2021 7th International Workshop on Data Analysis and Reduction for Big Scientific Data (DRBSD-7) | 978-1-7281-8672-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/DRBSD754563.2021.00011

Exploring Lossy Compressibility through Statistical Correlations of Scientific Datasets

David Krasowska*, Julie Bessac†, Robert Underwood‡, Jon C. Calhoun*, Sheng Di†, and Franck Cappello†
* *Holcombe Department of Electrical and Computing Engineering*, *Clemson University* Clemson, USA
† *Mathematics and Computer Science Division Argonne National Laboratory* Lemont, USA
‡ *School of Computing*, *Clemson University* Clemson, USA

***Abstract*—Lossy compression plays a growing role in scientific simulations where the cost of storing their output data can span terabytes. Using error bounded lossy compression reduces the amount of storage for each simulation; however, there is no known bound for the upper limit on lossy compressibility. Correlation structures in the data, choice of compressor and error bound are factors allowing larger compression ratios and improved quality metrics. Analyzing these three factors provides one direction towards quantifying lossy compressibility. As a first step, we explore statistical methods to characterize the correlation structures present in the data and their relationships, through functional regression models, to compression ratios. We observed a relationship between compression ratios and several statistics summarizing the correlation structure of the data, which is a first step towards evaluating the theoretical limits of lossy compressibility used to eventually predict compression performance and adapt compressors to correlation structures present in the data.**

***Index Terms*—compression, lossy compression, high performance computing, statistical correlation analysis**

I. Introduction

Scientific research increasingly uses error-bounded lossy compressors to achieve greater compression ratios in relation to lossless compressors [1]. This improved performance allows applications to run with larger and more frequently produced datasets due to faster I/O times and smaller I/O volumes. The theoretical limit on compressibilty of data using lossless compression is given by the entropy [2]. The entropy quantifies the information content present in a symbol from a source sequence based upon its probability of occurrence. Thus, for a given sequence of symbols the entropy enables computing the minimum number of bits, on average, needed to represent the data. For over 70 years, this concept has guided the development and evaluation of lossless compression algorithms. However, for lossy compression algorithms, there is currently no known bound for the maximum degree of compression that can be achieved for some specified point-wise error bound regardless of the compressor at stake. Establishing this compressor-free bound will allow researchers to anticipate compression performance alleviating manual assessments. This bound can be used to evaluate with respect to this compressor-free roofline and adapt existing compressors to correlation structures of the data ensuring they get the best compression ratio possible. Eventually, establishing the limit for lossy compression allows for the maximum efficiency for storing large scientific datasets.

One possible direction to establish an *entropy-like* bound for lossy compression is studying the impact on compressibility of correlation structures (correlations in space, time, or other dimensions) of the data, compressor types and error bounds. We refer to "compressibility" as the maximum compression ratio associated with a given error bound. As many compressors implicitly or explicitly exploit correlation structures of the data (e.g. SZ [3], ZFP [4]), analyzing the relationships between correlation structures and compression performances will allow researchers to anticipate compression performances. Ultimately, we seek to establish an entropy-like metric for lossy compression algorithms which can guide the lossy compression community to optimal development and usage by adapting compressors to correlation structures present in the data.

In this work, we focus on correlations in datasets and their link to compression ratios for several compressors. The goal of the research paper is to explore:

1) statistical methods to characterize the correlation structures of the data and
2) their relationships, through functional models, to compression ratios.

These models will form the first step into evaluating the theoretical limits of lossy compressibility used to eventually predict compression performance. In particular, we focus on estimated correlation ranges through variograms and its effect on compressibility. The variogram is commonly used in geostatistics to estimate second-order dependence and more precisely how data are correlated with distance. It can be applied to regularly and irregularly spaced data. Under stationary conditions, the variogram and covariance function have a direct correspondence. However, in practice the variogram can be estimated under more relaxed assumptions than the covariance and thus is preferred by practitioners. To characterize compressibility, we use the compression ratio, which is an important statistic within lossy compression due to its informing link to storing and processing as much data as possible.

In this study, we focus on 2D-gridded datasets and through variogram analysis, perform a characterization of the correlation between grid-points along with a compression analysis (Section IV-C) with several compressors (Section IV-B). We

978-1-7281-8672-6/21/$31.00 ©2021 IEEE
DOI 10.1109/DRBSD754563.2021.00011

consider two types of data (Section IV-A), synthetic datasets consisting of correlated stochastic Gaussian fields with known correlation structures and another one consisting of simulations from a hydrodynamical model Miranda [5].

II. Background

A. Compressors

SZ [3], ZFP [4], and MGARD [6] are some of the leading error bounded lossy compressors. In this section we explain from an algorithmic perspective how these compressors exploit correlations in data.

