

A Statistical Analysis of Compressed Climate Model Data

Andrew Poppick*, Joseph Nardi*, Noah Feldman*, Allison H. Baker† and Dorit M. Hammerling†

* Carleton College, Northfield, MN 55057

apoppick@carleton.edu, nardij@carleton.edu, feldmann2@carleton.edu

† National Center for Atmospheric Research, Boulder, CO 80305

abaker@ucar.edu, dorith@ucar.edu

Abstract—The data storage burden resulting from large climate model simulations continues to grow. While lossy data compression methods can alleviate this burden, they introduce the possibility that key climate variables could be altered to the point of affecting scientific conclusions. Therefore, developing a detailed understanding of how compressed model output differs from the original is important. Here, we evaluate the effects of two leading compression algorithms, SZ and ZFP, on daily surface temperature and precipitation rate data from a popular climate model. While both algorithms show promising fidelity with the original output, detectable artifacts are introduced even at relatively low error tolerances. This study highlights the need for evaluation methods that are sensitive to errors at different spatiotemporal scales and specific to the particular climate variable of interest, with the ultimate goal to improve lossy compression collaboratively with the algorithm development teams.

I. INTRODUCTION

Earth system models, such as the popular Community Earth System Model (CESM) [1], have been generating increasingly larger volumes of data as the models take advantage of advances in supercomputing resources. While data compression for climate data has been the subject of several recent studies (e.g., [2]–[8]), more analysis is needed to satisfy concerns in the climate community regarding data compression-induced artifacts.

Climate data is typically output by time slice, meaning that variables are stored by spatial fields. Even for the so-called CESM “time-series” files, the multidimensional arrays are laid out such that time is the outer array dimension and the most efficient access is by spatial slice. For this reason, compression algorithms applied to CESM data thus far (e.g., [4], [7], [8]) have been applied to spatial fields, and any dependence on time has been ignored. It is therefore natural to investigate whether compression applied to spatial fields independently introduces any artifacts that change either the temporal or spatial characteristics of the model output. These questions are worth investigating in detail: examining climate characteristics over time is a critical component to most climate data analyses, as is examining coherent spatial features of climate variables.

The effects of lossy compression are routinely evaluated with simple metrics such as root mean squared error or maximum point-wise error, which are quite appropriate for many applications (e.g., visualization), but less so for floating-point climate simulation data. Capitalizing on our familiarity with climate data, we thoroughly explore the artifacts of lossy compression on CESM [9], a summary of which is presented here. Our goal is to work with compression algorithm development teams to address the specific issues that arise.

II. EXPERIMENT DETAILS

A. CESM Data

We use data from the publicly available CESM Large Ensemble (CESM-LE) project [10]. The project includes a set of 40 ensemble runs for the period 1920–2100. All simulations use the fully coupled one degree latitude-longitude version of CESM with the Community Atmosphere Model (CAM) v5. We focus on the atmospheric model output and use the historical forcing period (1920–2005) for ensemble

member 30. Each CAM variable (159 total) is stored in a time-series file, and variable output is at temporal frequencies of monthly, daily, or 6-hourly. The CAM grid corresponds to 192×288 grid points per vertical level, such that the 192 rows indicate the latitude (90° to -90°) and the 288 columns indicate the longitude (0° to 360°). We study the following 2D time-series variables, which consist of 31,390 time slices (file size in parentheses):

- TS: Daily average surface temperature, in $^\circ\text{K}$ (3.8 GB).
- PRECT: Daily average precipitation rate, in m/s (5.4 GB).

These variables were chosen as TS is smooth with a relatively small data range (“easy” to compress), while precipitation variables change more abruptly and have vastly larger ranges (“hard” to compress). Note that CESM writes data to file in single precision (32 bits).

B. Compression Methods

Lossy methods give an approximation of the original data upon reconstruction, and the quality of approximation is controlled by algorithm-dependent parameters. This is in contrast to lossless methods, which exactly preserve the original data. Our interest is in lossy methods, as they offer the most meaningful data reduction for floating-point simulation data. While many lossy methods have been advocated for use on floating-point data recently (e.g., see methods in [11]), we experiment with two popular methods: SZ and ZFP.

The SZ compressor ([12], [13]) is a predictive method that uses adaptive error-controlled quantization and variable-length encoding to optimize compression. We use SZ 1.4.13 in fixed-accuracy mode (errorBoundMode = ABS). Given an absolute error tolerance, ϵ , SZ defines a size 2ϵ interval that is centered on the predicted value, as well as $(2m - 2)$ additional adjacent intervals. The interval that the actual value falls into determines the identifying index (size m bits). If the actual value does not lie in any of the intervals, it is given an index that flags it as unpredictable and uses an alternative coding scheme (with longer codes). All predicted values are then subjected to Huffman encoding and compressed with GZIP. In this study, we use single layer prediction (layers = 1), optimized auto-selection of quantization intervals with a maximum number of 65,536 (i.e., $m = 16$), the SZ default compression mode, and no offset. Note that SZ offers several additional error modes: fixed relative error (normalized by the data range), fixed PSNR (peak signal-to-noise ratio), and fixed point-wise relative error.

