Reply to Referee 1

We thank referee 1 for the constructive, helpful criticism and the suggestion for revision. We have thoroughly revised the manuscript based on the comments given by the referees. A detailed point-by-point response to the comments by referee 1 are given below.

The manuscript investigates the lossy compression of ERA5 reanalysis data and its impact on trajectory calculations using the Lagrangian model for Massive-Parallel Trajectory Calculations (MPTRAC). The topic is timely and relevant, with clear implications for data storage optimisation, I/O performance, and reproducibility in geoscience workflows. ERA5 is one of the most widely used reanalysis datasets in the atmospheric sciences, with applications ranging from climate research to operational forecasting, and it is increasingly important as a training and validation source for AI models. The authors assess two lossy compression methods (ZFP and Layer Packing, PCK) and one lossless compressor (ZSTD), examining their influence on 10-day forward trajectories distributed globally in the free troposphere and stratosphere.

The study addresses an important problem and contributes to the relatively underexplored area of quantifying the impact of lossy compression on scientific analyses. Work along these lines can help the community better understand these impacts and support the adoption of compression methods that offer substantial data reduction without compromising scientific integrity.

However, the current version omits substantial parts of the relevant literature and does not sufficiently engage with the state of the art in scientific lossy compression. Several methodological choices also weaken the strength of the conclusions. The following points require major attention:

1. Omission of relevant literature – Key works are missing, including Tinto et al. (2024, GMD, (https://doi.org/10.5194/gmd-17-8909-2024)), which also deals with the impact of lossy compression of geoscientific data, and other publications that define the current state of the art.

Thanks a lot for this comment. We have added the work of Tinto (2024) as well as the works of Baker (2016); Dueben (2019); Delaunay (2019); Zender (2016) and Poppick (2020). These are cited in the introduction and the paragraph on P2, L59 has been changed as follows:

"Geoscientific data as e.g meteorological reanalysis data, climate simulation data and satellite data have increased immensely in size and their application in full resolution has become quite challenging for users. This is a known problem and has been in the focus of several previous studies. The efficiency of compressing climate simulation data was tested by e.g. Baker (2016); Dueben (2019) and Poppick (2020) and the compression of satellite data sets was tested by e.g. Delauney et al. (2019). Compression of meteorological reanalysis data was the focus of the study by Tinto (2024). In the geoscientific community the netCDF4 or HDF5 formats are widely used and thus compression of data sets in these formats was the focus of the studies by Delaunay (2019) and Zender (2016). All these studies came to the conclusion that lossy data compression is promising for reducing storage requirements. However, Poppick (2020) point out that it is important to evaluate the quality of compression in order to ensure that minimal scientific information is lost due to compression."

2. Ignoring state-of-the-art compressors – The evaluation is limited to ZFP, PCK, and ZSTD,

omitting widely recognised high-performance compressors such as SZ and MGARD. Without including these methods, the results cannot be considered representative of current capabilities of lossy compression for ERA5 data compression.

We agree that including also SZ and MGARD would be quite valuable, however this is beyond the scope of this study. Our intention for this study is not to test which of all available compression methods is the most efficient compressor for the ERA5 data, but to understand what impact the compression of the meteorological reanalysis data has on the trajectory calculations with MPTRAC. Further, to our knowledge, previous studies have also typically focused on a selection of compression methods and setups, rather than attempting to exhaustively compare all possible options. We agree that some of our comments made concerning PCK were inadequate since we are not considering the full range of possible compression techniques available and set-ups of the respective techniques that are possible. We will remove the misleading statements we made.

3. Suboptimal use of ZFP – ZFP is applied in precision mode, which the literature reports as less efficient and with poorer rate–distortion performance than accuracy mode. This disadvantages ZFP in the comparisons and may bias the conclusions.

We thank the reviewer for this important comment. We acknowledge that there are studies reporting better rate–distortion performance for ZFP in accuracy (absolute tolerance) mode compared to precision mode. We agree that the optimal choice of mode and error bounds may depend strongly on the variable considered: for parameters that vary by orders of magnitude with altitude, relative precision can be advantageous, whereas for parameters that remain within a similar order of magnitude across levels, absolute tolerance may provide better compression efficiency. A systematic evaluation of these options would indeed be valuable, but it is unfortunately beyond the scope of the present work. Our study builds on a setup for ZFP that was previously tested within our group and found to perform reasonably well for ERA5 data, and our focus here is on understanding the implications of compression for trajectory simulations rather than on identifying the optimal ZFP configuration.