SZ scans through the data block by block, with a block size of 16×16 for 2D data. For each block, a prediction is made mimicking the data in each block via the Lorenzo predictor or the regression predictor. The Lorenzo predictor uses the neighboring points to estimate the value at the current position. The regression predictor fits a hyper-plane through the block, and uses the fitted hyper-plane to interpolate the values within each block. If these predictions are linearly quantized, and with sufficient accuracy according to the error bound, the quantized values are stored. Otherwise, the values are stored exactly. Finally, the entire sequence is passed first through a Huffman encoding, and then through the Zstd lossless compressor to exploit patterns in the quantized sequence. However, since the predictor does not observe values outside of its block, it cannot exploit global correlation structures easily.

ZFP likewise uses local correlations, but uses a different compression principle based on near orthogonal transforms – a similar approach to JPEG image compression. ZFP first partitions 2D data into blocks of size 4×4, then it converts each block of floating point data into a common fixed point representation, performs the near orthogonal transform, applies an embedded encoding that orders bits from most significant to least significant and finally truncates to archive a desired tolerance. Again, since blocks are compressed independently, the compressor cannot acquire a global knowledge of the data correlation structures.

MGARD, however, is a newer compressor and uses an approach that can account for global correlation structures. MGARD relies on the mathematical theory of multi-grid methods in order to compress the data. It operates by decomposing the data into multi-level coefficients which represent recursively defined sub-regions until the block is represented within the allowed tolerance. These multi-level coefficients are then quantized and compressed with either Zlib (in older versions) or Zstd (in the newest unreleased version). Because these multi-level coefficients can represent regions of differing sizes and potentially the entire dataset, MGARD can capture multi-level effects in a way that SZ and ZFP cannot, making it an important comparison for our paper.

B. Variogram

The variogram is a function describing the degree of dependence of a correlated spatial field [7], [8], it gives a measure of how much two points of the same field correlate depending on the distance h between those points. The variogram is a function of the distance h and is typically characterized by several parameters: nugget (microscale variability), sill (variance of the studied field), range (distance at which autocorrelation vanishes), see Figure 1. In the following, we focus solely on the variogram range a as it corresponds to the distance (h) where the variogram (γ) plateaus (at the sill value) and indicates the distance beyond which the spatial correlation among grid-points vanishes. Intuitively, the larger the range is the stronger the correlation is across grid-points.

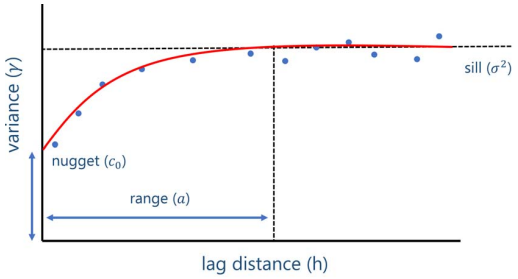

Fig. 1. Example of a variogram as a function of the distance h between points. [9]

In practice, the empirical semi-variogram is computed on the data via Equation (1):

$$\gamma(h) = \frac{1}{2N(h)} \sum_{|x_i - x_j| = h}^{N(h)} \left(z(x_i) - z(x_j)\right)^2, \qquad (1)$$

where z is the studied field of interest (e.g. *velocityx* from the Miranda dataset), x_i and x_j are grid-points coordinates or indexes, $N(h)$ is the number of points at distance h from each other. The variogram corresponds to $2\gamma(h)$, in practice the terms semi-variogram and variogram are used interchangeably. In the following, we will refer to variogram as $\gamma(h)$. Note that this formula is valid in the general context of datasets that are accompanied by coordinates or for which coordinates could be attributed that represent a notion of proximity (structured meshes, unstructured meshes or even irregularly sampled spatial points). Finally, to estimate the variogram range a we fit by least-squares a parametric squared-exponential variogram ($\gamma(h) = c_0(1 - \exp(-h^2/a))$) to the empirical variogram range estimated via Equation (1).

C. Complex correlation patterns

Statistical tools exist beyond the variogram to quantify and extract complex correlation structures of datasets for instance in order to analyze long-range dependencies [10], to detect change-point in time series [11], to cluster and identify correlation regimes [12]. Identifying multiscale components of scientific datasets mostly relies on eigen or a basis-function decompositions such as singular value decomposition (SVD) or wavelet decomposition [13], [14]. However due to the often extreme complexity of correlations and dependencies in data, developing methods to extract spatial and spatio-temporal heterogeneity or non-stationarity is still an on-going research. Although the detailed use of these techniques is outside the scope of the current work and left for future efforts, we provide

preliminary results and future work directions with SVD at the end of the Section V.