The ZFP compressor [14] was designed to facilitate random data access, but can be used for error-bounded sequential compression, depending upon the specified parameters. ZFP partitions d -dimensional arrays into blocks of 4^d values and compresses each block independently via a floating-point representation with a single common exponent per block, an orthogonal block transform, and embedded encoding. We use ZFP 0.5.3 in fixed-accuracy mode. Specifying an absolute error tolerance of 0 indicates that the compressor should achieve lossless (if possible) or near lossless compression. While we use fixed-accuracy mode in this study, ZFP can also be used in a fixed-rate mode (required for random access) or fixed-precision mode.

C. Data for Analysis

We applied both SZ and ZFP to the 2D time-series data for TS and PRECT. For each file, we call the compressor on each time slice sequentially over time, meaning that each is compressed independently. We note that both SZ and ZFP can typically achieve more data reduction (i.e., a smaller compression ratio) when applied to “larger” fields. For SZ, the overhead associated with the Huffman tree is better mitigated with more data, and for ZFP, higher dimensional data is better suited to hiding the per block overhead. Indeed, for these 2D time-series, we could compress multiple time slices at once so as to operate on 3D data. However, this choice introduces complexity due to memory constraints: while we could compress several slices at once for this grid, it is unlikely that we could compress an entire time-series file with one call to the compressor for high resolution, long duration, or frequent temporal output. Therefore, some dividing between time slices is necessary, and to simplify our analysis, we compress each time slice independently. The data that we analyze results from applying compression, followed by reconstruction, to each variable’s time-series file. To simplify the comparison of SZ and ZFP, we use both compressors in their fixed-accuracy modes as those are essentially equivalent (with the caveat that ϵ for ZFP is converted to the nearest power of 2). We selected a range of ϵ , where the characteristics of PRECT led us to extend the range to much smaller tolerances than needed for TS.

III. EXPLORATORY ANALYSIS

A. Daily Surface Temperature (TS)

Table I summarizes global characteristics of the quality of compression for both SZ and ZFP. We show mean errors (to capture any systematic biases in the sign of the error), along with mean absolute errors (MAEs) and root mean square errors (RMSEs), where the latter is more sensitive to extreme errors than the former. We also list the achieved compression ratio (CR), which is the ratio of the size of the compressed data to the original data. Overall, the simple metrics indicate that the quality of compression scales with ϵ . The SZ output becomes lossless at $\epsilon = 10^{-4}$ and ZFP at $\epsilon = 10^{-5}$. For a fixed ϵ , ZFP shows substantially smaller MAEs and RMSEs than SZ; however, ZFP shows larger magnitude mean errors, indicating that ZFP errors have a stronger systematic mean bias than SZ errors. The average ZFP errors are positive at ϵ of 1.0, 0.1, 10^{-3} , and 10^{-4} and negative for 0.5 and 0.01. CR values for SZ are smaller than ZFP for $\epsilon \geq 10^{-2}$, but larger for $\epsilon \geq 10^{-3}$. The overall ϵ are much smaller than the scale of variation in the temperatures themselves. For example, the globally pooled TS standard deviation is about 8.6°K, whereas the maximum global RMSE is about 0.37°K (for SZ $\epsilon = 1.0$). This implies that pointwise comparisons of the original and compressed output will appear very similar relative to the scale of variability in the data (i.e., the correlation between the original output and compressed output is nearly one when averaged globally and across time).

B. Daily Average Rainfall Rate (PRECT)

Table II summarizes global characteristics of compression for PRECT. In addition to the CR and global error metrics (MAE and RMSE), we list the mean PRECT value in the compressed output and the global probability of positive and negative rainfall (i.e., the proportion of days with positive or negative rainfall across all days and gridcells). Even at the smallest error tolerances, MAEs and RMSEs are of the same order of magnitude as the mean PRECT values. Globally, for $\epsilon \geq 10^{-3}$, ZFP sets PRECT to zero everywhere. For smaller ϵ , ZFP preserves some rainy days but with a negative bias in the number, and both methods produce substantially more

