Madhuchhanda Bhattacharjee
School of Mathematics and Statistics,
University of Hyderabad,
Hyderabad-500046, India

and

Snigdhansu Chatterjee
School of Statistics,
University of Minnesota,
313 Ford Hall, Minneapolis MN 55455, USA

On Bayesian spatio-temporal modeling of oceanographic climate characteristics

1

On Bayesian spatio-temporal modeling of oceanographic climate characteristics

Madhuchhanda Bhattacharjee and Snigdhansu Chatterjee

CONTENTS

1.1 Introduction

Recent change in the planet's climate conditions are a matter of great concern, owing to its potential devastating effects for all life forms. A thorough study of climate variables should inform us about possible patterns of the climate of this planet, which further goes on to inform and guide policy decisions relating to adaptation and mitigation strategies to counter climate change, better management of resources, risk management, and for better quality of life for all. Many of these aspects relate to extremes as well as typical values of climate variables. Consequently, it is of interest to obtain the joint distribution of the several climate variables. Owing to the complex dependence patterns possible in such joint distributions, a Bayesian modeling of the data is needed. Moreover, using Bayesian methodologies in climate field also allows for systematic, coherent and simple treatment of Physics-driven known relations and constraints, combining multiple sources of data of varying size, dimensions and precision.

However, attempting to understand the patterns in data on climate variables presents unique challenges for Bayesian modeling. Note that many of the software tools that are available to the practitioner of Bayesian analysis are tailor-made to handle data supported on two-dimensional space, or for the analysis of a univariate variable of interest, or allow for very limited

forms of spatial or temporal dependencies. As examples of such tools, consider several of the contributed packages in the statistical computing platform `R`, which have been developed primarily for epidemiological, forestry mapping and other application domains.

Unlike typical datasets from these domains, climate data is typically available over three dimensional space and time, on multiple variables, and generally form a non-stationary random field. In this paper, we present an analysis of such multivariate, multiple-indexed (by three dimensional space as well as time) climate data using the example of Arctic Ocean seawater data. We consider the dataset on global seawater ([26]), accessible from the repository (http://data.giss.nasa.gov) of the Godard Institute of Space Studies, which will be the focus for the rest of this paper. We discuss details of the dataset later.

In this paper, we demonstrate how to capture the spatial variability of the response variables, allowing for smooth change over space without using specific fixed spatial relational function. For this we employ a technique similar to the *distributed lag model* in economics or *normal dynamic linear models* in clinical trails. In these comparable models, the lag would be considered that of time. Here, we generalize this framework to use the geographical space, and consider a spatial version of a dynamic model. We require additional modifications to account for the circular nature of earth surface and for the resulting cyclic dependence in the data. By adding constraints on the effect parameters we are able to achieve this and retain the model in DAG () framework.

We consider several climate variables together as the response. These response variables are known to have interdependence in this case. While one option would be to use known explicit functional form for this, note that these are only approximate relations, and not necessarily shared or common between for geographic locations and depths of the sea. Thus we would rather like to keep possibilities to explore such relationship open and study it a-posteriori. We modeled the response variables are jointly as multivariate normal random vector where the parameters depend of the three dimensional geographic co-ordinates.

As a component of climate related research, analysis of oceanographic data is less commonly observed compared to that of atmospheric variables. Mention must be made of the GLODAP project ([16]) and research that it generated. A concern for the planet's climate springs from possible acidification of the oceans, see [21], [24], [25] and others. Other studies include [12], [10], and various others. However, we have not come across a comprehensive Bayesian analysis of multiple variables relating to climatic properties of the Earth's oceans. It is challenging to analyze multiple response data indexed by multiple state variables and potentially having complex non-linear and non-stationary patterns, and in the context of climate data such cases may form a , as presented in [4]. In view of this, this paper potentially contributes in two ways. First, we present a framework that allows for the posterior to "talk" outside the strictures of stylized parametric dependency structures (simple examples of which

are temporal autoregression or spatial conditional autoregression) which may be useful for analyzing M-open problems. Second, we demonstrate an applied Bayesian analysis exercise in the context of that uses such a framework.

1.2 A description of the data set

We obtain the data from the repository http://data.giss.nasa.gov. This dataset includes measurements of temperature, salinity, deuterium, the ratio of the O-18 and O-16 isotopes of oxygen; co-variable information about the depth of the sea, the latitude and longitude, and the month and year at which the data was collected; along with references and notes. It is a compilation of data gathered by various teams of researchers at different points of time and location. Calibrations are carried out to correct for the difference in standards, techniques and instruments used by these teams, and such corrections are flagged. Missing values are present. Information on the time at which the data is collected is available in terms of month and years. Further technical minutiae relating to the dataset is available in the aforementioned website. Exploratory analysis of an earlier edition of this dataset have been carried case of [5]. A small-area type predictive analysis of this dataset has been presented in [23].

We access the data records from this source that correspond to Latitude values 60 degree North and higher. We further limit that data to be from 1975 or more recent times. The region from which the data has been gathered is depicted in Figure 1.1. This map has been produced by the afore-mentioned website.

1.3 A description of the methodology

We consider the variables *temperature*, *salinity*, and the ratio of Oxygen-16 and Oxygen-18 isotopes, called *Oxygen-18* in the sequel. We might have considered modeling for the hydrogen and deuterium isotope ratio, but only 8 records out of a total of 11800 contained values for these, the rest were missing. Consequently deuterium isotope ratio was not considered as a variable in this study. There were some missing values in each of the other variables as well, but not as overwhelming.

We consider observations from 60 North and above latitudes, which largely correspond to the Arctic Ocean region. The data is gathered at various depths, at different longitudes, over different months of the year, from 1975 onwards. Since data collection is sparse and uneven outside the summer months, we

restrict our attention to data from July-September only. We consider spatial blocks of data, in order to model spatial smoothness and coherence. In the analysis presented below, we consider four bands of latitude values (60-70, 70-75, 75-80 and 80-90 North), ten levels for longitude values produced by the following break points (-160, -80, -40, -5, 0, 20, 40, 90, 130, 180), and eleven levels of depths (bins separated at depth values of 0, 5, 15, 30, 50, 100, 150, 300, 500, 1000, 10000). These choices were made keeping data availability in mind. Note that the latitude, longitude and depth bands of values have a natural ordering each. Thus, spatial dependence may be modeled by considering neighboring bands. In view of the circular nature of longitudes, an additional constraint is imposed.

