Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-177-AC2, 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 2

1.1 General comments

1. *This paper describes a variant of lossy encoding which leverages the multi-dimensional nature of many scientific datasets that have greater data variances along different axes. The axes with small variations in data values are labeled "thin dimensions" and the axes with large variations in values are labeled "thick dimensions". The datasets are then "layer packed" with a linear scaling algorithm in the thin dimensions, recording a scale & offset value for each coordinate in the thick dimension.*
   I think the insights into the "thick" and "thin" dimensions are the primary value of this paper, with the actual compression algorithm and results being less important, overall.

   Yes – one of the main things we are trying to do here is to assess whether treating different dimensions differently during gives much benefit over and above other methods that can be easily applied to such datasets. This is essentially trying to combine the best elements of GRIB and netCDF/HDF5.

1.2 Specific comments

1. *Applying the idea of thick & thin dimensions appropriately to other compression methods (such as the JPEG-2000 algorithm used in GRIB2) would be more valuable than just the idea of the simple scale & offset compression chosen.*

   We agree, and we spent a large amount of time trying to get this to work while preparing these revisions.
   In revising this work, we were able to run (after many technical hiccups) the same

set of tests for GRIB/JPEG2000 compression as well (using 8, 12, 16 and 20 bits to represent the data). Our preliminary results showed that the JPEG2000 algorithm yields greater compression compared to the methods presented here for the same level of error; this echoes the findings of Caron (2014, www.ecmwf.int/sites/default/files/elibrary/2014/13711-converting-grib-netcdf-4.pdf), which describe the efficient compression achieved. However like bit-grooming or layer-packing, JPEG2000 does not offer clear controls about the resultant errors and thus some experimentation (in setting the number of bits per value) is needed to avoid excessive loss of precision. We found that there was a large spread in the magnitude of the relative errors compared to the other methods considered.
However the technical challenges required to convert a general netCDF field into GRIB format to be far in excess of what may be recommended to the average practitioner of geoscientific modelling. For this reason, and for the large spread in the compression and error result in the GRIB-compressed fields, and in order to keep the manuscript as focussed as possible, we chose not to include these results.

2. *Near the bottom of page 6, "for simplicity will have" should be corrected to "for simplicity we have".*

   Yes, well spotted. We have fixed this.

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