Received: 14 July 2020 | Accepted: 17 July 2020

DOI: 10.1002/env.2649

DISCUSSION PAPER

WILEY

Discussion on A high-resolution bilevel skew-*t* stochastic generator for assessing Saudi Arabia's wind energy resources

Andrew Zammit-Mangion

School of Mathematics and Applied Statistics, University of Wollongong, Australia

Correspondence
Andrew Zammit-Mangion, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW, Australia.
Email: azm@uow.edu.au

Funding information
Australian Research Council, Grant/Award Numbers: DE180100203, DP190100180

Abstract

Statistical spatiotemporal environmental data analysis is rarely straightforward, with one having to face challenges relating to big data, non-Gaussianity, nonstationarity, multiple scales of behavior, deterministic (numerical) model output, and more. One often has to rely heavily on good statistical parallel computing skills and sound knowledge of the application domain. The work of Tagle et al. (2020) overcomes all of these challenges, and is an excellent example of the tangible contributions spatiotemporal modeling and distribution theory can make to the environmental sciences at the policy level. In this discussion piece I focus on a few high-level concepts in the paper of Tagle et al. (2020) that are relevant to related application domains. I also provide some technical suggestions that could be used to facilitate inference.

KEYWORDS

non-Gaussian models, skew-*t* distribution, spatial statistics, stochastic generator, variational Bayes

1 | INTRODUCTION

This discussion paper is broadly divided into two parts. In the first part I discuss some aspects of the work by Tagle et al. (2020) at a conceptual level; this part includes a brief discussion on the implications of treating numerical model output as "data," certain features of the spatio-temporal model used, and the authors' handling of non-Gaussianity. In the second part I propose some alternative computational strategies that could be used to facilitate inference.

2 | THE IMPLICATIONS OF USING NUMERICAL MODEL OUTPUT AS "DATA"

Tagle et al. (2020) use high-resolution data from the Weather Research and Forecasting (WRF) model to construct a surrogate statistical model that can, in effect, replace WRF over the region and time horizon of interest. Such a strategy of replacing the deterministic model with a statistical one serves several purposes. One can use the statistical model to elicit a prior distribution when doing data assimilation (e.g., Zammit-Mangion, Cressie, & Ganesan, 2016; Zammit-Mangion, Rougier, Schön, Lindgren, & Bamber, 2015), for statistical compression (e.g., Castruccio & Genton, 2016), or, as in Tagle

Abbreviations: ANA, anti-nuclear antibodies; APC, antigen-presenting cells; IRF, interferon regulatory factor

et al. (2020), to play the role of a stochastic weather generator that can be simulated from. In such cases one may execute a "Turing test" to ensure that prior predictive realizations from the fitted model cannot be distinguished from a numerical model output by a field expert (see Majumder, Hofmann, & Cook, 2013, for an example in the context of linear models). The maps in this paper certainly seem convincing, but is granularity from the imposed regionalization apparent at the daily scale? If yes, this may be problematic if one is interested in analyzing the output from the stochastic wind generator at a fine temporal resolution.

A common issue when developing a product from a numerical model is that the said numerical model acts as a single point of failure. WRF has a good track record of being fairly accurate but, like any numerical model, is imperfect (e.g., Carvalho, Rocha, Gómez-Gesteira, & Santos, 2014). In global carbon dioxide flux inversion, for example, the lack of consensus between research groups worldwide is largely due to the disparity between atmospheric transport models (McNorton et al., 2020; Schuh et al., 2019). A related issue is that output from atmospheric transport models are generally not provided with a measure of uncertainty, which is often substituted with the notion of variability. One should, however, differentiate between the (prior or posterior) uncertainty on the physical process, and the system variability. When working with deterministic models, uncertainty quantification necessitates the consideration of the numerical model parameters (plausible configurations), uncertainties in the boundary conditions (in Tagle et al.'s case, the boundary conditions were obtained using an analysis product from the European Centre for Medium-Range Weather Forecasts ECMWF), and model-data discrepancy (if real data is available). In flux inversion, a compromise is generally made by imposing and inflating uncertainties associated with the model outputs until inferences across model groups tend to be in agreement. Which leads me to the question: How would one account for uncertainty associated with the numerical model in this framework?

