Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-177-RC4, 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 #3

Received and published: 1 September 2016

—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.

—General Comments—

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.

–Specific Comments and Technical Corrections—

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.

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

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.

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)..."

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.

Page 4, line 11: run-on sentence

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

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..."

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

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

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?

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.

Page 9, line 1: run-on sentence

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

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

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