Discussion on “Saving Storage in Climate Ensembles: A Model-Based Stochastic Approach”

Andrew Poppick

1. INTRODUCTION

The authors are to be congratulated for a thought-provoking paper on the use of statistical models for a form of data compression for climate model output. This paper continues the authors’ many contributions to the literature on climate model emulation and compression and highlights that these two goals, often thought of as separate within the climate community, may for some purposes be more similar than has been typically appreciated. The statistical models proposed here are based on one originally proposed in Castruccio et al. (2014) and further refined and expanded upon in subsequent papers. While much of the authors’ prior work in this area has been for the purpose of climate model emulation, following Castruccio and Genton (2016) the idea here is to view the statistical model as a data compression to address the increasing and unsustainable storage burden for large climate model experiments. The proposed “unconditional” data compression is not a data compression in the sense usually meant (termed “conditional” by the authors) because it does not allow for the recovery of an approximation of the actual values produced by the climate model; rather, it provides an approximation of the unconditional spatiotemporal distribution of the temperature values. Since climate is by definition a distribution, such an unconditional simulation may be sufficient for some purposes related to studying projected climate change and its impacts.

While there is much to value in the ideas in this paper, I believe that the approach to compression described here will not be without controversy within the community of researchers who use climate model output. Since the original data (or a good approximation

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

A. Poppick (✉) Department of Mathematics and Statistics, Carleton College, Northfield, MN, USA
(E-mail: *apoppick@carleton.edu*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 28, Number 2, Pages 345–348
https://doi.org/10.1007/s13253-023-00537-2

thereof) are not recoverable after unconditional compression, I believe that the primary concern is whether all distributional features of potential interest to an end user have been modeled with fidelity. It is hard to know how someone else might want to use a climate model's output in a future study, and it is also hard to check that a statistical model is fully adequate in capturing complex dependencies and other distributional features. It therefore makes sense to me that the primary approach to data compression for climate models has instead been the traditional (conditional) approach to data compression (e.g., as investigated in Baker et al. (2016, 2017), Poppick et al. (2020), and others).

That said, if one is willing to accept an emulation of a climate model in the place of an actual run that could not be completed due to computational limitations, why not accept the same such approximation to a run that was completed but could not be saved due to storage limitations? Given the demonstrated utility (albeit imperfect) of climate model emulators, I am therefore cautiously sympathetic to the unconditional compression approach advocated here, especially in circumstances where storage limitations are such that the original data or a good (conditional) lossy compression of it simply cannot be saved. I do, however, believe that if this approach were adopted, then we as statisticians would have an important role to play in articulating which features of the original output have been modeled with fidelity and which features have not. For example, while the authors have made efforts to partially capture temperature variability, I am concerned that their model may not yet be adequate for this purpose. There are also important hurdles to consider when attempting to adapt this approach to a multivariate setting, which the authors briefly but intriguingly describe in their paper.

Below I elaborate on some of the issues I have alluded to above. I focus my discussion on the compression of daily, gridcell-level data, since that is the case explored where there is the most compelling reason to apply a data compression method.

2. VARIABILITY AND EXTREMES

The topic of daily temperature variability and extremes, and changes thereof, has received a substantial amount of attention within the climate literature, in part because climate change impacts can depend disproportionately on changes in extreme events. There is evidence both from observations and climate model output that temperature variability and extremes have changed over the historical record and are projected to change in the future, but changes are both spatially and seasonally heterogeneous (Fischer and Schär 2009; Huang et al. 2016; Poppick et al. 2016; Rhines et al. 2017; Haugen et al. 2018; Bathiany et al. 2018; Dunn et al. 2019 and many others). None of the models proposed by the authors allow for changes in variability over time, as they note. While it may be possible to add a model for variability changes to their existing model, this is not necessarily a straightforward task, given the heterogeneity in changes, and highlights a general drawback to the unconditional compression approach: No distributional feature of the original data will be retained unless it is explicitly modeled.

Likewise, as the authors note, their model would not be expected to be a good description of the extremal behavior of temperatures produced by the climate model (as illustrated in

Figure 6 (D) especially), despite their effort to account for some elements of non-Gaussianity using the Tukey g-and-h transformation. While the authors argue that a separate model that captures the extremal behavior could be used, this would not be straightforward either: Evidence suggests that the standard approach of modeling annual extrema in climate model output using the GEV distribution may not be adequate due to the fact that an annual block size may not be long enough for the GEV approximation to be reasonable (Huang et al. 2016; McKinnon and Simpson 2022) and modeling spatially dependent extremes (relevant for understanding, e.g., the behavior of heat events that affect whole regions) is a notoriously challenging statistical modeling and computational problem that remains an active area of research. On the one hand, this points to perhaps fertile ground for the statistical modeling community to explore; on the other hand, I worry that we are not currently at a point where straightforward extensions of the work described here would give an adequate compression for users interested in extremes.

