Exploring Lossy Compressibility through Statistical Correlations of Scientific Datasets

DRBSB-7 SC21 St. Louis, MO
Sunday, November 13, 2021

Argonne NATIONAL LABORATORY

CLEMSON UNIVERSITY

David Krasowska*, Julie Bessac‡ , Robert Underwood𝕎, Jon 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
𝕎 School of Computing, Clemson University - Clemson, USA

Motivation

- Scientific research increasingly uses error-bounded lossy compressors to achieve greater compression ratios in relation to lossless compressor
- Entropy[1]: theoretical limit on compressibility of data using lossless compression
    - There is no current limit for lossy compression
- Establishing the limit for lossy compression allows for the maximum efficiency for storing large scientific datasets

[1] C. E. Shannon, “A mathematical theory of communication,”Bell SystemTechnical Journal, vol. 27, no. 3, pp. 379–423, Jul. 1948.

Goal

1. Explore statistical methods to characterize the correlation structures of the data
2. Explore 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 this presentation:

- To characterize compressibility, we use the compression ratio
- Relationship between compression ratios and statistics summarizing the correlation structure of the data

This 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

Background: Compressors

SZ

- Scans block by block, with a block size of 16 ×16 for 2D data
- Predicting the data in each block uses:
  - *Lorenzo predictor* which the neighboring points to estimate the value at the current position
  - *Regression predictor* which fits a hyper-plane through the block and uses the fitted hyper-plane to interpolate the values within each block
- Passed first through a Huffman encoding, then passed to Zstd lossless compressor to exploit patterns in the quantized sequence
- Cannot exploit global correlation structures easily
  - Does not observe values outside of its block

[1] 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.2018.862252

ZFP

- Partitions 2D data into 4×4 block size
- Compression principle based on near orthogonal transforms
  - 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
  - Truncates to achieve a desired tolerance.
- Cannot exploit global correlation structures easily
  - Does not observe values outside of its block

[1] 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

MGARD

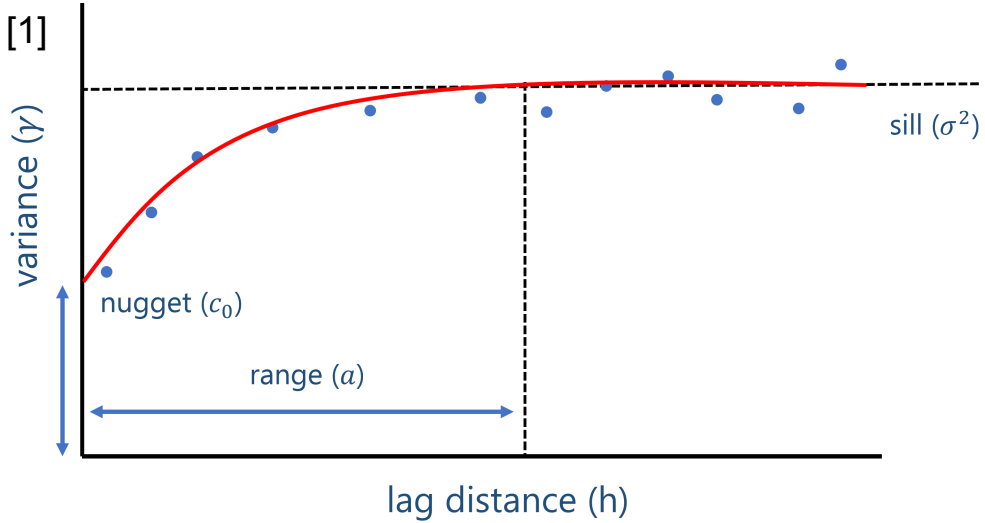
- Decomposes data into multi-level coefficients which represent:
  - Recursively defined sub-regions until block is within error bound
- Multi-level coefficients are quantized and compressed with either Zlib (older versions) or Zstd (newest unreleased version).
- Can exploit global correlation structures easily
  - Multi-level coefficients can represent regions of differing sizes

[1] 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

Background: Global Variogram

Representation of the variogram

- Range (*a)* as it corresponds to the distance (*h*) where the variogram (*γ*) plateaus
- Indicates the distance which the spatial correlation among grid-points vanishes
- The larger the range is the stronger the correlation is across grid-points



[1] https://aegis4048.github.io/spatial-simulation-1-basics-of-variograms

Formulation

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

Field	Description
z	field of interest
x_i and x_j	grid-point coordinates / indexes
N(h)	number of points at distance h from each other

- 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.