A generic notation for a data point may be $Y(\ell_1, \ell_2, d, t, j)$, where ℓ_1 specifies the latitude level, ℓ_2 the longitude level, d the depth level, t the time, and j an oceanic feature (temperature, salinity, O-18 ratio). We might further collapse the first four indices (latitude, longitude, depth and time), and consider a generic observation $y_i \in \mathbb{R}^3$ as a feature vector indexed by space-time.

For each observed sample point $i = 1, \ldots, n$ (n = 6795 number of samples) and $N(= 3)$ number of response variables, we assume the following model:

$$[y_i|\eta_i, \mathbf{\Theta}_i] \sim N_3\left(\eta_i, \mathbf{\Theta}_i\right),$$

where η_i is the mean vector $\in \mathbb{R}^N$, and $\mathbf{\Theta}_i$ is the corresponding precision matrix. Notice that these vary with the observed sample points.

We assume that the mean function η_i are affected by the spatial interdependence between the response variables, and not the variance-covariance structures.

In the above the the precision matrices $\mathbf{\Theta}_i$ are allowed to be different for different observations, with identical distributions for different levels of longitude and depth. Suppose

$$\tau_1 \sim Wishart(\mathbb{I}_N, N),$$

is a $N \times N$ random matrix. Here $\mathbb{I}_N$ is the identity matrix of dimension N. The degrees of freedom parameter is set at N as this corresponds to the least informative proper prior. For observations at a given level of longitude and depth values, we use the prior that the precision matrices $\mathbf{\Theta}_i$ are distributionally identical copies of the corresponding τ_1 matrices. Thus, geographic locations were assumed to have exchangeable prior distributions which in this case expressed for the precision matrices as Wishart distributions with prefixed hyper-parameters.

We use a spatially smoothed dynamic linear model for the mean vectors. The response variables are assumed to have individual overall means with variation around it depending upon geographical location given by the triplet latitude, longitude and depth. Initial data exploration suggested that the behavior of the response variables at various depths is possibly non-smooth. Thus smoothness on the remaining 2-dimensional surface were explored and the fol-

lowing model describes possibilities of using lags in both (latitude, longitude) and (longitude).

The additive model for the mean vectors are given by, for $i = 1, \ldots, n$ and $j = 1, 2, 3(= N)$,

$$\eta_{i,j} = \mu_{0j} + \mu_{lat_i, long_i, depth_i, j}.$$

Here the overall effect vector is modeled with multivariate normal distribution as follows:

$$\mu_0 \sim N_N\left(\mathbf{0}_{3\times 1}, \mathbb{I}_N\right).$$

We have a more complex modeling structure for the space-dependent mean structure, as follows:

$$\mu\left(k, \ell, m\right) = \begin{cases} \mathbf{0}, & \text{if } \ell = 1, \text{ for all } k, m, \\ \sim & N_N\left(\mu\left(k, l-1, m, 1:N\right), \tau_{N\times N}\right) \\ & \text{if } k = 1, \ell = 2, \ldots, 10 \text{ and all } m \\ \sim & N_3\left(0.5\mu(k-1, \ell, m) + 0.5\mu(k, l-1, m), \tau_{N\times N}\right) \\ & \text{for all } m,\ k = 2, \ldots, 4 \text{ and } \ell = 2, \ldots, 10. \end{cases}$$

The hyper-parameter τ appearing above is given a Wishart distribution in this hierarchical structure.

$$\tau \sim Wishart\left(\mathbb{I}_{N\times N}, N\right).$$

In the above framework, the spatial dependence within a given block of (latitude, longitude, depth)-level is captured by common parameters, while between block dependencies are captured by shared hyper-parameters. Temporal dependencies are captured by the built-in dependence structure for each longitude and depth block in the precision matrices $\mathbf{\Theta}_i$'s, and in the shared dependencies in the η_i's. Note that it is guaranteed that we have a proper posterior distribution, since all the priors were chosen to be probability density measures. This model can be enhanced with more complex features. However, our analysis showed that adding more complexity either by more complex mean and dispersion structures, or with additional spatial or temporal dependency measurements, does not enhance the quality of the statistical model. This is at least partially because of the nature of the data, and in keeping with the results of earlier, non-Bayesian attempts at analyzing this data, see [5, 23]. In addition, this feature of additional complexity not leading to model improvement strongly suggest that this problem is M-open. A Bayesian analysis problem is considered M-open if the data generating mechanism is not within the collection of models used for analysis, and there is no prior belief on how the data related to the "true model". One of the most important reason for analyzing climate data is for predicting future climate patterns, especially in a climate-change regime. It may be noted that prediction presents unique challenges in M-open problems, see [6, 7, 8] for detailed discussion on such issues.

1.4 Results

We run a Markov Chain Monte Carlo procedure according to the model specified above, to generate an approximate posterior distribution of the parameters of interest. We used 2 parallel chains, and a burn-in of 10,000 for each chain, followed by a further sequence of 50,000 iterations, which we thinned by a factor of 10. This allowed us to generate posteriors summaries based on (2 x 50000 / 10 =) 10000 samples, which are presented below.

In Figure 1.2, we show how the three response variables, (temperature, salinity and O-18 isotope ratio) varies with depth and longitude. Figure 1.3 show how the elements of the precision matrix, denoting the joint variability of the three responses, vary across depth and longitude. The figures for the corresponding posterior variance matrix over these three responses is given in Figure 1.4. In Figure 1.5, Figure 1.6 and Figure 1.7, we present the posterior distributions of these three responses as various latitude values.

