Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2022-82-RC1, 2022

Comment on gmd-2022-82

Anonymous Referee #1

Referee comment on "The impact of altering emission data precision on compression efficiency and accuracy of simulations of the community multiscale air quality model" by Michael S. Walters and David C. Wong, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-82-RC1, 2022

"The Impact of Altering Emission Data Precision on Compression Efficiency and Accuracy of Simulations of the Community Multiscale Air Quality Model" by Michael S. Walters and David C. Wong

GENERAL COMMENTS

Air quality models are a vital tool for air quality research and management, but these models require the use of large input data sets and they generate even larger output data sets. Moreover, the size of these data sets is expected to grow with time, so the management and archival of such large data sets is an ongoing challenge for the air quality modeling community.

This paper describes a new and useful approach to this issue: an overall "lossy" compression algorithm that reduces the size of both input and output files while preserving their important features. This is done by combining a "lossy" precision-reduction conditioning of such data sets followed by lossless data compression. The paper describes this algorithm and then assesses the costs and benefits of this new data-set management approach in terms of disk space savings, model run times, and model accuracy. Both model-to-observations and model-to-model comparisons are considered for a 2016 annual simulation with the CMAQ air quality model along with a comparison of the efficacy of two well-known lossless data compression utilities.

I found this to be a well-structured and reasonably well-written paper that would be suitable for publication in GMD. I recommend its acceptance after a number of minor revisions, including some to improve its clarity. To this end I have made a number of specific comments and suggestions below that I believe will improve the final version and that I hope the authors will consider.

SPECIFIC COMMENTS

1. The paper considers both input emissions files and CMAQ output files but the title only references input files. Should the title be expanded slightly along the lines of "The Impact of Altering Emission **and Output** Data Precision on ..."?

2. Terminology matters, and the manuscript is not always clear about exactly what is being done. The word "altered" is used a lot (32 times), but using a compound modifier like "altered-precision" or "reduced-precision" instead might be clearer. The terms "pre-processed" and "post-processed" are also used (lines 24, 25, 28, 105, 110, 117, 167) in what seems to me to be a confusing way, in that "pre-processing" is used in connection with emission files, which are input files upstream of the CMAQ model, whereas "post-processing" is used for CMAQ-generated output files. However, emission files are also output files that are generated by an emissions processing system like SMOKE, and for both types of files the precision reduction is applied **after** the files have been generated. Again, referring to "reduced-precision emission files" and "reduced-precision CMAQ output files" might make it clearer that a transformation or conditioning step, namely precision reduction, has been applied to files after they have been generated.

3. It would be helpful to the reader if a bit more detail were given in the Methodology section about (a) the characteristics of the emissions input files and CMAQ output files and (b) the 2016 base annual simulation ("orig"):

(a) The manuscript does not give any information about the temporal resolution of the emissions files or CMAQ output files -- do they contain hourly fields or more frequent or less frequent fields? How many different types of fields are there (e.g., different species, emissions vs. concentration vs. deposition fields)? What is the horizontal grid size? These details would help the reader to understand the size of the file sets.

(b) I know the details of the "orig" simulation are not directly relevant to the subject of this paper, but if the 'orig' simulation has already been documented in previous publications or reports, it would be helpful to have one sentence referring the reader to that additional information. For example, Table 6 states that the maximum daily PM2.5 bias was 512 ug/m3, a very large value, which left me wondering whether the 'orig' run included wildfire emissions (line 101 does mention "ptfire", but it is not clear whether this file was used in the 'orig' run).

4. References to daily measurements from the Ammonia Monitoring Network (AMoN) are

incorrect. The sampling duration of the AMoN measurements is two weeks (e.g., https://www3.epa.gov/castnet/docs/AMoNfactsheet.pdf). This error should be corrected. It also raises concerns about whether the model predicted values were properly aligned with the observed values, that is, were they two-week averages. And since Tables 5 and 6 and Figure 4 present metrics for NH3 in units of μg/m3, are the units of ppbv presented in Figures 5 and 8 for NH3 correct?

5. Table 1 is used to illustrate the output of the precision-reduction algorithm. However, it does not fully describe how rounding is handled. For example, two lines could be added to show the A05, A04, and A03 values of 100150.0 and 100250.0.

6. The definition of the statistical metrics on page 5, specifically on line 132, is not quite correct. The mean and standard deviations of the distribution are unknown, but they are estimated from the sample of model-measurement pairs, so the metrics are based on sample means and sample standard deviations.

7. In the description of the Burrows-Wheeler algorithm, the clause "which chronologically reduces sequences of datasets by processing sequences through multiple layers of compression algorithms" (p. 5, l. 138) is not clear to me, especially the use of "chronologically". I know the Burrows-Wheeler transform is not easy to explain, but is some rewording or new wording possible?

