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

ALEXANDRA POULOS, Clemson University, USA
JON C. CALHOUN (ADVISOR), Clemson University, USA

Compression ratio estimation is an important optimization of I/O workflows processing terabytes of data. Applications such as compression auto-tuning or lossy compressor selection require a high-throughput, accurate estimation. Prior works that utilize sampling are fast but inaccurate, while approaches which do not use sampling are highly accurate but slow. Through sensitivity analysis we show that sampling a small number of moderately sized data blocks maintains fast data transfer and yields superior estimation accuracy compared to existing sampling approaches, and we use this to construct a new fast and accurate sampling method. In relation to non-sampling techniques, our method results in less than 10% degradation in estimation accuracy, while still maintaining the high throughput of the less accurate sampling methods.

Additional Key Words and Phrases: compression, lossy compression, high performance computing, statistical correlation analysis

ACM Reference Format:
Alexandra Poulos and Jon C. Calhoun (Advisor). 2023. Sensitivity of Black-Box Statistical Prediction of Lossy Compression Ratios for 3D Scientific Data. 1, 1 (August 2023), 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION AND BACKGROUND

Scientific simulations are increasingly reliant on lossy compression, as output data is on the scale of terabytes. There are many lossy compression algorithms, and the performance of a lossy compressor can vary greatly under different circumstances. Fast and accurate compression ratio (CR) estimation is necessary for many use cases, such as compression auto-tuning and fast compressor selection. Prior work has driven estimation techniques to be highly accurate at the cost of performance, but some HPC applications are better optimized by high speed compression and can accept a higher degree of error in estimation. Using data sampling to estimate CR yields up to a 3X improvement in throughput, but existing sampling approaches are accurate less than 80% of the time. In this work, we conduct a sensitivity analysis of block sampling techniques using data from three real-world scientific applications [6] and four state of the art lossy compressors, and present a hybrid lightweight compressor-agnostic sampling method that bridges the gap between CR estimation performance and accuracy.

2 METHODOLOGY

Sampling has been successful in the past, but it was very inaccurate. We develop a new method that leverages the advantages of sampling while maintaining the accuracy of approaches which utilize the entire data buffer. We develop the linear regression model shown in Equation 1 which uses the standard deviation and compression ratio of each

Authors' addresses: Alexandra Poulos, Clemson University, USA, alpoulo@clemson.edu; Jon C. Calhoun (Advisor), Clemson University, USA, jonccal@clemson.edu.

Manuscript submitted to ACM

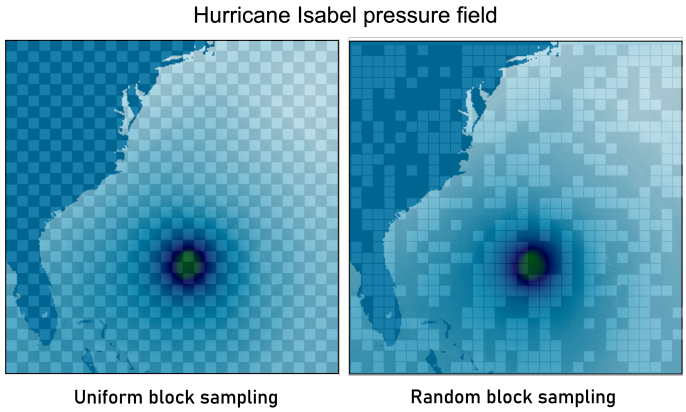

Fig. 1. We use two grid-aligned sampling approaches: uniform and random block sampling.

sampled block, along with a locality metric which allows the model to incorporate spatial correlation between sampled points. To assess the accuracy, we conduct a sensitivity analysis with respect to the sampling design.

We consider two grid-aligned block sampling approaches: uniform sampling and random sampling. Uniform sampling allows for equal coverage of all areas in a data buffer with no bias. Random sampling does not treat all blocks of the data buffer as equal, which leads to gaps in coverage in some areas and clusters of sampled blocks in others. This can be seen in Figure 1. The block sampling method takes a 3D tensor and splits it into $M \times M \times M$ 3D blocks. The number of block samples N is varied from 16 to 128. The size of the blocks M is varied from 4 to 32.

$$\log(\text{CR}) = a + b \times \sum \sigma_i \times \text{cr\_local}_i + c \times \sum \text{loc}_i \times \text{cr\_local}_i + d \times \sum \sigma_i \times \text{loc}_i \times \text{cr\_local}_i \,, \tag{1}$$

where $\text{loc}_i = \frac{\sum_j |x_{ii} - y_{ij}|}{\sqrt{\sum_j (x_{ii} - y_{ij})}}$.

