Rejoinder on 'Saving Storage in Climate Ensembles: A Model-Based Stochastic Approach'

Huang HUANG, Stefano CASTRUCCIO, Allison H. BAKER, and Marc G. GENTON

We thank all the discussants for their valuable comments. Throughout this rejoinder, we denote the discussants by D = Datta, P = Poppick, BA = Banerjee, BU = Burr, BUD = Bessac, Underwood and Di. We will also use the same acronyms as in the discussion, i.e., VAE = Variational Autoencoders and DNN = Deep Neural Networks. We organize the rejoinder around four main themes.

1. CHALLENGES OF UNCONDITIONAL COMPRESSION

There are several comments raised about challenges associated with unconditional compression. P argued that no distributional feature of the original data will be retained unless it is explicitly modeled. We believe this is ultimately the main limitation of our modeling approach. We have discussed validations in this work and more diagnostics could certainly be developed in order to understand 'which features of the original output have been modeled with fidelity and which features have not' (more in the next section on diagnostics). While we agree that 'it is hard to know how someone else might want to use a climate model's output in a future study', we are ultimately aligned with P's stance that our approach would help in 'preserving models for future comparison or reproducibility and replicability of results, especially in validation of climate models as used in policy'. In this regard, we argue then

This article is response to commentaries for https://doi.org/10.1007/s13253-023-00537-2, https://doi.org/10.1007/s13253-023-00538-1, https://doi.org/10.1007/s13253-023-00539-0, https://doi.org/10.1007/s13253-023-00540-7, https://doi.org/10.1007/s13253-023-00541-6.

H. Huang · M. G. Genton (✉), Statistics Program, King Abdullah University of Science and Technology, 23955-6900 Thuwal, Saudi Arabia (E-mail: *marc.genton@kaust.edu.sa*).
S. Castruccio, Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, USA.
A. H. Baker, Computational and Information Systems Lab, National Center for Atmospheric Research, Boulder, CO 80305, USA.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 28, Number 2, Pages 370–374
https://doi.org/10.1007/s13253-023-00542-5

that storing the parameters and algorithms to decompress is infinitely better than simply losing the data to make storage space for more advanced simulations in the future.

The possibility of modeling multiple variables simultaneously is also arguably poised to be one of the main barriers to a widespread usage of this approach, and to address the remark by P, the second question posed by BUD as well as the second part of BA's discussion, our current stance is that a joint modeling approach for multiple variables is likely not going to lead to competitive and flexible enough modeling. Instead, we advocate for a conditional modeling approach, where some less well-behaved variables are either not compressed at all or (more likely) compressed with some standard algorithms such as the ones proposed by BUD, while the remaining variables are modeled with an unconditional approach. This approach would be especially appealing for high resolution variables such as precipitation, whose modeling would require hierarchical Bayesian models and latent Gaussian processes.

Our model could also be improved by allowing for a temporally changing harmonic trend or dependence structure, as highlighted by BU and P. If this approach is to become in fact used for actual compression of CMIP6 or other future ensembles, climate scientists and statisticians need to establish in the documentation of such approach what is preserved and what is not.

A further comprehensive comparison of the proposed unconditional compression approach with more standard compression algorithms would also help put our modeling framework better in the context on the current research on the topic. In this regard, we appreciate the comparison performed by BUD and we are especially pleased to see that the right panel of their Figure 2 confirms that, if we choose a compression algorithm such as SZ3 with the same compression rate as our approach (within 1%), a model-based approach such as the one we proposed achieves better results, as shown by comparing the first and fourth boxplot. Clearly, specifying a smaller compression ratio would result in a more accurate representation of the original data, as can be seen by the second and third boxplot, but this would come at the cost of an increase in storage that may not be sustainable in the long term.

In their third question, BUD also mentioned the possibility of framing stochastic emulators as alternatives to Earth system models for reducing the carbon footprint. While this is certainly true for traditional emulators (less computations for sensitivity analysis imply a reduction in emissions), in the context of compression this is perhaps less apparent. Indeed, while our proposed method can be used to save storage, and hence energy, the current space limitations are bound to pressure users of climate cyberinfrastructures to immediately use the available space for new simulations. As such, we believe that the link between sustainability and compression via statistical modelling is bound to be open to discussion.

In summary, we believe that, while there are proved modeling and computational limitations, we also think that some, if not most, of the hurdles to disseminate our idea in the climate community will be dictated by accepting a conceptual change of framework. On this topic, we hope that the decades of collaborating history between the climate and statistics community in developing emulators would help. P has well summarized our general attitude towards this emerging challenge: '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?'.

2. DIAGNOSTICS

In the previous section, we have argued that diagnostics are going to play a key role in the widespread adoption of the practice of compression in climate science. Traditional compression algorithms are currently limited, in that they rely on some parameters to control the aggressiveness (or amount) of compression and are generally limited to absolute, relative or pointwise error tolerance in space or time. The interest, however, lies in ensuring that a compressed dataset would result in the same scientific conclusion that one would draw from the original data. This is problematic since (1) the relationship between compression error and scientific conclusion varies by model, meteorological variable and resolution; and (2) it is not usually known a priori what type of scientific investigation one would be interested in performing. As such, traditional compression approaches are intrinsically limited as they require knowledge of the features that are desired to be preserved.