Our major conclusion from this quite extensive analysis is that ocean variables are co-dependent, their mean and covariance structures seem to be strongly related to the spatial and temporal frames from which the observations have been gathered. There are naturally some differences between the variables, and some scope of bringing in Physics-guided knowledge among the response variables we studied. The patterns we found in the data suggest that no simple relationship would possibly suffice to explain the nature of any one of the response variables over space and time, or for the nature of co-dependence among the variables themselves. The figures we have presented show that the lack of a simple relationship should not be confused with a lack of a relationship; there is very strong suggestion of a pattern, see Figure 1.2 for example. One common theme that emerges from these graphics is that the posterior is nether simple, nor smooth over space, and standard summary measures like posterior location or scale parameters may be misleading.

We now present some evidence about the performance of the Markov Chain Monte Carlo procedure we adopted. In Figure 1.8 we show the posterior histogram of the μ-parameters for the three responses. The precision τ-parameters have posterior histograms as shown in Figure 1.9. Various diagnostic results, for example convergence graphs, posterior deviance results, auto-correlation plots are presented in Figure 1.10.

These results strongly suggest that the extracted samples from the MCMC runs may resemble a sample from the true posterior distribution of the various parameters under consideration. We have performed some robustness studies, whereby our conclusions do not seem to be altered by a choice of hyper-parameter values.

1.5 Discussion

The existing literature on climate modeling is essentially one of modeling nature using knowledge from a variety of scientific disciplines. From a Physics-based perspective, it is typical to consider the variable under study as a deterministic, but extremely complex, function of other physical variables. This kind of modeling retains a high degree of fidelity to the true process by which the data is generated. However, it is inevitable that not all features of a system as complex as climate will be measured or retained in various forms of data records. Also, our present state of knowledge about how different physical, atmospheric, geophysical and other variables interplay is limited, as is natural in any scientific discipline. A very partial and incomplete review of natural scientific modeling of climate may be obtained from [14], [22], [9], [11] and several references therein. A Bayesian framework where such Physics-based approach may be considered may be obtained from [17].

Climate models are used for several purposes. Of these, some of the more important ones are *detection* of climate change, *attribution* of climate change to a cause, *forecasting* of future climate scenarios. A study of climate in the pre-historic past, using data and proxies based on tree-rings, ice-core samples and other geophysical and fossilized sources, forms the topic of *paleoclimate*, and is useful as a reference for climate as of today and in future. Examples of paleoclimate studies may be found in [20], [19] [1], [15], [13] and other sources.

Climate research has progressed beyond change detection and attribution. Forecasts, and quantifying errors of forecasts relating to future climate scenarios, predicting possible consequences of climate change; combination of outputs from several AOGCM models, and other important studies now form a part of climate research. Several works using Bayesian ideas have contributed to this research, see, for example [3], [18], [2], [27] and references therein. In much of the above cited literature, oceanographic variables have not been considered. This is partially because the atmosphere is better understood than the hydrosphere in Physics, partially because it was felt that modeling the atmosphere was of greater importance for understanding climate change. However, in recent times, the hydrosphere, cyrosphere, biosphere are being studied with greater vigor.

In this context, this paper attempts to present a Bayesian analysis of a three-dimensional response on ocean water. We illustrated the complexity of patterns in oceanographic variables, and suggested a possible approach towards understanding the statistical properties of these variables. While a more detailed and thorough study needs to be done, our MCMC results suggest that a Bayesian approach may provide answers to several interesting questions in this domain.

In addition, the possibility of analyzing data from M-complete and M-open problems using the broad framework we adopted here: namely, con-

structing blocks of indexing variables to reduce or eliminate data sparsity, using least informative proper priors, and assuming minimal structure otherwise, should be explored.

Acknowledgment: We thank two anonymous referees for their comments, which greatly improved this paper. The second author's research was partially funded by NSF grants # SES-0851705 and # IIS-1029711, and research grants from the Institute on the Environment and College of Liberal Arts, University of Minnesota.

Bibliography

[1] Allen, M. R. and Tett, S. F. B. (1999) Checking for model consistency in optimal fingerprinting. *Clim. Dyn.* **15**, 419–434

[2] Berliner, L. M., Levine, R. A. and Shea, D. J. (1999) Bayesian Climate Change Assessment, *J. Climate*, **13**, 3805-3820.

[3] Berliner, L. M., Nychka, D. and Hoar, T. (2000) *Studies in the Atmospheric Sciences* (2000) edited by Mark L. Berliner, Douglas Nychka, Timothy Hoar; Springer, New York.

[4] Bernardo, J. and Smith, A. (2000) *Bayesian Theory*, John Wiley, Chichester.

[5] Chatterjee, S., Deng, Q., and Xu, J. (2009) The statistical evidence of climate change: an analysis of global seawater data technical report #677, School of Statistics, University of Minnesota.

[6] Clarke, B. (2010) Desiderata for a predictive theory of statistics *Bayesian Analysis*, **5**, (2), 283-318.

[7] Clarke, J. L., Clarke, B. and Yu, C.-W. (2013) Prediction in M-complete problems with limited sample size. *Bayesian Analysis*, **8**, (3), 647-690.

[8] Clyde, M. and Iversen, E. S. (2013) Bayesian model averaging in the M-open framework. In *Bayesian Theory and Applications*, ed. P. Damien, P. Dellaportas, N. G. Polson, and D. A. Stephens, pages 483-498, Oxford University Press.

[9] Cox, P. M., Betts, R. A., Bunton, C. B., Essery, R. L. H., Rowntree, P. and Smith, J. (1999) The impact of new land surface physics on the GCM simulation of climate and climate sensitivity. *Clim. Dynam.* **15**, 183–203.

[10] Durack, P. J., Wijffels, S. E., and Matear, R. J. (2012). Ocean salinities reveal strong global water cycle intensification during 1950 to 2000. *Science*, 336(6080), 455-458.

[11] Giorgi, F. and Mearns, L. O. (2002) Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM

Simulations via the "Reliability Ensemble Averaging" (REA) Method, *J. Climate*, 10, 1141-1158.

[12] Gouretski, V., and Reseghetti, F. (2010). On depth and temperature biases in bathythermograph data: Development of a new correction scheme based on analysis of a global ocean database. Deep Sea Research Part I: *Oceanographic Research Papers*, 57(6), 812-833.