	Mean Error	MAE	RMSE	CR
sz 1.0	-1.45e-03	1.35e-01	3.68e-01	.06
sz 0.5	-7.18e-04	6.77e-02	1.84e-01	.09
sz 1e-1	-1.48e-06	1.69e-02	4.60e-02	.16
sz 1e-2	4.47e-06	2.11e-03	4.06e-03	.29
sz 1e-3	2.72e-06	1.38e-04	3.66e-04	.60
sz 1e-4	0.00e+00	0.00e+00	0.00e+00	.78
zfp 1.0	7.71e-03	3.42e-02	6.75e-02	.23
zfp 0.5	-3.86e-03	1.71e-02	3.43e-02	.26
zfp 1e-1	4.82e-04	2.16e-03	4.38e-03	.36
zfp 1e-2	-6.01e-05	2.69e-04	5.48e-04	.45
zfp 1e-3	7.76e-06	3.35e-05	6.93e-05	.54
zfp 1e-4	9.17e-08	5.49e-07	3.20e-06	.67
zfp 1e-5	0.00e+00	0.00e+00	0.00e+00	.76

TABLE I

SELECTED SUMMARY STATISTICS FOR TS BY {SZ, ZFP} AND ϵ .

	Mean	MAE	RMSE	Prob >0	Prob <0	CR
Original	2.82e-08	—	—	0.985	0.000209	—
sz 1e-1	1.95e-09	2.76e-08	6.51e-08	0.992	6.37e-05	.0002
sz 1e-4	1.95e-09	2.76e-08	6.51e-08	0.992	6.37e-05	.0002
sz 1e-5	2.06e-07	1.91e-07	2.50e-07	0.999	0	.0002
sz 1e-6	4.36e-08	4.62e-08	8.31e-08	1	5.65e-05	.0009
sz 1e-7	2.87e-08	3.95e-08	8.07e-08	0.875	0.125	.01
sz 1e-9	2.82e-08	3.86e-08	8.06e-08	0.951	0.0488	.15
sz 1e-11	2.82e-08	3.86e-08	8.06e-08	0.964	0.0358	.61
zfp 1e-1	0	2.82e-08	6.58e-08	0	0	.002
zfp 1e-3	0	2.82e-08	6.58e-08	0	0	.002
zfp 1e-5	2.50e-09	3.15e-08	7.98e-08	0.00647	0.00181	.002
zfp 1e-6	2.40e-08	4.35e-08	8.55e-08	0.277	0.069	.002
zfp 1e-7	2.81e-08	3.91e-08	8.07e-08	0.76	0.118	.11
zfp 1e-9	2.82e-08	3.86e-08	8.06e-08	0.938	0.0538	.28
zfp 1e-11	2.82e-08	3.86e-08	8.06e-08	0.969	0.0294	.50

TABLE II

SELECTED SUMMARY STATISTICS FOR PRECT BY {SZ, ZFP} AND ϵ .

negative PRECT values than in the original output. Figure 1 shows the percentage of gridcells with positive rainfall by day of the year, aggregating globally and across years, and the log-odds ratios comparing these percentages in the compressed output to the original output. The odds of rain are defined as $\omega = \frac{p}{1-p}$, where p is the probability of rain on that day. Denoting the odds of rain under the compressed output as $\tilde{\omega}$, the odds ratio is $\tilde{\omega}/\omega$. Figure 1 indicates a discernible seasonal pattern in global probabilities of rainfall, as well as in errors in these probabilities (as measured by the odds ratios). In general, rainfall is more likely in the first half of the year, which is also the period when the odds ratios are largest in magnitude.

IV. GRIDCELL-LEVEL ANALYSIS: TS

Gridcell-level errors can show detectable artifacts that are not readily apparent in global summaries. Patterns in gridcell-level errors can themselves produce artifacts in quantities derived from the compressed output. In this section and the next, we examine compression errors at finer spatial and temporal scales. We first look at TS, noting that we generally show results for error tolerances 1.0, 0.5, 0.1, and

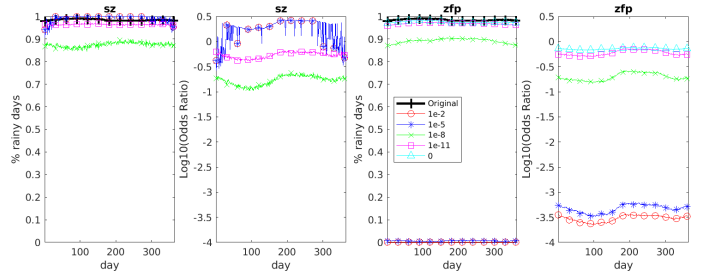


Fig. 1. Percentage of gridcells with positive rainfall by day and log-odds ratios (base 10) comparing the compressed to the original output percentages (e.g., positive ratios mean that it rains more often in the compressed output).