We test our estimation model with data from three scientific applications from different domains and four state of the art lossy compressors, fpzip, sz2, sz3, and zfp. Each lossy compressor is run with absolute error bounds of 1E-5, 1E-4, 1E-3, and 1E-2. All experiments are run on Clemson's Palmetto cluster using a node with 40 core Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz and 372 GB of RAM. The OS is Linux CentOS 8 with compiler GCC 9.5.0.

3 RESULTS

3.1 Accuracy

To measure the accuracy of our model, we use the median absolute percentage error (MAPE) between the predicted CR and the true CR. The MAPE is calculated as $\frac{|CR_{true} - CR_{pred}|}{CR_{true}}$ and offers a robust estimate of the accuracy of a model as it is not affected by outliers such as extremely accurate or extremely inaccurate predictions. Each of the following heatmaps shows the MAPE for different compressors with various configurations. As a measure of uncertainty, each block is also annotated with the range between the 10% and 90% quantile of the estimation error.

We determined that the number of blocks sampled had little effect on the estimation accuracy. As shown in Figures 2 and 3, it is the size of the sampled blocks that has the biggest impact on the predicted CR accuracy. We also determine that the sampling approach utilized does not appear to impact the CR estimation. Figure 2 shows the MAPE across

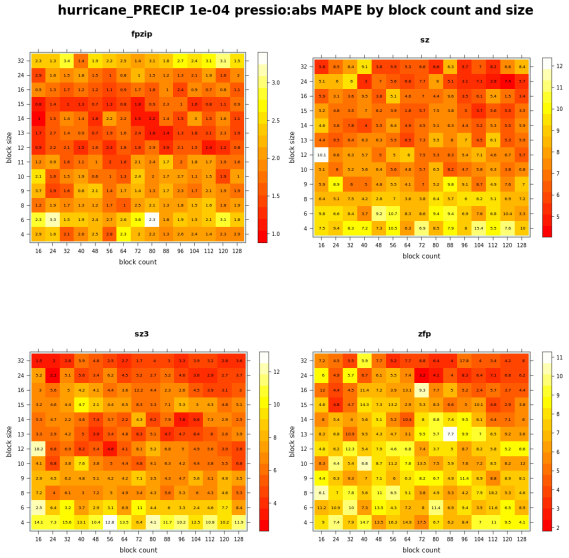

Fig. 2. Hurricane Isabel PRECIP field MAPE with *uniform* block sampling

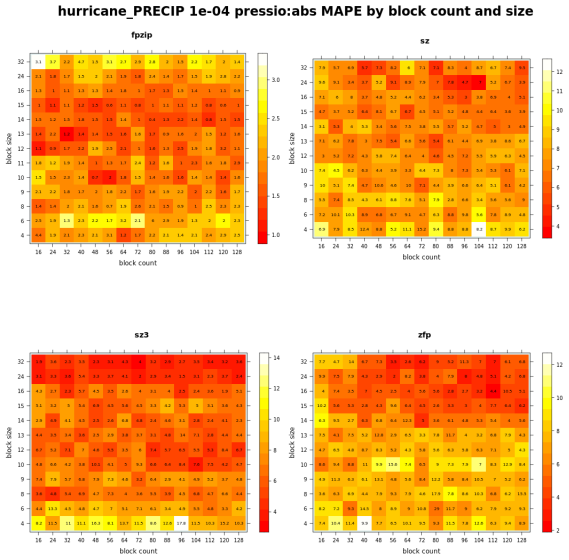

Fig. 3. Hurricane Isabel PRECIP field MAPE with *random* block sampling

all lossy compressors using uniform block sampling on the Hurricane Isabel PRECIP field with a fixed error bound of $1e-04$, while Figure 3 shows those results for random block sampling. Additionally, the error bound does not seem to affect prediction capability. This can be seen in Figure 4 which shows the prediction results of the SZ3 compressor on the

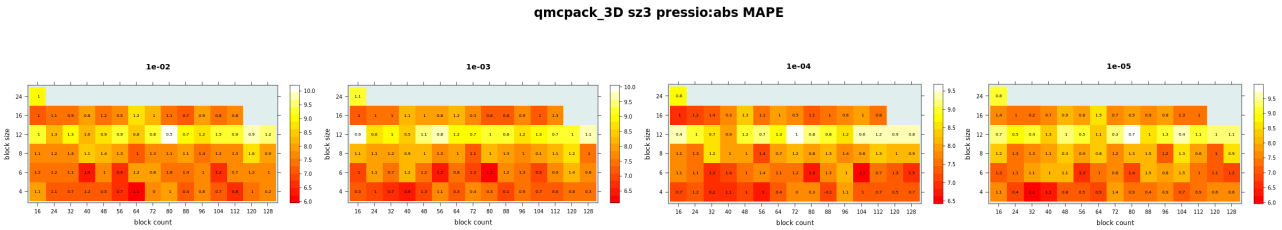

Fig. 4. QMCPACK MAPE with the SZ3 compressor across all tested error bounds. The accuracy of our model is not impacted by the lossy compressor error bound. Note that the dimensionality of this dataset ($115 \times 69 \times 69$) does not allow for testing of all block configurations.

