Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-124-RC1, 2020

Geoscientific Model Development Discussions

Interactive comment on "Adaptive lossy compression of climate model data based on hierarchical tensor with Adaptive-HGFDR (v1.0)" *by* Zhaoyuan Yu et al.

Anonymous Referee #1

Received and published: 1 July 2020

The authors describe Adaptive-HGFDR, a lossy compression method that adapts to the data such that error distributions are uniform. They apply the method to several atmospheric variables. The writing in the manuscript needs significant improvement. Also many details seem to have been left out (addressed in comments below), leaving the reader with many questions.

———————— Comments/suggestions/questions ————————-

-This manuscript contains many grammar issues and needs to be written better. Many phrases are awkward or do not make sense.

-Abstract: does "stable compression error" mean uniform distribution or error? This is

odd terminology. (Also in line 155 and 231)

-line 26: Not all ESMD data is high-precision. In fact, it is typical that calculations are done in double precision, but that data is output in single precision (e.g., for CESM). What is the precision of the data that you are compressing?

- line 30-32: "lossless compression has an upper limit of compression ratios" - I'm not clear what is meant by this. It really depends on the data so I don't know how you'd define an upper limit. Loseless compression works quite well on smooth data, for example. Also I don't get "which grows much slower than the velocity of data volume grows". Is the upper limit growing? What is meant by the velocity of the data volume?

-line 38: I'm not sure that I agree with "the error cannot be controlled in a data-driven way". I would say that the SZ method in Liang at al 2019 ("Significantly Improving Lossy Compression Quality Based on an Optimized Hybrid Prediction Model") is data driven.

-line 39: what is meant by "the data compression parameters of common data files are relatively fixed". Different variables within a file could be compressed different amounts.

-the distinction between file-based and encoding-based compression is not clear to me. Are they meant to be mutually exclusive? Why can't one use an encoding based method on the entire file. Please explain this more clearly.

- the references need to be carefully checked. There are a lot of errors! Many are incomplete due to missing information (page numbers, journal names, ...) Also I noticed incomplete/incorrect author lists, multiple entries for the same article, ...

-line 60: "intercepting the floating-point precision" does not make sense

-lines 61-62: "As the distribution of floating-point precision of data is not uniform, the compression errors may also distribute unevenly. Therefore, it is difficult to control the distribution of data compression error." This statement needs to be clarified (it is not true for all data).

-line 66: does zfp really use "feature prediction -based" encoding? It is a transform method - is that what you mean? Please clarify what is meant by "feature prediction-based encoding" ...

-Intro: consider including SZ compression - it's very popular

-line 74: "values in neighboring ranges tend to be numerically close to each other" This is not true for all data in ESMD - a number of variables have abrupt changes (e.g., clouds).

-section 2: "the partial block data is not only much smaller than the original geographic spatial tensor"- you need to explain what is meant by the partial block data. The discussion just starts talking about blocks, which can mean a lot of different things, and we don't want to assume the reader is familiar with previous work.

-Figure 1: "Judging the rank meets the error" is awkward...

-Section 2 was not that useful in terms of getting the big picture. It should be written as a more general overview as title implies (Basic Overview) - so avoid (or define) undefined terms (e.g., the fast search method, block of divided data, target error, ...) Consider beginning with explaining the spatial tensor as in the flow chart.

-line 139: what is meant by making the blocks an "ideal size"?

-Please motivate why uniform error distribution is important? In some cases, there may be a need for more accuracy in some regions than others.

-line 163: It's not clear what alpha and beta are...

-Section 4.1.: what model is this data from? More specific info on the data is needed (or refer to the section at the end), and in line 189: this reference does not make sense for NetCDF: Springer, U. S.: Community Earth System Model (CESM), Encycl. Parallel Comput., 351, 2011. (Also I'm not sure what this is referring to). Even in the data availability section is doesn't say what model was used. Also is there a doi for the data

or how do I find the data on Data Cloud? (Now I see that this is CESM data - it's only written in the abstract. Also why the choice of data from 2013?)

-Figure 4; can you better explain how the block size is affecting the compression ratio (resulting in this v-shape)? Is this behavior "typical" or expected?

-section 4.2 - how does the block number relate to the block size? please clarify the distinction.

-section 4.2 - is 1e-4 relative error or absolute error? Also is this a max error or an average error (e.g., rmse)?

-Figure 3: The caption is on a different page than the figure. Also it is hard to see what is going on here. Consider plotting the errors instead. (I assume you are plotting temperature in K, though it doesn't say that).

-section 4.3: How is zfp being applied to the data? It's effectiveness quite depends on the spatial locality. Also why did you choose zfp?

-section 4.3: Why would the compression ratio of the HGFDR go up so much with a change in error tolerance from 1e-4 to 5e-4? This needs to be explained as it is hard to believe.

-I don't quite understand what is being plotted in figure 6. But the regular pattern in zfp error is likely explained by its block size. (e.g. Hammerling et al 2019 "A Collaborative Effort to Improve Lossy Compression Methods for Climate Data"). It would be helpful to discuss this. ZFP block patterns are also evident in figure 7a.

-section 4.4: Why did you pick .01 for the error limit? Is this a relative error? Errors limits were smaller in the previous section.

-section 4.4: Why did you pick those particular variables? Do they have a good representation of the different types? For example, temperature variables are "easy" to compress as compared to variables with discontinuities and large dynamic ranges (like

precipitation). So having a variety of test variables is important. There are hundreds of variables in the CESM atmospheric component.

-line 249: Please clarify: "maintains the maximum compression ratio under the constraints of the same compression error"

-line 251: "removes data redundancy" - please clarify what you mean (all compression methods are trying to model the data so as to remove redundancy).

-section 4.4: More discussion is needed to explain why the compression ratios vary so much in figure 8. Some of these CRs are very high and I question what the error looks like. The .01 threshold is large, but a 600x reduction is pretty shocking really. Also instead of sharing the std dev - shouldn't we see some sort of average or max error instead?

——————————- Minor items: ——————————-

- "Earth" should be capitalized everywhere in "Earth system model"

- the ZFP method should be referenced by Lindstrom 2014 (line 65)

- line 88, 97: What is the citation for the first HGFDR work? (I think it's Yuan 2015, but you have not made that clear here.)

-line 171: compression is misspelled....

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-124, 2020.