Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-177-AC5, 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 find attached a PDF document that collates the revised draft, the new supplementary material document and the full point-by-point reply to the reviewers.

Main changes

- Compression, errors and complexity are assessed at the variable-level, rather than the dataset-level (i.e. for a number of variables combined).
- We calculated a range of statistics on the individual variables, in order to improve our understanding of why certain variables compress well with one method or another.
- Some material was moved to a supplementary document.
- The introduction has been abbreviated as recommended.
- The Methods section was expanded to provide a clearer description of the layer-packing method.
- The discussion includes a brief review of related work.
- Additional description of the deflate and shuffle compression algorithms were added to the Methods section.
- All figures have been reworked.

Minor changes

- Variables are now chunked in a consistent manner for the different methods to improve comparability across compression methods.
- A minor error was found and corrected in the calculation of file sizes. The differences would have been very minor for the results in the original manuscript, since the file sizes were much larger than when doing the analysis on individual variables, but became apparent when working with the single-variable data files.

The error was that the results were calculated based on “resident” rather than “actual” file size.

- Minor improvements were made to the layer-packing code, resulting in more stable treatment of non-finite values, avoiding rare cases of floating-point overflow, and more stable handling of dimensions.
- We ran the test suite on a 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).

Please also note the supplement to this comment:
http://www.geosci-model-dev-discuss.net/gmd-2016-177/gmd-2016-177-AC5-supplement.pdf

---

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