Sensitivity of Black-Box Statistical Prediction of Lossy Compression Ratios for 3D Scientific Data

Alexandra Poulos[1], Jon C. Calhoun[1] (advisor), Robert Underwood[2] (advisor), Sheng Di[2] (advisor), Franck Cappello[2] (advisor)
[1]Holcombe Department of Electrical and Computer Engineering – Clemson University; [2]Mathematics and Computer Science Division – Argonne National Laboratory

BACKGROUND

Lossy compression bridges the gap between compute and I/O. Compression ratio (CR) estimation optimizes I/O workflows processing terabytes of data. Use cases such as CR auto-tuning or fast lossy compressor selection require high-throughput, accurate estimations.

Prior Sampling Approaches	Fast	Accurate
Krasowska et. al. [1]	✗	✓
Tao et. al. [2]	✓	✗
This work	✓	✓

We show that sampling a small number of moderately sized data points maintains fast data transfer and yields superior estimation accuracy when compared to existing sampling approaches. In relation to non-sampling techniques, our method results in less than 10% degradation in estimation accuracy with speedup comparable to sampling approaches.

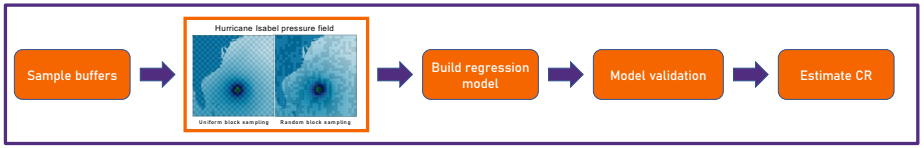
USE CASES

- (1) Auto-tuning for lossy compressor parameter optimization:: Determine a configuration of a lossy compressor that achieves a specified CR
- (2) Fast lossy compressor selection: Determine which set of compressors achieves the best CR

CONTRIBUTIONS

- Conduct a sensitivity analysis of block sampling approaches to determine their impact on CR prediction accuracy using multiple state-of-the-art lossy compressors
- Develop a linear regression model using statistical and spatial properties of dataset samples to accurately estimate CR performance
- We present a lightweight estimation method that spans the gap between runtime performance and CR estimation accuracy
- Compared to non-sampling techniques, our method results in less than ~10% degradation in estimation accuracy with similar runtime of less accurate sampling approaches.

METHODOLOGY



Proposed workflow: 1) Sample N chunks of size MxMxM from the dataset using uniform or random block sampling; 2) Calculate the CR, standard deviation, and locality metric for each sampled block and use to determine regression model; 3) Obtain confidence interval of model using k-fold validation to calculate the median average percentage error (MAPE); 4) Use CR estimations for fast lossy compressor selection, parameter optimization, etc.

MODELING CR

$$\log(CR\_est) = a + b \times \sum \sigma_i \times cr\_local_i + c \times \sum loc_i \times cr\_local_i + d \times \sum loc_i \times \sigma_i \times cr\_local_i,$$

where $loc_i = \left(\frac{\sum_j |x_{ii} - y_{ij}|}{\sqrt{\sum_j (x_{ii} - y_{ij})}}\right) / \sum_j |x_{ii} - y_{ij}|$

Linear regression model estimates global compression ratio (CR_est) using the local compression ratio (cr_local_i) and standard deviation (σ_i) of each sampled block. Locality metrics (loc_i) are calculated using the location of a sampled block relative to each other sampled block.