III. Related Work

Beyond the classical efforts to estimate compressibility using entropy, there has been relatively little attention afforded to the topic of lossy compressibility. As a parallel in data reduction techniques, [15] investigated the determination of thresholds for singular value decomposition of large matrices based on some optimality loss criteria. [16] identified several factors that affect compression ratios for SZ and ZFP. They relied on a block-based sampling approach which considered individual data points tailored to each compressor that they consider and used Gaussian models to estimate the subsequent compressibility. For SZ, they considered a number of predictors such as the number of elements in the dataset, the quantization interval, information about the Huffman tree constructed, and the number of points that are unpredictable by SZ's predictor. For ZFP, they offered a proof of their sampling methods estimated accuracy and empirically show it to be 99% accurate across many datasets. In the same vain, [17], based on their prior work, used deep neural networks to estimate the compression ratio instead of a Gaussian model. However the built neural network may not generalize to other applications and could over-learn the training data while the testing data would not prevent the over-learning [18]. [19] designed an automated methodology to switch between SZ and ZFP based on which compressor is estimated to have a greater compression ratio. Compression ratios are estimated in a block-based sampling approach using Shannon entropy [2] of the sampled quantized blocks to investigate SZ's behavior. Most of these works have limited generalizability because of their reliance on algorithmic details of each studied compressor or reduction technique.

Finally, little effort has been directed to explore explicit links of correlation structures in the data to reduction and compression techniques and their performance. [20] has explored the decorrelation efficiency of specific reduction methods on scientific datasets in order to identify trade-offs for parameterization. [21] lead an evaluation and comparison of several lossy reduction techniques that are based on basis-function decompositions adapted to temporal and spatial data. [22] developed an adaptive hierarchical geospatial field data representation (Adaptive-HGFDR) based on blocked hierarchical tensor decomposition to exploit multidimensional correlations of the data. However, none of these works systematically investigate the explicit link between correlation structures and compressibility. Our work goes beyond these approaches to consider a direction that is compressor independent by looking only at local spatial relationships in the data. Contrary to [16], [18], [19], in the following study as an initial step to establish a compressibility limit, we do not assess the computation overhead of our methodology.

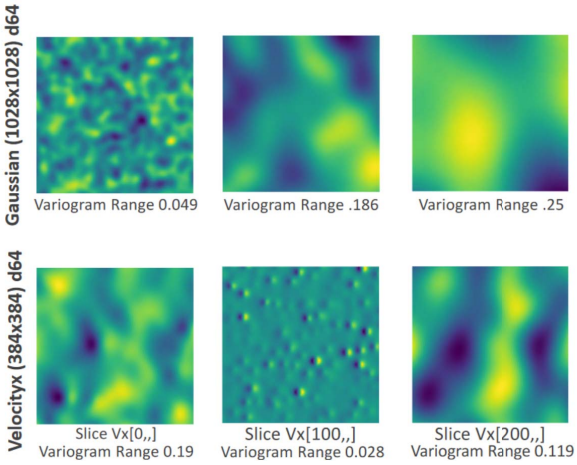

Fig. 2. Original images of 2D Gaussian fields (top) and Miranda dataset 2D-slices (bottom)

IV. Methodology

A. Datasets

In the following, we refer to datasets as a particular field at a particular time when in an application (e.g. a single continuous variable in memory). In particular, we work with a single temporal snapshot of the studied data.

The first dataset consists of synthetic 2D Gaussian fields with a controllable correlation structure following a squared-exponential correlation model. These 2D fields are 1028×1028 grid-points. We consider these fields as "ideal" as the correlation range is known and varied to create multiple correlated fields. Gaussian fields z over a grid defined by indexes x_i are generated using the following probability distribution f

$$f(z(x_1),\ldots,z(x_k)) = \frac{\exp\left(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}(\mathbf{x})^{-1}(\mathbf{z}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}(\mathbf{x})|}} \tag{2}$$

with $\mathbf{z} = (z(x_1),\ldots,z(x_k)) \in \mathbb{R}^k$, the mean $\boldsymbol{\mu} = 0 \in \mathbb{R}^k$ in this study, and a squared-exponential correlation $\boldsymbol{\Sigma}(x_i, x_j) = \sigma^2 \exp(-|x_i - x_j|^2/a^2)$, where the variance σ^2 is set to 1, a is the correlation range that is known and varied, and x_i are spatial grid-points of the 2D field images. In our evaluation, we consider two types of synthetic Gaussian datasets: single correlation range fields and multiple correlation range fields. We consider single correlation range Gaussian fields as a proof of concept that gives us a high degree of control over correlation structures in the data. However, because application datasets are likely to exhibit complex correlation structure as multiple correlation ranges, we consider also multiple range Gaussian fields. For the multi-range correlation, we generate Gaussian fields with two distinct correlation ranges contributing equally to the total field. This provides a controlled case with increased complexity.

