Discussion of "Saving Storage in Climate Ensembles: A Model-Based Stochastic Approach" by Huang Huang, Stefano Castruccio, Allison H. Baker and Marc Genton

Sudipto BANERJEE

I congratulate the authors on this insightful article that is very relevant to statisticians and the broader community of climate scientists. The article addresses stochastic approximations of climate models from the perspectives of storage and computational burden by analyzing an ensemble of simulation output using a range of statistical models of increasing complexity designed to capture spatial–temporal variability at increasingly high spatial and temporal resolutions. The challenges statisticians encounter in analyzing output from climate models have been neatly articulated in the manuscript. I enjoyed reading the manuscript and find myself largely in agreement with the overall approach pursued by the authors here. Nevertheless, I take this opportunity to present some ideas pertaining to Bayesian learning of mechanistic models that may be relevant to inference for climate models.

First, it is worth pointing out a rather substantial literature in statistical modeling of mechanistic systems including, but not limited to, climate models. Different approaches have been developed and recommended depending upon the analytic tractability of the mechanistic systems as well as the computational resources available to analysts. If the mechanistic system is governed by a system of differential equations (ordinary or partial), as is the custom in climate modeling, partial analytical tractability of the system may be available by building a possibly nonlinear dynamic model using finite difference approximations of the system (see, e.g., Wikle and Hooten 2010, for an excellent exposition and further references). This approach, when applied to a mechanistic system such as in Figure 2 of the article being discussed, would take the finite difference approximations of all partial derivatives with respect

This article is a commentary for https://doi.org/10.1007/s13253-022-00518-x.

S. Banerjee (✉) UCLA Department of Biostatistics, Los Angeles, CA 90095-1772, USA
(E-mail: *sudipto@ucla.edu*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 28, Number 2, Pages 365–369
https://doi.org/10.1007/s13253-023-00541-6

to time and recast the system into a hierarchical dynamical or state-space framework with the process evolving as $W_t = \mathcal{M}(W_{t-1}; \theta_t, x_t, \eta_t)$, where $\mathcal{M}(\cdot)$ is the evolution operator, θ_t is a collection of, possibly time-varying, mechanistic parameters, x_t is a collection of design inputs to the mechanistic system at time t, and η_t is a stochastic process that can accommodate other disturbances such as white noise at the process stage as well as other structured dependencies, such as spatial–temporal associations, as appropriate.

The above approach is attractive in that the evolution operator is directly derived from the mechanistic system and, hence, yields a learning framework that can accommodate the scientific parameters. Furthermore, if field observations are also available, then the evolution model for the process is conveniently embedded within a hierarchical framework, where the first stage is a probability model $Y_t = \mathcal{H}(W_t; \beta_t, \epsilon_t)$, where $\mathcal{H}(\cdot)$ maps the latent process to the observations Y_t and specifies the likelihood function for Y_t, β_t is, possibly, time-varying, statistical model parameters and ϵ_t captures measurement errors. If $\mathcal{M}(\cdot)$ and $\mathcal{H}(\cdot)$ are linear and the stochastic processes are Gaussian, or can be reasonably approximated as such, then we arrive at the familiar Kalman-filtering approaches for state-space models including ensemble methods for large data sets (see, e.g., Katzfuss et al. 2020, and references therein). Unfortunately, the evolution operator $\mathcal{M}(\cdot)$ may not be analytically or computationally very tractable for very complex scientific models such as the one presented in the article under discussion. For example, Chapter 3 in Collins et al. (2004) offers the mathematical model describing the dynamics of the NCAR Community Atmosphere Model and Section 3.1.7, in particular, develops the finite-difference calculations. Even a cursory look at this material reveals the complexity of $\mathcal{M}(\cdot)$ which impedes developing a hierarchical dynamical framework. This, from my vantage point, is a strong motivation for the approach proposed in the article under discussion. The underlying justification is to let the specialized computing environments produce the output from the mechanistic system and let the statisticians analyze the results to facilitate probabilistic learning and inference for the underlying scientific processes at play.