[13] Hegerl, G. C. and others (2006) Climate change detection and attribution: beyond mean temperature signals, *J. Climate*, **19**, 5058-5077.

[14] IPCC (2008) *Climate Change 2007*, Cambridge University Press, U.K.

[15] The International ad hoc detection and attribution group (2005) Detecting and attributing external influences on the climate system: a review of recent advances *J. Climate*, **18**, 1291-1314.

[16] Key, R.M., Kozyr, A., Sabine, C.L., Lee, K., Wanninkhof, R., Bullister, J., Feely, R.A., Millero, F., Mordy, C. and Peng, T.-H. (2004). A global ocean carbon climatology: Results from GLODAP. Global Biogeochemical Cycles 18, GB4031.

[17] Kennedy, M. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *J. Royal Statist. Soc., Series B.* **63**, 425-464.

[18] Levine, R. A. and Berliner, L. M. (1999) Statistical Principles for Climate Change Studies, *J. Climate*, **12**, 564-574.

[19] Li, B., Nychka, W. D. and Ammann, C. M. (2007) The "Hockey Stick" and the 1990s: A statistical perspective on reconstructing hemispheric temperatures *Tellus*, **59**, 591-598.

[20] Mann, M. E., Bradley, R. S. and Hughes, M. K. (1998) Global-scale temperature patterns and climate forcing over the past six centuries. *Nature* **392**, 779–787.

[21] Matsumoto, K.; Gruber, N. (2005). How accurate is the estimation of anthropogenic carbon in the ocean? An evaluation of the DC* method. *Global Biogeochem. Cycles* 19. doi:10.1029/2004GB002397.

[22] Mearns, L. O., Hulme, M., Carter, T. R., Leemans, R., Lal, M. and Whetton, P. (2001) Climate scenario development. In *Climate change 2001: the scientific basis* (eds J. T. Houghton et al.). Contribution of working group I to the third assessment report of the Intergovernmental Panel on Climate Change, pp. 739–768. Cambridge, UK: Cambridge University Press.

[23] Mukherjee, U. and Chatterjee, S. (2013) A Fay-Herriot type approach for better prediction in multi-indexed response with application to Arctic seawater data analysis, *preprint*.

[24] Orr, J. C. et al. (2005). Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms. *Nature* 437, 681-686.

[25] Raven, J. A. et al. (2005). Ocean acidification due to increasing atmospheric carbon dioxide. Royal Society, London, UK

[26] Schmidt, G.A., G. R. Bigg and E. J. Rohling (1999) "Global Seawater Oxygen-18 Database," http://data.giss.nasa.gov/o18data/

[27] Tebaldi, C., Smith, R. W., Nychka, D. and Mearns, L. O. (2005) Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multi-model ensembles. *J. Clim.* **18**, 1524–1540.

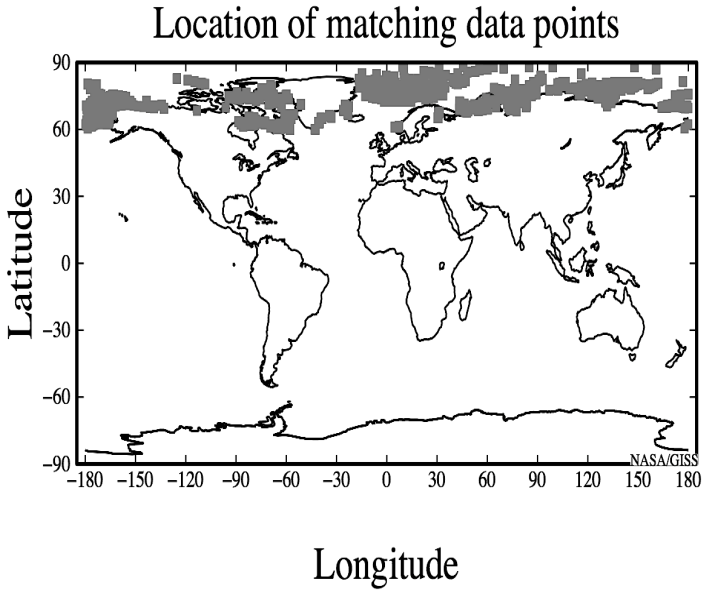

FIGURE 1.1
The spatial region from where the data has been gathered. This figure is generated from the same website from which the data is obtained.

