2 \). The prior distributions for the regression parameters \( \{\alpha_j\} \) and \( \{\beta_j\} \) were taken to be mean-zero Gaussian distributions with covariance matrices proportional to the identity and with large variances. The prior distributions for \( \{h_j\} \) were also taken to be mean-zero Gaussian distributions with covariance matrices proportional to the identity and with large variances. Finally, the prior specification for the joint distribution of \( \rho, \phi_{11}, \phi_{22}, \phi_{12}, \) and \( \phi_{21} \) was taken to be uniform over the range of values that yield a positive-definite covariance matrix. This parameter space is, in general, a complex region that is difficult to define analytically. Hence, the prior was further simplified to be uniform over the region where the sparse Cholesky decomposition did not fail; that is, it was determined numerically using a rejection method for sampling from it.
5.2 Results for the Winter Season

Posterior distributions were obtained using MCMC algorithms, and considerable care was taken to ensure the convergence of the parameters in the MCMC. This is especially true with respect to the conditional-dependence parameters where our experience has shown that straightforward approaches can lead to disappointing performance (i.e. very slow mixing and convergence). Ten chains were run, each with random starting values. The conditional-dependence parameters were chosen uniformly across the space of values that yield a positive-definite covariance matrix.

A Gibbs sampler was implemented that involved three distinct regimes. In the first regime (2500 iterations), each of the conditional-dependence parameters was updated one-at-a-time using Metropolis-Hastings. Gaussian proposal distributions were used, with periodic updates of the proposal variance in an attempt to hit an approximate 20% acceptance rate. In the second regime (the next 10,000 iterations), $\rho$, $\phi_{12}$, and $\phi_{21}$ were updated simultaneously using Metropolis-Hastings with a multivariate Gaussian proposal distribution. Again, the proposal covariance matrix was updated periodically to ensure an approximate 20% acceptance rate. Other conditional-dependence parameters were still updated using univariate Metropolis-Hastings. Finally, in the third regime (the last 10,000 iterations), $\rho$, $\phi_{12}$, and $\phi_{21}$ were again updated simultaneously, but no further updates of the proposal distribution were made. Convergence of the posterior distributions of the parameters in the MCMC was monitored using both graphical and numerical methods (e.g., Gelman, 1996). Posterior distributions were then estimated by sampling from the third regime.

Of particular interest are the conditional-dependence parameters since these control the nature and degree of the spatial correlation in the model. Figure 3 shows scatterplots and kernel estimates of the distribution of $\phi_{11}$ (temperature) and $\phi_{22}$ (precipitation), the parameters that control the conditional dependence between lattice points within a layer (Figure 1, left panel). The distributions show that there is considerable (conditional) spatial dependence within each variable as the distributions tend to be concentrated near the positive boundary of possible values for $\phi_{11}$ and $\phi_{22}$. There is evidence of a slightly
Figure 3: Left frame shows scatterplot of a random sample of 10,000 values of $\phi_{11}$ and $\phi_{22}$ (1,000 from each of the 10 chains). Contours represent approximate 25, 50 and 75% contours of a kernel density estimate. Right frame shows kernel estimates of the marginals for $\phi_{11}$ (blue) and $\phi_{22}$ (red).

stronger dependence for temperature ($\phi_{11}$).

Trace plots and other diagnostics for $\rho$, $\phi_{12}$, and $\phi_{21}$ suggest convergence after about 10,000 iterations, which corresponds to the end of the second sampling regime. These three parameters control the dependence structure across variables; $\rho$ summarizes the within-location dependence (Figure 1, middle panel) and $\phi_{12}$, $\phi_{21}$ summarize the cross-variable dependence (Figure 1, right panel). The estimated posterior mean and standard deviation for $\rho$ is $-0.12$ and $0.014$, respectively. The negative values for $\rho$ suggests that an increasing temperature is (conditionally) associated with a decreasing precipitation.

