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

Julie BESSAC, Robert UNDERWOOD, and Sheng DI

We thank the authors for this interesting paper that highlights important ideas and concepts for the future of climate model ensembles and their storage, as well as future uses of stochastic emulators. Stochastic emulators are particularly relevant because of the statistical nature of climate model ensembles, as discussed in previous work of the authors (Castruccio et al. in J Clim 32:8511–8522, 2019; Hu and Castruccio in J Clim 34:8409–8418, 2021). We thank the authors for sharing of some of their data with us in order to illustrate this discussion. In the following, in Sect. 1 we discuss alternative techniques currently used and studied, namely lossy compression and ideas emerging from the climate modeling community, that could feed the discussion on ensemble and storage. In that section, we also present numerical results of compression performed on the data shared by the authors. In Sect. 2, we discuss the current statistical model proposed by the authors and its context. We discuss other potential uses of stochastic emulators in climate and Earth modeling.

Key Words: Lossy Compression; Stochastic Emulators; Climate model outputs.

1. LOSSY COMPRESSION, STORAGE, AND STATISTICS

1.1. LOSSY COMPRESSORS

Lossy compressors are increasingly adopted in scientific research, tackling volumes of data from experiments or parallel numerical simulations and facilitating data storage and movement. Lossy compression enables significantly reducing the data size without sacrificing data integrity, which is a concerning research problem for many of today's scientific projects. Lossy compressors typically comprise (i) a decorrelation step that exploits correlations present in the dataset to transform the data into a more compressible version, (ii)

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

J. Bessac (✉) · R. Underwood · S. Di
National Renewable Energy Laboratory, Computational Science Center, Golden, CO, USA
(E-mail: *julie.bessac@nrel.gov*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 28, Number 2, Pages 358–364
https://doi.org/10.1007/s13253-023-00540-7

an approximation/quantization step that reduces the precision of the input data bringing the lossyness in the compression pipeline, and (iii) an encoding step that minimizes the number of bits used to represent the approximation step outcomes. Different lossy compressors leverage different decorrelation, quantization, and encoding methods. Leading lossy compressors include SZ3 (cubic-spline interpolation-based prediction Zhao et al. 2021), SPERR (wavelet based Li and Clyne 2022), ZFP (near-orthogonal transform Lindstrom 2014), TThresh (singular value decomposition based Ballester-Ripoll et al. 2020, and Big Grooming Zender 2016). In the following, we focus on SZ3 because it provides the best compression quality based on previous studies (Underwood et al. 2022a).

1.2. LOSSY COMPRESSION IN EARTH SYSTEM MODELING

Earth system model outputs are particularly relevant to lossy compression because they generate large outputs that present correlations. Recent studies have quantified the influence of spatial correlation on lossy compression outputs. In Klöwer et al. (2021), the concept of bitwise real information (BIR) is introduced as the mutual information of bits in adjacent grid points. In particular, the stronger the association with neighboring bits, the greater the BIR. In Krasowska et al. (2021), global and local measures of spatial correlation via variogram estimation are introduced as explanatory variables of compression ratios (ratio of the original data size to the compressed data size).

In lossy compression, the question of evaluating the quality of reconstructed data naturally brings in statistics. For instance, Baker et al. (2017) address the issue of striking a balance between meaningfully reducing data volume and preserving the integrity of the simulation data via a series of quality assessment metrics: the structural similarity image metric (SSIM/d-SSIM), the p-value of the Kolmogorov–Smirnov test, the Pearson correlation coefficient of determination (R^2), and the spatial relative error. In their paper, Baker et al. accompany each metric with an acceptable threshold. In a recent work, Underwood et al. (2022a) propose an extensive comparison of 11 lossy compressors for Community Earth Systems Model (CESM) simulations and suggest the use of different metrics based on probability distributions including the Wasserstein distance.

1.3. NUMERICAL RESULTS

In this section, we collect results of some compression algorithms that have been performed on the original data provided by the authors, and we compare some of the decompressed data with the simulated ensemble proposed by the authors. We perform the analysis in the monthly regional context as this being the largest data provided by the authors.

