6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

COVARIANCE STRUCTURE ANALYSIS OF CLIMATE MODEL OUTPUT

Chintan Dalal[1], Doug Nychka[2], Claudia Tebaldi[3]

***Abstract*—To understand future climate change, different Earth system models from groups worldwide simulate projections of future climates. However, results from these simulations are computationally very expensive, often requiring several months on a supercomputer. In this paper, we provide a new statistical emulation method that may allow a realization of future climate projections within a day rather than several months. Specifically, we analyze the structure of several existing outputs from various climate models on a manifold of covariance matrices. The manifold covariance structure provides a method to compare existing climate model outputs, as well as to sample a new realization of future climate projections. We validated our climate model output comparison method using known dependencies between various climate models. Additionally, we showed, using semi-variogram plots, that the distribution of our realizations lie within the distribution of existing climate model outputs. The proposed statistical emulator could find its use in future climate impact assessment.**

I. INTRODUCTION

Our understanding of future climate changes can improve by analyzing various plausible realizations of future climate projections. However, generating a climate simulation from an Earth System model is computationally very expensive since the model captures the complex interactions among the many components of the Earth's climate system (see [1]). The Coupled Model Inter-comparison Project (CMIP [2]) coordinates efforts between various groups developing Earth system models to create a database of multi-model ensembles of climate simulations. For example, Fig. 1 shows changes in precipitation for North America from two separate Earth system models that are part of the CMIP Phase 5 (CMIP5) multi-model ensemble. Both of these models show a plausible, yet, different view of future climate changes. Hence, a thorough assessment of future climate impact requires a framework that can capture the variability across all climate model outputs.

Corresponding author: C. Dalal, chintan.dalal@rutgers.edu, [1] Department of Computer Science, Rutgers University, NJ, [2]National Center for Atmospheric Research, Boulder, CO, [3]Climate Central, NJ

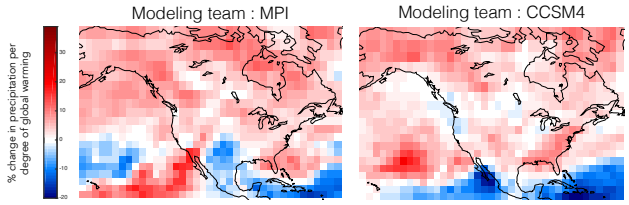

Fig. 1. Projections (2090) of percent change in precipitation per degree of change in the global mean temperature for North America from the CMIP5 multi-model ensemble. Shown here are projections from the Max Planck Inst. (MPI, Germany) and Community Earth System Model (CCSM4, USA).

An overview of methodologies that can capture the variability among climate model outputs is given in [3], along with the limitations of these approaches. For example, some of the climate models share common physical representation and numerical methods, and, thereby, cannot be considered as independent simulations. Additionally, the dependencies in climate models reduces the spread of future climate projections. To address the inter-model dependency issue, a Bayesian hierarchical framework has been suggested by [4], [5], [6], [7]. However, the proposed Bayesian framework faces difficulties in robustly modeling the inter-dependencies because of its sensitivity to prior assumptions.

Recent work by [8] shares similar methodological goals as ours in that the authors address issues of model dependencies and sampling in a non-parameteric set-up. The authors use a standard Euclidean metric on a low dimensional space by fixing the modes of variance within the available ensemble. Thus, limiting the amount of variability information that is present in the climate model outputs.

This paper presents an approach that allows for the variability information from the climate model outputs to be estimated. Specifically, we assume that the ensemble of well fitted covariance matrices provides sufficient information to characterize a distance measure. One application is to sample new realizations from an existing ensemble of climate model outputs.

II. METHOD

To capture the variability of future climate projections at locations around the globe and across the ensemble members, one requires a multivariate model framework. If $\tilde{\mathbf{y}}$ is a mutlivariate normal ($\mathcal{N}(\mu, \Sigma)$) then the standard multivariate normal sampling method (sMVN) is given by

$$\tilde{\mathbf{y}} = \mu + \Sigma^{\frac{1}{2}}\epsilon, \tag{1}$$

where $\tilde{\mathbf{y}}$ is the new multivariate sample representing the future climate projection, μ is the ensemble mean , Σ is the ensemble covariance matrix, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the standard normal random vector.