Background: Correlation Structures

- Statistical tools exist beyond the variogram to quantify and extract complex correlation structures of datasets
- Identifying multiscale components of scientific datasets mostly relies on eigen or a basis-function decompositions
  - Singular value decomposition (SVD) or wavelet decomposition [1, 2]
- Developing methods to extract spatial and spatiotemporal heterogeneity is still an on-going research
  - Due to the complexity of correlations and dependencies in data

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

[2] . 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

Related Work

- Little attention afforded to the topic of lossy compressiblity
- [1] investigated the determination of thresholds for singular value decomposition of large matrices based on some optimality loss criteria
- [2] identified several factors that affect compression ratios for SZ and ZFP
  - Complex interplay between compressor design, data features and compression performance
  - Compression ratios are estimated in a block-based sampling approach using Shannon entropy [3] of the sampled quantized blocks to investigate SZ's behavior

[1] M. Gavish and D. L. Donoho, “The optimal hard threshold for singular values is 4\√3,” pp. 5040 – 5053, 2014

[2] T. Lu, Q. Liu, X. He, H. Luo, E. Suchyta, J. Choi, N. Podhorszki,S. Klasky, M. Wolf, T. Liuet 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

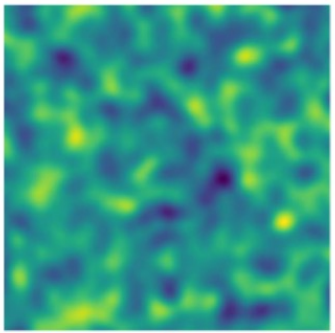
[3] C. E. Shannon, “A mathematical theory of communication,” Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, Jul. 1948.

Methodology: Datasets

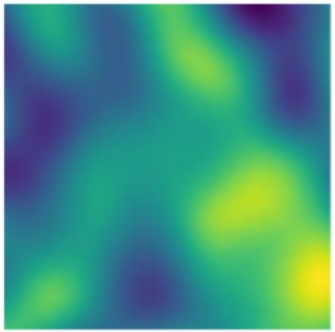
Gaussian 2D Generated Datasets

- 2D Gaussian fields with a controllable correlation structure following a squared-exponential correlation model
- We consider these fields as ‘ideal’ as the correlation range is known and varied to create multiple correlated fields
- 1028x1028 data dimensions

