2022 IEEE/ACM 8th International Workshop on Data Analysis and Reduction for Big Scientific Data (DRBSD) | 978-1-6654-6337-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/DRBSD56682.2022.00011

Understanding Impact of Lossy Compression on Derivative-related Metrics in Scientific Datasets

Zhaoyuan Su*, Sheng Di†, Ali Murat Gok, Yue Cheng*, Franck Cappello†
* University of Virginia, Charlottesville, VA
† Argonne National Laboratory, Lemont, IL, USA
acf7ea@virginia.edu, sdi@anl.gov, alimuratgok2017@u.northwestern.edu, yuecheng@virginia.edu, cappello@mcs.anl.gov

***Abstract*—Today's scientific simulations are producing extremely large amount of data everyday, which induces grand challenges in transferring and storing the data efficiently. Error-bounded lossy compression has been thought of as the most promising solution to the bigdata issue, however, it would cause data distortion that has to be controlled carefully for user's post-hoc analysis. Recently, the preservation of quantities of interest has become a priority. Derivative-related metrics are critical quantities of interest for many applications across domains. However, no prior research explored the impact of lossy compression on derivative-related metrics in particular. In this paper, we focus on understanding the impact of various error-controlled lossy compressors on multiple derivative-related metrics commonly concerned by users. We perform solid experiments that involve 5 state-of-the-art lossy compressors and 4 real-world application datasets. We summarize 5 valuable takeaways, which can shed some light in understanding the impact of lossy compression on derivative-related metrics.**

I. Introduction

With the ever-increasing execution scale of high performance computing (HPC) simulations across different scientific domains, too large amounts of data are being produced every day. Even though memory and storage capabilities of modern machines have been improved immensely, the sheer amount of data still puts significant pressure on the memory and storage systems, so that the limited memory/storage capacity has become a significant concern for the scientific simulations.

Significantly reducing the data volume at runtime has become one of the most promising solutions to the big scientific data issue for these large-scale applications. Although lossless compression can have the reconstructed data identical to the original data, it suffers from very low compression ratios (generally 2:1 according to existing studies [1]–[3]), which is far from the expected level (generally 10:1 or higher [4]). In comparison, error-bounded lossy compression [5]–[8] allows a certain distortion of data during the compression such that the compression ratio could reach a very high level (10:1, 100:1, or even higher), while the data distortion can also be controlled strictly based on user-specified error bound.

The existing state-of-the-art error-bounded lossy compressors mainly offer specific types of error bounds (such as absolute error bound and relative error bound) for users to control the data distortion, which still brings out a significant gap to users' de-facto requirement on data fidelity for their post hoc analysis.

In this paper, we focus on derivative-related metrics, which has a wide range of usage in post hoc analysis across different scientific domains. In practice, the derivatives in a dataset are generally approximated by calculating the difference (or delta) of the adjacent values in the dataset in different directions.

In what follows, we give several examples to demonstrate the significance of the derivative-related metrics used by the practical analysis in different applications. Laminar/turbulent flow is a very significant topic that is closely related to a wide range of research domains such as air flow going around a vehicle's windshield or an aircraft's wings, fluid flow passing by a ship's body, and combustion airflow. The post hoc analysis based on these fluid dynamics simulations is closely related to or driven by the gradients (derivative vector) of datasets. For instance, velocity gradient [9] is a critical metric in analyzing turbulent flow and its stress. Climate/weather simulation such as Hurricane simulation is also closely related to the gradients of air flows during the simulation. In fact, hurricanes could be thought of as heat engines, which are driven by two different types of thermal gradients [10]. In addition to gradients, other derivative-related metrics such as Laplacian and Sobolev norm are also closely related to scientific research, which will be detailed later.

The error-bounding nature offered by the existing error-controlled lossy compressors [3], [11]–[14] is far less than enough for preserving the derivative-metric values in the dataset [15]. In fact, raw data and the derivatives calculated based on them are fundamentally different in nature. For example, if a dataset contains the position of an asteroid near Earth, the first-order derivative of the asteroid's movement indicates the speed, and the second-order derivative indicates its acceleration (and in turn, the forces exerted on that asteroid, due to F=ma, where F and m refer to force and mass, respectively). Suppose one simulation needs to determine whether the asteroid would hit the earth, a relatively small error on the raw data may not have a big effect on the predictions of the asteroid's trace, while the similar errors introduced to the derivatives, which correspond to the speed and acceleration, may distort the trace significantly since these errors considerably affect the asteroid's behavior as the simulation time passes.

Assessing the impact of lossy compression on derivative-related metrics faces two grand challenges. On the one hand, the derivative metric is much more complicated and diverse

978-1-6654-6337-9/22/$31.00 ©2022 IEEE
DOI 10.1109/DRBSD56682.2022.00011

than the traditional commonly-used lossy compression quality metric such as mean squared error (MSE), Peak Signal to Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM), in that it may involve different spatial directions and different orders. On the other hand, the derivative-related metric values are generally more volatile than the original raw datasets. Specifically, the higher order of the derivative, the more volatile the derivative metric tends to be.

In this paper, we provide an in-depth analysis of the impact of lossy compression on multiple derivative-based metrics. The derivative metrics include the point-wise derivatives generated along each dimension, the statistics (such as PSNR and SSIM) of the derivative dataset, Laplacian, gradient, and Sobolev norms. Not only does our analysis include theoretical derivation of the upper bound for derivative metric's distortion under error-bounded lossy compression, but we also perform careful evaluation running state-of-the-art lossy compressors with real-world scientific datasets across from a wide range of domains including cosmological simulation, climate simulation, and turbulence simulation. We summarize a series of valuable characteristics of gradient metrics and takeaways regarding error-bounded lossy compression quality from our comprehensive characterization work.

The remaining of the paper is organized as follows. Section II discusses the related work. Section III describes our analysis methodology. In Section IV, we present our characterization results with an in-depth analysis. In Section V, we conclude the paper with a discussion of the future work.

II. Related Work

To the best of our knowledge, this paper is the first systematic study about the impact of error-controlled lossy compressors on the derivative-related metrics (a kind of critical quantity of interest (QoI) to many applications in practice) for scientific datasets. In what follows, we discuss the existing related work in the regard of controlling errors based on different QoI metrics with error-bounded lossy compressors.

Most of the existing state-of-the-art error-bounded lossy compressors support multiple simple metrics other than absolute error bound. SZ [16]–[20], for example, support absolute error bound control, point-wise relative error bound control, and peak signal to noise ratio (PSNR). ZFP [11] offers three compression modes: absolute error bound, precision (an integer), and fix-rate (or fix compression ratio). MGARD [13], [21] supports different types of norms such as infinity norm (i.e., absolute error bound) and more bounded linear QoIs.

There are also some existing studies focused on how to preserve the specific QoIs by error-bounded lossy compressors. Liang et al. [22] proposed a novel strategy to preserve the topological features in 2D and 3D vector fields. More specifically, they developed a new compression method by leveraging SZ compression pipeline, which can keep each critical point in its original cell and retain the type of each critical point (e.g., saddle and attracting node). Liu et al. [23]–[25] developed a compression method that can optimize the compression ratio and quality, according to diverse user-specified constraints (such as preserving data points' sign, supporting diverse error bounds according to regions of interest). Recently, a quality metric oriented lossy compressor was proposed [26]: it can automatically adjust the compressor to adapt to user's diverse requirement in optimizing the rate-distortion measure based on different quality metrics (such as PSNR, SSIM, and compression ratio).