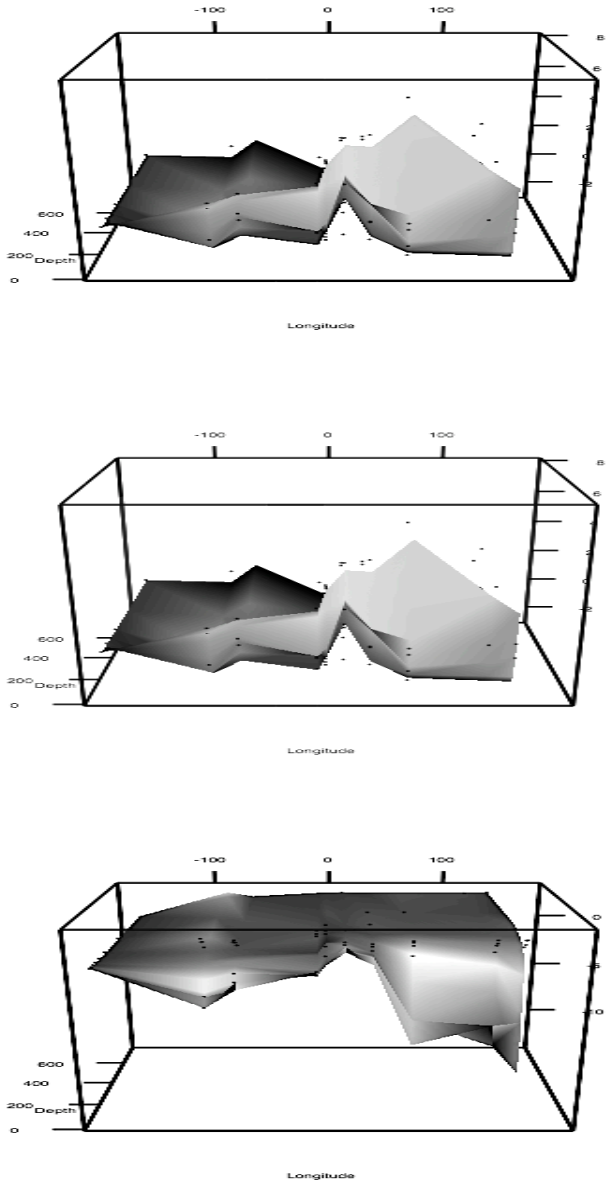

FIGURE 1.2
The pattern of posterior means *temperature* (top), *salinity* (middle) and *O-18 isotope ratio* (bottom) at various longitudes and depths of sea in the Arctic Ocean region.

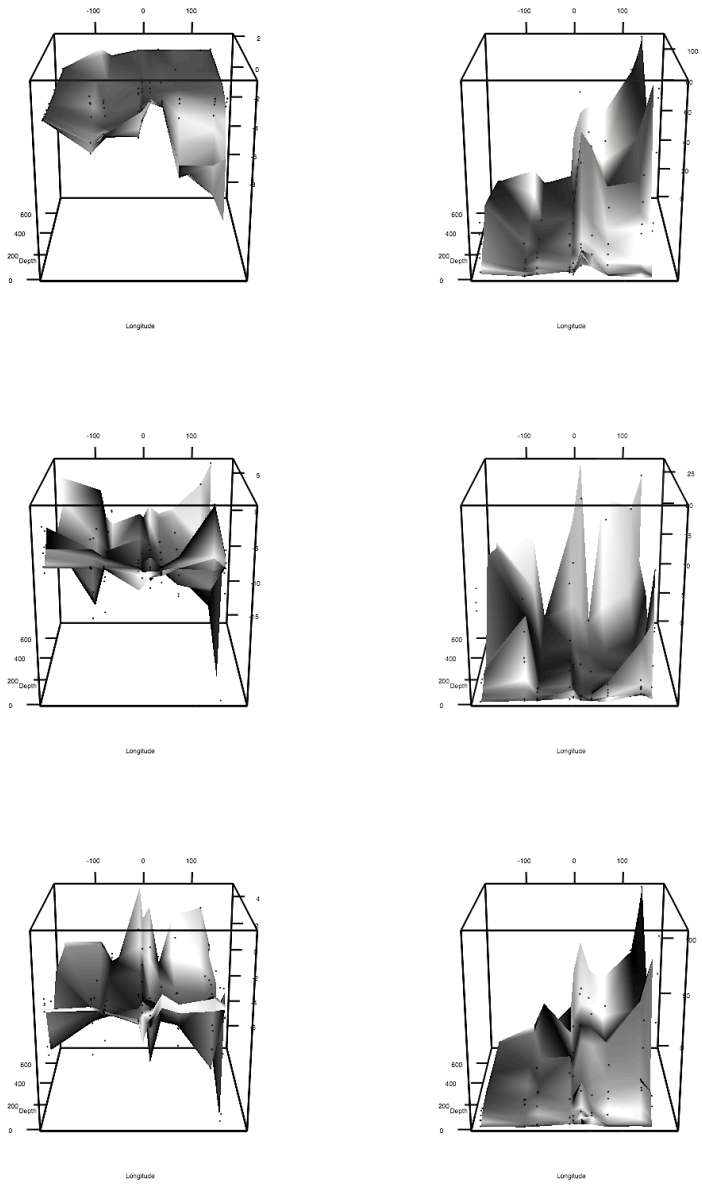

FIGURE 1.3
The top row figures are the (1,1) and (1,2) elements, middle row figures are the (1,3) and (2,2) elements, and bottom row figures are the (2,3) and (3,3) elements in the *posterior precision matrix* across *temperature*, *salinity* and *O-18 isotope ratio* at various longitudes and depths of sea in the Arctic Ocean region.