QMCPACK dataset across all error bounds. Table 1 shows the accuracy performance on the QMCPACK dataset across all compressors and error bounds with a fixed block size of 16. Similar results were obtained for other compressors and datasets.

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

Table 1. Prediction accuracy metrics for fpzip, sz2, sz3, and zfp's CR estimation with a block size of 16. MAPE with 10% and 90% quantiles are reported for the QMCPACK dataset.

3.2 Runtime Performance

CR estimation methods which utilize the entire dataset are significantly slower than sampling approaches. Because our approach does not rely on the entire data buffer, our method is significantly faster than non-sampling approaches. For that reason we compare our runtime performance to sampling methods. Figure 5 shows our timing performance by compressor on the SCALE-LetKF dataset compared to the work in [3]. Our method is comparable in runtime performance while offering superior prediction accuracy.

4 CONCLUSION AND FUTURE WORK

We present a lightweight estimation method which is flexible across compressors, error bounds, and datasets. Our method only requires a small number of moderately sized samples from a data buffer in order to achieve a high prediction accuracy while maintaining high data throughput. This work leads to the next step in advancing compression auto-tuning and fast lossy compressor selection. Future work includes i) exploring the use of importance-based and other sampling methods to improve model accuracy and ii) incorporating the use of mixture modeling to increase the robustness of our CR estimation across datasets.

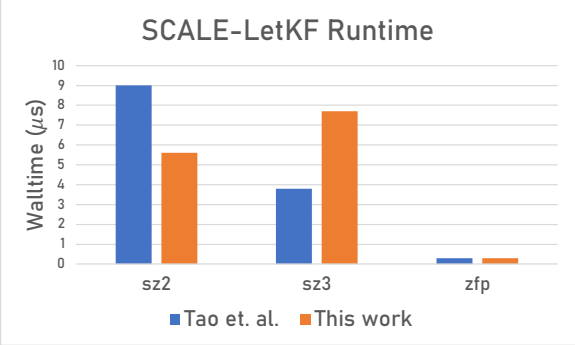

Fig. 5. Runtime of lossy compressors on the SCALE-LETKF rainfall simulation compared with the work from [3].

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-1943114 and the 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.

REFERENCES

[1] David Krasowska, Julie Bessac, Robert Underwood, Jon C. Calhoun, Sheng Di, and Franck Cappello. 2021. Exploring Lossy Compressibility through Statistical Correlations of Scientific Datasets. arXiv:2111.13789

[2] Tao Lu, Qing Liu, Xubin He, Huizhang Luo, Eric Suchyta, Jong Choi, Norbert Podhorszki, Scott Klasky, Mathew Wolf, Tong Liu, and Zhenbo Qiao. 2018. Understanding and Modeling Lossy Compression Schemes on HPC Scientific Data. In *2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 348–357. https://doi.org/10.1109/IPDPS.2018.00044

[3] Dingwen Tao, Sheng Di, Xin Liang, Zizhong Chen, and Franck Cappello. 2019. Optimizing Lossy Compression Rate-Distortion from Automatic Online Selection between SZ and ZFP. *IEEE Transactions on Parallel and Distributed Systems* 30, 8 (2019), 1857–1871. https://doi.org/10.1109/TPDS.2019.2894404

[4] Robert Underwood, Julie Bessac, David Krasowska, Jon C. Calhoun, Sheng Di, and Franck Cappello. 2023. Black-Box Statistical Prediction of Lossy Compression Ratios for Scientific Data. arXiv:2305.08801 [cs.DC]

[5] Robert Underwood, Jon C. Calhoun, Sheng Di, Amy Apon, and Franck Cappello. 2022. OptZConfig: Efficient Parallel Optimization of Lossy Compression Configuration. *IEEE Transactions on Parallel and Distributed Systems* 33, 12 (2022), 3505–3519. https://doi.org/10.1109/TPDS.2022.3154096

[6] Kai Zhao, Sheng Di, Xin Lian, Sihuan Li, Dingwen Tao, Julie Bessac, Zizhong Chen, and Franck Cappello. 2020. SDRBench: Scientific data reduction benchmark for lossy compressors. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2716–2724. https://sdrbench.github.io