Turning to the specific approach in the article, modeling the response (temperature) $Y(\cdot)$ and the radioactive forcing $X(\cdot)$ as stochastic processes is attractive and offers substantial richness and flexibility. While modeling the radiative forcing as autoregressive processes is not unreasonable, and in fact may even offer computational benefits, it raises my curiosity about enriching the model by jointly modeling temperature and forcing as a continuous space-time process. Furthermore, from the scientific perspective, it may be worth exploring the feasibility for allowing radiative forcing to vary spatially as well as temporally. To be specific, writing $\ell = (s, t)$ as the space-time coordinate, we can consider

$$\begin{aligned} Y_r(\ell) &= \beta_0(\ell) + \beta_1(\ell)X(\ell) + \eta_{y,r}(\ell); \\ X(\ell) &= \mu(\ell) + \eta_z(\ell), \end{aligned} \tag{1}$$

where $\eta_r(\ell) = (\eta_{y,r}(\ell), \eta_z(\ell))^{\mathrm{T}}$ is a bivariate spatial–temporal process specific to the r-th realization of the scientific model. If there is no spatial information on radiative forcing, then we can adapt the above model by setting $X(\ell) = X(t)$.

Instead of assuming an autoregressive structure for radiative forcing, we can model $\eta_r(\ell)$ as bivariate Gaussian processes for each realization of the climate model or, if deemed

appropriate, even accommodate dependence among the realizations. In any case, the process $\eta_r(\ell)$ can be specified using space-time cross-covariance functions see, e.g., Genton and Kleiber (2015), for a comprehensive review. Here, I understand that the authors' prefer an unstructured covariance matrix for capturing spatial associations, which corresponds to a coregionalized model with $\eta_r(t) = A\omega_r(t)$, where $\eta_r(t)$ is the $S \times 1$ process obtained by collecting $\eta_r(\ell)$ over S spatial regions, $\omega_r(t) = (\omega_{r,1}(t), \ldots, \omega_{r,S}(t))^{\mathrm{T}}$ is an $S \times 1$ vector whose elements, $\omega_{r,j}(t)$ for $j = 1, 2, \ldots, S$, are independent latent temporal processes with unit variance corresponding to each region, and A is the Cholesky factor of any spatial covariance matrix. If the model needs to operate at large scales over either space or time, then familiar ideas such as dynamic nearest-neighbor Gaussian processes (Datta et al. 2016b) can be used (although scalability does not seem to be an issue with $S = 58$ regions). Furthermore, if we have a balanced design for the realizations of the climate model, where we can treat $Y(t)$ as an $S \times R$ matrix whose (s, r)-th elements are $Y_r(s, t)$ for $s = 1, \ldots, S$ and $r = 1, \ldots, R$, then computationally efficient rich Bayesian modeling frameworks based on matrix-variate multivariate regression models for temporally evolving $Y(t)$ can be constructed by extending (Zhang and Banerjee 2022) to temporal processes.

A pertinent issue is whether the above joint modeling approach will deliver substantial benefits to compensate for perhaps some additional levels of complexity. I will admit that at this point, it is difficult to see such a benefit unless the learning exercise also involves predictive interpolation of radiative forcing at arbitrary regions. In that case, it is worth remarking that joint modeling can offer benefits over conditional modeling. For example, consider any fixed realization of the scientific model and let Y_r be the $ST \times 1$ vector of $Y_r(\ell)$ over the S regions and T time points where the climate model output has been recorded for the r-th realization. Let X be the corresponding vector collecting $X(\ell)$ over the ST space-time coordinates and let ℓ_0 be a new space-time coordinate where we wish to predict $X(\ell_0)$ and $Y(\ell_0)$. In the conditional model, as pursued by the authors', the joint distribution is obtained after extending a probability law to the radiative forcing as

$$\begin{aligned}[X(\ell_0), Y(\ell_0), X, Y_r \mid \Omega] = [X \mid \Omega] \times [X(\ell_0) \mid X, \Omega] \\ \times [Y_r \mid X, \Omega] \times [Y(\ell_0) \mid X(\ell_0), \Omega],\end{aligned}$$

where Ω is a generic notation for all parameters in the model. Then, the predictive distribution of $X(\ell_0)$ and $Y(\ell_0)$ given the data $\{X, Y_r\}$ is

$$[X(\ell_0), Y(\ell_0) \mid X, Y_r, \Omega] \propto [X(\ell_0) \mid X, \Omega] \times [Y(\ell_0) \mid X(\ell_0), \Omega]$$

which implies that, conditional on the information on Ω, $[X(\ell_0) \mid X, Y_r, \Omega] \propto [X(\ell_0) \mid X, \Omega]$ so the process $X(\cdot)$ at a new location ℓ_0 will not learn from the associated process realizations Y_r, although $Y(\cdot)$ will learn from $X(\cdot)$. This apparent asymmetry in predictive learning may be deemed unintuitive in scientific applications.