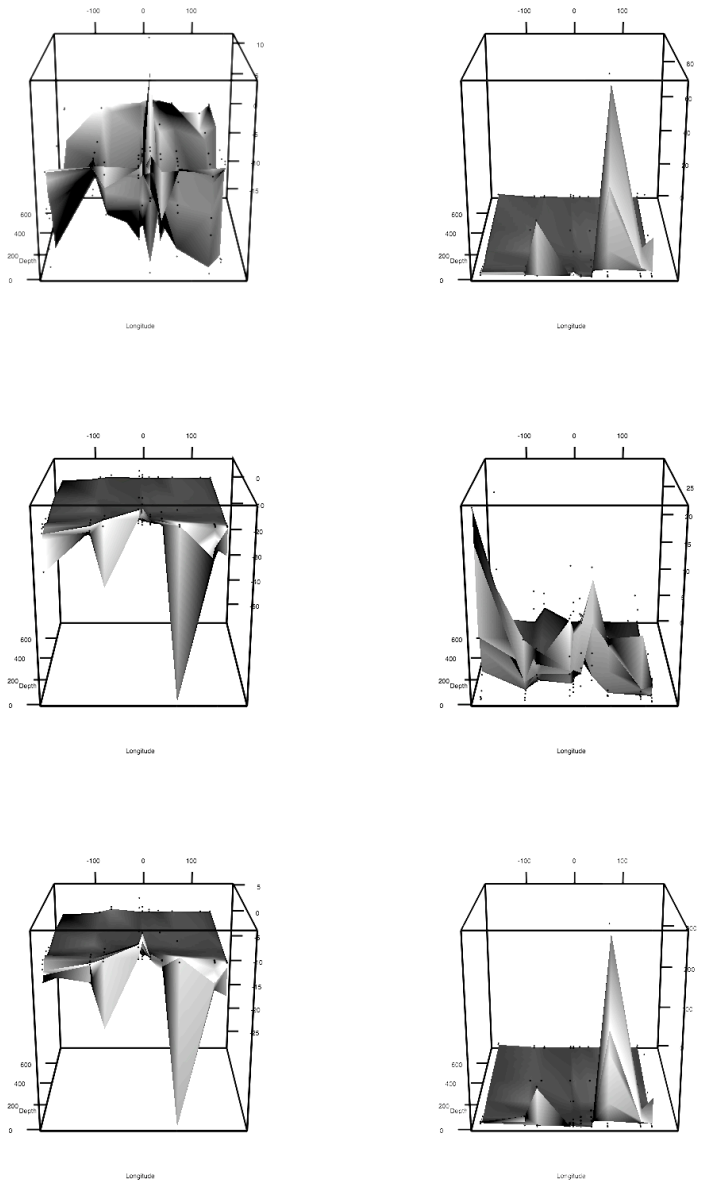

FIGURE 1.4
The top row figures are the (1,1) and (1,2) elements, middle row figures are the (1,3) and (2,2) elements, and bottom row figures are the (2,3) and (3,3) elements in the *posterior variance matrix* across *temperature*, *salinity* and *O-18 isotope ratio* at various longitudes and depths of sea in the Arctic Ocean region

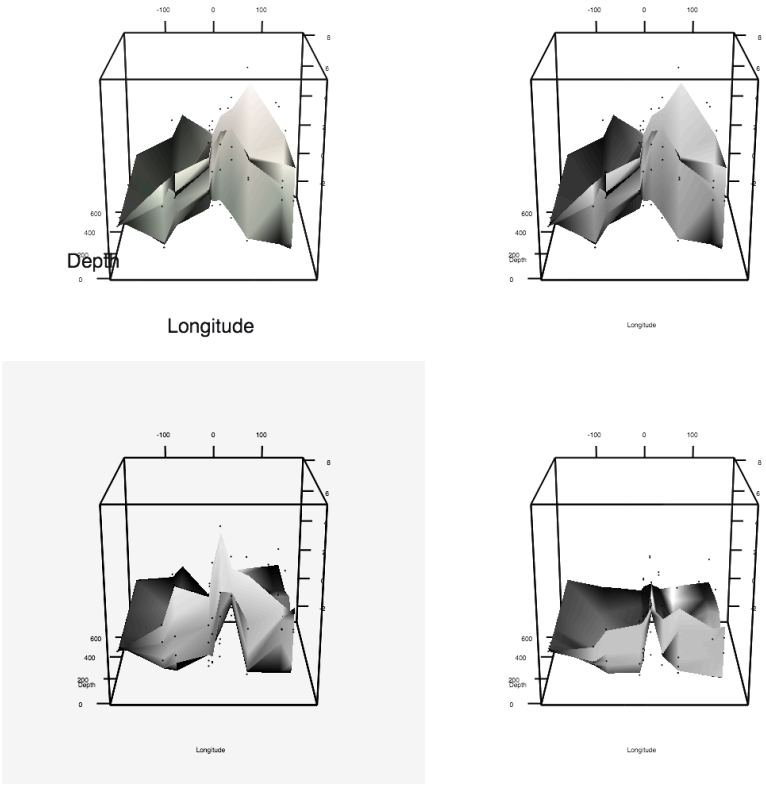

FIGURE 1.5
The top row figures are the patterns in the posterior mean of *temperature* of sea water, at various longitude and depth values, over Latitudes 60-70 North, and 70-75 North. The bottom row contains corresponding figures for Latitudes 75-80 North, and 80-90 North.

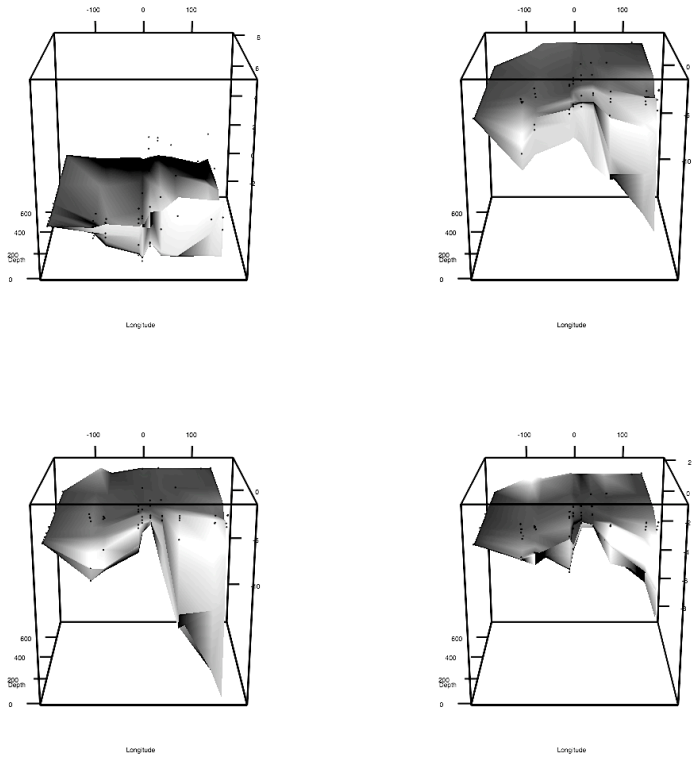

FIGURE 1.6
The top row figures are the patterns in the posterior mean of *salinity* of sea water, at various longitude and depth values, over Latitudes 60-70 North, and 70-75 North. The bottom row contains corresponding figures for Latitudes 75-80 North, and 80-90 North.