Another point worth mentioning is that the WRF output used is from a 6-year period in the past. Recent studies show that climate change is causing wind speed patterns to change over relatively short time horizons (on the order of a few decades) that span the planning, construction, and production phases of wind energy plants (Zeng et al., 2019). Could this be an issue, and is it possible to somehow adapt the framework to reliably generate wind demand patterns over the next few decades?

3 | THE SPATIOTEMPORAL MODEL

Wind speed, when treated as a stochastic process, is likely to be non-separable in space and time, and the stencil neighbourhood scheme used in Tagle et al. (2020) does allow for nonseparability. The model also allows for spatially varying dynamics, which is important since dynamics do tend to vary spatially. Figure S3 appears to indicate that temporally varying dynamics do not need to be modeled, however this is very uncommon in environmental applications of this nature, especially at the daily scale and over large time horizons. Catering for time varying dynamics is sometimes difficult, but possible. For example, Zammit-Mangion and Wikle (2020) considered spatially varying parameters in a filtering context using convolution neural networks combined with the integro-difference equation. There, sea-surface temperature was modeled daily, and the dynamics were clearly changing on a daily scale. I suspect that modeling temporally varying dynamics would be important if Tagle et al. were to downscale their model to hourly time intervals to yield prior predictive realizations that are more realistic.

In addition, it is somewhat unusual for models of this kind to not incorporate covariates. Elevation, or terrain gradient, seem to be obvious candidates here. They may be used as fixed effects in $\eta(\mathbf{s})$, or in the covariance function (Schmidt, Guttorp, & O'Hagan, 2011). Alternatively, they may be used to parameterize other parameters appearing in the model, such as ν_r. One can expect that certain properties like tail behavior are dependent on the local topography. Do Tagle et al. (2020) see any scope for adding covariates to their model, and in what way?

4 | MODELING NON-GAUSSIANITY AND THE GAME OF COMPROMISES

The use of a non-Gaussian process, and particularly the skew-t distribution, is elegant. However, moving from a latent Gaussian specification comes at a cost. The first casualty here appears to be the spatial continuity when modeling the residuals. The second casualty is the relative computational ease of making inference; a rather involved Metropolis-within-Gibbs-within-expectation-maximization scheme is used, which I believe is feasible only because of the relatively severe modeling assumptions that have been adopted. While inference does not appear to have been an issue

here, single-site Gibbs updating in conjunction with Metropolis–Hastings is notorious for poor mixing. Further, good mixing and convergence must be difficult to assess/guarantee automatically for each expectation-maximization (EM) iteration. The approach taken is pragmatic, but in other applications one might need to take a different stance when weighing the use of a pure non-Gaussian model against that of a latent Gaussian model.

Figure 2a in Tagle et al. (2020) depicts only mild departures from Gaussianity, so a latent Gaussian model might indeed be appropriate here. Cseke, Zammit-Mangion, Heskes, and Sanguinetti (2016) fitted a spatiotemporal log-Gaussian Cox process model where the spatial dimension was approximately 10,000, and where the data spanned a few hundred time points. The neighboring interactions were also modeled using a stencil-like approach, with a spike and slab prior on the interaction terms. A laptop was used for fitting, which suggests to me that the scales Tagle et al. are considering can be tackled within a large computing environment. This would allow one to do end-to-end estimation in one stage, and avoid the problem of regionalization. However this comes at the cost of using a latent Gaussian approach. Tagle et al. provide some discussion on this compromise in the Introduction, but is there compelling evidence in this application that a latent Gaussian approach would have been inadequate?