Despite the above in-depth related lossy compression works on QoIs, the impact of these error-bounded lossy compressors on the derivative-related metrics are still unknown. This raises a significant gap to derivative-dependent applications, which will be filled in this paper.

III. Analysis Methodology

In this section, we provide an overview of our analysis method, including how we select state-of-the-art lossy compressors and their error configurations, description of various derivative metrics, description of real-world scientific datasets used in our study, and how we execute the compressors and open source libraries/tools such as Z-checker [27], [28], QCAT [29] and Paraview [30].

A. Error-bounded Lossy Compressors

In our study, we selected 5 error-bounded lossy compressors – ZFP [11], SZ2 [16], SZ3 [17], SZx [3] and FPZIP [12], which are all state-of-the-arts because of their high execution performance, excellent compression ratios, or great reconstructed data quality. We describe them in Appendix A.

Note that in our study, we focus only on the error-bounded lossy compressors that have relatively high execution throughput (generally 200+MB/s on a single CPU-core). Some other lossy compressors such as TTHRESH [31], [32] and Auto-encoder based compressor [33] can get fairly high compression ratios, but they all suffer from very low speed/throughput (generally 10s of MB/s or even lower), which may cause very low efficiency in the evaluation because of many datasets across multiple scientific domains involved in our study.

B. Derivative Metrics in Our Study

We describe the derivative metrics investigated in our study as follows. All these metrics are fundamental and essential to many application users across different scientific domains. In the following, we use x, y, z to represent the three different dimensions in the dataset. We use f to denote the corresponding function based on the dataset and we use D to denote the values of the dataset: e.g., $D(x,y)$ means the value at the data point (x,y) in the given dataset.

1) Partial Derivatives: Partial derivative (denoted by $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, $\frac{\partial f}{\partial z}$) is the basic metric that represents the rate of data variation in space along one coordinate dimension, which is also a fundamental measure for computing all other derivative metrics. Intuitively, for a data point, its partial derivative with respect to a coordinate dimension (such as x) is the difference between adjacent values on this data point. Basically, there

are two ways to calculate the partial derivative – forward difference and central difference.

Forward difference is an intuitive approach to compute derivatives on a scalar filed:

$$\frac{\partial f(i,j)}{\partial x} = D[i, j+1] - D[i, j] \tag{1}$$

$$\frac{\partial f(i,j)}{\partial y} = D[i+1, j] - D[i, j] \tag{2}$$

Central difference is an accurate and most commonly used method to calculate derivatives in the visualization community:

$$\frac{\partial f(i,j)}{\partial x} = \frac{(D[i,j+1]-D[i,j])+(D[i,j]-D[i,j-1])}{2} = \frac{D[i,j+1]-D[i,j-1]}{2} \tag{3}$$

$$\frac{\partial f(i,j)}{\partial y} = \frac{(D[i+1,j]-D[i,j])+(D[i,j]-D[i-1,j])}{2} = \frac{D[i+1,j]-D[i-1,j]}{2} \tag{4}$$

Similarly, we can calculate the second-order derivative along the dimension x and mixed partial derivative on the data point (x,y) based on the following formulas, respectively:

$$\frac{\partial f^2(i,j)}{\partial x^2} = (D[i, j+1] - D[i, j]) - (D[i, j] - D[i, j-1]) \tag{5}$$

$$\frac{\partial f^2(i,j)}{\partial xy} = \frac{D[i-1,j-1]+D[i+1,j+1]-D[i-1,j+1]-D[i+1,j-1]}{4} \tag{6}$$

Figure 1 illustrates how we calculate the first-order and second-order derivatives on the point (i,j) by an example with a 2D dataset. The above derivative formulas can be extended easily for higher-dimensional cases such as 3D datasets.

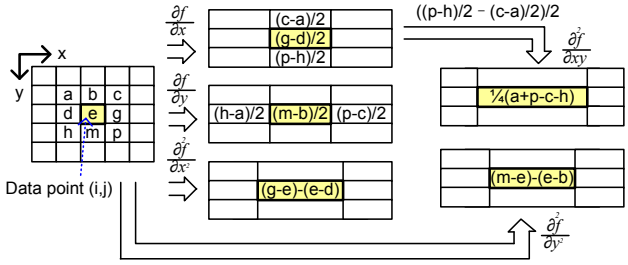


Fig. 1: Illustration of calculating 1st/2nd-order derivatives on the point (i,j).

2) Statistics of Partial Derivatives: In order to measure the distortion of partial derivatives generated based on the lossy reconstructed dataset, we propose to apply three statistical measures (PSNR, SSIM and point-wise maximum errors) on the distortion of the derivative datasets. Specifically, we calculate the partial derivative datasets based on the original dataset (denoted by f') and the decompressed dataset (denoted by $\widehat{f'}$), respectively. The range of f' is denoted by R. And then, we compute the PSNR, SSIM and the point-wise errors based on the two derivative datasets (f' and $\widehat{f'}$), respectively. As an example, we demonstrate how to calculate the derivative PSNR in the following formula.

$$PSNR(f', \widehat{f'}) = 20 \log_{10} \left(\frac{R(f')}{\sqrt{MSE(f', \widehat{f'})}} \right) \tag{7}$$

*3) Gradient Length (a.k.a., Gradient Magnitude, denoted by **GL**):* Gradient is the vector with the direction and rate of fastest increase at a data point: more specifically, it is the vector sum of the partial derivatives along all the coordinate variables of a scalar quantity, as shown in the Formula (8) (using 2D dataset as an example).

$$|\nabla f(i,j)| = \sqrt{\left(\frac{\partial f(i,j)}{\partial x}\right)^2 + \left(\frac{\partial f(i,j)}{\partial y}\right)^2} \tag{8}$$

The length of the gradient is a critical metric to measure the local intensity of the data point, which has a very wide range of usage, such as characterizing turbulent flows in fluid dynamics simulation [9], calculating halos by pressure gradients in cosmology simulation [34] and detection of edges in visualization [35]. It is also known as gradient magnitude (measuring the intensity/magnitude of local change at a point) in visualization or image processing domain.

*4) Laplacian (a.k.a., Laplace operator, denoted by ∇^2 or **Lap**):* Laplacian is a differential operator provided by the divergence of a scalar function's gradient on Euclidean space. Laplacian is achieved by calculating the sum of the second-order partial derivatives of the function with respect to each coordinate variable, as shown in the Formula (9) (using 2D dataset as an example).

$$\nabla^2 f(i,j) = \frac{\partial^2 f(i,j)}{\partial x^2} + \frac{\partial^2 f(i,j)}{\partial y^2} \tag{9}$$

Laplacian is an operator widely used in many differential equations describing physical phenomena, such as Poisson's equation that describes electric and gravitational potentials, the diffusion equation which describes heat and fluid flow, the wave equation that describes wave propagation, and the Schrödinger equation in quantum mechanics.

5) Sobolev Norm: Sobolev norm [36] (denoted by $||f||_{k,p}$) is the natural norm admitted in Sobolev space, which is a vector space of functions with a norm combining L^p-norms of the function with its derivatives up to a specific order (denoted by k). Sobolev spaces [36] are very effective for studying the partial differential equations (PDEs) since they are Banach spaces, bringing out a powerful tool of functional analysis.

Formula (10) gives the definition of the Sobolev norm based on the p-norm.

$$||f||_{k,p} = \left(\sum_{i=0}^{k} ||f^{(i)}||_p^p \right)^{\frac{1}{p}} \tag{10}$$

where $f^{(i)}$ represents the i-th partial derivative. In our study we focus only on the 2-norm cases in that the corresponding Sobolev spaces form a Hilbert space, which is the core concept of many research problems/domains such as functional analysis and quantum mechanics.