The example data are provided as a 3-dimensional dataset in which the first and third dimensions, representing the regions and the ensemble members, are numerically uncorrelated. This structure would be difficult for most compressors to capture because they are designed to exploit local (typically less than 16 elements away) spatial correlations within the dataset dimensions. When data are presented as 1-dimensional arrays to the compressors, however, the compressors are able to better capture this local behavior. SZ3 with its dynamic cubic-spline predictor captures this behavior very accurately at competitive

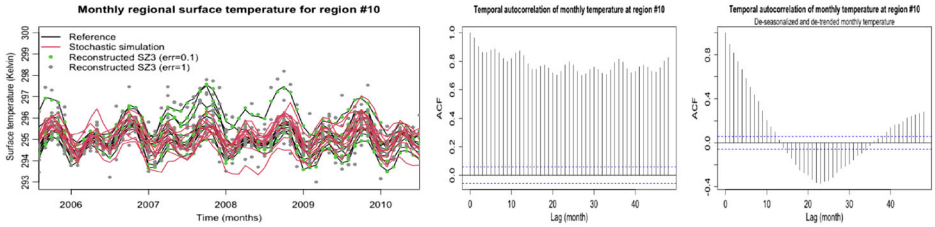

Figure 1. Left: Time series of each monthly and regional signal for 4 consecutive years of the study (2006-2010). Solid black line: original data; solid red lines: stochastic samples proposed by the authors; and dots: reconstructed data from compression with SZ3 (green: absolute error bound of 10^{-1}, grey: absolute error bound of 1). Central and right: temporal autocorrelation of monthly regional temperature at region #10, respectively, for the raw data (typically input in a compressor) and detrended and deseasonalized data (typically operated by statisticians) (Color figure online).

compression ratios. Figure 1 shows the temporal autocorrelation of the monthly average temperature (that is input as such in a compressor) at region #10 (chosen arbitrarily) and their detrended and deseasonalized (accounting for 6- and 12-month periods) counterparts (typically operated by statisticians). In this figure, we highlight the high correlation ranges present in the data, in both raw and detrended and deseasonalized data, as well as periodicities (6- and 12-month periods), which vastly exceed what most compressors can observe and exploit to increase compression ratios. To find a configuration of the compressor that is the closest to matching the compression ratio of the proposed statistical model, we use OptZConfig, developed by Underwood et al. (2022b), which finds an error bound enabling a given compression ratio. In our case, the absolute error bound is sought between 1 and 100. SZ3 was able to match within 1% the compression ratio achieved by the authors by using an absolute error bound of 12.9.

The left panel of Fig. 1 shows the monthly regional (region #10) temperature and its counterparts coming from the stochastic model proposed by the authors and from a reconstruction after compression with SZ3. We note that the overall trends, periodicities, and ensemble spread are captured by most methods. We also observe the trade-off between compression techniques and the stochastic simulations. Stochastic simulations reproduce the statistical behaviors of the reference data and as many samples as can be generated. On the other hand, compression methods capture pointwise behaviors of each ensemble member and provide a tool applicable in a general context (i.e., other variables); however, they cannot generate new scenarios. We also exemplify the various quality of reconstructed data from compression by showing two different absolute error bounds, 10^{-1} and 1. As the error bound becomes more permissive, the accuracy of the reconstruction decreases; however, the compression ratio increases. In the left panel of Fig. 2, we compare the distributions of the original data, stochastic simulations from the authors, and decompressed data. We note that simulation and compression methods both perform well but do not capture the same parts of the distribution. The right panel of Fig. 2 shows the distributions of the Wasserstein distance computed between the reference data and the simulated and decompressed data, following (Underwood et al. 2022a). Each boxplot data point represents the Wasserstein distance computed for a region. Stochastic samples, reconstructions from SZ3 at the 10^{-1} and 1 error level, exhibit a low median Wasserstein distance. However, the most permissive

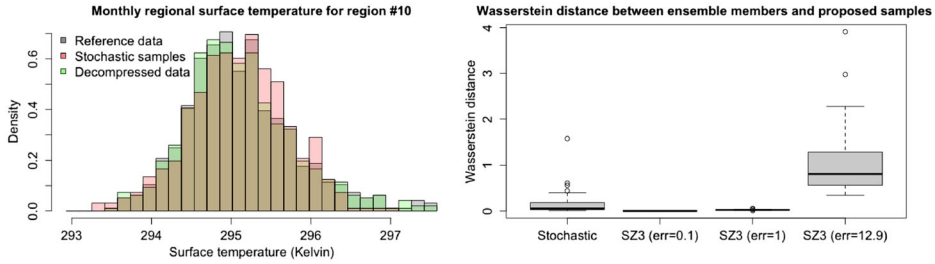

Figure 2. Left: histograms of the original monthly temperature data (region #10), corresponding stochastic samples from the authors, and decompressed data from SZ3. Right: Wasserstein distance between the original data and, from left to right, stochastic samples from the authors and SZ3 reconstruction with error = 0.1, 1, and 12.9 .

compression with tolerated errors at 12.9, corresponding to the compression ratio matching the authors' one, does not capture well the distribution of the reference data. As pointed out by the authors, the stochastic samples may not recover the highest points of the distributions. This situation is common with statistical models and can explain the variability in the corresponding Wasserstein distances.