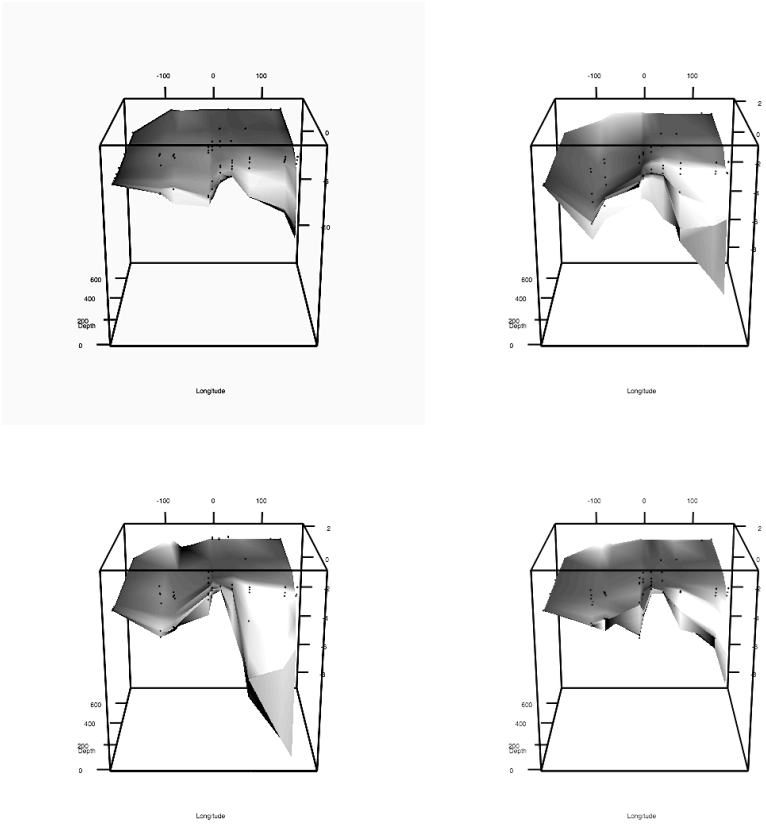

FIGURE 1.7
The top row figures are the patterns in the posterior mean of *Oxygen-18 isotope ratio* of sea water, at various longitude and depth values, over Latitudes 60-70 North, and 70-75 North. The bottom row contains corresponding figures for Latitudes 75-80 North, and 80-90 North.

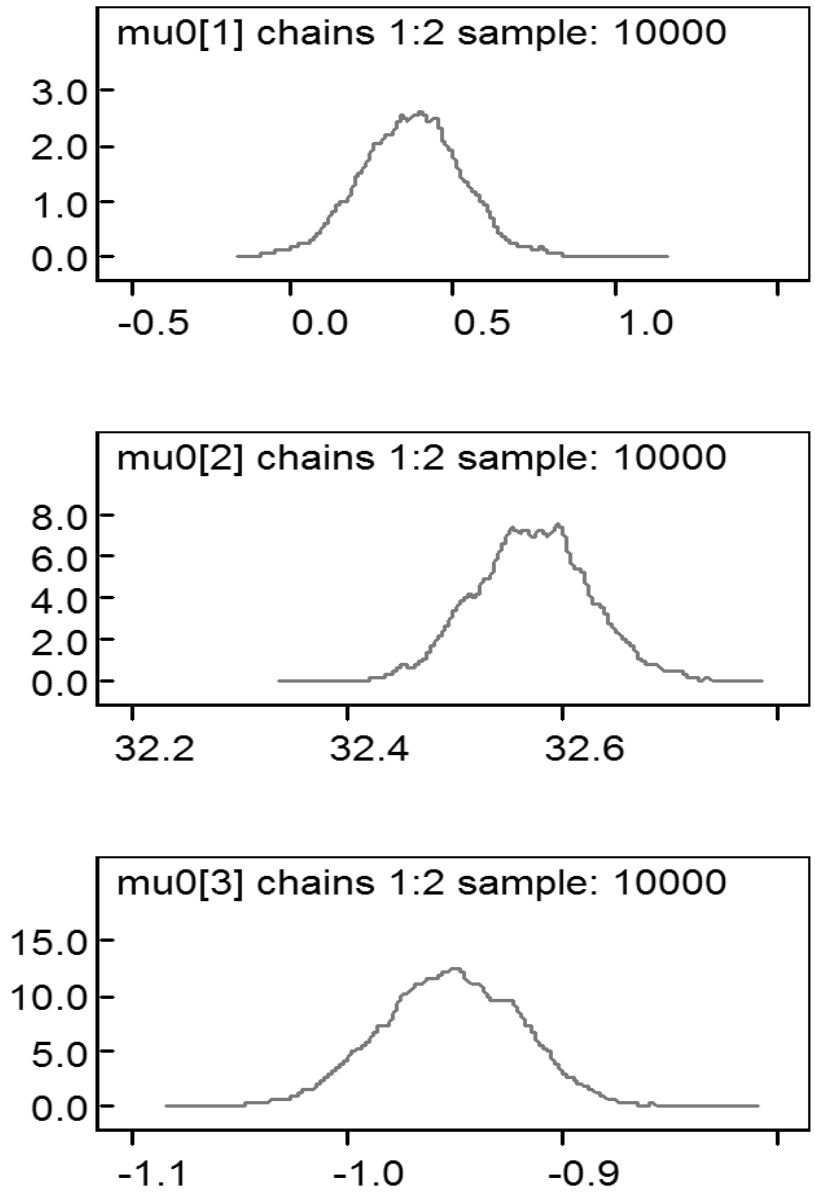

FIGURE 1.8
The histograms depicting the posterior distribution of the μ parameters across *temperature*, *salinity* and *O-18 isotope ratio*.