Our investigation involves three k values (0,1,2), and the corresponding Sobolev norms are shown in the following formulas (in a two dimension dataset D, where M and N represents the total number of data points in the direction of first and second dimension, respectively).

$$||f(i,j)||_{0,2} = \sqrt{||f||_2^2} = \sqrt{\frac{\sum D[i,j]^2}{MN}} \tag{11}$$

$$||f(i,j)||_{1,2} = \sqrt{||f||_2^2 + ||f^{(1)}||_2^2} = \sqrt{\frac{1}{MN}\sum\left(D[i,j]^2 + f'_x(i,j)^2 + f'_y(i,j)^2\right)} \quad (12)$$

$$||f(i,j)||_{2,2} = \sqrt{||f||_2^2 + ||f^{(1)}||_2^2 + ||f^{(2)}||_2^2} = \sqrt{\frac{\sum\left(D[i,j]^2 + f'_x(i,j)^2 + f'_y(i,j)^2 + f''_x(i,j)^2 + f''_y(i,j)^2 + f''_{xy}(i,j)^2\right)}{MN}} \quad (13)$$

where $f'_x(i,j)=\frac{\partial f(i,j)}{\partial x}$, $f'_y(i,j)=\frac{\partial f(i,j)}{\partial y}$, $f''_x(i,j)=\frac{\partial^2 f(i,j)}{\partial x^2}$, $f''_y(i,j)=\frac{\partial^2 f(i,j)}{\partial y^2}$, and $f''_{xy}(i,j)=\frac{\partial^2 f(i,j)}{\partial xy}$.

C. Theoretical Analysis of Derivative Metric Value Distortion under Error-bounded Lossy Compression

In the following text, we present the theoretical analyses mainly focusing on the maximum errors of the derivative metrics. We denote the raw dataset by D, lossy reconstructed dataset by $\hat{D}$, and the user-specified absolute error bound by e.

Proposition 1: The maximum error (or difference) of the derivatives calculated based on the reconstructed dataset and the raw dataset can be derived as Formula (14):

$$\left|\frac{\partial f(i,j)}{\partial x} - \frac{\partial \hat{f}(i,j)}{\partial x}\right| \leq 2e \quad (14)$$

Proof: According to Formula (1), the 1st-order partial derivative's error at the point(i, j) based on the reconstructed dataset and the raw dataset can be written as $\left|\left|(D[i,j+1]-D[i,j])\right| - \left|(\hat{D}[i,j+1]-\hat{D}[i,j])\right|\right|$. Due to $|D[i,j+1]-\hat{D}[i,j+1]| \leq e$ and $|D[i,j]-\hat{D}[i,j]| \leq e$, the total 1st-order partial derivative's distortion at the point(i, j) would be bounded by 2e. ■

Proposition 2: The PSNR of the 1st-order partial derivative datasets can be derived as Formula (15):

$$PSNR(f', \widehat{f'}) \geq 20\log_{10}\left(\frac{R(f')}{2e}\right) \quad (15)$$

Proof: According to Formula (7), the PSNR of the 1st-order derivative datasets calculated over the reconstructed dataset and the raw dataset is determined by the range of the raw derivative dataset and the MSE between the two derivative datasets, which can be expressed as Formula (16).

$$MSE(f', \widehat{f'}) = \frac{\sum (D(i,j)-\hat{D}(i,j))^2}{MN} \quad (16)$$

In the worst case, the maximum MSE is up to $4e^2$ when the maximum point-wise error between the raw dataset and reconstructed dataset is e. So, the PSNR should be no smaller than $20\log_{10}\left(\frac{R(f')}{2e}\right)$. ■

Proposition 3: The point-wise error of gradient length datasets calculated based on the raw data and lossy reconstructed data can be derived as Formula (17):

$$\left|\left|\nabla f(i,j)\right| - \left|\nabla \hat{f}(i,j)\right|\right| \leq 2\sqrt{e\left(\left|\frac{\partial f(i,j)}{\partial x}\right| + \left|\frac{\partial f(i,j)}{\partial y}\right|\right) + 2e^2} \quad (17)$$

Proof: Based on the Formula (8), the physical meaning of the gradient length is the length of the vector with a x-direction derivative and y-direction derivative as subvectors, and the calculation follows the Pythagorean theorem. After squaring the difference of the gradient lengths from the reconstructed dataset and the raw dataset, the equation can be presented as $\left(\left|\nabla f(i,j)\right| - \left|\nabla \hat{f}(i,j)\right|\right)^2 = \nabla f(i,j)^2 - 2 \times \nabla f(i,j) \times \nabla \hat{f}(i,j) + \nabla \hat{f}(i,j)^2$. In the worst case, each element of the reconstructed dataset is larger than the element of the raw dataset by e. It would be true that $\nabla \hat{f}(i,j) > \nabla f(i,j)$ and $\left(\left|\nabla f(i,j)\right| - \left|\nabla \hat{f}(i,j)\right|\right)^2 \leq \nabla \hat{f}(i,j)^2 - \nabla f(i,j)^2$. After simplifying and square rooting, the proposition is proved. ■

Proposition 4: The point-wise error of the Laplacian dataset calculated based on the raw and the reconstructed dataset can be derived as Formula (18):

$$|\nabla^2 f(i,j) - \nabla^2 \hat{f}(i,j)| \leq 8e \quad (18)$$

Proof: Based on the Formula (9), the value at point(i, j) of the Laplacian dataset is calculated by the sum of the second-order partial derivatives as $\nabla^2 f(i,j) = D(i+1,j) + D(i-1,j) + D(i,j+1) + D(i,j-1) - 4D(i,j)$. Since the error of every point from the reconstructed dataset and the raw dataset is up to e, the absolute value of the point-wise error of the Laplacian dataset is up to 8e. And the same proposition extends to 3D datasets with a maximum error of 12e. The proposition is proved. ■

Proposition 5: The error of the Sobolev norm 2 order 0 calculated based on the raw dataset and the reconstructed dataset can be derived as Formula (19):

$$||f||_{0,2} - ||\hat{f}_{0,2}|| \leq \sqrt{2e \times Avg + e^2} \quad (19)$$

where $Avg=\frac{\sum D(i,j)}{MN}$

Proof: Based on the Formula (11), the Sobolev Nomr 2 order 0 is calculated by the value of points of raw datasets. In the worst case, each element of the reconstructed dataset is larger than the element of the raw dataset by e. It would be true that $||f||_{0,2} \leq ||\hat{f}_{0,2}||$. After squaring the difference of the order 0 of Sobolev Norm 2, the equation can be represented as $(||f||_{0,2} - ||\hat{f}_{0,2}||)^2 \leq (||\hat{f}||_{0,2})^2 - (||f_{0,2}||)^2$. After simplifying and square rooting, the proposition is proved. ■

Due to the complication of the Sobolev Norm 2 Order 1 and Order 2, the proofs of the maximum error of them will be represented in our future work.

Takeaway 1: The maximum error regarding several derivative metrics such as Laplacian and derivative maximum error is only determined by the user-specified error bound e. The point-wise error of derivative datasets has an upper bound of $2e$, and the point-wise error of Laplacian datasets has the upper bound of $8e$ and $12e$ in 2D and 3D, respectively.

Takeaway 2: For some derivative metrics such as derivative' PSNR, maximum error of gratident length, and 0-ordered Sobolev norm 2, the deviation is determined not only by e, but also by the raw dataset values or derivative metric values. The derivative' PSNR depends on the value range of the first-order partial derivative datasets; the maximum error of the gradient length depends on the value of first order derivatives of the raw dataset in

respective directions; the maximum error of the Sobolev norm 2 order 0 relies on the average of the raw dataset.