One upside of regionalization is that it leads to straightforward modeling of nonstationarity, which in this case extends to orders higher than the conventional first and second orders. With latent Gaussian models it is possible that this regionalization could be avoided while still accommodating nonstationarity. One possible way forward is to consider a combined parallel model and parallel computation strategy such as that employed by Zammit-Mangion and Rougier (2020), wherein finer scales are associated with increasing degrees of non-stationarity.

5 | SOME THOUGHTS ON THE INFERENCE PROCEDURE

As noted above, assessing convergence of the Gibbs sampler for each EM iteration is not a trivial problem. For example, the conditional distribution of each $U_{r,t}$ is likely to be bimodal since its prior distribution is symmetric around zero and only its absolute value appears in the likelihood. I suspect that, since a Metropolis–Hastings algorithm was used, only one mode was explored by the Gibbs sampler. This may not be a problem in practice, but one that should be borne in mind.

Since each $U_{r,t}$ and $Z_{r,t}$ is associated with a few hundred data points (since one has 84,000 spatial locations and 200 regions at each time point), it is likely that the posterior distributions over these variables (ignoring any bimodality) are quite tight around the posterior mean, and Gaussian-like. If this is indeed the case, one can possibly justify using Laplace approximations in the EM algorithm instead of a Gibbs sampler, as in Sengupta and Cressie (2013). This would lead to a considerable simplification in the inferential algorithm with potentially little loss of accuracy. If one is wedded to having a non-Gaussian treatment of the conditional distributions over $\{U_{r,t}\}$ and $\{Z_{r,t}\}$, then another alternative is to resort to an approximate inference scheme such as stochastic variational Bayes (SVB), which can be useful in a highly non-Gaussian setting, as I show next.

Consider the simple model

$$Y_i = \theta + \epsilon_i; \quad i = 1, \ldots, m,$$

where $\epsilon_i \sim SN(0, 1, \lambda)$ (i.e., ϵ_i is skew-normally distributed with location zero, scale one, and slant λ), and $\theta \sim N(0, 1)$, and assume that one wishes to make inference on θ from the data $\mathbf{Y} \equiv (Y_1, \ldots, Y_m)'$. In variational Bayes, one seeks to maximize a lower bound $\mathcal{L}$ on $p(\mathbf{Y})$, given by

$$\mathcal{L} = \mathbb{E}_{q(\theta)}\left[\log \frac{p(\mathbf{Y}|\theta)p(\theta)}{q(\theta)}\right], \tag{1}$$

where $q(\theta)$ is a variational posterior distribution (in this case an approximation to the posterior distribution). If m is small (less than 30, say), the true conditional distribution of θ, when conditioned on the data $\mathbf{Y}$, is likely to be highly non-Gaussian. It is therefore reasonable to let $q(\theta)$ be a skew-normal distribution,

$$q(\theta) = SN(\xi, \omega^2, \alpha),$$

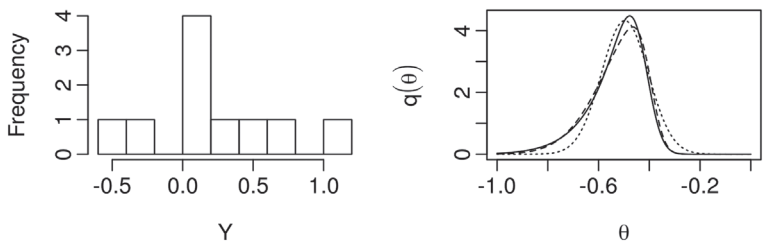

FIGURE 1 Left panel: Histogram of the simulated data. Right panel: The true posterior distribution $p(\theta|\mathbf{Y})$ (solid line), the variational posterior distribution when $q(\theta)$ is assumed to be skew-normal (dashed line), and the variational posterior distribution when $q(\theta)$ is assumed to be Gaussian (dotted line)

where ξ is the unknown location parameter, ω is the unknown scale parameter, and α is the unknown slant parameter of the variational posterior distribution. There have been attempts in the past to use skew-normal distributions in a classical VB framework (e.g., Ormerod, 2011).

