Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-146-RC1, 2016

Geoscientific
Model Development
Discussions
Open Access

Interactive comment on "Evaluating Lossy Data Compression on Climate Simulation Data within a Large Ensemble" *by* Allison H. Baker et al.

Anonymous Referee #1

Received and published: 12 September 2016

The topic of the paper is of high importance for data intensive computational sciences, in particular here for climate research. It discusses the possibility to use lossy compression for data volume reduction without harming the quality of scientific validity.

Lossy compression achieves higher compression rates than lossless compression and thus allows storage costs (investment and operational) to be reduced. For a single compute center that concentrates on weather and climate computations these savings will be a two-digit million USD amount with every new computer generation.

The paper written by a group of authers investigates the question to what extent lossyly compressed data can be detected in the subsequent steps of a scientific workflow. It does not discuss the question whether scientific results will be invalid.

The main idea is to introduce new data in ensemble computations where two new

ensemble members went through lossy compression and reconstruction. The study looks into whether these two ensemble members can be identified.

The authors group investigates 7 ways of how to potentially detect the new members that went through lossy compression. All 7 ways are different and sophisticated, thus a single reviewer will not be able to check the validity of each of the seven analyses.

In summary the authors present 4 lessons learned from the analyses. So, lossy compression seems to be possible and beneficial when certain conditions are met and the approach follows certain rules.

The paper presents a new research strategy for a highly important and exciting question. So there are several questions to the authors from a theory of science point of view and from a more meta science perspective.

1. When reading the paper it was not clear that the individual analyses are conducted by different authors who are specialists in their fields. This should be highlighted. Did they know of each other? Where they confronted with the research question independently? Please explain a bit the methodology of how the cooperation went.

2. Why exactly these 7 analyses? Do they somehow cover in a representative way what is done with the data or can be done with the data? Are there further analyses that should be added? Perhaps add some text at the end of Chapter 3 and describe the methodology of your approach (also for question 1).

3. There is not really a related works section with respect to this particular research. I assume in fact that there is not much related work. So: who of the community is looking into lossy compression and its effects on scientific validity of results? At least the GRIB people might do that? Any comparable efforts at other centers like NOOA, ECMWF, MetOffice, etc.? Please report. If there is no related effort, please confirm that in your paper.

4. What about the state-of-the-art with respect to algorithms? Google queries with

"lossy compression of climate data" and "lossy compression of medical image data" show that others also conduct research here. Lossy compression might be an issue in several other areas. Please give some details.

5. The analyses are a mixture of mathematical and visual approach: at first you apply a mathematical operation onto the data, then visualise it, then say that the effects of compression are discernable or not. Sometimes you mean: with the naked eye when looking at the charts? Is this a methodologically correct approach? Discernable depends on the person who looks at it.

6. fpzip seems to be old? Are there other algorithms with different approaches. Just like as for audio files there is quite some progress with lossy compression.

I completely agree with your conclusion at the end of page 4. Unfortunately, this is not clear to all researchers! You need to repeat it whenever there is a chance for it.

Now as we see, that lossy compression should be possible but is technically complicated because you would e.g. have to inspect all variables and decide upon what to do with them the questions are:

7. Should we proceed with looking into lossy compression as the advantage over lossless might only be a factor of 3 and with lossless there is no further problem?

8. What will be the extra costs in order to support lossy compression correctly with respect to human resources for e.g. variables analysis? Also there might be costs for additional hardware to do the compression efficiently and for power to operate this hardware. Of course, the benefits will be easier to quantify. However, please make a comment on the potential cost-benefit-ratio of introducing lossy compression into the science workflow.

A technical question to table 2:

9. Who/what defined the level that was used for each variable. Based on what consideration? Was this explained?

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-146, 2016.