Figure 4 highlights the distribution of $\phi_{12}$ and $\phi_{21}$. The strong correlation between these two conditional cross-correlation parameters is clearly shown in the left frame of Figure 4. However, there is another feature of note. In particular, it appears that there is evidence of an asymmetry in these two parameters, suggesting that there is higher conditional dependence between precipitation values and neighboring temperatures than the conditional dependence between temperature values and neighboring precipitation values. The evidence for asymmetry is compelling in the left panel, and somewhat masked in the right panel due to marginalization.
Figure 4: Left frame shows scatterplot of a random sample of 10,000 values of $\phi_{12}$ and $\phi_{21}$ (1,000 from each of the 10 chains). Contours represent approximate 25, 50 and 75% contours of a kernel density estimate. Right frame shows kernel estimates of the marginals for $\phi_{12}$ (blue) and $\phi_{21}$ (red).

Figure 5 shows posterior means for the regression components (left column), the spatial random effects (middle column), and the sum (right column) for the change in winter average temperature (top row) and change in total winter precipitation (bottom row). The fixed effects show a clear latitudinal effect as well as an east-to-west gradient. For precipitation, there is a more dominant east-to-west gradient. The spatial random effects for the change in temperature seem to follow the features of the topography, and are, in general, of smaller magnitude than the fixed effects. The spatial effects for the change in precipitation also follow the features of the topography, but there are additional strong local features, for example in northern California. In contrast to temperature, the spatial effects are larger relative to the fixed effects.

The sum of the fixed effects and the spatial random effects for temperature shows a consistent pattern of winter warming on average throughout the west, while the sum for precipitation shows patterns that are much more localized. The most dominant features for precipitation are the regions of sharp decline in winter precipitation in northern California. To aid in the identification of areas that might be most at risk for change, as projected by this particular regional climate model experiment, Figure 6 shows the results of a model-
Figure 5: Posterior means for the regression (left), the spatial effect (middle), and the sum (right) for the winter season. The top row represents the change in midpoint temperature (°K), while the bottom represents the change in total precipitation (inches).

Based clustering (Fraley and Raftery, 2002, 2008) performed on the posterior means of the change in temperature and precipitation. The scatterplot of the posterior means for each grid box shows considerable structure, which the clustering is able to distinguish. The spatial patterns associated with the clusters are also shown in Figure 6. The dark red areas, for example, highlight a region associated with a strong increase in temperature and decline in precipitation.

Of course, these representations are based on posterior means. It is also useful to consider some measure of uncertainty. Fields of standard deviations are one approach. We find it useful to consider other measures, in particular when considering the multivariate nature of this model output. For example, Figure 7 shows estimated pointwise probabilities of a simultaneous increase in temperature and decline in precipitation based on a sampling of the posterior distribution of the joint spatial fields. Again, on the basis of this model, we see evidence of widespread increase in temperature and decline in precipitation across
Figure 6: Results from clustering the posterior means of the change in temperature and precipitation. The left frame is a scatterplot with the clusters indicated through different colors. The right frame shows the clusters spatially.

Figure 7: Estimated pointwise probabilities of a simultaneous increase in temperature and decline in precipitation for the winter season.
the western US.

5.3 Results for the Summer Season

A slightly different scheme was used for the Gibbs sampler for the analysis of the summer model output. The three-regime sampling was still used, although with twice as many iterations (20,000) in the second regime. In addition, all five conditional dependence parameters ($\rho$, $\phi_{11}$, $\phi_{22}$, $\phi_{12}$, and $\phi_{21}$) were updated simultaneously. With that said, however, convergence of parameters for the summer season was similar to that for the winter season, although somewhat slower, and the specifics of those results are not discussed here. Distributions of the posteriors for the parameters were similar, with the exception of the cross-dependence parameters, and in particular $\rho$ (posterior mean of $-0.41$ versus a posterior mean of $-0.12$ for winter), suggesting a much stronger and more negative correlation between the change in average temperature and the change in total precipitation.

Figure 8 shows posterior means for the summer season with the same layout as in Figure 5 for the winter season. There now appears to be a west-to-east gradient in the fixed effects for temperature, and, again, the spatial random effects pick up more of the topography that is not accounted for in the model by the regression with elevation. Again, for temperature, the spatial effects are of smaller magnitude than the fixed effects.