The final dataset is generated from the Miranda [5] code, designed for hydrodynamical large turbulence simulations. These data are more complex than the Gaussian fields due to multiple correlation ranges and complex dependencies. These original 3D data dimensions are $256\times384\times384$. The

Software	Version	Purpose
ZFP [3]	@2.1.11.1	lossy compressor
ZFP [4]	@0.5.5	lossy compressor
MGARD [6]	@0.1.0	lossy compressor
gstat [23]	@2.0-7	obtain variogram range
numpy [24]	@1.21.1	polyfit function to graph the curves
Libpressio [25]	@0.70.0	compress and measure the data

TABLE I
COMPRESSORS AND SOFTWARE USED FOR THE STUDY

3D datasets are split into separate 2D slices based on equally spaced slices along the first dimension. In this paper, we use slices of the *velocityx* variable as shown in Figure 2.

B. Compressors and Software

Each lossy compressor in Table I is run with the following absolute error bounds: 1E-5, 1E-4, 1E-3, and 1E-2. We choose the software and compressor versions from the latest available on Spack from a selection of leading error-bounded lossy compressors. We use the absolute error bound because it is supported by each of the considered error-bounded lossy compressors. Additionally, there are formal equivalences between the absolute error bound mode and other error bounds modes such as the value range relative error bound mode which are used by compressors such as SZ [3]. Therefore, considering the compressiblity given an absolute error bound is generalizable to other kinds of bound. All experiments are run on Clemson's Palmetto cluster using a node with two 32 core Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz and 384 GB of RAM. The OS is Linux CentOS 8 with compiler GCC 8.4.1. Additional software and packages used in the study are listed in Table I.

C. Compression Statistics and Statistical Methods

1) Compression Statistics: Compression ratio, the ratio of the uncompressed data size by the compressed data size, is used to compare the different compressors and their efficiency. Compression ratio depends on: error bound, compressor used, and correlation structures within the data. The compression ratio is comparable between different compressors and error bounds. In the following, this quantity is computed on the studied datasets from Section IV-A for different compressors and error bounds, and investigated as function of a measure of several correlation statistics of data computed through the variogram range described below.

2) Variogram study: In the following section, we compute the empirical variogram of each 2D data-slice from the datasets described in Section IV-A and based on the Euclidean distance between grid-points. The corresponding range a is then estimated and reported in the following section as estimated variogram range. In particular, we estimate the variogram ranges on the entire 2D field in order to assess the overall correlation structure of the fields. However, this is insufficient to characterize local heterogeneity in datasets, hence we compute the variogram ranges in windows of a given size that cover the entire 2D field in a tiled fashion [26]. More specifically, we compute and report the standard deviation of variogram ranges estimated over the windows covering the entire field. This statistic provides information on the spatial diversity and spread of local correlations present in the data.

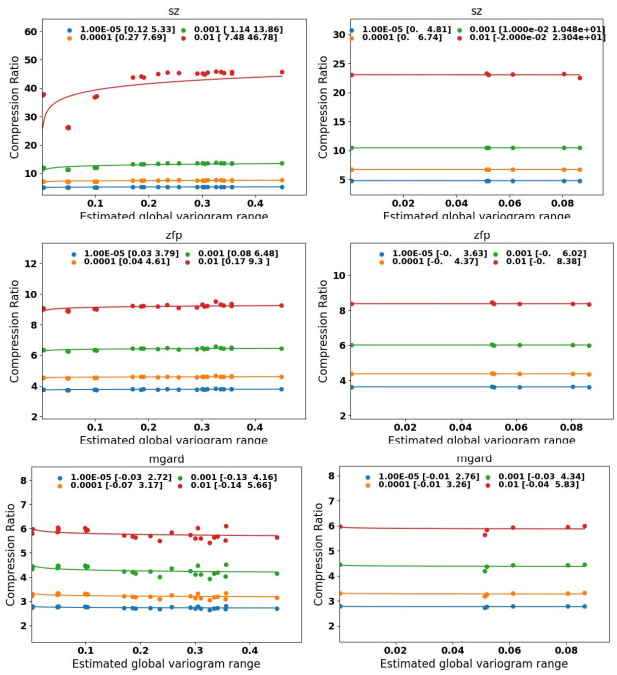

Fig. 3. Compression ratios against estimated variogram range for 2D Gaussian fields with single correlation range (left) and multi-range correlation (right). Each color is associated with an error bound. Empirical calculations are depicted with dots and fitted logarithmic regressions are in solid lines. Estimated logarithmic regression coefficients α and β are given in the legend.