8. There may be a discrepancy between the description of data set sizes in the Introduction (lines 57-58) and the discussion of data storage results in Section 3.1 (lines 159-165). If the CMAQ emissions files for one day are about 7 GB in size, how can the compression utilities reduce their size, again for a day, by 111 GB or 241 GB? Or if Section 3.1 is considering the emissions files for one entire year, then 7 x 365 = 2,555 GB and 21% and 48% of that number are 537 GB and 1,226 GB, respectively. Can some clarification be provided.

9. Section 3.2 would benefit from the addition of some discussion. Why should reducing the precision of the input emissions result in an overall reduction in run time but also cause increases in run times for some days? Why should there be a seasonal dependence in runtime differences, with larger differences for the first half of the annual simulation? Why are the run-time differences larger for the A05 and A03 simulations than for the A04 simulation? Are the decreases in overall run time expected and are they significant? Even if these questions cannot be answered with certainty, they should at least be raised and any possible explanations offered. The first paragraph of the Conclusion section (lines 263-264) also gives a questionable summary of Section 3.2, stating that "the A05, A04 and A03 simulations (were) consistently faster than the 'orig' simulation in a undedicated HPC system". This statement is contradicted by Figure 3, which shows run-time increases for at least some days of all three simulations run with precision-reduced emissions.

10. In Section 3.3 a comparison is made between the ranges of maximum absolute bias,

that is, an analysis of model errors, and the concentration values of the air quality standards (lines 202-203) but the reason for this comparison was unclear to me? Second, in the methodology description in line 206 was RMSE really first calculated for all hourly grid-grid pairs in 2016, or do you mean that grid-grid pair **differences** were first calculated? Otherwise, it is not clear how average hourly RMSE could then be calculated for each season and region. And when you state "all grid-grid pairs", do these include grid cells located outside the contiguous U.S.? And in lines 209-211 where you state that "total accumulative RMSE for PM2.5, O3, and NH3 (sum of all region's RMSE) did not exceed 0.1 μg/m3, 0.4 ppbV, and 0.1 ppbV, respectively for all cases and for all seasons", based on Figure 5 could these values not be smaller (perhaps 0.04 μg/m3, 0.3 ppbV, and 0.05 ppbV)?

11. The analysis of accuracy for hourly deposition rates at the end of Section 3.3 is helpful, but it does not directly address the main impacts of deposition, which are cumulative. Could something be said about the accuracy of seasonal or annual deposition values for the simulations and cases?

TECHNICAL CORRECTIONS/SUGGESTIONS

p. 1, l. 13: Perhaps "... archive input and output data sets" or "... archive input and output files"

p. 1, l. 16: "desired post-processing **of the** output (e.g., for evaluations or graphics)"?

p. 1, l. 21, 23: Change "losslessness compression" to "lossless compression"

p. 1, l. 22: Don't you mean "before" rather than "after"?

p. 1, l. 27: Would this be more accurate: "To enhance the analysis of disk space efficiency, **the output from the** altered emissions CMAQ simulations ..."?

p. 1, l. 30: Would this be more clear: "Thus, in total, 13 gridded **output** products (four simulations and nine **altered output** cases) were ..."? (cf. line 108)

p. 3, l. 70: Perhaps "... by manipulating the mantissa of individual floating-point numbers".

p. 3, l. 90: "This study proceeds as **follows**:"

p. 4, l. 105: "Emission input and CMAQ output data were **then** compressed" -- this addition would emphasize that the lossy algorithm discussed in this manuscript consists of two steps.

p. 4, l. 110: Perhaps "...(see Table 2 for **a full list of** simulations and cases)"

p. 5, l. 142: Perhaps "Examples of precision-reducing transformation of floating-point numbers from their ..."

p. 7, l. 157: Perhaps "... throughout the entirety of the 2016 simulation"

p. 7, l. 161: Perhaps "... when applied to reduced-precision CMAQ output files"

p. 7, l. 162: Should this be "FX05, FX04, and FX03"?

p. 9, l. 183: Perhaps "... all available model-measurement pairs throughout 2016".

p. 10, l. 20: Perhaps "... maximum absolute bias for 24-h PM2.5 and MDA8 O3 do not ...".

p. 19, l. 269: Perhaps change "mimic" to "are similar to those for".

p. 19, l. 281, 285: Rather than "performance" and "quality", do you really mean "accuracy"?

Fig. 4 and 5 captions: Perhaps expand "(x)" and "(y)" to "(x-axis)" and "(y-axis)".

Table 7 caption: Perhaps "... with respect to the 'orig' simulation across all grid cells".