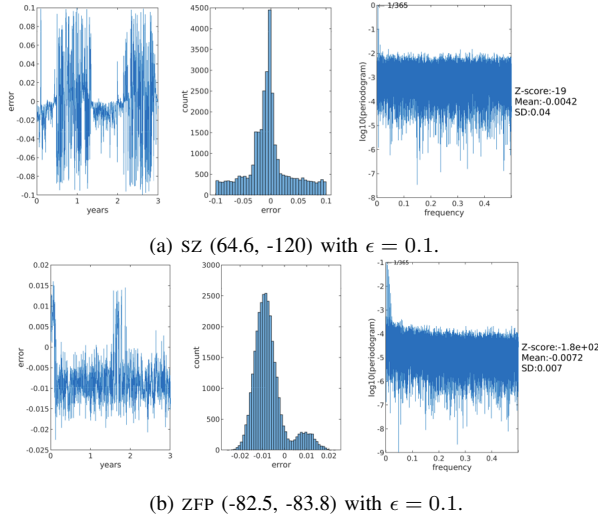


Fig. 2. TS: Left, error time-series for first 3 years; middle, histogram of entire time-series; right, periodogram of entire time-series.

0.01, as these highest tolerances show the clearest artifacts. Unless otherwise specified, the behavior of the compressed output at smaller error tolerances is qualitatively similar to the behavior at $\epsilon = 0.01$, until the output becomes lossless.

We first discuss the nature of some temporal artifacts created by SZ and ZFP ($\epsilon = 0.1$) with example locations in Canada (64.6, -120) and Antarctica (-82.5, -83.8), respectively. Figure 2 contains the error time-series, the histogram of daily errors, and their periodogram¹. We also list Z-statistic and mean and standard deviation of the error. Large Z-statistics indicate that the error is not mean zero and/or is not independent and identically distributed across time. Both locations show an overall negative bias, indicated by a negative mean error and large Z-statistic. A mean seasonal cycle is detectable in the errors at both locations (i.e., the spike in the periodogram at frequency 1/365), and errors are positively correlated (i.e., the decay in the periodogram in frequency). Errors for SZ are bounded by 0.1 and frequently achieve values close to that bound, whereas ZFP errors are smaller and their distribution tails decay more rapidly. The SZ distribution is unimodal, centered near zero; the ZFP error distribution is bimodal, with a large negative mode and a smaller positive mode. Both error time series show evidence of temporal nonstationarity, with periods of increased variability (particularly for the location chosen for SZ).

A. Mean Errors

Figure 3 shows the mean compression errors at the gridcell level at several error tolerances for daily TS data. In the SZ output, larger mean errors occur in bands at longitudes -120, 0 and 120, at the South Pole, in parts of the Atlantic, and in Oceania. At ϵ of 0.5 and 1, mean errors over the ocean appear to be strongly spatially correlated. In contrast, ZFP mean errors show much stronger patterns. Notably, a 4×4 gridding pattern is apparent in the mean errors, with periodic mean errors that are large in magnitude. The gridding pattern does not align with the positions of the 4×4 partition used in the ZFP algorithm across the entire globe: there is misalignment at particular longitudes (again at -120, 0, and 120 degrees), as particularly evident in Figure 4. The large mean errors are mostly positive for ϵ of 0.1 and

¹The periodogram is the squared magnitude of the Fourier transform of the time-series and represents the amplitude of variability associated with oscillations with the corresponding frequency.

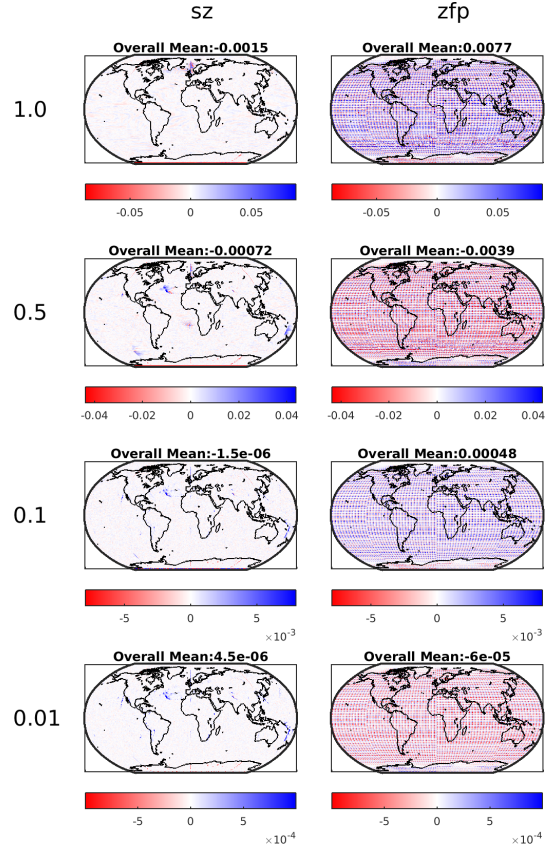


Fig. 3. Mean errors for TS. Left, SZ; right, ZFP, with error tolerances indicated on the left. Global mean errors are indicated in the plot titles.