D. Tool Chain and Development

The analysis workflow that involves different open-source tools and our codes is described in Appendix B.

IV. Evaluation and Analysis

In this section, we conduct several experiments and analyze the evaluation results regarding several derivative-related metrics.

A. Scientific Application Datasets in Our Study

We perform the evaluation using 4 real-world scientific applications from different domains. All the datasets were downloaded from SDRBench [37], [38], which are described in Appendix C in detail.

B. Characterization of The Impact of Lossy Compressors on Derivative-related Metrics

First, we check the data quality of gradient length calculated based on the reconstructed dataset under 5 cutting-edge compressors (denoted by *Cmprs*) with different compression error bounds. Because of space limits, we demonstrate the compression ratio (denoted by *CR*), maximum error, average error and PSNR using only the NYX dataset (baryon_density_log_2), as shown in Table I (The large numbers have be bolded). We can observe that the maximum error of gradient length dataset calculated based on the raw and reconstructed dataset is larger than the real maximum compression error under 3 compressors, even up to 2.69× lager than the compression error in some cases. Whereas, the maximum errors demonstrated in the gradient length are smaller than the compression errors on the raw data with respect to the other two compressors probably due to different compression mechanisms. Due to the small average value of the elements in the gradient length dataset, the point-wise average errors of gradient length are always smaller than the compression errors on the raw dataset. The similar observation also goes to the PSNR of gradient length. **Takeaway 3: For a relatively smooth dataset with the lossy compression, the point-wise average error of the gradient length dataset calculated from the raw and the reconstructed dataset tends to be smaller than the compression error, but the maximum error tends to be larger than that.**

Figure 2 represents the visual comparison of the data distortion of the raw dataset and the gradient length dataset calculated based on it under 2 compressors. Figure 2a and Figure 2b show the 3D rendering and 2D top view of the raw dataset, respectively. Figure 2c and Figure 2d show the reconstructed data from SZ3 and ZFP with compression ratio CR=117, respectively, and we can observe that the reconstructed data's visual quality is good for both compressors. Regarding gradient length, the visual difference of gradient length calculated based on the raw dataset and the reconstructed dataset under SZ3 (Figure 2g) is also slight, but the data distortion of gradient length from reconstructed data under ZFP (Figure 2h) is significant. **Takeaway 4: The compressor SZ3 can preserve the gradient length very well at a compression ratio of 100 with slight distortion of visual quality, but ZFP may substantially affect visual quality of the gradient length dataset at the same compression ratio on the NYX dataset.**

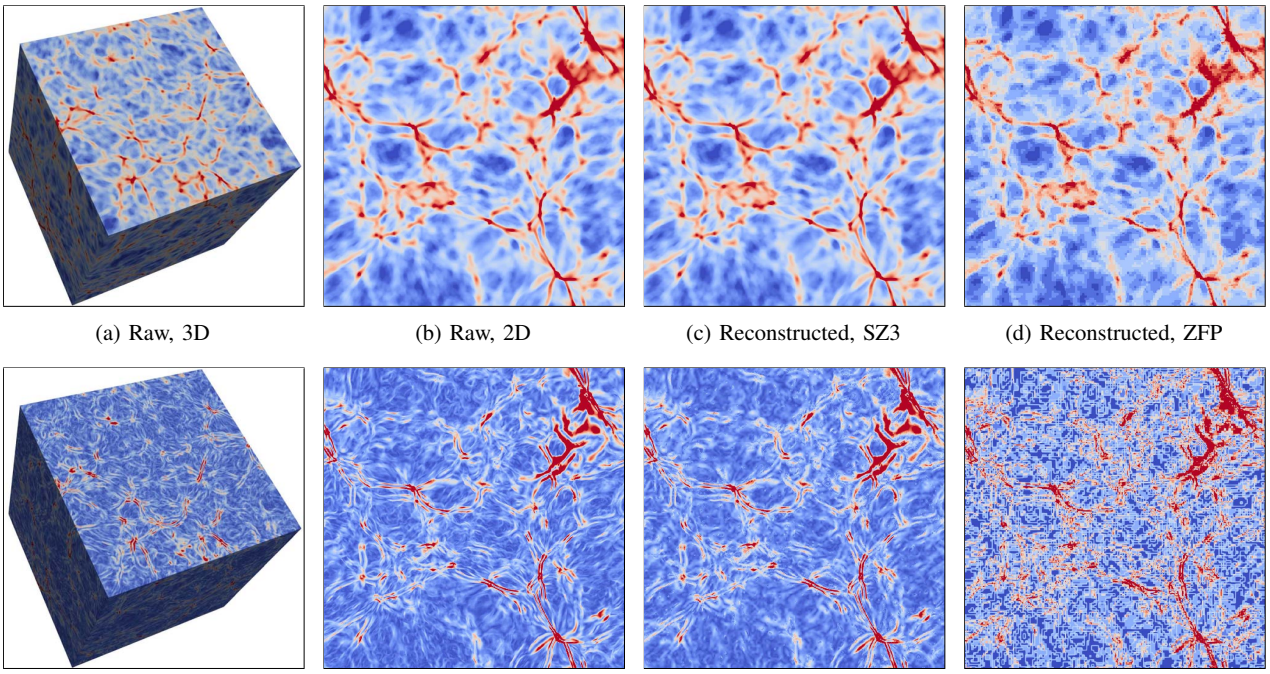
TABLE I: Raw, Reconstructed and Gradient Length Datasets.

baryon density log2 From EXASKY-NYX								
Cmprs	e	CR	Maximum Error		Average Error		PSNR	
			Raw & Rec	GL	Raw & Rec	GL	Raw & Rec	GL
SZx	1	4.97	1.0000	**1.1508**	**1.23E-1**	5.27E-2	**41.5377**	36.8293
	1E-1	3.09	0.0313	**0.0375**	**7.85E-3**	3.51E-3	**65.7024**	61.0256
	1E-2	2.40	0.0039	**0.0048**	**9.86E-4**	4.43E-4	**83.7226**	78.9964
	1E-3	1.89	0.0005	**0.0007**	**1.23E-4**	5.55E-5	**101.7827**	97.0500
SZ2	1	217.09	1.0000	**2.4658**	**1.18E-1**	7.88E-2	**41.3460**	32.3355
	1E-1	21.65	0.1000	**0.2560**	**4.83E-2**	3.29E-2	**51.4113**	42.2389
	1E-2	10.48	0.0100	**0.0250**	**4.98E-3**	3.31E-3	**71.2081**	62.2314
	1E-3	5.27	0.0010	**0.0023**	**4.97E-4**	3.31E-4	**91.2144**	82.2380
SZ3	1	1034.23	1.0000	**2.6937**	**1.92E-1**	7.63E-2	**38.7435**	33.4168
	1E-1	59.20	0.1000	**0.2287**	**3.67E-2**	2.39E-2	**53.3757**	44.8030
	1E-2	11.79	0.0100	**0.0237**	**4.91E-3**	3.28E-3	**71.3038**	62.3314
	1E-3	5.37	0.0010	**0.0024**	**5.00E-4**	3.32E-4	**91.1865**	82.2100
ZFP	1	24.20	**0.2641**	0.2266	**2.25E-2**	1.63E-2	**57.1644**	48.0335
	1E-1	6.71	**0.0193**	0.0156	**2.10E-3**	1.60E-3	**77.7409**	68.2671
	1E-2	4.12	**0.0024**	0.0019	**2.63E-4**	2.01E-4	**95.7543**	86.2719
	1E-3	2.97	**0.0003**	**0.0003**	**3.29E-5**	2.51E-5	**113.8156**	104.3333
FPZIP	p=10	17.14	**3.9850**	2.6553	**1.92E-1**	9.34E-2	**38.0284**	32.4843
	p=12	12.76	**1.0000**	0.6572	**5.33E-2**	2.41E-2	**48.9581**	43.6705
	p=14	9.12	**0.3630**	0.1648	**1.37E-2**	5.75E-3	**60.7513**	56.3279
	p=16	6.24	**0.1130**	0.0615	**3.44E-3**	1.45E-3	**72.7325**	68.2820

TABLE II: Raw, Reconstructed and Laplacian Datasets.