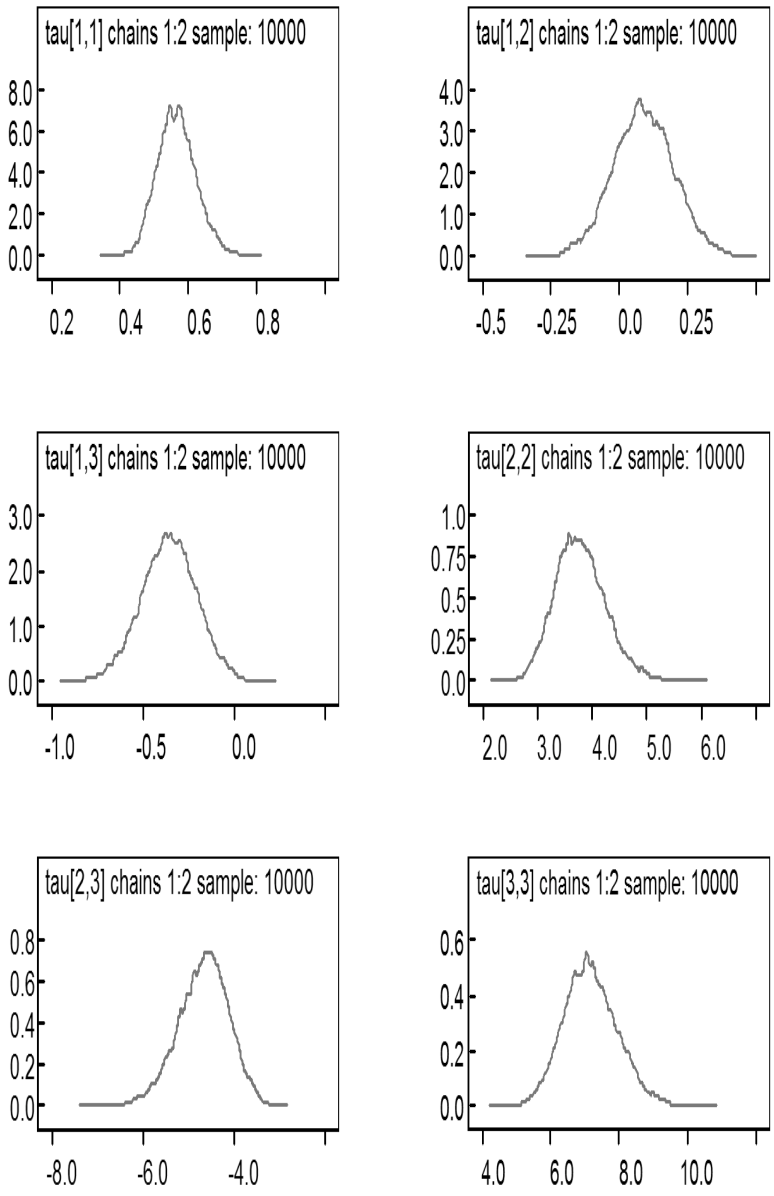

FIGURE 1.9
The top row figures are the histograms of the (1,1) and (1,2) elements, middle row figures are the histograms of the (1,3) and (2,2) elements, and bottom row figures are the histograms of the (2,3) and (3,3) elements depicting the posterior distribution of the *precision matrix* across *temperature*, *salinity* and *O-18 isotope ratio*.

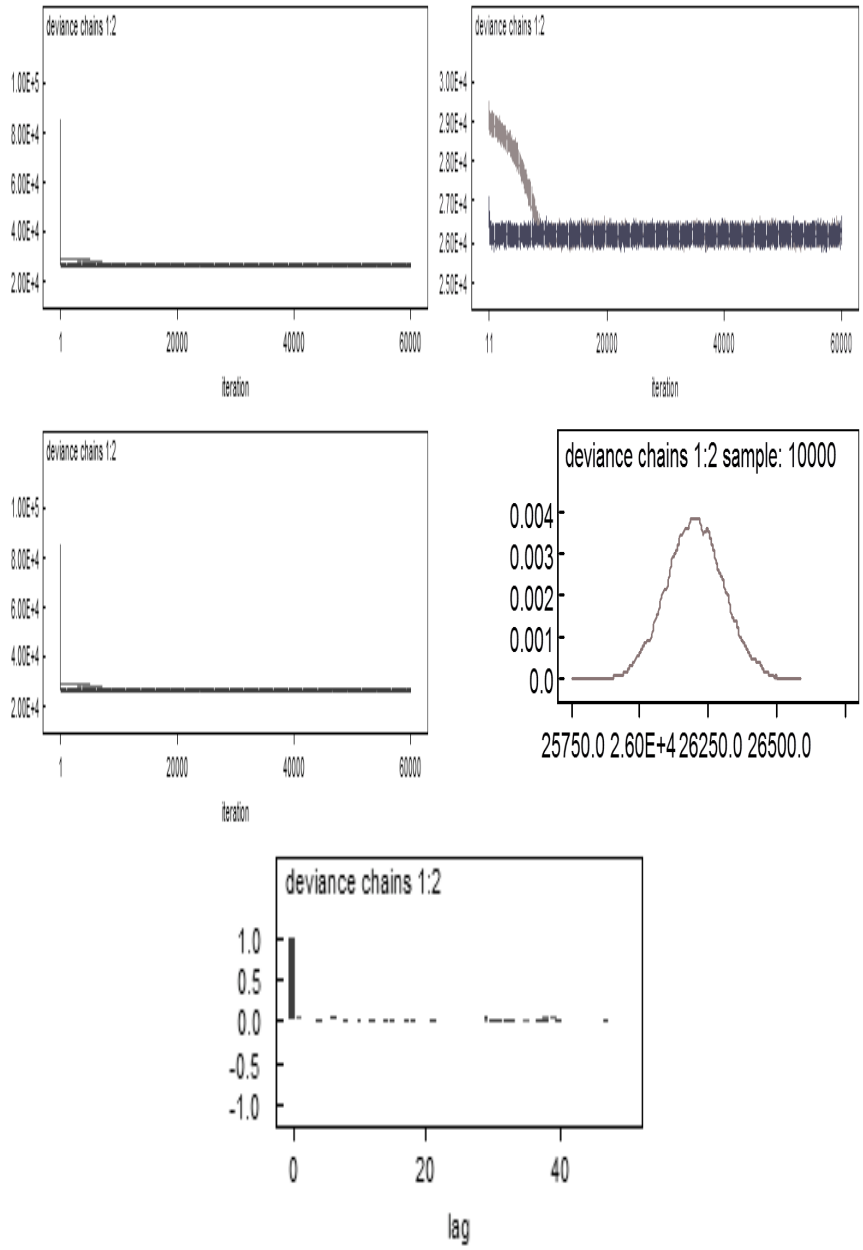

FIGURE 1.10
Graphs denoting convergence properties, deviance, and autocorrelation structures in the MCMC runs.