We have highlighted some differences between both methods. Since each technique has its own advantages and drawbacks, these differences could be leveraged as synergies in future work depending on the user's priority, which could be storage, time performance, or generative aspects.

Questions: Trust and confidence are topics often raised in data compression. Have the authors thought of statistics that could be associated with their model in that regard?

2. STOCHASTIC EMULATORS

2.1. PROPOSED STOCHASTIC GENERATORS

As noted by the authors, the stochastic simulations of different fields can represent a challenge (Ailliot et al. 2015a). Wind fields comprising wind intensity and direction, which necessitate a circular treatment, require regime switching because both variables are linked to weather types (Ailliot and Monbet 2012; Ailliot et al. 2015b). Precipitation is another challenging variable since it combines an occurrence and a rainfall amount variable (Katz 1977; Thompson et al. 2007; Kleiber et al. 2012). Additionally, to mimic operational settings, one would need to simulate concurrently multiple variables together. This remains a challenging aspect of statistical models. Richardson (1981) and Parlange and Katz (2000) pioneered single-site models for multiple variables; however, the addition of a spatial component has rarely been attempted since then.

The authors mentioned the difficulty of modeling the bulk and tails of a distributions. Recent developments in that vein have been proposed for unidimensional probability distribution functions (PDFs); see (Tencaliec et al. 2020; Stein 2021a,b). These models seek to capture low, moderate, and high values of a variable of interest into a single PDF model.

However, the models require additional layers of modeling to account for spatial and temporal dependencies and their nonstationarities.

The study of the minimal number of training samples from the CESM-LENS is crucial and interesting. We expect that this number of training samples would increase if one were to look at other variables such as wind or precipitation and also at different temporal scales such as subdaily scales. Determining a reduced number of ensemble members that are representative of the ensemble could also be a way to think about storage saving, but it requires defining statistically and mathematically the concept of "representativeness."

Questions: We wonder whether the authors have thought how they would generalize their model to a multivariate setting. Do they have any thoughts about the representativeness of ensemble members as a way to save storage?

2.2. NEXT GENERATION OF WEATHER AND CLIMATE MODELS

As discussed by the authors, the need to rethink climate model simulations is critical as the increasing computational power and high-resolution modeling lead to larger and larger outputs. As an example discussed by Klöwer et al. (2021), the European Centre for Medium-Range Weather Forecasts produces 230 TB of data on a typical day, and this data production is expected to quadruple within the next decade because of the increased spatial resolution of the forecast model (Bauer et al. 2020). In that vein, recent works have studied the use of single and mixed precision in climate modeling in order to tackle large amounts of data while ensuring forecast quality (Váňa et al. 2017; Tintó Prims et al. 2019). In the meantime, in a world of changing climates, Loft (2020) pleads for greener Earth system modeling that reduces the carbon footprint of weather and climate models; he highlights the need to "develop machine learning algorithms to avoid unnecessary computations." This statement resonates with the current paper and further applications of stochastic emulators, in particular since the authors mention in their paper that statistical models are especially suited for fine spatiotemporal scales.

Questions: Have the authors thought of other future and alternative uses of stochastic emulators for future computing in Earth systems modeling? How might one reduce the footprint of statistical inference in general?

ACKNOWLEDGEMENTS

We thank Franck Cappello from Argonne National Laboratory for helpful discussions and comments.

Funding Open access funding provided by National Renewable Energy Laboratory Library

[Accepted March 2023. Published Online May 2023.]

REFERENCES

Ailliot P, Monbet V (2012) Markov-switching autoregressive models for wind time series. Environ Model Softw 30:92–101

Ailliot P, Allard D, Monbet V, Naveau P (2015) Stochastic weather generators: an overview of weather type models. J Soc Fr Stat 156(1):101–113

Ailliot P, Bessac J, Monbet V, Pene F (2015) Non-homogeneous hidden Markov-switching models for wind time series. J Stat Plan Inference 160:75–88

Baker AH, Xu H, Hammerling DM, Li S, Clyne JP (2017) Toward a multi-method approach: lossy data compression for climate simulation data. In: Kunkel JM, Yokota R, Taufer M, Shalf J (eds) High Performance Computing. Springer International Publishing, Cham, pp 30–42. https://doi.org/10.1007/978-3-319-67630-2_3

Ballester-Ripoll R, Lindstrom P, Pajarola R (2020) TTHRESH: tensor compression for multidimensional visual data. IEEE Trans Vis Comput Graph 26(9):2891–2903. https://doi.org/10.1109/TVCG.2019.2904063

