Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-177-RC1, 2016

Geoscientific Model Development Discussions

Interactive comment on "Finding the Goldilocks zone: Compression-error trade-off for large gridded datasets" *by* Jeremy D. Silver and Charles S. Zender

Anonymous Referee #1

Received and published: 22 August 2016

———————- Summary: ———————-

The variation in a gridded dataset may be notably different in each spatial direction. Taking this into account when applying (lossy) compression can improve the resulting precision. They propose a technique called layer packing that achieves better compression ratios than a lossless approach and preserves precision better than a comparable lossy approach.

———————- General comments: ———————-

(1) This paper addresses an important issue because data compression is very much needed to mitigate large data volumes in geophysical data.

(2) Treating the dimensions differently when applying lossy compression to gridded data makes a lot of sense.

(3) Section 1 and 2 need some rearranging and improvement (more details are given below in "specific comments") in terms of introducing the ideas and terminology. It could be better to shorten the introduction and then really explain the methods well in section 2.

(4) The audience for this work may not be too familiar with compression techniques other than just using defaults in netcdf, so improving the explanations for the techniques would be helpful. (For example, defining a "deflate and shuffle" algorithm).

(5) The paper's contribution should be clarified in the introduction (section 1). It is not clear to me whether "layer packing" is a new idea that is first presented here. (It is mentioned a bit more clearly in section 3).

(6) For this paper to really impact the broader geophysical data community, I feel that more details on the compression approaches need to be provided.

(7) More details on the datasets are needed to be able to understand why compression effects the each differently. Perhaps look at variables instead of multi-variable datasets?

———————- Specific comments: ———————-

(1) page 2, par. 1: For this audience, please give more explanation of the techniques. For example, please provide more explanation of how "deflate and shuffle" works (rather than just pointing to a reference).

(2) page 2, line 22: "Linear packing with a single scale-offset parameter" - is discussed here but not well-defined. Note that "packing" is later defined in line 32. Then "scalar linear packing" on p.3. line 2. In general, the terminology used and defined in this paragraph is hard to follow in that it is sometimes defined after being used. (Also, is "linear packing with a single scale-offset parameter" the same as "scalar linear packing"?)

(3) p.2, line 29: I'm not sure the audience will be familiar with "quantization" (like the audience for a CS publication would).

(4) section 2.1.1 (Layer packing)" Here I would suggest providing more detail (maybe an example) - particularly if this approach is the main contribution of the paper. Rather than providing syntax details, consider defining/explaining the parameters (the reader may not be familiar with what these are) here.

(5) section 2.1.2, line 15: Explain what "level" means in the algorithm.

(6) section 2.1.2, line 17: Explain a shuffle filter.

(7) section 2.3: Regarding the datasets listed, more information about the model source (other than acronym and reference) would be helpful - especially in interpreting the results. Without more details, I cannot really understand how the datasets differ and, therefore, why/how they would respond to compression differently. For example, the number of grid points are given - but does this number represent a domain on the entire globe for all datasets? The number of vertical levels is listed, but do all models simulate to the same height? What is the time dimension? Hourly? Monthly averages? Is the time dimension the same for each data et?

(8) Fig 1: For compression results, I think it would be more intuitive/standard to compare to the uncompressed size (and have all ratios below 1.0). Also I don't understand the meaning of the comp./decomp. time in the left panel for uncompressed data.

(9) page 6, line 30: The paper could be much stronger with specific examples of individual variables and how affected by compression approach and choice of metric (e.g. by std. dev. or mean normalization). Since all results are averaged across datasets, this information is not available.

(10) Section 3: This section contains some useful information (and examples) about linear scaling and layer packing that would have been good to explain earlier in the paper when the concepts/algorithms are first introduced (and before the results are

given).

(11) More related lossy compression work on geophysical data should be mentioned for better context, for example:

Hubbe, Wegener et al., ISC '13 (http://link.springer.com/chapter/10.1007%2F978-3-642-38750-0_26)

Baker, et al., HPDC '14 (http://dl.acm.org/citation.cfm?id=2600217)

Woodring et al., LDAV '11 (http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6092314&ta

(12) Other competitive lossy compression algorithms for scientific data should probably be mentioned as many may be affected by differences in the variation across spatial dimensions for gridded data - this could be really interesting. Also many lossy compression methods for scientific data could eventually by incorporated into netcdf.

(12) Fig. 2: Because the differences between the datasets are not more thoroughly addressed, then it's unclear what conclusion to draw by comparing the SD and mean normalizations in Figure 2 (e.g., what is the takeaway point?). Basically, it seems that the two plots are quantitatively similar enough that both should be included only to illustrate a point, which I am not seeing. Can you clarify?

(13) fig 3: Same comment as above, plus I am not sure what conclusion to draw given that some datasets compress better than others without a more clear understanding of dataset differences. I think looking at individual variables, rather than entire datasets would make it easier for the reader to understand the differences in the approaches.

———————— Final thoughts ————————-

I like the idea of treating spatial dimensions differently with lossy compression, and I think the authors could have really taken off with this concept and it explored it much more thoroughly. I question whether the contributions in this particular version are significant enough for a GMD paper.