In sMVN, the estimation of the parameters of the distribution of the climate model outputs (Σ and μ) is of the forms

$$\Sigma(\theta)(i.e.\ \hat{\Sigma}) = \psi\mathbf{I} + \sigma^2\mathrm{H}(\phi), \quad \hat{\mu} = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\mathbf{y_i}. \tag{2}$$

In this preliminary study, H is selected as a stationary anisotropic matérn covariance function. Here, stationarity is selected for simplicity, and the anisotropic matérn covariance function is a standard choice in geostatistics. Finally, $\hat{\mu}$ is estimated as an equally weighted average.

The parameters $\theta = \{\phi, \sigma, \psi\}$ are also known as range, sill, and nugget, resp., in the geostatistics literature. They are estimated by maximizing the likelihood function, which is of the form $\ell(\theta|\mathbf{y}_1, \ldots, \mathbf{y}_\mathrm{N}) \propto |\Sigma(\theta)|^{-\frac{\mathrm{N}}{2}} \prod_{i=1}^{\mathrm{N}} \exp(-\frac{1}{2}(\mathbf{y_i} - \hat{\mu})^\mathsf{T}\Sigma(\theta)^{-1}(\mathbf{y_i} - \hat{\mu})))$, where N is the number of ensemble members, and $\mathbf{y}_i$ is a vector field of climate model outputs.

Our statistical emulation method, which we call the information geometric multivariate normal sampling method (igMVN), is depicted in Fig. 2. In igMSN, the estimate of the parameters Σ and μ is of the form

$$\hat{\Sigma} = \bar{\Sigma} + \Lambda^{\frac{1}{2}}\epsilon, \quad \hat{\mu} = \sum_{j=1}^{n}\sum_{k=1}^{m_j}\frac{1}{nm_j}\mathbf{y}_{j,k}, \tag{3}$$

where $\hat{\Sigma}$ is sampled from a normal distribution on a manifold of covariance matrices, i.e., $\Sigma(\theta) \sim \mathcal{N}(\bar{\Sigma}, \Lambda|\Sigma(\theta_1), \ldots, \Sigma(\theta_\mathrm{N}))$. Finally, $\hat{\mu}$ is estimated as a weighted average. Here, n is the number of clusters of covariance matrices, and m_j is the number of covariance matrices in each cluster.

The parameters $\bar{\Sigma}$ and Λ are the mean and variance, resp., of the ensemble of covariance matrices. Each $\Sigma(\theta_i)$(i.e. Σ_i) is a covariance matrix of individual ensemble members, and θ_i is learned by maximizing a likelihood function.

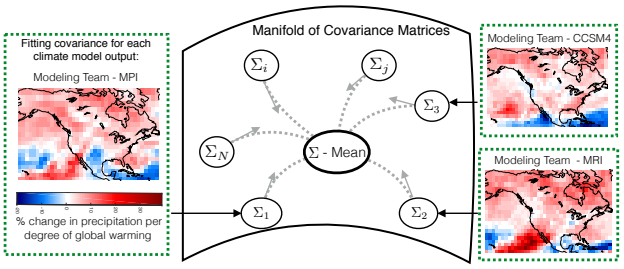

Fig. 2. Our inter-model comparison and sampling method: A Manifold view of the covariance structure of climate model outputs from various modeling teams (e.g. MPI, MRI, CCSM4)

A theoretical background for statistical distributions of symmetric positive definite matrices on a manifold can be found in [9], and the computational form to estimate $\bar{\Sigma}$ and Λ is given in [10], [11], [12], [13].

In order to estimate $\hat{\mu}$ using the weighted average, we first cluster the covariance matrices on a manifold using a standard hierarchical clustering method. The criteria for a cluster is $\max\{D(\Sigma_1, \Sigma_2) : \Sigma_1 \in S_1(\Sigma_i), \Sigma_2 \in S_2(\Sigma_i)\} < \text{threshold}$. The choice of the clustering method and the criteria for clustering are chosen for simplicity. The threshold is empirically chosen as 2 in our experiments, S_1 and S_2 are two sets of clusters of Σ_i's, and the distance metric (geodesic) on the manifold of covariance matrices is of the form $\mathrm{D}^2(\Sigma_1, \Sigma_2) = \frac{1}{2}\mathrm{Tr}(\log^2(\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}))$.