4. Unsupported claim that PCK is the "best choice" – The conclusion that PCK is the most suitable compressor lacks supporting evidence, as state-of-the-art methods are not included and ZFP is used in a suboptimal configuration. This risks misleading readers about PCK's competitiveness.

We totally agree and we once again would like to apologize for our misleading comments. For us it was the best choice because it is easy to apply since no user specific set-ups are needed and it results in a 50% reduction of the files with additionally maintaining a high accuracy for the trajectory calculations. In addition, the PCK method offers the advantage of extremely fast decompression, requiring only the application of scaling and offset factors, and significantly reduced runtime needed for file input. Overall, it was the only method among those tested that reduced both file size and runtime requirements simultaneously. We make this point much more clear now in the manuscript and have checked our manuscript for misleading statements and removed/revised these. E.g. a misleading statement was made in the abstract. The sentence there has been changed to:

"Thus, our study shows that all compression methods considered here (ZSTD, PCK and

ZFP) would be valuable for application in atmospheric sciences and that with compression of the ERA5 meteorological reanalyses data one can overcome the challenges of high demand of disk space from this data set."

We would like to keep the sentences concerning PCK in the discussion and conclusion since these refer to our personal choices we made concerning compression of ERA5 data and MPTRAC trajectory calculations. However, we adjusted these as follows to be more clear:

"For us, PCK is the best choice, since it is also for inexperienced users easy to apply and is quite efficient on our current supercomputer system at compressing the ERA5 data (CR=2) while at the same time keeping the accuracy of the data, resulting in low transport deviations ($<$ 40 km). ZSTD has the advantage of being a lossless compression method, so there is no loss, and storage requirements are reduced by 30%. ZFP has the advantage of allowing users to specify themselves the level of compression."

"For our needs PCK is the best choice and has thus been implemented in MPTRAC as the default compression method. However, ZFP and ZSTD can be also used by enabling them when compiling MPTRAC. In the future, we plan to add other widely used compression methods, such as SZ3 and MGARD."

Recommendation: I recommend major revisions. The authors should (1) expand the literature review to include key recent works and provide proper context on the state of the art, (2) revisit the ZFP configuration to use competitive modes reported in the literature, and (3) either include tests with state-of-the-art compressors such as SZ or explicitly limit their claims, providing a clear justification for the exclusion of these methods.

Thanks again for your valuable comments. We have taken all your comments into account and hope that the revised version covers now all these points to your satisfaction.

References

Baker, A. H., Hammerling, D. M., Mickelson, S. A., Xu, H., Stolpe, M. B., Naveau, P., Sanderson, B., Ebert-Uphoff, I., Samarasinghe, S., De Simone, F., Carbone, F., Gencarelli, C. N., Dennis, J. M., Kay, J. E., and Lindstrom, P.: Evaluating lossy data compression on climate simulation data within a large ensemble, Geosci. Model Dev., 9, 4381–4403, https://doi.org/10.5194/gmd-9-4381-2016, 2016.

Delaunay, X., Courtois, A., and Gouillon, F.: Evaluation of lossless and lossy algorithms for the compression of scientific datasets in netCDF-4 or HDF5 files, https://doi.org/10.5194/gmd-12-4099-2019, 2019.

Düben, P. D., Leutbecher, M., and Bauer, P.: New Methods for Data Storage of Model Output from Ensemble Simulations, Mon. Weather Rev., 147, 677–689, https://doi.org/10.1175/mwr-d-18-0170.1, 2019.

Poppick, A., Nardi, J., Feldmann, N., Baker, A., Pinard, A., and Hammerling, D. M.: A statistical analysis of lossily compressed climate model data, Comput. Geosc., 145, 104 599, https://doi.org/10.1016/j.cageo.2020.104599, 2020.

Tintó Prims, O., Redl, R., Rautenhaus, M., Selz, T., Matsunobu, T., Modali, K. R., and Craig, G.: The effect of lossy compression of numerical weather prediction data on data analysis: a case study using enstools-compression 2023.11, Geosci. Model Dev., 17, 8909–8925, https://doi.org/10.5194/gmd-17-8909-2024, 2024.

Zender, C. S.: Bit Grooming: statistically accurate precision-preserving quantization with compression, evaluated in the netCDF Operators (NCO, v4.4.8+), Geosci. Model Dev., 9, 3199–3211, https://doi.org/10.5194/gmd-9-3199-2016, 2016.