Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-177-AC1, 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

Jeremy D. Silver and Charles S. Zender

jeremy.silver@unimelb.edu.au

Received and published: 28 October 2016

We wish to thank the reviewers to taking the time to read the manuscript and provide feedback. We note that we have taken the challenge of major revision seriously and reworked the analysis to a much more fine-grained level, included a range of new and interesting results, remade all the figures, and restructured and rewritten much of the text. We believe that the reviewers' comments have helped to improve the manuscript and strengthen our findings.

Please see the other replies which include the revised manuscript and a summary of changes.

1 Reviewer 1

1.1 General comments

1. *This paper addresses an important issue because data compression is very much needed to mitigate large data volumes in geophysical data. Treating the dimensions differently when applying lossy compression to gridded data makes a lot of sense.*

   We agree.

2. *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.*

   We have rearranged material in these sections given the feedback provided.

3. *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).*

   We have provided additional details as suggested.

4. *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).*

   Layer packing per say is not a new idea, and is the foundation for compression in the GRIB data format. However the idea of layer-packing is generalised here

beyond two-dimensional slices. The work presented here is a test-of-concept for combining some of the better aspects of both GRIB and netCDF/HDF5 formats. The introduction and discussion reiterate these points.

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

   We have provided more details as recommended.

6. *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?*

   This is an excellent suggestion and one that we have adopted. One of the main changes to the manuscript between the initial submission and this revision is that we examine compression in a variable-by-variable approach rather than as a whole-dataset approach. This allows us to look at individual variables in terms of their compressibility, the "complexity" of the variable and error resulting from the lossy compression; this fine-grained approach allows for greater insight and a much larger sample size. As such the results section has been heavily revised.

1.2 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).*

   We have introduced additional detail about these methods as recommended.

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"?)

   We have reviewed how the notation is introduced in order to improve readability.

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

   This has been clarified

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.*

   In hindsight we agree that details about the algorithm itself are required, rather than syntax. We have moved the syntax to a supplementary section. The algorithm itself is outlined in the methods section.

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

   This has been explained.

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

   We have added additional details.

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 set?

The original description of these datasets was deliberately kept short, as this was not the main focus of the paper. We have compromised by abbreviating the description of the datasets to a table and moving the full descriptions of these datasets to the Supplementary Material section.
Regarding the question about why variables respond differently to compression, we believe that this has been solidly addressed in the analysis of the entropy of the data and exponent fields, which was made possibly by following the suggestion to shift the focus of the paper from compressing entire datasets to compressing individual variables.

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.*

   The compression ratios are now defined in terms of the uncompressed size as suggested, and we have also moved to a more standard definition of the compression ratio (i.e. uncompressed size / compressed size, so that larger values represent greater compression). The compression times represent the time taken from the original data to the compressed file, whereas the decompression time is to unpack the layer-packed data. This has been clarified

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.

We agree and we have adopted the variable-level rather than dataset-level approach. We included examples of six variables (among a total of 255) in the Supplementary Material document as illustrations of the errors induced by the six lossy compression methods considered.

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).*

    We have given additional details about linear scaling and layer packing in the Methods section. Additional examples for illustrative variables appear in the Results section.

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&tag=1)*

    We have given more details about related lossy compression work in this field. We thank the reviewer for the suggested citations and have included some in the manuscript.

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.*

We have expanded the discussion to refer to other lossy compression algorithms for scientific data, formats beyond netCDF (e.g. based on image- and video-compression).

13. *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?*

    Both plots were included in order avoid the perception of a biased interpretation of the results. Normalization by the SD or the mean advantages one method or the other, however the conclusions are the same regardless of the normalization. We agree that including both plots does not add much value to the paper. We note that all the figures have been completely reworked.

14. *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.*

    As noted previously, we agree with the reviewer's comment and have redone the analysis to examine variables separately, rather than groups of variables clustered together as datasets.

1.3 Final thoughts

1. *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.

The purpose of this study was to test the concept of layer-packing, in an attempt to combine some of the best aspects of the GRIB and netCDF/HDF5 data formats. We acknowledge that the results have not been conclusively in favour of the layer-packing with respect to bit-grooming, however we would argue that this is worth publishing all the same. This partly relates to the discussion of publishing "positive" versus "negative" results; if only "positive" findings are published, this will result in a great deal of time and effort being wasted within the scientific community in repeating superficially appealing experiments. As such, transferring this knowledge to the public domain has value. The geoscientific modelling and measurement community (e.g. the volume of data generated by satellite retrievals) relies heavily on these data formats, and it is important that their refinement is an ongoing process.

Regardless of any ambiguity between the choice of bit-grooming or layer-packing, one clear result from this study is that simple linear packing typically results in *much* greater loss of precision than either of the two lossy methods discussed here. This is despite its widespread use.

Other useful contributions include the focus on the error-compression trade-off, the finding that the normalized entropy of the exponent field can be used to help determine which compression method is most appropriate, and the idea (introduced in the discussion) that the changes in the normalized entropy of the data could be used to determine how many significant figures should be retained.