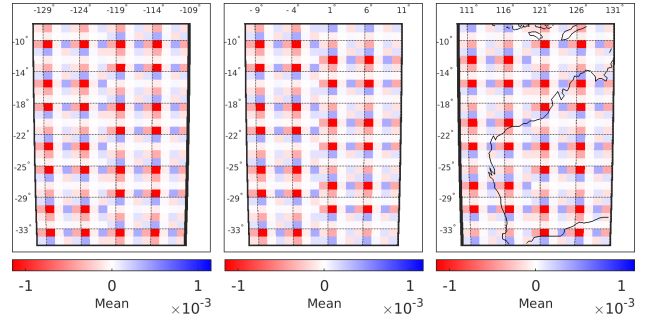


Fig. 4. Zoomed-in plot of the mean errors of ZFP at $\epsilon = .01$ for TS. Note the misalignment at -120, 0, and 120 degrees longitude.

1 and mostly negative for ϵ of 0.01 and 0.5. Whatever the dominant sign for ϵ of interest, there is a shift towards the South Pole where the mean errors become close to zero and then reverse in sign. We note that the artifacts in mean errors are much larger than would be expected if the compression errors were mean zero and independent and identically distributed in time, as measured by the corresponding Z-statistics, which show patterns similar to those for mean errors and large in magnitude values, particularly for ZFP errors (not shown).

B. Contrast Variances

Contrast variances, i.e. average squared gradients, are natural quantities to inspect for artifacts of compression because differences in temperatures between adjacent gridcells are typically small. (In comparison, raw temperatures are typically more variable than ϵ , which can mask artifacts of compression in the raw temperature

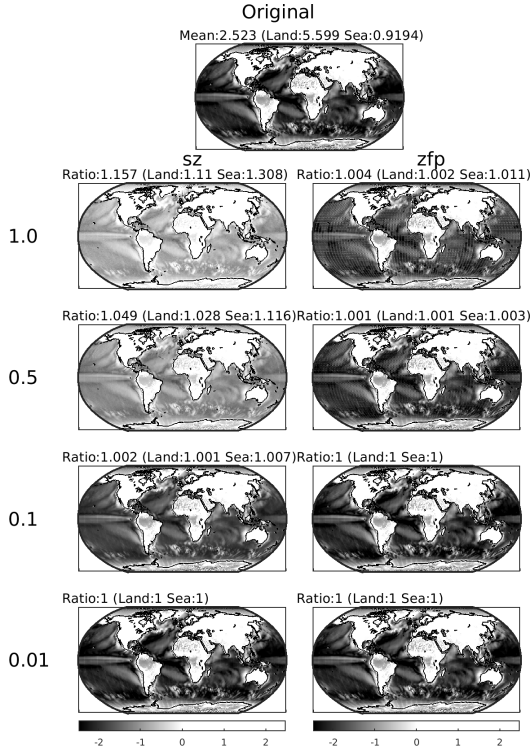


Fig. 5. January E-W TS contrast variances (log scale) for the original (top), SZ (left), and ZFP (right) by ϵ . The global mean contrast variance is provided for the original, and its ratio to that of the original is listed for compressed data; land and ocean averages are also given.

fields.) Contrast variances were also used in [15] to assess quality of compression. Artifacts in contrast variances indicate biases in the fine-scale spatial dependence structure of the compressed output. East-West (E-W) contrast variances for January are shown in Fig. 5. Some of the artifacts shown in the mean errors (Fig. 3) are apparent in the contrast variances, particularly at high ϵ . For high ϵ , SZ compression shows an overall positive bias in contrast variances (indicating increased fine-scale spatial variation), and ZFP shows a gridding pattern in addition to a smaller overall positive bias. For both, biases are larger at higher ϵ and over oceans.

C. Biases in Seasonality

Variation in errors over time can be due in part to biases in the mean seasonal cycle (e.g., Fig. 2). Figure 6 shows the mean seasonal cycles at the two gridcells in Fig. 2 along with the standardized seasonal cycle error. For both examples, the absolute bias in the seasonal cycle is small because the ϵ is much smaller than the scale of the original seasonal cycle; however, the biases are strongly detectable relative to the error variability. In the gridcell where ZFP produces a bias, the sign of the bias is inconsistent across the varying ϵ (as also seen in the overall mean errors in Fig. 3). In the location where SZ produces a bias, the seasonal cycle is slightly enhanced in the SZ output. We also looked for biases in mean seasonal cycles globally by examining the amplitude of the first seasonal harmonic of the error at each grid cell. Error seasonality with SZ is less common than in ZFP (for any given tolerance, the percentage of gridcells with significant correlation is lower), but is apparent in a few longitudinal bands and in some regions of the ocean. With ZFP, error seasonality is most consistently seen in the Arctic and Antarctic.