The estimates of $\hat{\mu}$ and $\hat{\Sigma}$ in igMVN incorporate extra information about the structure of covariance matrices that the sMVN fails to consider. This extra information is enabled using statistics on the structure of the covariance matrices in order to detect dependencies in climate model outputs and, thereby, incorporate known limitations in the ensemble members.

III. EVALUATION

To gain insight into the applicability of our proposed statistical emulation method, we used the ensemble of climate model outputs from CMIP5 experiments of future projections under RCP scenarios (see [1]). In order to test our method against various patterns in climate model outputs, we selected the climate variable of percent change in precipitation per degree of change in the global mean temperature.

In this paper, we restrict our study to the spatial dataset of the North American region in order to analyze the regional spatial variability aspect of the climate model outputs. Additionally, we have included single simulation runs from each of the Earth System Models (ESMs), rather than multiple simulation runs, in order to reduce biases in the ensemble.

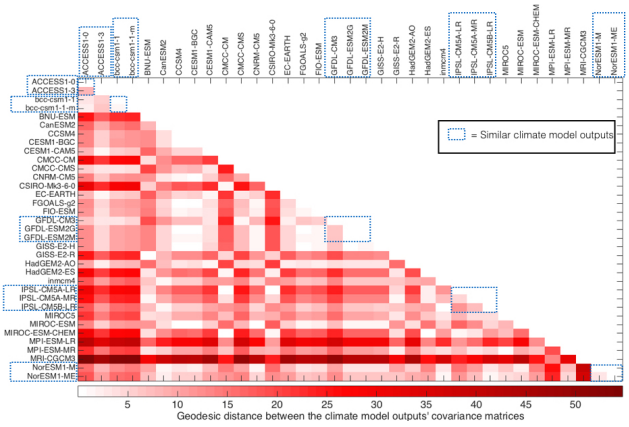

Fig. 3. A representation of the similarity measure between climate models outputs of the CMIP5 ensemble members. The similarity measure that we designed is a geodesic distance between the fitted covariance matrix of individual climate model outputs. Rows and columns of the above plot represent various climate model outputs, lighter shades of red represent higher similarity between models, and boxes represent climate models that are validated to have high inter-model dependencies.

Fig. 3 shows the values of our distance metric between each of the ensemble members. In this figure, lighter shades of red represent higher similarity in the covariance matrices of the climate model outputs, and, in turn, imply higher dependencies between the models. The climate model outputs from the same Earth system modeling group are highlighted by the blue boxes and are known to have high inter-model dependencies for reasons that include code and data sharing (see [14]). The highlighted blue boxes show lighter shades of red, and, in turn, demonstrate that the chosen geodesic distance metric can be used to compare and cluster climate model outputs in a non-parametric fashion.

Fig. 4 shows the experimental semi-variogram plots of climate model outputs and statistically generated samples from a number of methods. Given the semi-variogram function, one can estimate the parameters (range, sill, and nugget) of the covariance function. Hence, semi-variogram plot, explained in detail in [15], is a good tool in spatial statistics to visualize the differences in covariance matrices.

The climate model outputs (as shown by the red lines in Fig.4)) in the RCP2.6 and 4.5 ensembles (Fig.4) (a), (b), (c), (d)) has higher inter-model variability in its semi-variogram plots than the RCP8.5 ensemble (Fig.4) (e) and (f)). Hence, the spread of the climate model outputs realizations (as shown by the blue lines in Fig. 4) using the igMVN method (Fig.4) (b) and (d)) is better than the sMVN method (Fig.4) (a) and (b)) in representing the underlying spread of the climate model outputs. The wide spread in the igMVN samples