V. EXPERIMENTAL RESULTS

A. Compressibility and global correlation

In the following figures, the variogram range estimated on entire 2D fields is referred to as "Estimated global variogram range", the standard deviation of locally estimated variogram range on 32×32 windows is denoted as "Std estimated of local variogram range (H=32)". Finally, an additional statistic is considered to illustrate future research directions. It consists in the standard deviation of locally extracted SVD thresholds on 32×32-windows. Thresholds correspond to the number of required singular modes to recover 99% of the variance of the initial field. This statistic is referred to as "Std of truncation level of local SVD (H=32)".

In Figures 3 and 4 , the compression ratios for SZ [3] and ZFP [4] (top panels) are plotted against the variogram range estimated on entire 2D fields. These statistics are respectively estimated on the synthetic Gaussian fields (left: single-range correlation and right: multi-range correlation) and Miranda datasets. As the estimated global variogram range increases indicating stronger dependence between spatial points, dataset variability decreases yielding smoother, more compressible data. This increasing relationship between compression ratio and variogram range exhibits a plateau for highly correlated data (large variogram ranges) suggesting a limit in compressibility of the data for a given error bound and compressor. Note that this trend is less visible on the multi-range correlation

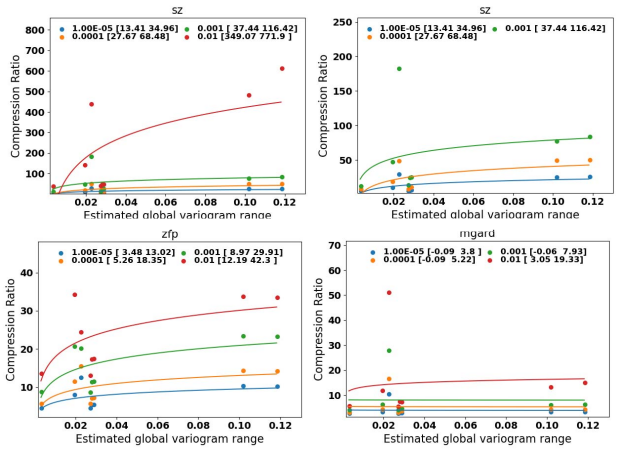

Fig. 4. Compression ratios against estimated variogram range for Miranda *velocityx*. Empirical calculations are depicted with dots and fitted logarithmic regressions are in solid lines. Each color is associated with an error bound. Estimated logarithmic regression coefficients α and β are given in the legend. Due to the large spread of compression ratios for SZ (top left panel), the top right panel corresponds to results for error bounds strictly below 1E-2 in order to ease the reading.

Gaussian fields (right column of Figure 3) due to the equal contributions of each correlation to the total field, preventing any dominant correlation pattern to prevail and thus to be characterized properly by the variogram range of the entire field. Compression ratios for MGARD [6] are less sensitive to the dependencies to correlation ranges present in the data which is likely due to the global scope of the compressor.

Finally, in order to quantify these relationships and compare them across different compressors and error bounds, the compression ratios are fitted by least-squares to logarithmic regressions of the estimated variogram range a: $CR = \alpha + \beta \log(a) + \epsilon$, where CR is the compression ratio, a the estimated variogram range, and α and β are estimated coefficients and reported in the legend box of each panel. The fitted logarithmic regressions show a good match to the datapoints indicating in most cases a logarithmic dependence of the compression ratios to the estimated variogram ranges. Lower compression error bounds exhibit lower variance of datapoints around the fitted curve and fewer outliers. The outliers present in the top-left corner of Figure 4 are due to datasets having similar variogram ranges however different variances and hence different compression ratios. This suggests the need to add information to the variogram range as an exploratory variable. Regressions fit better the single-scale correlation Gaussian fields (Figure 3 left column) than multi-range correlation Gaussian fields and the Miranda data due to their lower complexity that is captured reasonably well by the global variogram range. In particular, the fitting on datapoints from Miranda data show more dispersion around the fitted curves but a matching trend.

B. Compressibility and local correlation

The variogram range estimated on each entire 2D field represents an average correlation range observed on the entire field and is not suited to characterize spatial local heterogeneity nor multiple scales that can be present in complex scientific datasets, such as in Miranda datasets or multi-range Gaussian fields. Hence, we estimate the variogram ranges on local windows (32×32) tiling each entire 2D-field. Additionally, exploring local heterogeneity is important since many error-bounded lossy compressors exploit some notions of locality in their algorithms as discussed in Section II-A. In Figure 3, as the global variogram range increases, the compression ratio fails to have a slope of greater than absolute value of 1E-2 for mutli-range Gaussian fields (right column). This illustrates the shortcomings of the global variogram range statistic while dealing with multi-range correlation datasets. Figure 5 shows the compression ratios computed on single-range correlation Gaussian fields (left) and multi-range correlation fields (right) as a function of the local variogram ranges and the left column of Figure 7 shows the statistics computed on the Miranda fields. Both figures corroborate and illustrate the benefit of considering local information for complex datasets to better explain compression ratios by local variograms rather than global ones. In particular, we observe that Figures 3 and 4 depict several close values of compression ratios for close variogram ranges, whereas this effect is less visible in Figures 5 and 7 indicating a stronger explanatory skill of the local statistic to the compression ratios. However, results for the single-range correlation Gaussian fields show a weaker sensitivity of the compression ratios to this local statistic. This might suggest the need to use several statistics to provide