For precipitation, there is also a west-to-east gradient in the fixed effects. The spatial effects for the change in precipitation also follow the features of topography, but there are strong local features, now occurring in the eastern part of the domain. In comparison to temperature, the spatial effects are larger relative to the fixed effects.

The sum of the fixed effects and the spatial random effects for temperature shows a consistent pattern of summer warming on average throughout the west, while the sum for precipitation shows patterns that are much more localized, just as in the winter season. However, the most dominant features for precipitation are the regions of decline in summer precipitation in the eastern part of the domain. Again, a model-based clustering was performed on the posterior means, which is summarized in Figure 9. There appears to be
Figure 8: Posterior means for the regression (left), the spatial effect (middle), and the sum (right) for the summer season. The top row represents the change in midpoint temperature (°K), while the bottom represents the change in total precipitation (in).

more widespread warming and declining precipitation during the summer months, but the clustering again highlights structure in the joint distribution.

Figure 10 shows estimated pointwise probabilities of a simultaneous increase temperature and decline in precipitation based on a sampling of the posterior distribution of the spatial fields. The summer warming and declining precipitation is widespread, even more so than in the winter, and this is highlighted in the figure.

6 Concluding Remarks

Climate models have become an important tool in the study of climate and climate change. Ensemble experiments of climate model output, be they comprised of perturbed initial conditions, perturbed physics, or multiple models, have also become important in studying and quantifying the uncertainty in climate model output. However, there are typically only a limited number of runs that can be produced due to the time and expense of running these
models, even on modern supercomputers. Hence, statistical methods become necessary to quantify the distribution and the breadth of variation in the model output.

With this idea in mind, we have introduced a hierarchical statistical model designed primarily for the analysis of regional climate model output on the basis of a simple ensemble (perturbed initial conditions). This model is multivariate and has the capacity to simultaneously characterize multiple model outputs, for example, the change in average temperature and the change in total precipitation. While analysis of the individual model outputs might yield estimates of marginal distributions, the strong correlations across variables, such as those uncovered here in the analysis of the changes in summer average temperature and total precipitation, make a multivariate analysis, such as that considered here, crucial for joint inference.

The statistical model also captures the spatial variation in the model output through a novel implementation of a multivariate MRF model. In addition to the computational benefits arising from using MRF models, this formulation of a multivariate MRF has a great deal of flexibility in modeling the conditional-dependence structure and is easily extendable. For example, more complex neighborhood structures can easily be considered (Sain et al.,
and it is not difficult to conceptualize how one might even consider modeling the joint distribution of multiple variables that are on different lattices.

There is great interest in more complex ensembles such as perturbed physics experiments and multi-model ensembles. We have not considered such ensembles here as we have focused on the multivariate aspect of the analysis of simple ensembles of regional climate model output. However, the model presented here can be extended to consider such ensembles by straightforward modifications to the process model, in particular Eqn. (8), or through a functional analysis of variance similar to that of Kaufman and Sain (2007).

Aside from the obvious computational challenges to simply fitting such models in the multivariate setting, there is, of course, much work that would need to be done to quantify the variation associated with different model physics or different models. However, there is a rich background in multivariate analysis of variance to build from.

Finally, there is also much interest, for example from people examining the impacts of climate change, in combining model output in order to obtain improved projections of climate change or to span the variation across a climate model experiment. Of course, with the more complex climate model experiments, there is the issue of model-to-model correlations. We believe that the inherent multivariate nature of this model provides an
excellent starting place to consider such correlations and this is also the focus of future work.

Acknowledgments

The National Center for Atmospheric Research is sponsored by the National Science Foundation (NSF). The research of the first author was also supported, in part, by NSF grants DMS-0355474, DMS-0707069, ATM-0502977 and ATM-0534173. The research of the second author was partially supported by NSF grant DMS-0621118. The research of the third author was partially supported by the Office of Naval Research under award N00014-08-1-0464.

References