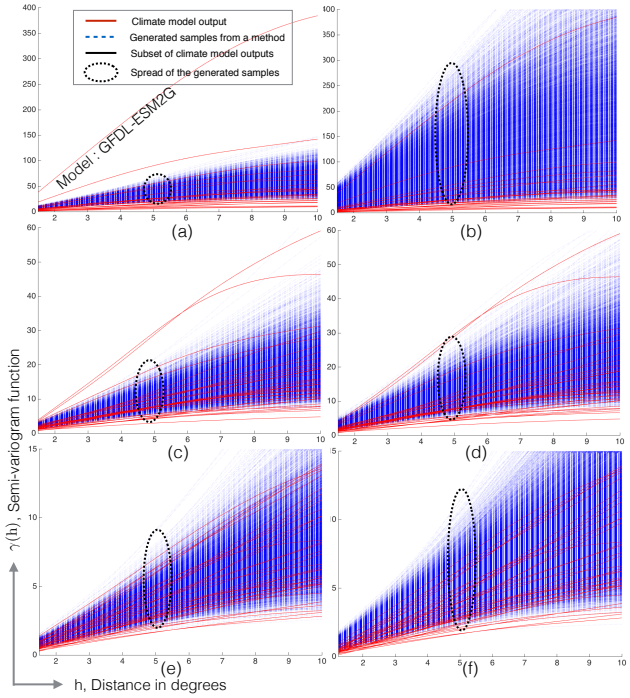

Fig. 4. Diagnostic plots showing the experimental semi-variogram function for various climate model outputs from CMIP5 ensembles (red lines) and the statistically generated samples (blue lines) from the standard multi-variate normal sampling method (sMVN) for the (a) RCP2.6 ensemble, (c) RCP4.5 ensemble, and (e) RCP 8.5 ensemble. Realizations from our sampling method (igMVN) are shown for the (b) RCP2.6 ensemble, (d) RCP4.5 ensemble, and (f) RCP8.5 ensemble. The ellipse in each plot focuses on the spread of the generated samples from each sampling method.

could be attributed to the sampling of the covariance matrices from a manifold.

Fig.5 (a) shows the spatial field of the GFDL-ESM2G model output (a member in the CMIP5-RCP2.6 ensemble) overlaying the North American region. From the semi-variogram plots in Fig.4 (a) and (b) we see that the realizations from the sMSV method does not emulate the climate data well, when compared to the igMVN method, for the GFDL-ESM2G model. Similarly, in Fig.5 (b) and (c) we see that there are more matching pixels (as shown by the grey colored boxes) in the realizations from the igMVN method (c) than the sMVN method (a). Therefore, the igMVN method may have some advantages over more traditional approaches; hence, it would be worth pursuing this method to compare and sample climate model outputs.

IV. DISCUSSION

In this paper, we have shown a non-parametric statistical emulator that can potentially mimic the existing ensemble of climate model outputs for projections of precipitation changes over North America and under the RCP scenarios. Additionally, we have provided