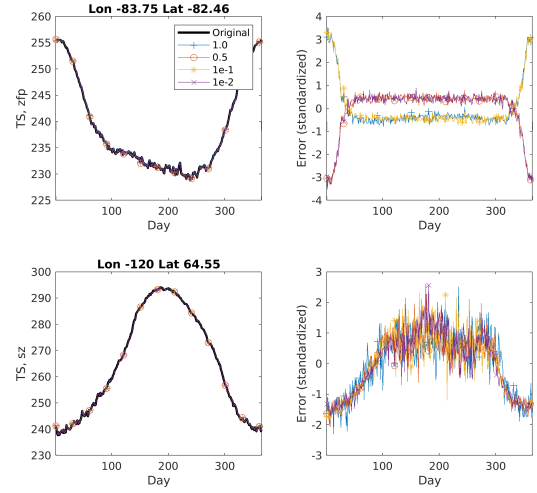


Fig. 6. Top: seasonal cycles (left) for TS at a location in Antarctica where ZFP errors show a strong mean seasonal cycle and the error seasonal cycles (right) at this location (standardized). Bottom: for SZ in Canada.

D. Temporal Correlations

Finally, we investigate whether compression affects the temporal correlation structure of the TS data at the gridcell level. Figure 7 compares the temporal correlation structure in the original output to that in the compressed output. We show the lag-1 autocorrelations of the first differences of the de-seasonalized TS values. (We first-difference and de-seasonalize the data as a naive way to remove temporal trends and seasonal cycles that would otherwise be confounded with other sources of temporal correlation.) Both algorithms strongly suppress temporal correlation at high ϵ , particularly over the ocean (where temporal correlation in the original output is strongest). The suppression is stronger for SZ than for ZFP; however, ZFP also produces odd artifacts in the Southern Hemisphere oceans. Even at $\epsilon = 0.1$, the SZ TS field shows temporal correlations over oceans that are much weaker than in the original. At $\epsilon = 0.01$, the temporal correlations appear to match those in the original.

V. GRIDCELL-LEVEL ANALYSIS: PRECT

The metrics we use to evaluate the quality of compression for PRECT are necessarily different from those for TS, in part because rainfall has a positive probability of being zero. Recall that the compressed PRECT output shows global biases in the percentage of days with positive rainfall and also contains excess days with negative PRECT values. To examine biases at the gridcell level, Fig. 8 shows the first 50 days of PRECT values at two example locations on the globe in the original output and at various ϵ . Biases are readily apparent, particularly for the larger ϵ , where PRECT is set to zero in ZFP and positive but too small in SZ. Both algorithms also produce negative PRECT values at small ϵ ; however, ZFP appears to perform somewhat better at smaller ϵ at these two locations.

To assess the bias in the number of rainy days produced by compression, Fig. 9 shows the odds of daily rain and odds ratio comparing original and compressed output at each gridcell (as in Section III-B) for SZ and ZFP with multiple ϵ . There is a negative bias in the odds of rain under the compressed output at all locations at small ϵ . The odds of rainfall in SZ are qualitatively similar to those in ZFP at these ϵ , except that the 4×4 gridding used in the ZFP algorithm is also apparent in that output. The locations where the bias is strongest are equatorial ocean gridcells and off the western coasts of continents, where the odds of rain in the original output are

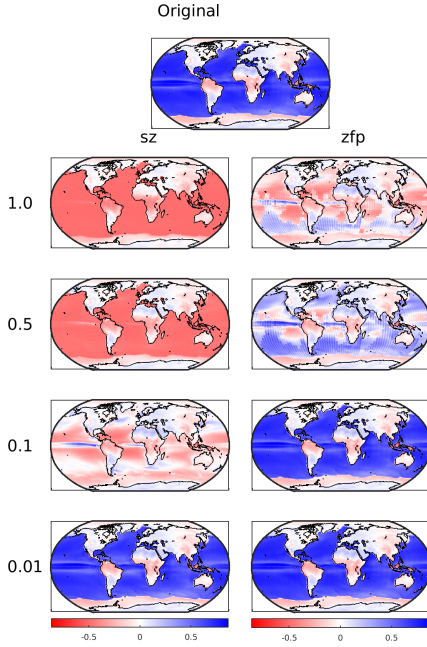


Fig. 7. Lag-1 correlations of the first differences of de-seasonalized TS for the original, SZ, and ZFP by ϵ . Negative values are a result of over-differencing and do not imply that the original series is negatively correlated in time.