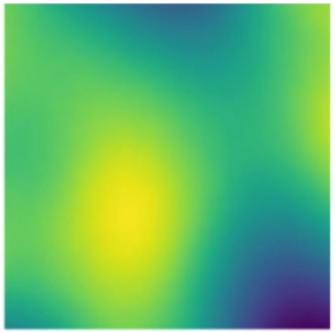
Gaussian (1028x1028) d64



Variogram Range 0.049



Variogram Range .186



Variogram Range .25

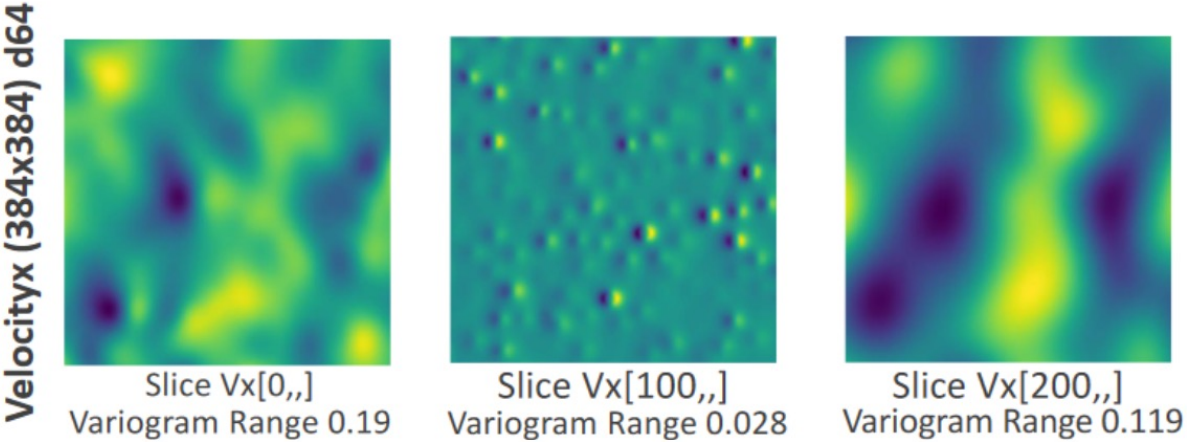
Gaussian 2D Probability Distribution

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

- $z = \left(z(x_1), \ldots, z(x_k)\right) \in R^k$
  - Gaussian fields z over a grid defined by indexes x_i
- $\mu = 0 \in R^k$
- $\Sigma(xi, xj) = \sigma^2 \exp(-|xi - xj|2/a^2)$
  - squared-exponential correlation
  - $\sigma^2 = 1$
  - a is the correlation range
  - x_i are spatial grid-points of the 2D field images

Miranda[1]

- Designed for hydrodynamical large turbulence simulations
- More complex than Gaussian fields due to multiple correlation ranges and complex dependencies
- 256x384x384 original data dimensions
  - Split along the first dimension into 384x384 slices
- *velocityx* was used for this paper



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

Methodology: Compressors and Software

Software	Version	Purpose
SZ	@2.1.11.1	lossy compressor
ZFP	@0.5.5	
MGARD	@0.1.0	
gstat	@2.0-7	obtain variogram range
numpy	@1.21.1	polyfit function to graph the curves
Libpressio	@0.70.0	compress and measure the data

- 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.

Methodology: Compression Statistics and Statistical Method

Compression Statistics: Compression ratio

- Impacted by: error bound, compressor used, and correlation structures within the data
- Comparable between different compressors and error bounds
- Computed on the studied datasets for different compressors and error bounds

Variogram Study

- Compute empirical variogram of each 2D data-slice from the datasets based on the Euclidean distance between grid-points.
- Estimate variogram ranges on the entire 2D field in order to assess the overall correlation structure of the fields
  - Corresponding range, *a*, is estimated
  - Referenced as the *estimated global variogram range*
  - Insufficient to characterize local heterogeneity
- Characterize local heterogeneity and spatial diversity
  - Compute the variogram ranges in windows of a given size that cover the entire 2D field in a tiled fashion [1].

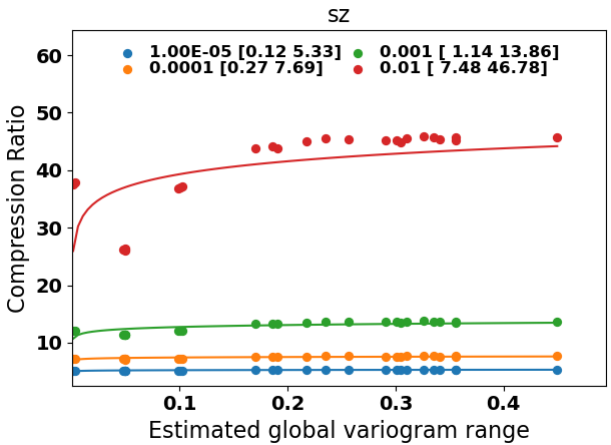
[1] J. Bessac, A. H. Monahan, H. M. Christensen, and N. Weitzel, “Stochastic parameterization of subgrid-scale velocity enhancement of sea surface fluxes,” Monthly Weather Review, vol. 147, no. 5, pp.1447–1469, 2019. [Online]. Available: https://doi.org/10.1175/MWR-D-18-0384.1