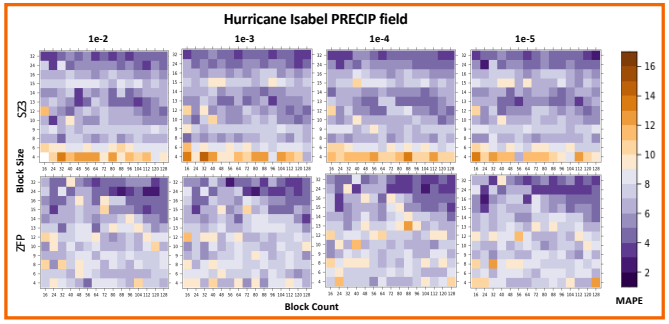
ACCURACY

Compressor	MAPE	10% quantile	90% quantile
fpzip	1.35	1.04	1.64
sz2	14.5	12.6	17.1
sz3	7.65	6.31	8.94
zfp	3.87	3.10	4.90

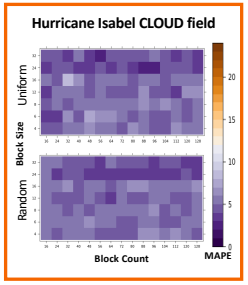
Prediction accuracy metrics for fpzip, sz2, sz3, and zfp's CR estimation. MAPE with 10% and 90% quantiles are reported for the QMCPACK dataset with a sample block size of 16.

Prior Work	MAPE	10% quantile	90% quantile
[2]	76	41	82
[4]	26	12	58
Block size = 24	4.2	2.5	6.8
Block size = 32	3.9	2.3	6.3

Prediction accuracy metrics for prior block sampling approaches on SZ2 CR estimation are reported for the Hurricane Isabel CLOUD field. We present our results for block sizes of 24 and 32.

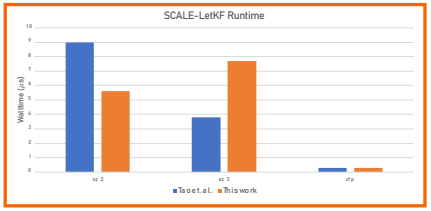


Both SZ3 and ZFP compressors exhibit low MAPE across error bounds on the Hurricane Isabel dataset when sufficiently large block sizes are used, independent of the lossy compressor error bound and the number of blocks sampled..



Prediction error using SZ3 compressor with absolute error bound 1e-4. .Our model is robust against grid-aligned block sampling approaches.

TIMING PERFORMANCE



Runtime of lossy compressors on the SCALE-LETKF rainfall simulation compared with the work from [2]. Our method is comparable in runtime performance while offering superior prediction accuracy.

ANALYSIS OF RESULTS

- Sampling a small number of moderately sized data points maintains fast data transfer and yields superior CR estimation accuracy compared to existing sampling approaches
- Grid-aligned uniform and random block sampling approaches have minimal impact on estimation accuracy
- Our method is flexible across lossy compressors, error bounds, and datasets

FUTURE WORK

- Use mixture modeling to increase robustness of CR estimation across datasets
- Explore the use of importance-based and other sampling methods to improve model accuracy

References

[1] D. Krasowska et. al., "Exploring Lossy Compressibility through Statistical Correlations of Scientific Datasets," in 2021 7th International Workshop on Data Analysis and Reduction for Big Scientific Data (DRBSD-7), Nov. 2021, pp. 47–53. doi: 10.1109/DRBSD7545632021.00011
[2] D. Tao et. al., "Optimizing Lossy Compression Rate-Distortion from Automatic Online Selection between SZ and ZFP," IEEE Transactions on Parallel and Distributed Systems, vol. 30, no. 8, Art. no. 8, Aug. 2019, doi: 10.1109/TPDS.2019.2894404
[3] R. Underwood et. al., "Black-box statistical prediction of lossy compression ratios for scientific data," The International Journal of High Performance Computing Applications, vol. 37, no. 3–4, pp. 412–433, Jul. 2023, doi: 10.1177/10943420231179417.
[4] R. Underwood, et. al., "OptZConfig: Efficient Parallel Optimization of Lossy Compression Configuration," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 12, pp. 3505-3519, 1 Dec. 2022, doi: 10.1109/TPDS.2022.3154096.
[5] https://sdrbench.github.io/

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-1943114 and Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration.