Comments on "Mission CO_2ntrol: A statistical scientist's role in remote sensing of atmospheric carbon dioxide"

Marc G. Genton[1] and Jaehong Jeong[1]

December 8, 2017

1 Introduction

Carbon dioxide (CO_2) is one of the major greenhouse gases in the Earth's atmosphere. With global warming and climate change being urgent issues, the mapping of CO_2 sources and sinks through time is essential, and satellites offer raw measurements of CO_2 at various spatial and temporal resolutions. Noël Cressie has provided a very informative overview of the statistical methodology needed for the analysis of data from the Orbiting Carbon Observatory 2 (OCO-2) satellite and discussed the statisticians' role in the study of carbon dioxide molecules in the atmosphere. The strength of the paper is its comprehensive review and commentary on satellite remote sensing of CO_2. At each level of the remote sensing mission, the author shared real experiences and examples. In the last section of the discussion paper (DP), he commented on future directions of statistical science for CO_2 including statistical models and decision-making under uncertainty. We extend the discussion of two points from a modeling side: non-Gaussian spatial models and nonstationary covariance functions on the sphere.

2 Beyond Gaussian random fields

Statistical inference and predictions for spatial data are often based on Gaussian random fields. Since CO_2 is a long-lived gas with sources and sinks, and a central-limit-type result for the column-averaged CO_2 (XCO2) values at different pressure levels makes it close to Gaussian, the proposed model in the DP is well suited for XCO2. The Gaussian assumption simplifies the structure of spatial models and facilitates statistical predictions, yet it is not always supported

[1] CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia. E-mail: marc.genton@kaust.edu.sa, jaehong.jeong@kaust.edu.sa.

by data on a global scale. For instance for methane, which is a relatively short-lived gas with sources, the author and collaborators relaxed this assumption by using a non-Gaussian model, such as the log-normal (Zammit-Mangion et al., 2015) or Box-Cox (Zammit-Mangion et al., 2016) spatial random fields. As mentioned in the discussion section of the DP, geophysical data are not always symmetric and bell-shaped, therefore non-Gaussian distributions may bring further improvements to the models.

One simple yet flexible way to construct non-Gaussian random fields is to use Tukey g-and-h distributions (Tukey, 1977), which can approximate many distributions, such as Student's t, Cauchy, and Weibull distributions (Xu and Genton, 2015). Tukey's g-and-h transformation function is $\tau_{g,h}(z) = g^{-1}\{\exp(gz) - 1\}\exp(hz^2/2)$, where $z \in \mathbb{R}$, $g \in \mathbb{R}$, and $\tau_{g,h}(z)$ is strictly monotone when $h \geq 0$. If $Z \sim N(0,1)$, then $Y = \xi + \omega\tau_{g,h}(Z)$ is said to have a Tukey g-and-h distribution. Here, ξ is a location parameter, ω is a scale parameter, g controls the skewness (i.e., $g > 0$ and $g < 0$ make the distribution right-skewed and left-skewed, respectively), and h governs the tail behavior. In a similar fashion, a general Tukey g-and-h random field, $Y(\mathbf{s})$, can be defined (Xu and Genton, 2017) as $Y(\mathbf{s}) = \xi + \mathbf{X}(\mathbf{s})^\top\boldsymbol{\beta} + \omega\tau_{g,h}\{Z(\mathbf{s})\}$, where $\mathbf{X}(\mathbf{s})$ represents the observed covariates at location $\mathbf{s}$ and $Z(\mathbf{s})$ is a standard Gaussian random field, with $\mathrm{E}\{Z(\mathbf{s})\} = 0$ and $\mathrm{var}\{Z(\mathbf{s})\} = 1$. For $h = 0$, $Y(\mathbf{s})$ is essentially a log-normal random field used by Zammit-Mangion et al. (2015) for methane. For more detailed properties about the Tukey g-and-h random field, such as its spatial mean and covariance function needed for kriging, see Xu and Genton (2017).

A significant advantage of Tukey g-and-h random fields is that they provide very flexible marginal distributions, allowing skewness and heavy tails to be adjusted. Moreover, if $Z(\mathbf{s})$ posesses properties such as second-order stationarity, mean-square continuity, and mean-square differentiability, then $\tau_{g,h}\{Z(\mathbf{s})\}$, $h < 1/2$, also retains such properties. This can lead to useful spatial dependence structures and second-order moments that are tailored to a particular application. Model inference can be performed similarly to the case of trans-Gaussian random fields, but for Tukey g-and-h random fields, which form a new class of trans-Gaussian random fields, the most suitable transformation for the dataset is estimated along with model parameters. One challenge in evaluating the likelihood function is that numerical evaluation can be slow for large data sets because the reciprocal transformation $\tau_{g,h}^{-1}(\cdot)$ does not have a closed form. To address this problem, Xu and Genton (2015, 2017) proposed a computationally efficient estimation method based on an approximated likelihood. A limitation of transformed Gaussian random fields is that the underlying dependence structure is still described by a Gaussian copula that lacks tail dependence and has a symmetric reflection regarding the joint dependence structure between the variables. Spatial and spatio-temporal random fields with flexible non-Gaussian copula structures have been proposed by Krupskii et al. (2017) and Krupskii and Genton (2017).

3 Nonstationary covariance models on the sphere

In the DP, there is a comment that the spatial random effects model has a spatially non-stationary covariance function that holds equally well on the surface of the sphere. Here we discuss some other works that developed nonstationary covariance models for global processes on the surface of a sphere. In a recent review, Jeong et al. (2017b) described a few available models that incorporate different construction approaches, such as differential operators (Jun and Stein, 2007; Jun, 2011, 2014), spherical harmonic representation (Stein, 2007), stochastic partial differential equations (SPDE) (Lindgren et al., 2011; Bolin and Lindgren, 2011), kernel convolution (Heaton et al., 2014), and deformation approaches (Das, 2000).

As a particular type of optimal spatial prediction for large data sets, the fixed rank kriging technique uses a covariance function that depends on a spatial random effects model (Cressie and Johannesson, 2008) to fill in gaps in global maps of XCO_2 measurements. This approach is computationally efficient regarding CPU time and memory storage (Bradley et al., 2015, 2016). Regarding scalable algorithms, the nested SPDE models (Bolin and Lindgren, 2011) are additionally appropriate for modeling global data. This class of models posesses desirable properties of the Markov random field framework, such as fast computation, adaptable extensions to nonstationarity, and applicability to general smooth manifolds. The nested SPDE models introduce nonstationarity via directional derivatives similar to Jun and Stein (2008), are computationally efficient via the Hilbert space approximation, hence are an appealing choice for large data sets.

For regularly spaced data in typical climate model outputs, multi-step spectrum models (Castruccio and Stein, 2013; Castruccio and Genton, 2014, 2016, 2018) are an option for inference on the full data set. This spectrum approach generalizes axially symmetric processes so that they are nonstationarity and have a flexible structure in the spectral domain while maintaining positive definiteness of the covariance functions. In particular, the multi-step spectrum model was designed to consider nonstationary covariance models across longitudes and to allow analysis of very large data sets by evaluating the likelihood with parallel and distributed computing. Castruccio and Guinness (2017) and Jeong et al. (2017a) showed how these models can be coupled with a land/ocean indicator and mountain ranges in the evolutionary spectrum. Since data from orbiting satellites have a particular observational location structure, the implementation of the above spectrum procedure on a non-gridded structure may be problematic. In this case, an interpolated likelihood approach could leverage on spectral methods (Horrell and Stein, 2015).

Given the ubiquity of big data in complex data structures such as satellite measurements evolving in space and time, computational methods for massive data sets have drawn a lot of attention in recent years; see Sun et al. (2012) and Bradley et al. (2016) for reviews. The dimension-reduction approaches that have been proposed for dealing with large data sets may lead to a loss of information when the spatial range is moderate to large, making them inadequate for space-time analysis (Stein, 2014). However, there is continued demand for computationally efficient methodologies that can handle massive data sets. Algorithms from other disciplines are appealing, e.g., one can consider approximations through parallel Cholesky decompositions of $\mathcal{H}$-matrices (Hackbusch, 1999, 2015) on different architectures (Litvinenko et al., 2017) and the integration of high-performance computing in exact likelihood inference and kriging (Abdulah et al., 2017) to handle large covariance matrices.

References

Abdulah, S., H. Ltaief, Y. Sun, M. G. Genton, and D. E. Keyes (2017). ExaGeoStat: A High Performance Unifed Framework for Geostatistics on Manycore Systems. *arXiv preprint arXiv:1708.02835*.

Bolin, D. and F. Lindgren (2011). Spatial Models Generated by Nested Stochastic Partial Differential Equations, With an Application to Global Ozone Mapping. *The Annals of Applied Statistics 5*, 523–550.

Bradley, J. R., N. Cressie, and T. Shi (2015). Comparing and Selecting Spatial Predictors Using Local Criteria. *TEST 24*(1), 1–28.

Bradley, J. R., N. Cressie, and T. Shi (2016). A Comparison of Spatial Predictors When Datasets Could be Very Large. *Statistics Surveys 10*, 100–131.

Castruccio, S. and M. G. Genton (2014). Beyond Axial Symmetry: An Improved Class of Models for Global Data. *Stat 3*(1), 48–55.

