Response to Reviewer #1

Lines 26–28: The example given can be justified conceptually but would benefit from clarification. Global mean temperature is statistically robust to small, spatially uncorrelated compression errors that may cancel upon averaging. In contrast, local wind power estimates depend on nonlinear relationships and fine-scale variability, making them potentially more sensitive to small local errors. Clarifying this reasoning would avoid confusion.

Response: We have added clarifications in the main text.

Lines 43–48: The term "variable characteristics" is vague. Please clarify which properties are meant (e.g., statistical distribution, intermittency/sparsity, smoothness vs. sharp gradients, spatial/temporal correlation scales, dynamic range, NaNs, extremes, etc.). A few examples would improve clarity.

Response: We have added specific examples of characteristics of temperature, precipitation, and sea surface temperature fields to clarify our points.

Line 70: "Actionable insights" sounds generic. Consider specifying the practical guidance provided (e.g., recommendations for particular variable types or error tolerances).

Response: We have added the following more specific recommendations: "Namely, compressors based on bit rounding or stochastic rounding achieve modest compression ratios of around 10 but they are very robust to failures. More advanced compressors can reach higher compression ratios but they are more prone to fail on edge cases so they require more careful use-case specific validation. Compressors generally have limited support for pointwise relative errors and different compressors might use slightly varying definitions of 'relative' which end users should be aware of."

Table 1: Cloud-related variables (e.g., liquid water content) are not included, although they represent a challenging case due to their 3-D structure, sharp gradients, and large near-zero regions. While sparsity and NaNs are partly represented by precipitation and SST, a brief comment on the exclusion of 3-D cloud condensates would be useful, possibly as a future extension.

Response: The goal of the benchmark was to be as concise as possible while still covering a wide range of different data characteristics. While cloud-related variables definitely have challenging characteristics, most of these characteristics are already covered in other variables: 3D structure (humidity, nitrogen dioxide, air temperature), sharp gradients (nitrogen dioxide, precipitation), large near-zero regions (precipitation). Based on your feedback we have also conducted some additional experiments with compressing the cloud ice water content from the IFS model and we found that the compressors do not behave qualitatively markedly different to the existing set of variables.

Table 1: Since V is mostly 1 here, variables are effectively treated independently. While reasonable, it would help to state explicitly that multivariate compression is beyond the current scope. In practice, many variables are physically correlated (e.g., atmospheric

chemistry tracers), and advanced methods may exploit such structure. A brief acknowledgment would strengthen the discussion.

Response: We have extended the text to highlight that we mainly focus on single variable compression. This is motivated by the fact that most large datasets store different variables in separate chunks so separate variables are commonly compressed separately.

Lines 79–88: The regridding discussion focuses on model output, but similar issues apply to satellite swath data provided in along-track/across-track coordinates. Regridding can alter statistical properties relevant for compression. A brief acknowledgment would clarify that restricting to regular grids simplifies comparability but does not reflect all real-world cases.

Response: We have adjusted the text to move a footnote into a main text highlight that the regridding can alter the compressor performance and make it explicit that data that does not lie on a regular structured grid is outside the scope of this work.

Line 101: "Linear packing" is mentioned in the context of the ERA5 data but not explained. A short definition or reference would be helpful.

Response: We have rephrased it as "linear quantization" which is a more common term and added a brief definition as well as a reference to more detailed explanations.

Lines 128–129: The manuscript states that the absence of standard error bounds is "partly" due to application dependence. "Largely" may be more appropriate, as tolerable error levels are typically driven by downstream applications.

Response: Adjustment made.

Lines 175–178 and Table 2: For several variables, the gap between the low (100th percentile) and mid (99th percentile) bounds exceeds that between mid (99th) and high (95th) significantly. This suggests sensitivity of the low bound to extreme outliers or heavy-tailed spread distributions. Please comment on this sensitivity and whether slightly lower percentiles (e.g., 99.9%) were considered.

Response: The low bound can indeed be sensitive to outliers. However, for our use case this is very much intended because the goal of the three error bounds is to span a plausible range of error bounds that might be used in practice. The relevant comparison here is really with the expert provided bounds. For some variables, like mean sea level pressure, the low bound is just above the expert provided bound which demonstrates that we really want to include the full range.

Lines 226–234: Instruction count is a useful reproducible metric, but wall-clock runtime remains highly relevant in practice. Parallelization (multi-threading, GPU support) and peak memory footprint can significantly affect scalability for large datasets. A brief discussion of these practical aspects would be valuable.

Response: We have added an extra paragraph to highlight that there are additional metrics that will be relevant to downstream users that we do not include in the benchmark because it is difficult to provide objective comparisons for them.

Figure 2: The scorecards are helpful. For example, SZ3 often achieves higher compression ratios but also larger error metrics and occasional bound violations. Although discussed later, more explicit guidance in the figure interpretation would help readers assess comparability across methods.

Response: We have added some extra discussion in the figure caption to provide some more guidance to the reader for how to read the scorecards.

Figure 7: This figure shows that compressors can produce markedly different error distributions, even under identical nominal absolute bounds. While discussed, this reinforces that methods are not strictly comparable under a single bound alone. Future work could consider complementing the current protocol with additional tail metrics (e.g., p99/p99.9) or joint criteria on maximum and distributional error properties.

Response: We have added a paragraph, in Section 3.3 at the end of the “Error Distributions” subsection, highlighting the fact that users should consider exploring different error bounds. In the benchmark codebase, users can also add their own metrics to the benchmark evaluation suite. However, because many metrics are very application specific we aimed to limit ourselves to a minimal subset of evaluation metrics.

Lines 468–471: Benchmarking full high-resolution model outputs (terabyte scale) would be highly valuable. Such tests would better reflect modern data volumes and assess compressors under realistic storage, I/O, and scalability constraints, complementing the current laptop-scale setup for HPC environments.

Response: One thing to highlight is that even if datasets have terabyte scale they are often stored in chunked representations that do not exceed 100MB and therefore a small scale benchmark like ours is still relevant. Of course, dealing with TB scale data creates its own set of engineering challenges which are key to unlocking good performance which we now specifically highlight in the text as well.

Technical corrections

Line 99: Remove extra “.”

Line 136: Index v runs from 0 to V, implying V+1 elements; is this intended?

Line 318: Rephrase as “This ensures …”

Figures 3 and 6: Adding distinct marker symbols alongside colors and linestyles would improve clarity.

Response: We have corrected all the technical comments.