CLOUDf48.log10.bin.f32 From Hurricane-ISABEL								
Cmprs	e	CR	Maximum Error		Average Error		PSNR	
			Raw & Rec	Lap	Raw & Rec	Lap	Raw & Rec	Lap
SZx	1	15.73	0.9821	**6.0603**	1.14E-1	**2.74E-1**	34.4480	**46.3344**
	1E-1	13.08	0.0995	**0.5472**	6.91E-3	**2.12E-2**	58.6824	**69.1870**
	1E-2	10.75	0.0039	**0.0388**	8.63E-4	**2.70E-3**	76.7594	**87.1710**
	1E-3	9.59	0.0005	**0.0048**	1.08E-4	**3.38E-4**	94.8283	**105.2177**
SZ2	1	102.83	1.0000	**11.5006**	1.50E-1	**3.43E-1**	30.4448	**39.6637**
	1E-1	40.50	0.1000	**1.1322**	1.35E-2	**6.96E-2**	50.7639	**55.6239**
	1E-2	26.81	0.0100	**0.1141**	1.38E-3	**6.61E-3**	70.9799	**75.9179**
	1E-3	18.56	0.0010	**0.0119**	1.30E-4	**6.84E-4**	91.0653	**95.8165**
SZ3	1	111.08	1.0000	**11.5480**	1.71E-1	**2.02E-1**	32.9973	**42.7558**
	1E-1	53.44	0.1000	**1.1247**	3.50E-2	**3.78E-2**	48.4059	**58.7284**
	1E-2	27.00	0.0100	**0.1146**	1.48E-3	**4.62E-3**	72.4406	**77.3927**
	1E-3	18.70	0.0010	**0.0111**	2.14E-4	**4.75E-4**	91.0997	**97.2605**
ZFP	1E	19.30	0.3112	**1.5371**	9.05E-3	**3.80E-2**	54.6258	**59.6900**
	1E-1	12.30	0.0201	**0.0996**	5.43E-4	**2.69E-3**	78.1210	**82.9645**
	1E-2	9.62	0.0025	**0.0121**	6.73E-5	**3.48E-4**	96.0244	**100.7967**
	1E-3	7.86	0.0003	**0.0015**	1.03E-5	**4.41E-5**	113.9136	**118.7569**
FPZIP	p=10	41.22	0.2500	**1.2114**	**8.67E-2**	2.29E-2	41.2149	**62.5481**
	p=12	30.83	0.0625	**0.3135**	**2.66E-2**	6.33E-3	51.4619	**73.9942**
	p=14	24.45	0.0015	**0.0772**	**1.15E-2**	2.17E-3	58.6077	**83.5911**
	p=16	19.94	0.0039	**0.0209**	**8.68E-4**	3.57E-4	81.1468	**98.5534**

We present the data distortion of the Laplacian datasets calculated based on the raw and the reconstructed datasets regarding several metrics in Table II. The results from the Hurricane-ISABEL dataset (CLOUDf48.log10.bin.f32) under different compressors show that the maximum error of Laplacian datasets are much larger than the maximum compression error in all cases with the peak around 11.55 times (which is held in the theoretical maximum error with 12e at Propositio 4). The average error of the Laplacian dataset calculated by summing up the second-order derivatives of the three dimensions grows rapidly with the dramatic variation change of the raw dataset. The experiment results show that only the average error of Laplacian datasets calculated based on the datasets recovered by FPZIP is smaller than the actual compression error, while the average error of Laplacian datasets caused by the other four compressors is larger than the actual compression error. We can observe that the PSNRs of the Laplacian datasets are all larger than that of the raw and



(a) Raw, 3D (b) Raw, 2D (c) Reconstructed, SZ3 (d) Reconstructed, ZFP

(e) GL from Raw, 3D (f) GL from Raw, 2D (g) GL from Reconstructed, SZ3 (h) GL from Reconstructed, ZFP

Fig. 2: Data Distortion of Raw and Gradient Length from SZ3 and ZFP, NYX, Compression Ratio 117.

reconstructed dataset probably due to the Laplacian datasets with a bigger range.

To demonstrate the influence of the Laplacian dataset caused by the lossy compression on the raw dataset, we present the visual distortion of the Laplacian data calculated based on the reconstructed dataset under SZ3 with different compression error bounds (see Appendix D: Figure 3). Figure 3a show the 3D rendering of the raw dataset, as well as the Figure 3b and Figure 3c represent the 3D rendering and 2D slicing pictures of the Laplacian dataset, respectively. We can observe that the Figure 3d and Figure 3e show little visual distortion under SZ3 with absolute error bound e=1E-2 and 1E-1 (value-range relative error bound around 1E-3 and 1E-2), respectively. While with the compression error bounds larger than that, the results have a non-negligible visual error of Laplacian datasets. **Takeaway 5: The Laplacian metric of raw datasets can be well preserved under SZ3 with the value-range relative error bound 1E-2 and a compression ratio of 53, while the compression with higher error bounds will cause a significant visual distortion on the Hurricane-ISABEL dataset**.

Through Table IV to Table VII (as presented in Appendix E), we can clearly observe the influence of lossy decompressed data on the Sobolev Norm 2 of the datasets based on different compression error bounds. The Sobolev Norm 2 is mainly calculated by the second order square of values and partial derivatives of the scalar field datasets, in which the larger the values and partial derivatives, the larger Sobolev Norm 2 values are. We note that the data distortion of Sobolev Norm 2 affected by the lossy compression exhibits similar trends on the 4 datasets. The lossy compression has nearly no impact on the 0th-, first-, and second-order of Sobolev Norm 2 under all compressors with e=1E-4, and even some lossy compressors can perform the same results with the error bound e=1E-3. By comparison, the lossy compression applied on raw datasets does not project a particular changing trend of the data distortion for different orders of Sobolev Norm 2.

V. Conclusion and Future Work

In this paper, we provide an in-depth understanding of the impact of lossy compressors on the derivative-related metrics. We theoretically derive the upper bounds of the distortion for different derivative-related metrics under the error-bounded lossy compression. In our experiments, we run 5 state-of-the-art lossy compressors with 4 real-world application datasets, and present the observed distortion of derivative metrics under the lossy compression compared with their original true values. We summarize 5 takeaways, which we believe are very helpful to scientific applications related to derivative calculation.

Acknowledgments

This research was supported by the ECP, Project Number: 17-SC-20-SC, a collaborative effort of two DOE organizations – the Office of Science and the National Nuclear Security Administration, responsible for the planning and preparation of a capable exascale ecosystem, including software, applications, hardware, advanced system engineering and early testbed platforms, to support the nation's exascale computing imperative. The material was supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research (ASCR), under contract DE-AC02-06CH11357, and supported by the National Science Foundation under Grant OAC-2003709,

OAC-2104023, CCF-1919075, CCF-1919113, OAC-2106446, and CMMI-2134689. We acknowledge the computing resources provided on Bebop (operated by Laboratory Computing Resource Center at Argonne) and on Theta and JLSE (operated by Argonne Leadership Computing Facility).