Despite the above cautions, there are likely some applications where the representation of variability provided in the authors' model would be sufficient for purpose, for example, in application areas where there is not a large sensitivity to temperature variability. It would perhaps be worthwhile to investigate the sensitivity of scientific conclusions to the use of this proposed model in order to better establish good use cases for this model.

3. MULTIVARIABLE RELATIONSHIPS

Similar issues arise when considering multivariate extensions of the method proposed by the authors. Many climate change impacts depend on both temperature and precipitation, for example, and, as the authors note, daily precipitation is much more challenging to model statistically (many zeros, highly skewed distribution on days with positive precipitation, and complicated spatiotemporal dependencies). The authors propose an interesting potential solution to this problem: do not trying to unconditionally compress precipitation, but then develop a statistical model for temperature conditional on precipitation as a compression of temperature. This idea may become more challenging with the introduction of more variables, but the idea to try to model conditional distributions rather than directly modeling joint distributions seems like a potentially more approachable and fruitful way to address multivariate extensions of this work.

4. CONCLUDING REMARKS

While I have discussed some challenges and concerns about the unconditional compression approach advocated by the authors, I think that the unconditional and conditional compression approaches can be viewed as complementary. Compared to the conditional approach, the unconditional approach may entail a lower storage burden and has the additional advantage that one can sample arbitrarily many times from the modeled distribution; however, for modest compression rates the conditional approach may have an easier time reproducing complex climate model distributional features. Assessing when any compression method (unconditional or conditional) produces data that is adequate for purpose will ideally require contributions from statisticians in addition to climate scientists and downstream users of climate model output (e.g., climate change impacts researchers), and I welcome the authors' encouragement that the statistics community be involved in this important work.

[Accepted March 2023. Published Online May 2023.]

REFERENCES

Baker AH, Hammerling DM, Mickelson SA, Xu H, Stolpe MB, Naveau P, Sanderson B, Ebert-Uphoff I, Samarasinghe S, De Simone F et al (2016) Evaluating lossy data compression on climate simulation data within a large ensemble. Geosci Model Develop 9(12):4381–4403

Baker AH, Xu H, Hammerling DM, Li S, Clyne JP (2017). Toward a multi-method approach: Lossy data compression for climate simulation data. In: International conference on high Performance computing, pp. 30–42. Springer

Bathiany S, Dakos V, Scheffer M, Lenton TM(2018). Climate models predict increasing temperature variability in poor countries. Sci Adv. 4(5), eaar5809

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

Castruccio S, McInerney DJ, Stein ML, Crouch FL, Jacob RL, Moyer EJ (2014) Statistical emulation of climate model projections based on precomputed gcm runs. J Clim 27(5):1829–1844

Dunn RJ, Willett KM, Parker DE (2019) Changes in statistical distributions of sub-daily surface temperatures and wind speed. Earth Syst Dynam 10(4):765–788

Fischer EM, Schär C (2009) Future changes in daily summer temperature variability: driving processes and role for temperature extremes. Clim Dynam 33(7):917–935

Haugen MA, Stein ML, Moyer EJ, Sriver RL (2018) Estimating changes in temperature distributions in a large ensemble of climate simulations using quantile regression. J Clim 31(20):8573–8588

Huang WK, Stein ML, McInerney DJ, Sun S, Moyer EJ (2016) Estimating changes in temperature extremes from millennial-scale climate simulations using generalized extreme value (gev) distributions. Adv Stat Climatol, Meteorol Oceanography 2(1):79–103

McKinnon KA, Simpson IR, (2022) . How unexpected was the, (2021) pacific northwest heatwave? Geophysical Research Letters. 49(18):100

Poppick A, McInerney DJ, Moyer EJ, Stein ML (2016) Temperatures in transient climates: improved methods for simulations with evolving temporal covariances. Annals Appl Stat 10(1):477–505

Poppick A, Nardi J, Feldman N, Baker AH, Pinard A, Hammerling DM (2020) A statistical analysis of lossily compressed climate model data. Comput Geosci 145:104599

Rhines A, McKinnon KA, Tingley MP, Huybers P (2017) Seasonally resolved distributional trends of north american temperatures show contraction of winter variability. J Clim 30(3):1139–1157