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

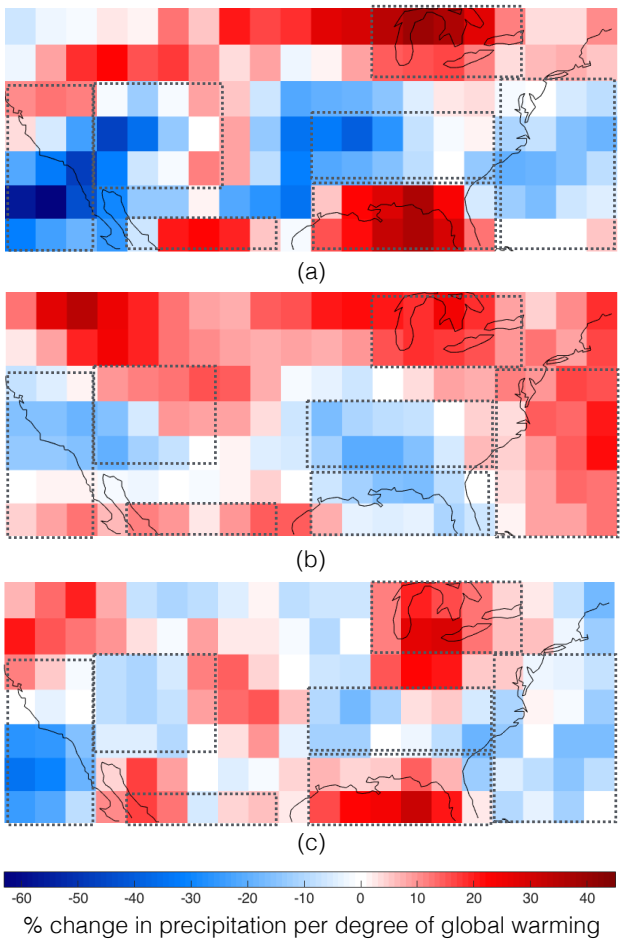

Fig. 5. Diagnostic plots showing the spatial field of climate variables from the Geophysical Fluid Dynamics Laboratorys climate model output of GFDL-ESM2G (a CMIP5-RCP2.6 ensemble member). The spatial field shown here is restricted to the North American region. (a) shows the climate model output, (b) shows one of closest realization (from Fig. 4(a)) using the sMVN method, and (c) shows one of closest realization (from Fig. 4(b)) using the igMVN method. The coast is represented by black lines, and the boxes represents patterns of similarity between the realizations and the climate model output.

a method to compare climate model outputs, which can be potentially used to investigate multi-model interdependencies in the CMIP5 ensembles.

By providing an emulator and a method for inter-model comparison, we can make the uncertainty in future climate projections more comprehensive and robust.

ACKNOWLEDGMENTS

This work is supported by the U.S. Department of Homeland Security under Grant Award Number 2012-ST-104-000044. We thank Vladimir Pavlovic, Dimitris Metaxas, Benjamin Sanderson, Matthew Edwards, Netta Gurari, and the anonymous reviewers for their feedback.

REFERENCES

[1] IPCC, "The physical science basis: Working group 1 contribution to the fifth assessment report of the intergovernmental panel on climate change," *New York: Cambridge University Press*, vol. 1, pp. 535–1, 2013.

[2] K. E. Taylor, R. J. Stouffer, and G. A. Meehl, "An overview of cmip5 and the experiment design," *Bulletin of the American Meteorological Society*, vol. 93, no. 4, p. 485, 2012.

[3] C. Tebaldi and R. Knutti, "The use of the multi-model ensemble in probabilistic climate projections," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 365, no. 1857, pp. 2053–2075, 2007.

[4] C. Tebaldi, R. L. Smith, D. Nychka, and L. O. Mearns, "Quantifying uncertainty in projections of regional climate change: A bayesian approach to the analysis of multimodel ensembles," *Journal of Climate*, vol. 18, no. 10, pp. 1524–1540, 2005.

[5] N. A. Leith and R. E. Chandler, "A framework for interpreting climate model outputs," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 59, no. 2, pp. 279–296, 2010.

[6] C. Tebaldi and B. Sansó, "Joint projections of temperature and precipitation change from multiple climate models: a hierarchical bayesian approach," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 172, no. 1, pp. 83–106, 2009.

[7] R. Furrer, S. R. Sain, D. Nychka, and G. A. Meehl, "Multivariate bayesian analysis of atmosphere–ocean general circulation models," *Environmental and ecological statistics*, vol. 14, no. 3, pp. 249–266, 2007.

[8] B. M. Sanderson, R. Knutti, and P. Caldwell, "Addressing interdependency in a multimodel ensemble by interpolation of model properties," *Journal of Climate*, vol. 28, no. 13, pp. 5150–5170, 2015.

[9] S. Amari, *Differential geometry in statistical inference*, vol. 10. Inst. of mathematical statistic, 1987.

[10] S.-I. Amari, "Information geometry on hierarchy of probability distributions," *IEEE transactions on information theory*, vol. 47, no. 5, pp. 1701–1711, 2001.

[11] X. Pennec, "Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements," *Journal of mathematical imaging and vision*, vol. 25, no. 1, pp. 127–154, 2006.

[12] T. Imai, A. Takaesu, and M. Wakayama, "Remarks on geodesics for multivariate normal models," *2011B-6*, 2011.

[13] H. Karcher, "Riemannian center of mass and mollifier smoothing," *Communications on pure and applied mathematics*, vol. 30, no. 5, pp. 509–541, 1977.

[14] R. Knutti, D. Masson, and A. Gettelman, "Climate model genealogy: Generation cmip5 and how we got there," *Geophysical Research Letters*, vol. 40, no. 6, pp. 1194–1199, 2013.

[15] N. Cressie, "Statistics for spatial data: Wiley series in probability and statistics," *Wiley-Interscience, New York*, vol. 15, pp. 105–209, 1993.