In SVB one approximates (1) via the Monte Carlo approximation

$$\mathcal{L} \approx \frac{1}{N}\sum_{j=1}^{N} \log \frac{p(\mathbf{Y}|\theta^{(j)})p(\theta^{(j)})}{q(\theta^{(j)})},$$

where $\theta^{(j)} \sim SN(\xi, \omega^2, \alpha)$. To obtain gradients with respect to ξ, ω, and α, these parameters need to be explicitly represented in the expression for $\mathcal{L}$. I therefore now draw on a connection between the distributions used by Tagle et al. (2020) and the reparameterization trick of Kingma and Welling (2014) by expressing the Monte Carlo samples via the parameterization of Azzalini and Capitanio (2013, Chapter 2),

$$\theta^{(j)} = \xi + \omega\gamma^{(j)}, \quad j = 1, \ldots, N,$$

where, for $\alpha \equiv \delta/\sqrt{1-\delta^2}$, $U_0^{(j)} \sim N(0,1)$, and $U_1^{(j)} \sim N(0,1)$,

$$\gamma^{(j)} = \delta|U_1^{(j)}| + \sqrt{1-\delta^2}U_0^{(j)} \sim SN(0,1,\alpha).$$

The variational distribution over θ is then found by solving the optimization problem

$$(\hat{\xi}, \hat{\omega}, \hat{\alpha}) = \arg\max_{\xi,\omega,\alpha} \frac{1}{N}\sum_{j=1}^{N} \frac{p(\mathbf{Y}|\theta^{(j)})p(\theta^{(j)})}{q(\theta^{(j)})}, \tag{2}$$

where an expression for $q(\theta)$ can be found in Azzalini and Capitanio (2013, Chapter 2). By virtue of the reparameterization, the optimization problem (2) can now be solved quickly and efficiently using parallel computing and automatic differentiation libraries.

For illustration, let $m = 10$, $\lambda = 20$ (assumed known), and $\theta = -0.5$. Figure 1 in this discussion piece, left panel, shows a histogram of the simulated data, while the right panel shows the true posterior distribution over θ (solid line), the approximate posterior distribution when $q(\theta)$ is assumed to be skew-normal (dashed line), and the case where $q(\theta)$ is assumed to be Gaussian (dotted line). SVB was run with $N = 500$ on a graphics processing unit and required only a few seconds to converge. Clearly, the approximated skew-normal variational posterior is reasonably representative of the true posterior distribution. I note that this reparameterization trick for skew-normal variational posterior distributions has been used recently in the context of copula models by Smith, Loaiza-Maya, and Nott (2020).

Some extensions would be needed to use SVB instead of a Gibbs sampler in Tagle et al. (2020)'s work since, for example, each $Z_{r,t}$ needs to be positive. Zammit-Mangion, Ng, Vu, and Filippone (2019) used trans-Gaussian variational posterior distributions to achieve a positivity constraint in the context of deep spatial models, while one may consider other functional forms for the approximate posterior distribution, such as the gamma distribution. An interesting avenue of future work is the development of skew-t variational posterior distributions, which would be suitable for approximating posterior distributions with heavy tails. Here, one could proceed by optimizing the lower bound for a selection of plausible integer-valued degrees of freedom; the optimal degrees of freedom would then be that which yields the largest variational lower bound.

ACKNOWLEDGEMENTS

AZ–M is supported by the Australian Research Council (ARC) Discovery Early Career Research Award, DE180100203, and the ARC Discovery Project, DP190100180.

DATA AVAILABILITY STATEMENT

N/A

REFERENCES

Azzalini, A., & Capitanio, A. (2013). *The Skew-normal and related families*. Cambridge, UK: Cambridge University Press.

Carvalho, D., Rocha, A., Gómez-Gesteira, M., & Santos, C. S. (2014). Comparison of reanalyzed, analyzed, satellite-retrieved and NWP modelled winds with buoy data along the Iberian Peninsula coast. *Remote Sensing of Environment*, *152*, 480–492.

