Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-124-RC2, 2020

Geoscientific Model Development Discussions

Interactive comment on "Adaptive lossy compression of climate model data based on hierarchical tensor with Adaptive-HGFDR (v1.0)" *by* Zhaoyuan Yu et al.

Anonymous Referee #2

Received and published: 27 July 2020

Summary:

The authors present a tensor-based lossy compression method that is based on HGFDR, which compressed netCDF files across all variables, rather than a single variable at a time which many other methods use. They cite most of the important references and compare their method to some state-of-the art methods. The main idea of this work is a good one and in parts, it's described in sufficient technical depth. However, the authors assume too much prior knowledge of the basics of this techniques, which should be introduced and defined clearly. This technique produced promising results, but the analysis of the results lacks some depth. Overall, the paper definitely

needs some work on clarity of writing, formatting, wording, and typos.

____ Detailed comments:

Missing definitions: - line 106: Explain the concept behind block data/partial block data. This is not common knowledge and it is an important basis of your method. - line 121: Explain what you mean by dichotomy (a simple definition will do for this one) - Section 4: define terms you use. What do you mean by slice, height, block number?

Clarify in writing: - Lines 23-27 are phrased in a somewhat confusing way (especially step 1). Turning this into a list may be helpful, and the Figure could be tied in more efficiently by referring to specific part of the flow chart. - Most of the Figures have insufficient captions. Please provide enough caption that the take-away message of each Figure is clear. E.g. which method performs best? Why should the reader care? - Figure 1: This flow chart seems incomplete. The iteration over different versions of the compression method is not represented adequately (e.g. there are multiple iterations until an optimal rank is found but the chart implies it's a single step) - lines 175-180: the list format is good but the style is inconsistent. Consider leading each row with a verb, and provide a natural language description of the steps which only have a formula. - line 181: O(log n) is claimed but is missing a justification (or proof). - line 189: If you only provide a single variable, put in a reference to Table 1. Furthermore, please explain why you chose this variable. A better way to put this may be to introduce the data as "this is a tensor with 23 attributes (full list later on in Table 1)." - line 191-192: what's the average memory occupancy/usage (not occupation!) of each variable? This would be much more valuable information than that of a single variable. - Section 4.2 and later: consider using "block count" instead of "block number", as the latter could also be a number (index) assigned to a single block. - line 204: these numbers are very likely not random... instead of claiming randomness, it would be better to provide a reason for these choices. Are you trying to look at different orders of magnitude? Furthermore, 256 is missing here although it is present in the picture and later descriptions - Figure 3/line 213: this may also be due to the colormap as the one chosen for this

Figure has very little color depth. The "hot" colormap in the same color family would provide better differentiation between values. The "viridis" colormap would be another good choice. - Figure 5: what type of error is this? Also, the scale on the x-axis is very hard to read. A side by side comparison with a more consistent scale may be helpful. Furthermore, readers who are not familiar with compression will appreciate some guidance on reading compression vs error charts. - line 228/229: This sentence does not make sense - line 228-230: This is not enough analysis for this chart. - line 232: again, not random. Why these numbers and not different ones? - Figure 6/analysis: blocked-HGFDR performs substantially better for a lot of slice numbers and despite some bigger changes, the error seems largely consistent for adjacent slices. Why is that? And since this method builds on blocked-HGFDR, why does this not happen for adaptive HGFDR? - Figure 6: Are those the numbers of the slices, or the count of slices overall? - Figure 7: What is "height"? Is that the number of layers? variables? What are the differences between different heights? How are you sub-selecting these layers/variables? - Figure 7: What type of error is this? - Figure 7: Color may be helpful. - line 243: "error limit" should probably be threshold"? Which type of error are you looking at? - Figure 8: insufficient discussion of compression time – why is this algorithm so much slower than the competition? Especially so much slower that Blocked-HGFDR which barely shows up on the chart and which provides similar compression ratios

Open questions: - Can this method work in situ with the simulation? Why/why not? - Under which conditions does the proposed algorithm perform better? Under which conditions are different algorithms better? - Can this algorithm be applied to other types of data? Why/why not?

References: - Overall, the authors provide a good overview of general compression work as well as some more domain-specific examples. - However, the formatting of these references needs some work. There is a lot of information that's missing, including author names (Anon, "Of and Acm", "None", "Matrices",) - I'm not sure if all references are summarized and categorized appropriately in the text. They're grouped

a bit weirdly

Minor issues: - various references are missing a space before the parenthesis (line 25, 28, and a couple of others) - line 53: vector quantization(VQ) (Vector Quantization) – remove duplicate - line 69: performance of the compression may *be* low - line 133: "dimenisonal" should be "dimensional" - line 137: missing space before Thus - Eq 2 and several other places: weird formatting of "res" – please use consistent font type - line 162: align alpha and beta with text - line 170: "conpression" should be "compression - line 172 and other places: remove space in "R_M ax" and "R_M in" - line 193: "in a MATLAB environment"? I would assume they were performed *by* the authors. - Figure 3: caption is not on the same page as the Figure. - Figure 6: "longitudanal" should be "longitudinal" - line 243: dimension should probably be resolution - Table 1: the alignment here is very confusing. Variable names are aligned to the bottom and variable descriptions are aligned to the top. This is very awkward to read. - line 253: replace square brackets with parentheses - line 254: "We pay more attention" should probably read "The main focus of this work"... - line 265: Missing "The" at the beginning of the sentence - line 266: "maintain" should be "maintains"

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-124, 2020.