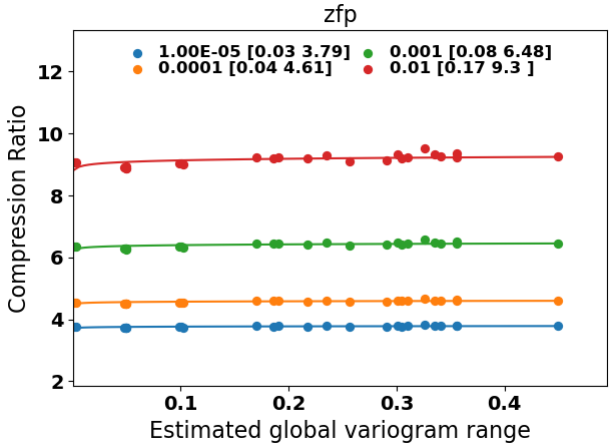
Methodology Summary

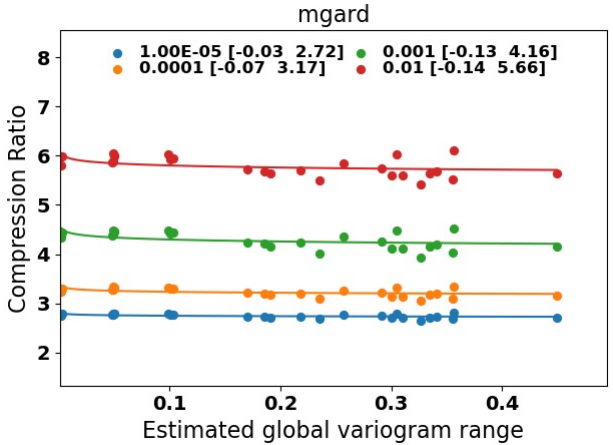
Compression ratio is investigated as function of a measure of several correlation statistics of data computed through the variogram range

Experimental Results: Compressibility and Global Correlation

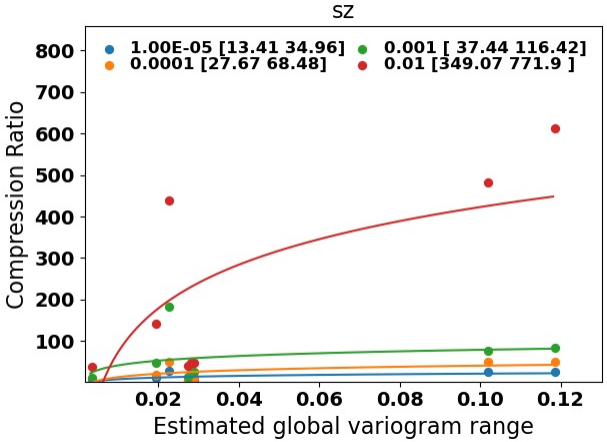
2D Gaussian fields with single correlation range