REFERENCES

[1] P. Lindstrom, "Error distributions of lossy floating-point compressors," Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), Tech. Rep., 2017.

[2] S. W. Son, Z. Chen, W. Hendrix, A. Agrawal, W.-k. Liao, and A. Choudhary, "Data compression for the exascale computing era-survey," *Supercomputing frontiers and innovations*, vol. 1, no. 2, pp. 76–88, 2014.

[3] X. Yu, S. Di, K. Zhao, J. Tian, D. Tao, X. Liang, and F. Cappello, "Ultrafast error-bounded lossy compression for scientific datasets," in *Proceedings of the 31st International Symposium on High-Performance Parallel and Distributed Computing*, ser. HPDC '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 159–171. [Online]. Available: https://doi.org/10.1145/3502181.3531473

[4] F. Cappello, S. Di, S. Li, X. Liang, G. M. Ali, D. Tao, C. Yoon Hong, X.-c. Wu, Y. Alexeev, and T. F. Chong, "Use cases of lossy compression for floating-point data in scientific datasets," *International Journal of High Performance Computing Applications (IJHPCA)*, vol. 33, pp. 1201–1220, 2019.

[5] X.-C. Wu, S. Di, E. M. Dasgupta, F. Cappello, H. Finkel, Y. Alexeev, and F. T. Chong, "Full-state quantum circuit simulation by using data compression," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3295500.3356155

[6] A. M. Gok, S. Di, Y. Alexeev, D. Tao, V. Mironov, X. Liang, and F. Cappello, "Pastri: Error-bounded lossy compression for two-electron integrals in quantum chemistry," in *2018 IEEE International Conference on Cluster Computing (CLUSTER)*, 2018, pp. 1–11.

[7] R. Shan, S. Di, J. C. Calhoun, and F. Cappello, "Towards combining error-bounded lossy compression and cryptography for scientific data," in *2021 IEEE High Performance Extreme Computing Conference (HPEC)*, 2021, pp. 1–7.

[8] F. Cappello, S. Di, and A. M. Gok, "Fulfilling the promises of lossy compression for scientific applications," in *Driving Scientific and Engineering Discoveries Through the Convergence of HPC, Big Data and AI*, J. Nichols, B. Verastegui, A. B. Maccabe, O. Hernandez, S. Parete-Koon, and T. Ahearn, Eds. Cham: Springer International Publishing, 2020, pp. 99–116.

[9] P. P. Wegener, *What Makes Airplanes Fly? History, Science, and Applications of Aerodynamics*. Springer, 1991.

[10] "Hurricane formation," http://ffden-2.phys.uaf.edu/212_spring2011.web.dir/Levi _Cowan/formation.html, 2022, online.

[11] P. Lindstrom, "Fixed-rate compressed floating-point arrays," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 2674–2683, 2014.

[12] P. G. Lindstrom *et al.*, "Fpzip," Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), Tech. Rep., 2017.

[13] M. Ainsworth, O. Tugluk, B. Whitney, and S. Klasky, "Multilevel techniques for compression and reduction of scientific data—the univariate case," *Computing and Visualization in Science*, vol. 19, no. 5, pp. 65–76, 2018.

[14] 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.

[15] F. Cappello, S. Di, and A. M. Gok, "Fulfilling the promises of lossy compression for scientific applications," in *Driving Scientific and Engineering Discoveries Through the Convergence of HPC, Big Data and AI*, J. Nichols, B. Verastegui, A. B. Maccabe, O. Hernandez, S. Parete-Koon, and T. Ahearn, Eds. Cham: Springer International Publishing, 2020, pp. 99–116.

[16] X. Liang *et al.*, "Error-controlled lossy compression optimized for high compression ratios of scientific datasets," *2018 IEEE International Conference on Big Data (Big Data)*, pp. 438–447, 2018.

[17] K. Zhao, S. Di, M. Dmitriev, T.-L. D. Tonellot, Z. Chen, and F. Cappello, "Optimizing error-bounded lossy compression for scientific data by dynamic spline interpolation," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2021, pp. 1643–1654.

[18] 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.

[19] X. Liang *et al.*, "SZ3: A modular framework for composing prediction-based error-bounded lossy compressors," https://arxiv.org/abs/2111.02925, 2021, online.

[20] J. Tian, S. Di, K. Zhao, C. Rivera, M. H. Fulp, R. Underwood, S. Jin, X. Liang, J. Calhoun, D. Tao, and F. Cappello, "Cusz: An efficient gpu-based error-bounded lossy compression framework for scientific data," in *Proceedings of the ACM International Conference on Parallel Architectures and Compilation Techniques*, ser. PACT '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 3–15. [Online]. Available: https://doi.org/10.1145/3410463.3414624

[21] X. Liang, B. Whitney, J. Chen, L. Wan, Q. Liu, D. Tao, J. Kress, D. R. Pugmire, M. Wolf, N. Podhorszki *et al.*, "Mgard+: Optimizing multilevel methods for error-bounded scientific data reduction," *IEEE Transactions on Computers*, 2021.

[22] X. Liang, H. Guo, S. Di, F. Cappello, M. Raj, C. Liu, K. Ono, Z. Chen, and T. Peterka, "Toward feature-preserving 2d and 3d vector field compression," in *2020 IEEE Pacific Visualization Symposium (PacificVis)*, 2020, pp. 81–90.

[23] Y. Liu, S. Di, K. Zhao, S. Jin, C. Wang, K. Chard, D. Tao, I. Foster, and F. Cappello, "Optimizing multi-range based error-bounded lossy compression for scientific datasets," in *2021 IEEE 28th International Conference on High Performance Computing, Data, and Analytics (HiPC)*, 2021, pp. 394–399.

[24] ——, "Understanding effectiveness of multi-error-bounded lossy compression for preserving ranges of interest in scientific analysis," in *2021 7th International Workshop on Data Analysis and Reduction for Big Scientific Data (DRBSD-7)*, 2021, pp. 40–46.

[25] ——, "Optimizing error-bounded lossy compression for scientific data with diverse constraints," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 12, pp. 4440–4457, 2022.

[26] J. Liu, S. Di, K. Zhao, X. Liang, Z. Chen, and F. Cappello, "Dynamic quality metric oriented error bounded lossy compression for scientific datasets," in *IEEE/ACM The International Conference for High Performance computing, Networking, Storage and Analysis (IEEE/ACM SC2022)*, 2022, pp. 1–11.

[27] D. Tao, S. Di, H. Guo, Z. Chen, and F. Cappello, "Z-checker: A framework for assessing lossy compression of scientific data," *The International Journal of High Performance Computing Applications*, vol. 33, no. 2, pp. 285–303, 2019.

[28] X. Yu, S. Di, A. M. Gok, D. Tao, and F. Cappello, "cuz-checker: A gpu-based ultra-fast assessment system for lossy compressions," in *2021 IEEE International Conference on Cluster Computing (CLUSTER)*, 2021, pp. 307–319.

[29] Q. C. A. T. (QCAT), https://github.com/szcompressor/qcat, 2022, online.

[30] "Paraview visualizatoin tool," https://www.paraview.org/, 2022, online.

[31] R. Ballester-Ripoll, P. Lindstrom, and R. Pajarola, "Tthresh: Tensor compression for multidimensional visual data," *IEEE Transaction on Visualization and Computer Graphics*, vol. 26, pp. 2891–2903, 2019.

