Response to: Saving Storage in Climate Ensembles

Wesley S. Burr

1. INTRODUCTION

I read this article by Huang, Castruccio, Baker and Genton with great interest. Access to computation has changed our profession in incredible ways, but as anyone familiar with architectures would note, the rapid growth in compute power has not typically been matched by growth in storage technology and media size, e.g., exponential growth in compute, while growth in storage is linear at best. The central idea of this paper seems to be that approximations, done carefully, are sufficient proxies for high computational demand model results with the extremely high storage requirements that come with saving high-dimensional models. As preserving models for future comparison is important for reproducibility and replicability of results, especially in validation of climate models as used in policy, the approximation concept is both timely and possibly very important.

The authors do a very nice job of summarizing the current state of climate model storage limitations, and although only presented as an example using NCAR, the proposed plateauing of storage mentioned is almost certainly becoming a universal constraint across numerous agencies and academic centers. The days of various federal agencies archiving and providing access to data, data products and simulation results (e.g., every bit ever recorded and stored by the Voyager satellites takes up only a small amount of storage, and is freely available by FTP) must be coming to a limited end due to the growth in total size per product. Additionally, although not mentioned explicitly in the early sections (but discussed in the last section of the paper) current technology for unpowered storage simply cannot grow to the levels needed for long-term archiving and storage of large-scale simulations: Individual physical storage media are largely constrained in total storage per unit by well-understood

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

W. S. Burr (✉) Department of Mathematics, Trent University, Peterborough, ON, Canada (E-mail: *wesleyburr@trentu.ca*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 28, Number 2, Pages 349–351
https://doi.org/10.1007/s13253-023-00538-1

physics at this point, and while theoretically we may eventually gain access to DNA storage, this is (like fusion) always somehow "40 years off."

The statistical model example in Section 3 was quite interesting, and I was convinced by the presentation that fitting such a model to the ensemble of climate models can provide a suitable proxy for the trend and uncertainty in the end-of-century projections. Similarly, in Section 4, the regional data example was well formulated and clear. The two-step process did present one concern, namely the question of trend and temporal parameter estimation. The idea that trend and temporal dependence can be estimated with high precision may hold for broad features of time series, but representing a time series only by, say, a polynomial trend and an autocorrelation of limited lag will almost certainly lose (possibly subtle, but still) important information. I would be concerned that subtle characteristics of the temporal dependency in these time series could be abstracted away by the representation, resulting in simulations that superficially match the broad features of the original ensemble, but with unaccounted-for loss. This may not be an issue at the temporal resolution examined at the start, but if this overall technique is to prove integral to climate efforts in the future, increased temporal resolution (even the level of daily) brings increased temporal information, often only obtainable through careful examination of better temporal statistics than the autocorrelation.

Section 5 touches on this idea, of needing "more" than just the autocorrelation, by including harmonics for the mean a month-specific variance for the error. This is at least an acknowledgment of the issue, which is reassuring, although if this process is to be done in practice (with possible commensurate permanent loss of the original ensemble of inputs due to recovery of storage), I would suggest that more care be taken in determining the exact harmonics of interest through robust estimation of both frequency and phase Thomson (1990a,b). This is especially relevant once the approximation process begins to be applied to daily data, as many interesting features exist at this temporal scale in climate data, and presumably should be captured and represented by at least parts of the models, and thus need to be further captured by the approximation first stage, in order to feed the spatial stage.

The authors conclusion and discussion match my feelings on this issue: Indeed, it is not a matter of if, but when, storage simply runs out. We may be able to buy some time with "cold storage" and tapes, in order to preserve previous ensemble runs, but lacking a breakthrough in storage technology which produces Moore's law levels of accelerated density gains, our computes (eyes) are now firmly larger than our storages (stomachs). I also agree with the authors that expecting a small amount of compute to reconstruct simulations is a perfectly reasonable expectation. Allowing for generation of data as needed, with minimal storage needs, is truly an excellent work-around, provided that the generators can be validated, as the authors suggest is a necessary future step, in the development of diagnostic tools.

In conclusion, I thank the authors for their very interesting paper, and for giving me the chance to respond (in a somewhat rambling fashion) to its points. I look forward to seeing where these ideas go if they are taken up by the broad community!

[Accepted March 2023. Published Online May 2023.]

REFERENCES

Thomson David J (1990) Quadratic-inverse spectrum estimates: applications to palaeoclimatology. Philosoph Trans Royal. Soci London Series A: Phys Eng Sci 332(1627):539–597

Thomson David J (1990) Time series analysis of Holocene climate data. Philosoph Trans Royal. Soci London Series A, Math Phys Sci 330(1615):601–616