The proposed model-based approach seems more suitable than traditional compression algorithms to preserve at least some climatological properties such as the annual cycle as they can be easily and explicitly specified inside the model. Despite these advantages, the unconditional compression approach only provides new realizations with similar statistical properties to the original data, so a point-by-point comparison with prediction metrics such as mean squared error are not meaningful. BUD have explicitly posed this question, and have also implicitly provided some answers by measuring the discrepancy between original and 'compressed' data in terms of Wasserstein distance. Along the same lines, similar metrics measuring the distributional distance between original and compressed data, such as Kullback–Leibler divergence and cross-entropy, could be used. Ultimately, however, these metrics are going to be less intuitive than the ones developed for traditional compression, as our approach has to be assessed in a probabilistic sense rather than with more intuitive point prediction metrics.

In addition to traditional metrics, visual assessment could also be performed to facilitate the acceptance of the proposed compression approach by climate scientists. In this regard some of our past work (Genton et al. 2015; Castruccio et al. 2019), and similarly some work performed at NCAR (Baker et al. 2016), advocate for the use of blind experiments to test if climate scientists are actually able to recognize real climate simulations from compressed versions, either with traditional algorithms or with unconditional compression. Clearly, such exercises are intrinsically limited, as scientific investigations are not performed with pictures, but we believe such endeavor could nevertheless be useful for facilitating the acceptance of this new framework within the climate community.

3. A SYSTEM'S PERSPECTIVE ON SAVING SPACE AND COMPUTATION

In the first part of BA's comment, it was pointed out how, from a broad perspective, climate models are actually just dynamical (albeit generally deterministic) space-time systems,

similarly to our approach. Indeed, using the same notation as BA, we could think of our approach as a stochastic state approximation

$$W_t = \widetilde{\mathcal{M}}(W_{t-1}; \tilde{\theta}_t, x_t, \tilde{\eta}_t),$$

where the structure of $\tilde{\eta}_t$ is nontrivial, but $\widetilde{\mathcal{M}}$ is considerably simpler in structure than the original $\mathcal{M}$, i.e., linear or almost linear. Instead, in the case of a climate model without stochastic parametrizations (as is usually the case) we would have $\tilde{\eta}_t = \emptyset$, but a considerably more complicated evolution $\mathcal{M}$, for example a non-linear system of PDEs such as the Navier–Stokes equations.

Under this perspective, the distinction between storage and computation is considerably more blurred. Indeed, both systems generate data, the only difference is that $\widetilde{\mathcal{M}}$ is a faster approximation of $\mathcal{M}$ with an additional stochastic structure. As such, one may argue that the true advantage of a stochastic approach actually lies in the convenience of a linear (or similarly computationally affordable) stochastic model compared to the original model. After all, if one could store the algorithm to generate the climate models, along with the physical parameters θ_t and the input x_t, how is that different from our proposed stochastic approach? Ultimately, we believe the answer is in the computational convenience of $\widetilde{\mathcal{M}}$, which can generate 'uncompressed' runs considerably faster than $\mathcal{M}$.

4. VARIATIONAL AUTOENCODERS AS ALTERNATIVE APPROACHES

We thank D for discussing the merits of an alternative nonparametric model comprising of variational autoencoders with deep neural networks modeling the mean and variance of the encoder and decoder. This is a promising direction since: (1) the representation of the mean and variance structure is indeed complex and is poised to be even more so for high resolutions; and (2) as D mentioned at the end of the discussion, storing parameters is bound not to be an issue, compared to storing the entire ensemble. D has also mentioned some of the potential shortcomings, and we will briefly elaborate further on them. Firstly, a fully nonparametric approach such as DNN could approximate some basic features such as the annual cycle or conservation of energy, but the lack of explicit modeling is bound to be a limitation. One solution would be to formulate semi-parametric models with some pre-specified features from the physics, and then perform the proposed VAE representation over the residuals. Secondly, for small datasets such a highly parametric approach may not be ideal. We agree with D on this point, and while for the application discussed in this work this may not be a major issue, this would certainly be the case for scientific investigations where only selected days or events are simulated on a small scale. We believe the DNN representation could be potentially still used with some appropriate modifications. Indeed, we could opt for some stochastic sparse network representation, along the lines of reservoir state space (e.g., echo state networks and liquid state machines in the context of time series) to still retain the flexibility implied by the DNN while reducing the parameter space.

[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, Carbone F, Gencarelli CN, Dennis JM, Kay JE, Lindstrom P (2016) Evaluating lossy data compression on climate simulation data within a large ensemble. Geosci Model Dev 9:4381–4403

Castruccio S, Genton MG, Sun Y (2019) Visualizing spatiotemporal models with virtual reality: from fully immersive environments to applications in stereoscopic view (with discussion). J R Stat Soc Ser A 182(2):379–387

Genton MG, Castruccio S, Crippa P, Dutta S, Huser R, Sun Y, Vettori S (2015) Visuanimation in statistics. Stat 4(1):81–96