The above are some points that should be worth attending to in these exercises. I will conclude by making a few remarks regarding statistical learning of mechanistic systems in the broader sense. It seems important from the statistical perspective that scientific inference on the climate should assimilate information from observed data in conjunction with

mechanistic systems manifested through climate models. Such inference should, ideally, assist in learning about the mechanistic system parameters from the data. This is possible with more tractable mechanistic systems using the approach described in Wikle and Hooten (2010) and also in the area of calibrating more complex mechanistic system parameters from field data using Gaussian process emulators (Kennedy and O'Hagan 2001; Higdon et al. 2008). In fact, in settings where the differential equations are analytically and computationally tractable, such as in industrial hygiene applications, learning or calibrating parameters using the state-space approaches of Wikle and Hooten (2010) and Gaussian processes have been shown to offer substantial inferential benefits (Monteiro et al. 2014; Abdalla et al. 2020).

However, the situation with climate models is more challenging for calibration. The climate model must be run over a well-designed range of input parameters. While computationally efficient Gaussian process emulators for computer models have been proposed (see, e.g., Gramacy et al. 2014; Gramacy 2016; Gramacy and Haaland 2016; Frankenburg and Banerjee 2022), the complexity of climate models still precludes calibration. In this regard, I feel one needs greater involvement of statisticians in constructing the climate models themselves. Rather than providing statisticians with a highly intricate climate model as an end product to analyze, smaller constituents of the climate model should be calibrated and validated with field observations and this calibrated learning, including quantified uncertainty, should be exploited in building other constituents. Such developments will require synergistic collaborations between climate scientists, applied and computational mathematicians with expertise in PDE's, and statisticians with expertise in mechanistic system calibration and emulation.

[Accepted March 2023. Published Online May 2023.]

REFERENCES

Abdalla N, Banerjee S, Ramachandran G, Arnold S (2020) Bayesian state space modeling of physical processes in industrial hygiene. Technometrics 62(2):147–160. https://doi.org/10.1080/00401706.2019.1630009

Collins WD, Rasch PJ, Boville BA, Hack JJ, McCaa JR, Williamson DL, Kiehl JT, Briegleb B (2004) Description of the NCAR Community Atmosphere Model (CAM 3.0). Technical Report TN-464+STR, National Center for Atmospheric Research. https://www.cesm.ucar.edu/models/atm-cam/docs/description/description.pdf

Datta A, Banerjee S, Finley AO, Hamm NAS, Schaap M (2016) Non-separable dynamic nearest-neighbor gaussian process models for large spatio-temporal data with an application to particulate matter analysis. Ann Appl Stat 10:1286–1316. https://doi.org/10.1214/16-AOAS931

Frankenburg I, Banerjee S (2022) Dynamic Bayesian learning and calibration of spatiotemporal mechanistic systems. arXiv: 2208.06528v2

Genton MG, Kleiber W (2015) Cross-covariance functions for multivariate geostatistics. Stat Sci, 147–163

Gramacy R (2016) laGP: large-scale spatial modeling via local approximate gaussian processes in R. J Stat Softw 72(1):1–46

Gramacy RB, Haaland B (2016) Speeding up neighborhood search in local Gaussian process prediction. Technometrics 58(3):294–303. https://doi.org/10.1080/00401706.2015.1027067

Gramacy RB, Niemi J, Weiss RM (2014) Massively parallel approximate Gaussian process regression. arxiv:1310.5182

Higdon D, Gattiker J, Williams B, Rightley M (2008) Computer model calibration using high-dimensional output. J Am Stat Assoc 103(482):570–583. https://doi.org/10.1198/016214507000000888

Katzfuss M, Stroud JR, Wikle CK (2020) Ensemble Kalman methods for high-dimensional hierarchical dynamic space-time models. J Am Stat Assoc 115(530):866–885. https://doi.org/10.1080/01621459.2019.1592753

Kennedy MC, O'Hagan A (2001) Bayesian calibration of computer models. J R Stat Soc Series B Stat Methodol 63(3):425–464. https://doi.org/10.1111/1467-9868.00294

Monteiro JVD, Banerjee S, Ramachandran G (2014) Bayesian modeling for physical processes in industrial hygiene using misaligned workplace data. Technometrics 56(2):238–247. https://doi.org/10.1080/00401706.2013.836988

Wikle CK, Hooten MB (2010) A general science-based framework for dynamical spatio-temporal models. TEST 19(3):417–451

Zhang L, Banerjee S (2022) Spatial factor modeling: a Bayesian matrix-normal approach for misaligned data. Biometrics 78(2):560–573. https://doi.org/10.1111/biom.13452