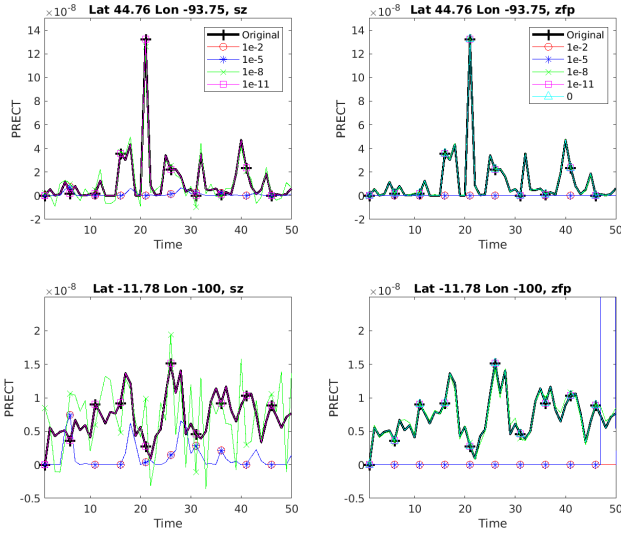


Fig. 8. The first 50 days of PRECT in Minnesota (top) and in the Pacific (bottom). Left, SZ; right, ZFP. Both methods produce absolute biases in PRECT as well as in number of rainy days, and can produce negative values.

on the order of 100 times greater than in the compressed output, even at the smallest ϵ . In the median over the globe, odds of rain in the original output is about 2.7 times greater than in the ZFP “lossless” output. Similarly, the odds of rain in the original output is about 11.5 times greater than in SZ at its lowest $\epsilon = 10^{-11}$. At larger ϵ , ZFP produces no rainfall whereas SZ produces positive rainfall at all locations on nearly every day (and the few nonpositive rain days occur at every location simultaneously). Both algorithms also produce days with negative rainfall at some locations; ZFP for $\epsilon < 10^{-5}$, and SZ for $\epsilon < 10^{-8}$. Negative PRECT values are most widespread at $\epsilon = 10^{-6}$ and $\epsilon = 10^{-7}$. While PRECT can be negative in the original output, negative values are much more common in compressed output with small ϵ . Negative rainfall appears to be most common in desert

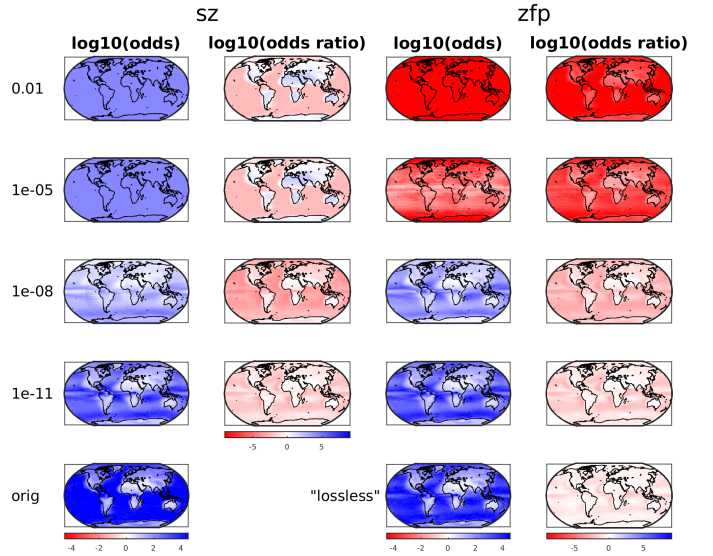


Fig. 9. Odds and odds ratio (compressed to original) for daily PRECT (base 10) for SZ and ZFP by ϵ . Negative log-odds ratios correspond to locations where daily rainfall is less likely in the compressed output.

regions at the smallest ϵ , although for other ϵ , negative rainfall can appear fairly widespread globally.

VI. CONCLUSION

We have explored the effects of SZ and ZFP compression on TS and PRECT from a historical run of CESM. For TS, it appears that both algorithms can achieve good fidelity with the original output for modest ϵ . However, for larger ϵ , both produce detectable artifacts that can impact important features of the spatiotemporal structure. ZFP appears to produce more artifacts in the temporal mean structure (e.g., mean biases and biases in seasonality), whereas SZ appears to produce larger biases in the spatiotemporal correlation structure (e.g., enhancement of contrast variances and reduction of temporal correlations). Even at the highest ϵ considered, errors are substantially smaller than natural temperature variability, which means that point-wise comparisons of the original and compressed output will make the compressed output look successful (i.e., high correlation between compressed and original output); however, small but detectable artifacts in errors can produce detectable (and possibly important) biases in fine-scale spatial and temporal correlations. These biases will typically not be captured by global measures, highlighting the importance of inspecting compressed output at multiple spatial and temporal scales. For PRECT, both algorithms inflate the number of days without positive rainfall even at very small ϵ . PRECT may be difficult to compress because rainfall has a positive probability of being zero on a given day, many very small values, but also a strongly skewed distribution of positive values with a large range.

