Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-177-AC4, 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 3

1.1 Summary

The paper introduces a "layer packing" lossy compression technique that takes advantage of the minimal horizontal variations in geoscience data relative to the larger variations across vertical dimensions. The layer packing technique is compared against many widely used lossless and lossy compression techniques and evaluated based on accuracy and time to solution. Layer packing is found to be beneficial in some cases while not in others, leading to the conclusion that care must be taken to evaluate whether lossy compression is worth the risk.

1.2 General Comments

1. *The paper makes a good first attempt to evaluate the layer packing technique, but the paper would benefit from an additional revision. First, it's not clear what the paper is contributing. The authors state that the technique is used in GRIB (page 7, section 3) but that the evaluation was not possible due to relative error not being reported. Since the technique is not new, then the only contributions of the paper are the announcement of the general availability of the new non-GRIB tools, as well as the modestly detailed evaluation of the many compression techniques.*

   The geoscientific modelling and remote-sensing community has to deal with the ever-growing volume of data generated. As such, it is important that the storage methods are reviewed in terms of the trade-off between compression, error and read/write times.
   We have tried to avoid debate about data formats. Both have an important roles; the geoscientific community relies heavily on netCDF/HDF5, and GRIB remains

the format of choice in many operational meteorological centres. Despite its excellent compression performance, GRIB can be regarded as less user-friendly.
The GRIB layer-packing is restricted to two-dimensional slices, whereas the layer-packing described here can operate on arbitrary hyperslices. The work presented in this manuscript aims to generate discussion about ways of incorporating the best of both methods.
With reference to the comment from Page 7, Section 3: "Caron (2014) estimated that GRIB2 files are on average 44% of the size of the equivalent deflate-compressed netCDF-4 files (n.b. relative errors were not reported, which limits the comparison)". The intended meaning was that the study of Caron (2014) reported the compression ratio, but not the relative errors, which makes it difficult to place the Caron (2014) results with those of this study.
The revisions to the original manuscript, focusing the analysis on the compressibility, errors and complexity of individual variables offers additional insights into these relationship and we believe adds substantially to the value of the paper.

1.3 Specific Comments and Technical Corrections

1. *The title, though catchy, is overloading the term "Goldilocks Zone" – the region around a star where perhaps liquid water might be found on a planet's surface. The title after the colon is clear on its own.*

   We have abbreviated the title, which as already been through several iterations.

2. *Page 2, line 3: "NetCDF" starts the last sentence on the line, though it should be "netCDF" for consistency.*

   We have revised for consistency of this term.

3. *Page 2, line 5: Why are three references necessary to describe the "deflate"*

   compression method? Throughout the paper, be consistent with terms. scale-offset vs scale and offset. linear-packing vs linear packing.

   Additional description of the deflate and shuffle algorithms has been added as suggested by Reviewer 1. We have reconsidered the references in this section. We have also reviewed the usage of the terms mentioned to improve the consistency of the manuscript.

4. *Page 3, line 30: I would suggest adding that ncdump is a command-line utility from the netCDF package because it might not be common knowledge. The paper introduces the "ncpacklayer" program and also uses other "nc"-prefixed tools from the NCO suite. For example, perhaps the following: "...(following the output format for the netCDF command-line utility ncdump)..."*

   Yes, this is correct, thanks for pointing this out. We have clarified this point.

5. *Page 3 (section 2 in general): More detail could be spent on the layer packing technique itself; the many monospaced examples of section 2 don't substantially add to the narrative and instead come across like a tutorial or README.*

   We have expanded the description of the algorithm itself. To keep the article short and concise, we have moved these details to an appendix.

6. *Page 4, line 11: run-on sentence*

   Thanks for pointing this out. This has been corrected.

7. *Page 4: The dollar symbol "$" is not explained, though I think you meant for it to refer to a shell variable syntax.*

   Yes, this is correct. This has been clarified.

8. *Page 5, Section 2.3: If I do the math correctly, the size of the datasets are (1) 962MB, (2) 267MB, (3) 68MB, (4) 613MB, (5) 30MB, and (6) 717MB. The rationale for the proposed compression is the growing volume of data in the geosciences, though none of these datasets are over a gigabyte in size. Compression of a multi-gigabyte dataset would make the argument more compelling, because datasets of such size will become more commonplace. Writing large datasets to disk as they are computed is a challenging problem and it would be nice to evaluate whether compressing large datasets is a viable option as they are generated. General comment about all Figures: Consider labeling the left and right panes of each figure as (a) and (b). For example, page 6, paragraphs starting on lines 9 and 17 sound too similar since Figure 1 is showing different things but is referred to in the text in the same way. It would be more clear to say something like "Figure 1A shows..." and "Figure 1B presents..."*

   The point about the magnitude of the file size is quite reasonable. We ran the test suite on variable of size 1.5 GB to examine the performance of the methods on larger datasets. This was included as an example referenced in the timing results, rather than adding it to the suite of variables presented in all results. This was mainly because, in the process of setting it up, the test suite was run many dozens of times; to accelerate the testing the variables considered were kept relatively small (the largest was about 65 MB).
   However by the same token, the analysis for the revised manuscript has been done on individual variables alone, so the basic unit of study has become much smaller. While this might not impress those working with terabyte-scale data, it allows for greater insight into the methodology itself.
   Regarding the figures, some of these have been moved to a supplementary material section. All panel plots now have labels (a), (b), etc., as suggested.

9. *Page 7: Starting on this page, for some reason all references to "figure 3" are lower case.*

   Thanks for pointing this out – it has been fixed.

10. *Page 8: Figure 1: The red and orange colors are too similar, though their position is clear from the legend.*

    All the figures have been thoroughly reworked. The color scheme in question no longer appears.

11. *Page 8, Figure 1, right panel: What does it mean to have the first column as "uncompressed" time since everything is normalized to DEFLATE? Was it the time to generate the data? Was it the time to copy the file?*

    Yes, in hindsight this wasn't very clear. It was effectively the time to copy the data. This bar is not included it in the revised manuscript. Thanks for drawing attention to it.

12. *Page 8, line 4: The reference to the HDF Group is used as an in-text citation as "(Group, 2016)". It would be best to fix your citation to not use HDF Group as a first/last name pair. See also your references on page 13, line 17.*

    Thanks for pointing this out. The default behaviour of the reference manager should have been over-ruled. This has been corrected.

13. *Page 9, line 1: run-on sentence*

    Thanks for pointing this out. It has been corrected.

14. *Page 10, Figure 3 caption: capitalize the Figure 1 and Figure 2 references.*

    This has been made more consistent.

15. *Page 11, line 6: misspelled "considered" – please consider a full spell check.*

This has been fixed and we will ensure to run the spell checker again before resubmitting.

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-177, 2016.