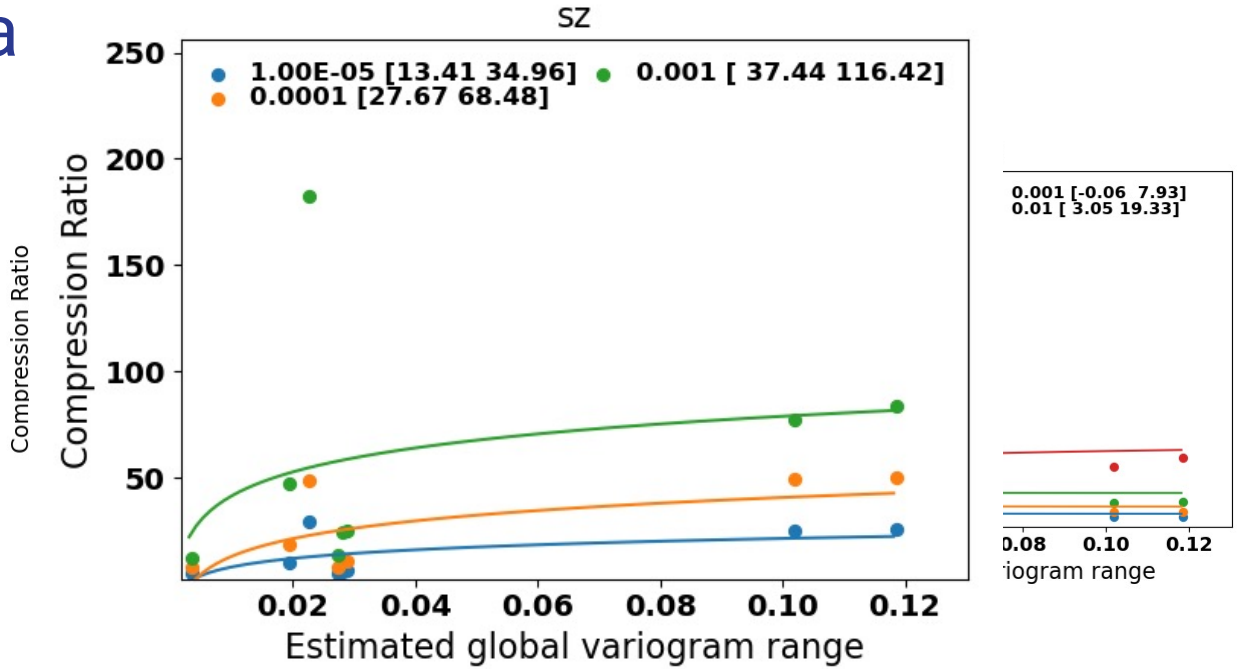




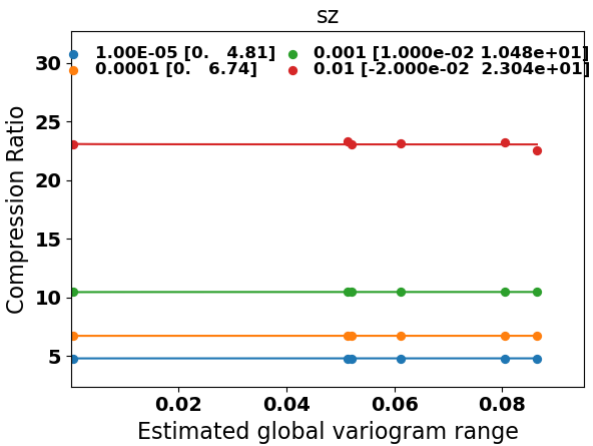


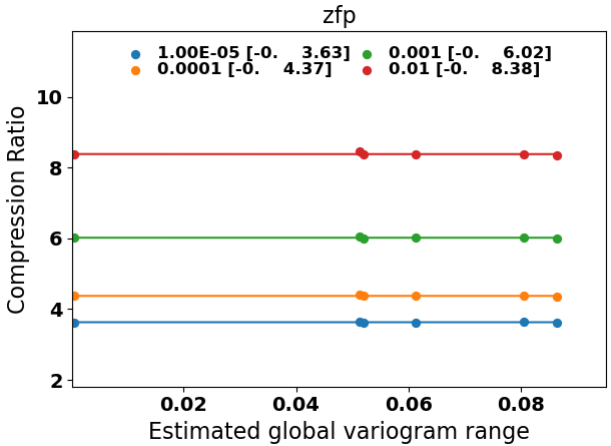
Velocityx Miranda

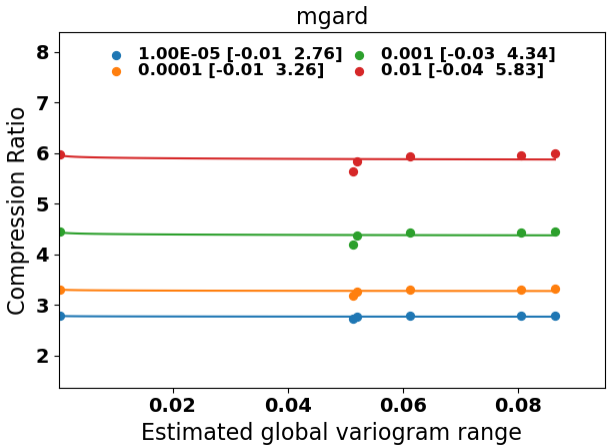




2D Gaussian fields with multi correlation range



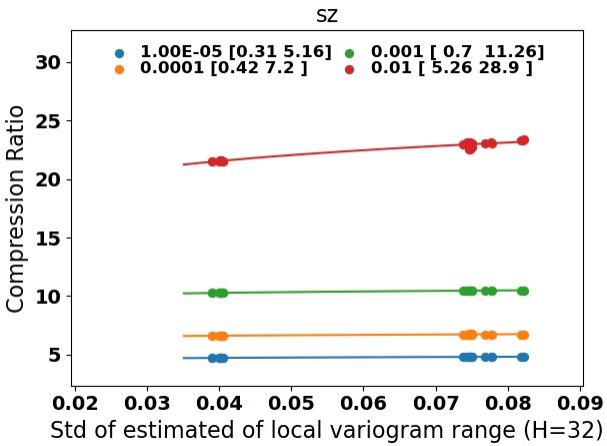


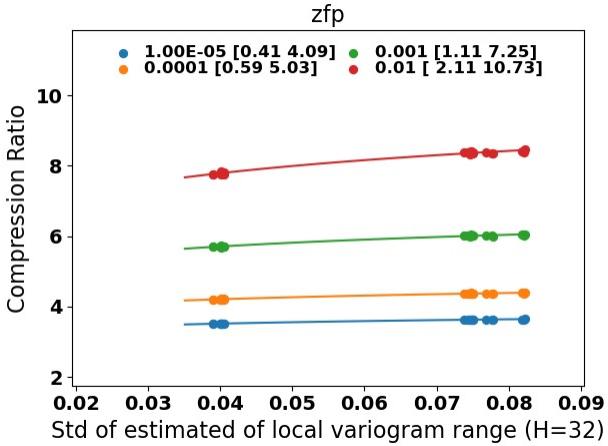


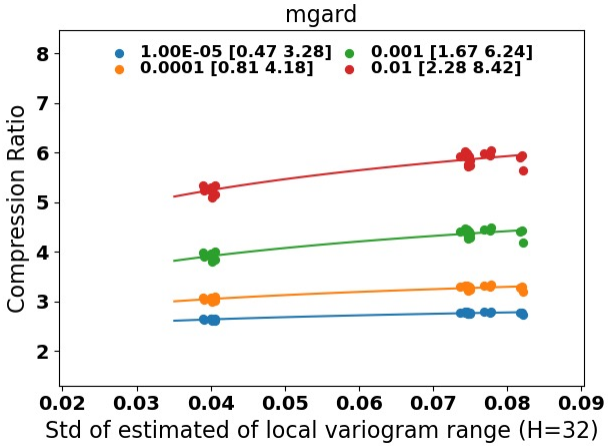
- Slope of each trendline for each compressor and bound is ~0
- Global variogram range is limited by multi-correlations
  - Indicates the need for local correlation statistical measurements

Experimental Results: Compressibility and local correlation

2D Gaussian fields with multi correlation range







- Slope of each trendline for each compressor and bound is >0
  - Characterizes spatial local heterogeneity

Conclusions and Future Work

Conclusions

- Our work represent a first step toward establishing the theoretical and compressor-free limit on lossy compressibility
- Estimated global and local variogram ranges can explain compression ratio in a logarithmic fashion for some compressors and given error bounds
  - SZ and ZFP seems to utilize the global and local spatial correlation ranges
  - MGARD seems less sensitive
- Heterogeneous (non-stationary) and multiscale correlations in the data may be mis-represented by the global spatial variogram

Future work

1. Explore more complex dependent variables (local correlation combined with multiscale statistics based on decomposition) as candidate predictors
2. Create more complex synthetic multiscale 2D Gaussian fields
3. Test the robustness of the proposed statistics and the method on other datasets
4. Create a model of compression ratio based on correlation metrics and error bound.

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.

Questions?

- Email: dkrasow@clemson.edu