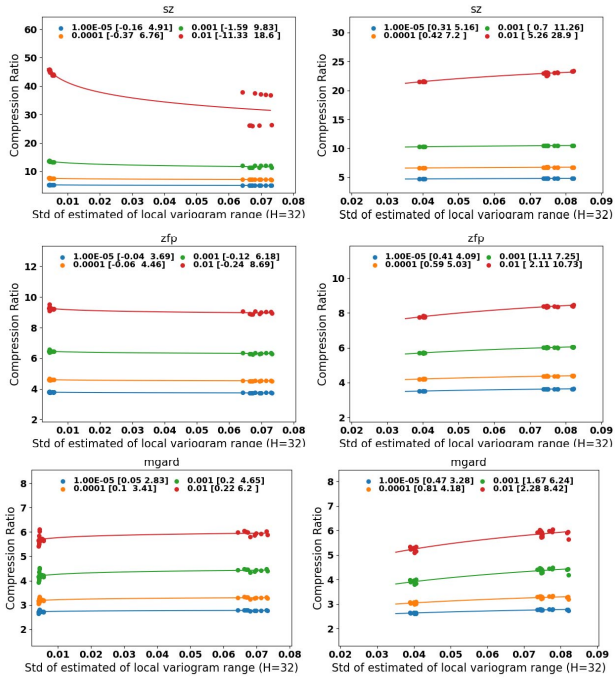

Fig. 5. Compression ratios against standard deviation of the local variogram range for single range correlation Gaussian fields (right) and for multi-range correlation Gaussian fields (right). Initial step towards the use of decomposition techniques to analysis fields with multiscale patterns. Empirical calculations are depicted with dots and fitted logarithmic regressions are shown with solid lines. Each color is associated with an error bound. Estimated logarithmic regression coefficients α and β are given in the legend.

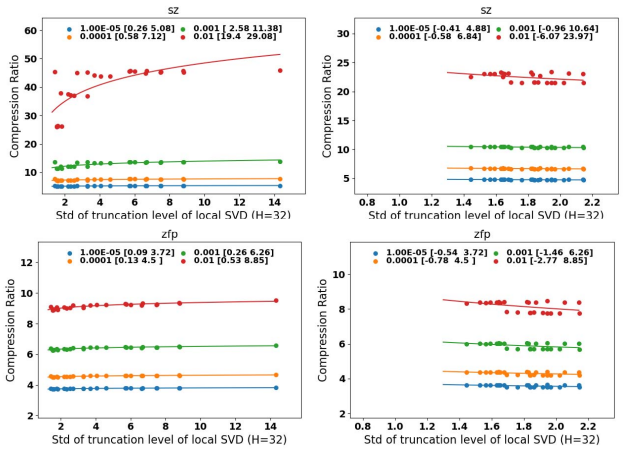

Fig. 6. Compression ratios against standard deviation of the truncation level of local SVD for single scale Gaussian fields (left) and for multiscale Gaussian datasets (right). Empirical calculations are depicted with dots and fitted logarithmic regressions are shown with solid lines. Each color is associated with an error bound. Estimated logarithmic regression coefficients α and β are given in the legend.

appropriate explanatory skills. Future work will explore this path.

C. Towards multiscale analysis and summary statistics

Figure 6 and the right column of Figure 7 illustrate an initial step towards the use of decomposition techniques to analyze fields with multiscale patterns. In this section, local SVDs are performed on the fields and summarized via the standard deviation of locally required numbers of singular modes to capture 99% of the variance of each local window. This local statistic, larger values of required singular modes are associated with less compressible fields so we expect mostly decreasing relationship of compression ratios to this local statistic. For brevity, as this section provides insights to future works, we do not show results for MGARD as it exhibits less sensitivity to the previous statistics and a weak link to the current statistic. Figure 6 and the right column of Figure 7 indicate that this local statistic provide a more diverse representation of the data as seen by the span of unique values over the x-axis, than the two other statistics based on variogram. This statistic tends to exhibit several relating trends to compression ratio, highlighting a need to refine it. Future work will explore transformation of this statistic or other representations based local SVDs and variograms, as both methods provided different and valuable information to explain compression ratios.