[32] R. Ballester-Ripoll and R. Pajarola, "Lossy volume compression using Tucker truncation and thresholding," *The Visual Computer*, pp. 1–14, 2015.

[33] J. Liu, S. Di, K. Zhao, S. Jin, D. Tao, X. Liang, Z. Chen, and F. Cappello, "Exploring autoencoder-based error-bounded compression for scientific data," pp. 294–306, 2021.

[34] B. Oppenheimer, "Cosmological simulations and dark matter," https://casa.colorado.edu/ boppenheimer/Lectures/Brazil_Lectures.pdf, online.

[35] D. Ziou and S. Tabbone, "Edge detection techniques - an overview," *INTERNATIONAL JOURNAL OF PATTERN RECOGNITION AND IMAGE ANALYSIS*, vol. 8, pp. 537–559, 1998.

[36] R. Adams, "Sobolev spaces," *Pure and Applied Mathematics*, vol. 65, 1975.

[37] "Sdrbench," https://sdrbench.github.io/, 2022, online.

[38] K. Zhao *et al.*, "SDRBench: Scientific data reduction benchmark for lossy compressors," https://arxiv.org/abs/2101.03201, 2021, online.

[39] Y. Collet, "Zstandard – real-time data compression algorithm," *http://facebook.github.io/zstd/*, 2015.

[40] X. Liang, S. Di, D. Tao, Z. Chen, and F. Cappello, "An efficient transformation scheme for lossy data compression with point-wise

relative error bound," in *2018 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE, 2018, pp. 179–189.
[41] X. Zou, T. Lu, W. Xia, X. Wang, W. Zhang, S. Di, D. Tao, and F. Cappello, "Accelerating relative-error bounded lossy compression for hpc datasets with precomputation-based mechanisms," in *2019 35th Symposium on Mass Storage Systems and Technologies (MSST)*, 2019, pp. 65–78.
[42] X. Zou, T. Lu, W. Xia, X. Wang, W. Zhang, H. Zhang, S. Di, D. Tao, and F. Cappello, "Performance optimization for relative-error-bounded lossy compression on scientific data," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 7, pp. 1665–1680, 2020.
[43] D. Tao, S. Di, X. Liang, Z. Chen, and F. Cappello, "Fixed-psnr lossy compression for scientific data," in *2018 IEEE International Conference on Cluster Computing (CLUSTER)*, 2018, pp. 314–318.
[44] "NYX simulation," https://amrex-astro.github.io/Nyx, 2019, online.
[45] J. E. Kay and et al., "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, no. 8, pp. 1333–1349, 2015.
[46] J. W. Hurrell, M. M. Holland, P. R. Gent, S. Ghan, J. E. Kay, P. J. Kushner, J.-F. Lamarque, W. G. Large, D. Lawrence, K. Lindsay *et al.*, "The community earth system model: a framework for collaborative research," *Bulletin of the American Meteorological Society*, vol. 94, no. 9, pp. 1339–1360, 2013.
[47] "Miranda application," https://wci.llnl.gov/simulation/computer-codes/miranda.

APPENDIX

A. Description of Error-bounded Lossy Compressors

- ZFP: ZFP is an outstanding error-bounded lossy compressor designed based on near-orthogonal data transform, which exhibits very high execution performance and also high data quality (especially for 3D datasets). ZFP supports three compression modes – accuracy (or absolute error bound), precision, and fix-rate. According to the prior studies (also as mentioned in ZFP's official website), the accuracy mode generally leads to the best rate distortion, so we use the accuracy mode in our investigation.
- SZ2 [16]: SZ2 is a classic prediction-based lossy compressor for scientific datasets, which is designed based on a prediction-based compression model. It has a hybrid predictor (Lorenzo and Linear-regression), followed by quantization, Huffman encoding and Zstd [39]. SZ2 supports compression based on absolute error bound (ABS), value-range based relative error bound (VR_REL), pointwise relative error bound (PW_REL) [40]–[42] and peak signal to noise ratio (PSNR) [43]. In our investigation, we choose the ABS mode because of the following reasons: (1) VR_REL and PSNR are equivalent to ABS in the regard of rate distortion, because SZ actually transforms them to corresponding absolute error bounds and then performs compression using the ABS mode [14], [43]; (2) PW_REL always suffers from very low compression ratios, as verified by prior works [40].
- SZ3 [17], [19]: SZ3 is a more advanced compressor over SZ2 especially in the situation with high compression ratios. In principle, the key difference between SZ3 and SZ2 is that SZ3 adopts a more effective predictor called dynamic spline interpolation [], which can obtain abut 2X higher compression ratios with the same PSNR. SImilar to SZ2, we also use the ABS mode to perform the compression for SZ3.
- SZx [3]: SZx (an ultra-fast error-bounded lossy compressor) is designed particularly for the high-speed requirement. SZx is about 3-4X as fast as SZ2/3 and ZFP on single CPU and about 10X faster on GPU, with a reasonably degraded compression ratios. SZx's compression ratio is generally around 5-30 [3], which is still much higher than the lossless compressors such as Zstd. SZx supports only absolute error bound (ABS) mode, so all the experiments with SZx in our study are based on ABS.
- FPZIP [12]: FPZIP is a classic lossy compressor which is developed based on Lorenzo predictor. FPZIP controls the data distortion via an integer (called precision), whose value is in the range of [1,32] (higher number means higher precision). A prior study about preserving vector-based critical points (mainly determined by gradients) in lossy compression showed that FPZIP outperforms both SZ2 and ZFP in rate distortion. As such, we include FPZIP in our study related to gradient metrics.

TABLE III: Data Fields of Applications Used by Evaluations

Datasets	of application	size per field	precision
baryon density log.f32	NYX	$512 \times 512 \times 512$	single
CLOUDf48.log10.bin.f32	Hurricane ISABEL	$100 \times 500 \times 500$	single
CLDHGH 1 1800 3600.dat	CESM-ATM	1800×3600	single
density.d64	Miranda	$256 \times 384 \times 384$	double

B. Analysis Workflow and Tool chain

1) We run all the error-bounded lossy compressors listed in Appendix A by their executables built from their packages.
2) Then, we perform the corresponding decompression for each compressed data file to get the lossy reconstructed data files, based on four absolute error bound (1E-1, 1E-2, 1E-3 and 1E-4). Since FPZIP does not have absolute error bound mode but the precision mode, we run it with different precision settings so as to obtain comparable compression ratios with other compressors.
3) Then, we generate all the derivative metrics or datasets which are listed in Section III-B. Specifically, we implemented all the derivative metrics in QCAT [29], and also integrated them in the Z-checker library/tool [27] such that our codes can also be used by other researchers if needed.
4) Finally, we use the analytical functions (such as *calculateSSIM*) provided by QCAT to perform an in-depth analysis, and also use the visualization tool (either *PlotSliceImage* offered by QCAT [29] or the vis panel offered by Paraview [30]) to plot the data for further understanding the impact of lossy compressors on these derivative metrics.

C. Application Datasets Used in The Paper

- NYX [44]: An adaptive mesh, cosmological hydrodynamics simulation code.
- Hurricane: A simulation of a hurricane from the National Center for Atmospheric Research in the United States.
- CESM-ATM [45], [46]: Community Earth System Model (CESM) is a fully-coupled global climate model providing computer simulations of the Earth's climate changes. CESM involves multiple sub-modules such as ICE, Atmosphere (ATM), Ocean (OCN), and Land (LND). We use ATM model in our experiments.
- Miranda [47]: A radiation hydrodynamics code designed for large-eddy simulation of multicomponent flows with turbulent mixing.