Bauer P, Quintino T, Wedi N, Bonanni A, Chrust M, Deconinck W, Diamantakis M, Düben P, English S, Flemming J, Gillies P, Hadade I, Hawkes J, Hawkins M, Iffrig O, Kühnlein C, Lange M, Lean P, Marsden O, Müller A, Saarinen S, Sarmany D, Sleigh M, Smart S, Smolarkiewicz D, annd Thiemert P, Tumolo G, Weihrauch C, Zanna C, Maciel P (2020) The ECMWF scalability programme: progress and plans. European Centre for Medium Range Weather Forecasts

Castruccio S, Hu Z, Sanderson B, Karspeck A, Hammerling D (2019) Reproducing internal variability with few ensemble runs. J Clim 32(24):8511–8522

Hu W, Castruccio S (2021) Approximating the internal variability of bias-corrected global temperature projections with spatial stochastic generators. J Clim 34(20):8409–8418

Katz RW (1977) Precipitation as a chain-dependent process. J Appl Meteorol 1962–1982:671–676

Kleiber W, Katz RW, Rajagopalan B (2012) Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. Water Resour Res 48(1):W01523. https://doi.org/10.1029/2011WR011105

Klöwer M, Razinger M, Dominguez JJ, Düben PD, Palmer TN (2021) Compressing atmospheric data into its real information content. Nat Comput Sci 1(11):713–724

Krasowska D, Bessac J, Calhoun J, Underwood R, Di S, Cappello F (2021) Exploring lossy compressibility through statistical correlations of scientific datasets. In: 7th International workshop on data analysis and reduction for big scientific data in conjunction with SC '21: the international conference for high performance computing, networking, storage and analysis, pp 47–53. https://arxiv.org/pdf/2111.13789.pdf

Li S, Clyne J (2022) Lossy scientific data compression with SPERR. Technical Report EGU22-946, Copernicus Meetings, March

Lindstrom P (2014) Fixed-rate compressed floating-point arrays. IEEE Trans Vis Comput Graph 20(12):2674–2683. https://doi.org/10.1109/TVCG.2014.2346458

Loft R (2020) Earth system modeling must become more energy efficient. EOS. https://doi.org/10.1029/2020EO147051

Parlange MB, Katz RW (2000) An extended version of the Richardson model for simulating daily weather variables. J Appl Meteorol 39(5):610–622

Richardson CW (1981) Stochastic simulation of daily precipitation, temperature, and solar radiation. Water Resour Res 17(1):182–190

Stein ML (2021) A parametric model for distributions with flexible behavior in both tails. Environmetrics 32(2):e2658. https://doi.org/10.1002/env.2658

Stein ML (2021) Parametric models for distributions when interest is in extremes with an application to daily temperature. Extremes 24(2):293–323. https://doi.org/10.1007/s10687-020-00378-z

Tencaliec P, Favre A-C, Naveau P, Prieur C, Nicolet G (2020) Flexible semiparametric generalized Pareto modeling of the entire range of rainfall amount. Environmetrics 31(2):e2582

Thompson CS, Thomson PJ, Zheng X (2007) Fitting a multisite daily rainfall model to New Zealand data. J Hydrol 340(1–2):25–39

Tintó Prims O, Acosta MC, Moore AM, Castrillo M, Serradell K, Cortés A, Doblas-Reyes FJ (2019) How to use mixed precision in ocean models: exploring a potential reduction of numerical precision in NEMO 4.0 and ROMS 3.6. Geosci Model Dev 12(7):3135–3148

Underwood R, Bessac J, Di S, Cappello F (2022a) Understanding the effects of modern compressors on the community earth science model. In: 8th International workshop on data analysis and reduction for big scientific data in conjunction with SC '22: the international conference for high performance computing, networking, storage and analysis

Underwood R, Calhoun JC, Di S, Apon A, Cappello F (2022) OptZConfig: efficient parallel optimization of lossy compression configuration. IEEE Trans Parallel Distrib Syst. https://doi.org/10.1109/TPDS.2022.3154096

Váňa F, Düben P, Lang S, Palmer T, Leutbecher M, Salmond D, Carver G (2017) Single precision in weather forecasting models: an evaluation with the IFS. Mon Weather Rev 145(2):495–502

Zender CS (2016) Bit Grooming: statistically accurate precision-preserving quantization with compression, evaluated in the netCDF operators (NCO, v4.4.8+). Geosci Model Dev 9(9):3199–3211. https://doi.org/10.5194/gmd-9-3199-2016

Zhao K, Di S, Dmitriev M, Tonellot T-LD, Chen Z, Cappello F (2021) Optimizing error-bounded lossy compression for scientific data by dynamic spline interpolation. In: 2021 IEEE 37th international conference on data engineering (ICDE). https://doi.org/10.1109/ICDE51399.2021.00145