VI. Conclusions and Future Work

Establishing the theoretical and compressor-free limit for lossy compression would allow for the maximum efficiency for compressing and storing large scientific datasets. Our work represent a first step toward this goal. We have demonstrated that estimated global and local variogram ranges can explain compression ratio in a logarithmic fashion for some compressors and given error bounds. This hypothesis was illustrated on synthetic Gaussian fields providing a proof of concept

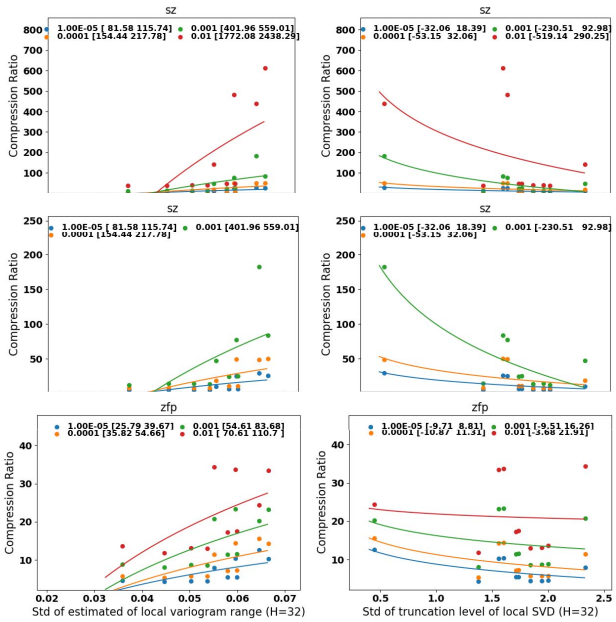

Fig. 7. Compression ratios against local statistics, standard deviation of the local variogram range (left) and standard deviation of the truncation level of local SVD (left) for Miranda *velocityx* 2D field. Empirical calculations are depicted with dots and fitted logarithmic regressions are shown with solid lines. Each color is associated with an error bound. Estimated logarithmic regression coefficients α and β are given in the legend. Due to the large spread of compression ratios for SZ, the central panels correspond to results for error bounds strictly below 1E-2 in order to ease the reading.

and corroborated by results on a user scientific dataset from Miranda. With the studied datasets, SZ and ZFP seem to utilize the global and local spatial correlation ranges in a logarithmic fashion with various coefficients showing the strength of the dependence. MGARD seems less sensitive to these trends in its compression ratios.

Since heterogeneous (non-stationary) and multiscale correlations in the data may be mis-represented by the global spatial variogram, other statistics, including local variograms and local SVD, have been studied to address these issues. In continuation of this research, there are a few goals for the future: i) explore more complex dependent variables (local correlation combined with multiscale statistics based on decomposition) as candidate predictors, ii) create more complex synthetic multiscale 2D Gaussian fields, iii) test the robustness of the proposed statistics and the method on other datasets, in particular from SDRBench, and iv) create a model of compression ratio based on correlation metrics and error bound. Future work will investigate the effects of correlation structures on quality metrics of reconstructed data such as PSNR along with a design of the statistics to a 3D context.

Another aspect for future work is how to quickly assess this metric. The current implementation relies on the singular value decomposition which is slow relative to modern compressors. We plan to leverage a sampling approach similar to prior work [16], [19]. We are hopeful that increasing levels of sampling by block can provide an increasingly accurate proxy for our metric.

ACKNOWLEDGMENTS

Clemson University is acknowledged for generous allotment of compute time on the Palmetto cluster. This material is based upon work supported by the National Science Foundation under Grant No. SHF-1910197, No. SHF-1617488. This material is based upon work supported in part by the Exascale Computing Project (17-SC-20-SC) of the U.S. Department of Energy (DOE), and by DOE's Advanced Scientific Research Office (ASCR) under contract DE-AC02-06CH11357.

REFERENCES

[1] F. Cappello, S. Di, S. Li, X. Liang, A. M. Gok, D. Tao, C. H. Yoon, X.-C. Wu, Y. Alexeev, and F. T. Chong, "Use cases of lossy compression for floating-point data in scientific data sets," *The International Journal of High Performance Computing Applications*, vol. 33, no. 6, pp. 1201–1220, Jul. 2019. [Online]. Available: https://doi.org/10.1177/1094342019853336

[2] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, Jul. 1948. [Online]. Available: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[3] X. Liang, S. Di, D. Tao, S. Li, S. Li, H. Guo, Z. Chen, and F. Cappello, "Error-controlled lossy compression optimized for high compression ratios of scientific datasets," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, Dec. 2018. [Online]. Available: https://doi.org/10.1109/bigdata.2018.8622520

[4] P. Lindstrom and M. Isenburg, "Fast and efficient compression of floating-point data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 1245–1250, Sep. 2006. [Online]. Available: https://doi.org/10.1109/tvcg.2006.143