Lossy data compression is promising for reducing the storage requirements resulting from large climate model experiments. However, in order to ensure that minimal scientific information is lost due to compression, it is important to evaluate the quality of compression with a sensitivity both to the characteristics of the climate variables of interest and also to climate variability at multiple spatiotemporal scales. These questions are worth investigating in detail so that we may then collaborate with compression algorithm development teams to address important issues that may arise for climate model data compression.

REFERENCES

- [1] J. Hurrell, M. Holland, P. Gent, S. Ghan, J. Kay, P. Kushner, J.-F. Lamarque, W. Large, D. Lawrence, K. Lindsay, W. Lipscomb, M. Long, N. Mahowald, D. Marsh, R. Neale, P. Rasch, S. Vavrus, M. Vertenstein, D. Bader, W. Collins, J. Hack, J. Kiehl, and S. Marshall, "The Community Earth System Model: a framework for collaborative research," *Bulletin of the American Meteorological Society*, vol. 94, pp. 1339–1360, 2013.
- [2] J. Woodring, S. M. Mniszewski, C. M. Brislaw, D. E. DeMarle, and J. P. Ahrens, "Revisiting wavelet compression for large-scale climate data using JPEG2000 and ensuring data precision," in *IEEE Symposium on Large Data Analysis and Visualization (LDAV)*, D. Rogers and C. T. Silva, Eds. IEEE, 2011, pp. 31–38.
- [3] N. Hübbe, A. Wegener, J. M. Kunkel, Y. Ling, and T. Ludwig, "Evaluating lossy compression on climate data," in *Proceedings of the International Supercomputing Conference (ISC '13)*, 2013, pp. 343–356.
- [4] A. Baker, H. Xu, J. Dennis, M. Levy, D. Nychka, S. Mickelson, J. Edwards, M. Vertenstein, and A. Wegener, "A methodology for evaluating the impact of data compression on climate simulation data," in *Proceedings of the 23rd International Symposium on High-performance Parallel and Distributed Computing*, ser. HPDC '14, 2014, pp. 203–214.
- [5] M. Kuhn, J. Kunkel, and T. Ludwig, "Data compression for climate data," *Supercomputing frontiers and innovations*, vol. 3, no. 1, pp. 75–94, 2016.
- [6] C. S. Zender, "Bit grooming: statistically accurate precision-preserving quantization with compression, evaluated in the netcdf operators (nco, v4.4.8+)," *Geoscientific Model Development*, vol. 9, no. 9, pp. 3199–3211, 2016. [Online]. Available: <http://www.geosci-model-dev.net/9/3199/2016/>
- [7] A. H. Baker, D. Hammerling, S. A. Michelson, H. Xu *et al.*, "Evaluating lossy data compression on climate simulation data within a large ensemble," *Geoscientific Model Development*, vol. 9, no. 12, 2016.
- [8] A. H. Baker, H. Xu, D. M. Hammerling, S. Li, and J. P. Clyne, "Toward a multi-method approach: Lossy data compression for climate simulation data," in *International Conference on High Performance Computing*. Springer, 2017, pp. 30–42.
- [9] J. Nardi, N. Feldman, A. Poppick, A. Baker, and D. Hammerling, "Statistical analysis of compressed climate data," National Center for Atmospheric Research, Tech. Rep. NCAR/TN-547+STR, 2018.
- [10] J. Kay, C. Deser, A. Phillips, A. Mai, C. Hannay, G. Strand, J. Arblaster, S. Bates, G. Danabasoglu, J. Edwards, M. Holland, P. Kushner, J.-F. Lamarque, D. Lawrence, K. Lindsay, A. Middleton, E. Munoz, R. Neale, K. Oleson, L. Polvani, and M. Vertenstein, "The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability," *Bulletin of the American Meteorological Society*, vol. 96, 2015.
- [11] P. Lindstrom, "Error distributions of lossy floating-point compressors," in *Joint Statistical Meetings 2017*, October 2017.
- [12] S. Di and F. Cappello, "Fast error-bounded lossy hpc data compression with sz," in *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2016, pp. 730–739.
- [13] D. Tao, S. Di, Z. Chen, and F. Cappello, "Significantly improving lossy compression for scientific data sets based on multidimensional prediction and error-controlled quantization," in *2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2017, pp. 1129–1139.
- [14] P. Lindstrom, "Fixed-rate compressed floating-point arrays," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2674–2683, 2014.
- [15] J. Guinness and D. Hammerling, "Compression and conditional emulation of climate model output," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 56–67, 2018.