Castruccio, S. and M. G. Genton (2016). Compressing an Ensemble with Statistical Models: An Algorithm for Global 3D Spatio-Temporal Temperature. *Technometrics 58*(3), 319–328.

Castruccio, S. and M. G. Genton (2018). Principles for Statistical Inference on Big Spatio-Temporal Data from Climate Models. *Statistics and Probability Letters*, to appear.

Castruccio, S. and J. Guinness (2017). An Evolutionary Spectrum Approach to Incorporate Large-scale Geographical Descriptors on Global Processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 66*(2), 329–344.

Castruccio, S. and M. L. Stein (2013). Global Space-Time Models for Climate Ensembles. *The Annals of Applied Statistics 7*(3), 1593–1611.

Cressie, N. and G. Johannesson (2008). Fixed Rank Kriging for Very Large Spatial Data Sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*, 209–226.

Das, B. (2000). *Global Covariance Modeling: A Deformation Approach to Anisotropy.* Ph.D. thesis, University of Washington.

Hackbusch, W. (1999). A Sparse Matrix Arithmetic Based on $\mathcal{H}$-Matrices. Part I: Introduction to $\mathcal{H}$-Matrices. *Computing 62*(2), 89–108.

Hackbusch, W. (2015). *Hierarchical Matrices: Algorithms and Analysis*, Volume 49. Springer.

Heaton, M., M. Katzfuss, C. Berrett, and D. Nychka (2014). Constructing Valid Spatial Processes on the Sphere Using Kernel Convolutions. *Environmetrics 25*(1), 2–15.

Horrell, M. T. and M. L. Stein (2015). A Covariance Parameter Estimation Method for Polar-Orbiting Satellite Data. *Statistica Sinica 25*, 41–59.

Jeong, J., S. Castruccio, P. Crippa, and M. G. Genton (2017a). Reducing Storage of Global Wind Ensembles with Stochastic Generators. *The Annals of Applied Statistics*, to appear.

Jeong, J., M. Jun, and M. G. Genton (2017b). Spherical Process Models for Global Spatial Statistics. *Statistical Science*, *32*, 501–513.

Jun, M. (2011). Non-stationary Cross-Covariance Models for Multivariate Processes on a Globe. *Scandinavian Journal of Statistics 38*(4), 726–747.

Jun, M. (2014). Matérn-Based Nonstationary Cross-Covariance Models for Global Processes. *Journal of Multivariate Analysis 128*, 134–146.

Jun, M. and M. L. Stein (2007). An Approach to Producing Space-Time Covariance Functions on Spheres. *Technometrics 49*, 468–479.

Jun, M. and M. L. Stein (2008). Nonstationary Covariance Models for Global Data. *The Annals of Applied Statistics 2*(4), 1271–1289.

Krupskii, P., and M. G. Genton (2017). Factor Copula Models for Data with Spatio-Temporal Dependence. *Spatial Statistics 22*, 180–195.

Krupskii, P., R. Huser, and M. G. Genton (2017). Factor Copula Models for Replicated Spatial Data. *Journal of the American Statistical Association 112*, xx–xx.

Lindgren, F., H. Rue, and J. Lindström (2011). An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(4), 423–498.

Litvinenko, A., Y. Sun, M. G. Genton, and D. E. Keyes (2017). Likelihood Approximation with Hierarchical Matrices for Large Spatial Datasets. *arXiv preprint arXiv:1709.04419*.

Stein, M. L. (2007). Spatial Variation of Total Column Ozone on a Global Scale. *The Annals of Applied Statistics 1*(1), 191–210.

Stein, M. L. (2014). Limitations on Low Rank Approximations for Covariance Matrices of Spatial Data. *Spatial Statistics 8*, 1–19.

Sun, Y., B. Li, and M. G. Genton (2012). Geostatistics for Large Datasets. In *Advances and Challenges in Space-Time Modelling of Natural Events*, E. Porcu, J. M. Montero, and M. Schlather (Eds.), pp. 55–77. Springer.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Xu, G. and M. G. Genton (2015). Efficient Maximum Approximated Likelihood Inference for Tukey's g-and-h Distribution. *Computational Statistics & Data Analysis 91*, 78–91.

Xu, G. and M. G. Genton (2017). Tukey g-and-h Random Fields. *Journal of the American Statistical Association 112*, 1236–1249.

Zammit-Mangion, A., N. Cressie, and A. L. Ganesan (2016). Non-Gaussian Bivariate Modelling with Application to Atmospheric Trace-Gas Inversion. *Spatial Statistics 18*, 194–220.

Zammit-Mangion, A., N. Cressie, A. L. Ganesan, S. O'Doherty, and A. J. Manning (2015). Spatio-Temporal Bivariate Statistical Models for Atmospheric Trace-Gas Inversion. *Chemometrics and Intelligent Laboratory Systems 149*, 227–241.