[5] F. Capello, K. Zhao, S. Di, D. Tao, J. Bessac, and Z. Chen, Jun 2018. [Online]. Available: https://sdrbench.github.io/

[6] M. Ainsworth, O. Tugluk, B. Whitney, and S. Klasky, "Multilevel techniques for compression and reduction of scientific data—the multivariate case," *SIAM Journal on Scientific Computing*, vol. 41, no. 2, pp. A1278–A1303, Jan. 2019. [Online]. Available: https://doi.org/10.1137/18m1166651

[7] G. Matheron, "Principles of geostatistics," *Economic geology*, vol. 58, no. 8, pp. 1246–1266, 1963.

[8] A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes, *Handbook of spatial statistics*. CRC press, 2010.

[9] E. Kim, "Spatial simulation 1: Basics of variograms," Feb 2019. [Online]. Available: https://aegis4048.github.io/spatial-simulation-1-basics-of-variograms

[10] P. Abry, D. Veitch, and P. Flandrin, "Long-range dependence: Revisiting aggregation with wavelets," *Journal of Time Series Analysis*, vol. 19, no. 3, pp. 253–266, 1998.

[11] J. Cabrieto, F. Tuerlinckx, P. Kuppens, M. Grassmann, and E. Ceulemans, "Detecting correlation changes in multivariate time series: A comparison of four non-parametric change point detection methods," *Behavior research methods*, vol. 49, no. 3, pp. 988–1005, 2017.

[12] J. Bessac, P. Ailliot, J. Cattiaux, and V. Monbet, "Comparison of hidden and observed regime-switching autoregressive models for (u,v)-components of wind fields in the Northeast Atlantic," *Advances in Statistical Climatology, Meteorology and Oceanography*, vol. 2, no. 1, pp. 1–16, 2016.

[13] P. S. Addison, *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance*. CRC press, 2017.

[14] A. Hannachi, I. T. Jolliffe, and D. B. Stephenson, "Empirical orthogonal functions and related techniques in atmospheric science: A review," *International Journal of Climatology: A Journal of the Royal Meteorological Society*, vol. 27, no. 9, pp. 1119–1152, 2007.

[15] M. Gavish and D. L. Donoho, "The optimal hard threshold for singular values is $4\sqrt{3}$," pp. 5040 – 5053, 2014.

[16] T. Lu, Q. Liu, X. He, H. Luo, E. Suchyta, J. Choi, N. Podhorszki, S. Klasky, M. Wolf, T. Liu *et al.*, "Understanding and modeling lossy compression schemes on hpc scientific data," in *2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2018, pp. 348–357.

[17] Z. Qin, J. Wang, Q. Liu, J. Chen, D. Pugmire, N. Podhorszki, and S. Klasky, "Estimating lossy compressibility of scientific data using deep neural networks," *IEEE Letters of the Computer Society*, vol. 3, no. 1, pp. 5–8, 2020.

[18] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Elsevier, 2011.

[19] D. Tao, S. Di, X. Liang, Z. Chen, and F. Cappello, "Optimizing Lossy Compression Rate-Distortion from Automatic Online Selection between SZ and ZFP," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 8, pp. 1857–1871, Aug. 2019.

[20] P. Lindstrom, "Fixed-rate compressed floating-point arrays," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2674–2683, 2014.

[21] A. Moon, J. Kim, J. Zhang, and S. W. Son, "Lossy compression on iot big data by exploiting spatiotemporal correlation," in *2017 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 2017, pp. 1–7.

[22] Z. Yu, D. Li, Z. Zhang, W. Luo, Y. Liu, Z. Wang, and L. Yuan, "Lossy compression of earth system model data based on a hierarchical tensor with adaptive-hgfdr (v1. 0)," *Geoscientific Model Development*, vol. 14, no. 2, pp. 875–887, 2021.

[23] E. J. Pebesma, "Multivariable geostatistics in s: the gstat package," *Computers & Geosciences*, vol. 30, no. 7, pp. 683–691, Aug. 2004. [Online]. Available: https://doi.org/10.1016/j.cageo.2004.03.012

[24] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

[25] R. Underwood, S. Di, J. C. Calhoun, and F. Cappello, "Fraz: A generic high-fidelity fixed-ratio lossy compression framework for scientific floating-point data," *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2020.

[26] J. Bessac, A. H. Monahan, H. M. Christensen, and N. Weitzel, "Stochastic parameterization of subgrid-scale velocity enhancement of sea surface fluxes," *Monthly Weather Review*, vol. 147, no. 5, pp. 1447–1469, 2019. [Online]. Available: https://doi.org/10.1175/MWR-D-18-0384.1