Castruccio, S., & Genton, M. G. (2016). Compressing an ensemble with statistical models: An algorithm for global 3D spatio-temporal temperature. *Technometrics*, *58*, 319–328.

Cseke, B., Zammit-Mangion, A., Heskes, T., & Sanguinetti, G. (2016). Sparse approximate inference for spatio-temporal point process models. *Journal of the American Statistical Association*, *111*, 1746–1763.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114.

Majumder, M., Hofmann, H., & Cook, D. (2013). Validation of visual statistical inference, applied to linear models. *Journal of the American Statistical Association*, *108*, 942–956.

McNorton, J. R., Bousserez, N., Agustí-Panareda, A., Balsamo, G., Choulga, M., Dawson, A. E. R., … Lang, S. (2020). Representing model uncertainty for global atmospheric CO_2 flux inversions using ECMWF-IFS-46R1. *Geoscientific Model Development*, *13*, 2297–2313.

Ormerod, J. (2011). *Skew-normal variational approximations for Bayesian inference*. Unpublished. Retrieved from. http://talus.maths.usyd.edu.au/u/jormerod/JTOpapers/snva.pdf

Schmidt, A. M., Guttorp, P., & O'Hagan, A. (2011). Considering covariates in the covariance structure of spatial processes. *Environmetrics*, *22*, 487–500.

Schuh, A. E., Jacobson, A. R., Basu, S., Weir, B., Baker, D., Bowman, K., … Deng, F. (2019). Quantifying the impact of atmospheric transport uncertainty on CO_2 surface flux estimates. *Global Biogeochemical Cycles*, *33*, 484–500.

Sengupta, A., & Cressie, N. (2013). Hierarchical statistical modeling of big spatial datasets using the exponential family of distributions. *Spatial Statistics*, *4*, 14–44.

Smith, M. S., Loaiza-Maya, R., & Nott, D. J. (2020). High-dimensional copula variational approximation through transformation. *Journal of Computational and Graphical Statistics.* in press. https://doi.org/10.1080/10618600.2020.1740097.

Tagle, F., Genton, M. G., Yip, A., Mostamandi, S., Stenchikov, G., & Castruccio, S. (2020). A high-resolution bilevel skew-*t* stochastic generator for assessing Saudi Arabia's wind energy resources. *Environmetrics*, e2628. https://doi.org/10.1002/env.2628.

Zammit-Mangion, A., Cressie, N., & Ganesan, A. L. (2016). Non-Gaussian bivariate modelling with application to atmospheric trace-gas inversion. *Spatial Statistics*, *18*, 194–220.

Zammit-Mangion, A., Ng, T. L. J., Vu, Q., & Filippone, M. (2019). Deep compositional spatial models. arXiv preprint arXiv:1906.02840.

Zammit-Mangion, A., & Rougier, J. (2020). Multi-scale process modelling and distributed computation for spatial data. *Statistics and Computing* in press. https://doi.org/10.1007/s11222-020-09962-6.

Zammit-Mangion, A., Rougier, J., Schön, N., Lindgren, F., & Bamber, J. (2015). Multivariate spatio-temporal modelling for assessing Antarctica's present-day contribution to sea-level rise. *Environmetrics*, *26*, 159–177.

Zammit-Mangion, A., & Wikle, C. K. (2020). Deep integro-difference equation models for spatio-temporal forecasting. *Spatial Statistics*, *37*, 100408.

Zeng, Z., Ziegler, A. D., Searchinger, T., Yang, L., Chen, A., Ju, K., … Wood, E. F. (2019). A reversal in global terrestrial stilling and its implications for wind energy production. *Nature Climate Change*, *9*, 979–985.

How to cite this article: Zammit-Mangion A. Discussion on A high-resolution bilevel skew-*t* stochastic generator for assessing Saudi Arabia's wind energy resources. *Environmetrics*. 2020;31:e2649. https://doi.org/10.1002/env.2649