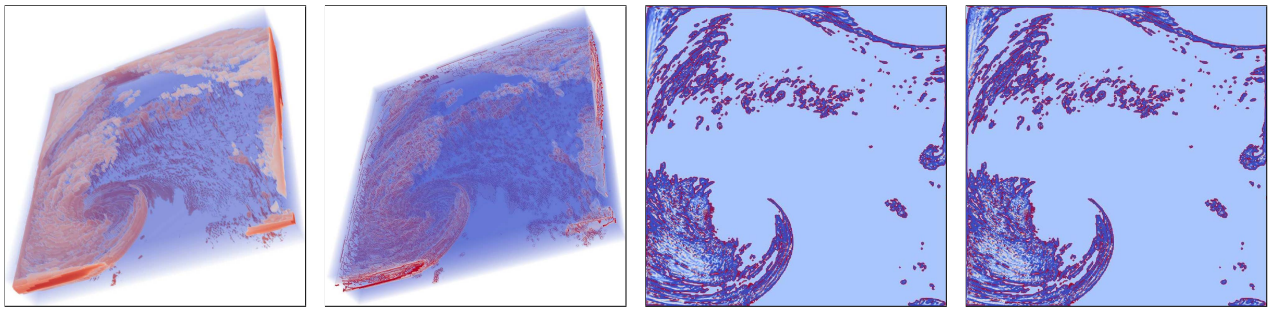
We present the details in Table III.

D. Visualizing Impact of Lossy Compression on Laplacian

Figure 3 visualizes the impact of SZ3 on Laplacian with Hurricane simulation dataset, based on different error bounds.

E. Evaluation Results about Sobolev Norms

The evaluation results about the impact of lossy compression on Sobolev norms are presented in Table IV through VII.



(a) RawD, 3D (b) LRaw, 3D (c) LRaw, 2D (d) LRec, SZ3, e=1E-2, CR=27

(e) LRec, SZ3, e=1E-1, CR=53 (f) LRec, SZ3, e=5E-1, CR=91 (g) LRec, SZ3, e=1, CR=111 (h) LRec, SZ3, e=3, CR=154

Fig. 3: Laplacian Distortion and Compression Error Bound in SZ3, Hurricane

TABLE IV: The Sobolev Norm 2, CESM-ATM

CLDHGH_1_1800_3600.dat								
Cmprs & e		CR	Order 0	Diff 0	Order 1	Diff 1	Order 2	Diff 2
Raw Data		1	0.251206	0	0.251428	0	0.251479	0
SZx	1E-1	39.57	0.250500	0.000706	0.250994	0.000434	0.252732	0.001253
	1E-2	4.70	0.251114	0.000092	0.251335	0.000093	0.251402	0.000077
	1E-3	3.19	0.251194	0.000012	0.251415	0.000013	0.251466	0.000013
	1E-4	2.30	0.251205	0.000001	0.251427	0.000001	0.251478	0.000001
SZ2	1E-1	809.42	0.251100	0.000106	0.251303	0.000125	0.251408	0.000071
	1E-2	60.18	0.251232	0.000026	0.251479	0.000051	0.251795	0.000316
	1E-3	17.15	0.251207	0.000001	0.251429	0.000001	0.251488	0.000009
	1E-4	7.04	0.251206	0	0.251428	0	0.251479	0
SZ3	1E-1	3079.48	0.251274	0.000068	0.251480	0.000052	0.251504	0.000025
	1E-2	137.22	0.251190	0.000016	0.251409	0.000019	0.251502	0.000023
	1E-3	19.77	0.251207	0.000001	0.251429	0.000001	0.251485	0.000006
	1E-4	7.08	0.251206	0	0.251428	0	0.251479	0
ZFP	1E-1	17.90	0.251283	0.000077	0.251515	0.000087	0.251646	0.000167
	1E-2	9.44	0.251199	0.000007	0.251421	0.000007	0.251476	0.000003
	1E-3	5.34	0.251207	0.000001	0.251429	0.000001	0.251480	0.000001
	1E-4	3.24	0.251206	0	0.251428	0	0.251479	0
FPZIP	p=14	21.94	0.248350	0.002856	0.248575	0.002853	0.248691	0.002788
	p=16	12.74	0.250488	0.000718	0.250709	0.000719	0.250765	0.000714
	p=18	7.64	0.251027	0.000179	0.251248	0.000180	0.251299	0.000180
	p=20	5.24	0.251161	0.000045	0.251383	0.000045	0.251434	0.000045

TABLE V: The Sobolev Norm 2, NYX

baryon_density_log.f32								
Cmprs & e		CR	Order 0	Diff 0	Order 1	Diff 1	Order2	Diff2
Raw Data		1.00	1.441198	0	1.473953	0	1.487979	0
SZx	1	4.97	1.390223	0.050975	1.417625	0.056328	1.418664	0.069315
	1E-1	3.09	1.434077	0.007121	1.466602	0.007351	1.480721	0.007258
	1E-2	2.40	1.440302	0.000896	1.473026	0.000927	1.487042	0.000937
	1E-3	1.89	1.441086	0.000112	1.473837	0.000116	1.487861	0.000118
SZ2	1	217.09	1.430359	0.010839	1.466008	0.007945	1.544127	0.056148
	1E-1	21.65	1.442209	0.001011	1.476344	0.002391	1.507392	0.019413
	1E-2	10.48	1.441210	0.000012	1.473981	0.000028	1.488215	0.000236
	1E-3	5.27	1.441199	0.000001	1.473953	0	1.487981	0.000002
SZ3	1	1034.23	1.390223	0.050975	1.417625	0.056328	1.441138	0.046841
	1E-1	59.20	1.442111	0.000913	1.475156	0.001203	1.493628	0.005649
	1E-2	11.79	1.441214	0.000016	1.473984	0.000031	1.488205	0.000226
	1E-3	5.37	1.441199	0.000001	1.473953	0	1.487981	0.000002
ZFP	1	24.20	1.441280	0.000082	1.474331	0.000378	1.490815	0.002836
	1E-1	6.71	1.441194	0.000004	1.473952	0.000001	1.488015	0.000036
	1E-2	4.12	1.441199	0.000001	1.473954	0.000001	1.487980	0.000001
	1E-3	2.97	1.441198	0	1.473953	0	1.487979	0
FPZIP	p=10	17.14	1.218766	0.222432	1.253661	0.220292	1.324954	0.163025
	p=12	12.76	1.378017	0.063181	1.410079	0.063874	1.431586	0.056393
	p=14	9.12	1.424908	0.016290	1.457346	0.016607	1.471807	0.016172
	p=16	6.24	1.437095	0.004103	1.469760	0.004193	1.483785	0.004194

TABLE VI: The Sobolev Norm 2, Miranda

density.d64								
Cmprs & e		CR	Order 0	Diff 0	Order 1	Diff 1	Order2	Diff2
Raw Data		1.00	1.923554	0	1.939679	0	2.138136	0
SZx	1E-1	38.05	1.920580	0.002974	1.935242	0.004437	2.119065	0.019071
	1E-2	27.82	1.923153	0.000401	1.939087	0.000592	2.135641	0.002495
	1E-3	21.51	1.923501	0.000053	1.939602	0.000077	2.137820	0.000316
	1E-4	15.85	1.923550	0.000004	1.939674	0.000005	2.138116	0.000020
SZ2	1E-1	603.41	1.922877	0.000677	1.938750	0.000929	2.134620	0.003516
	1E-2	188.50	1.923544	0.000010	1.939668	0.000011	2.138114	0.000022
	1E-3	122.13	1.923558	0.000004	1.939683	0.000004	2.138139	0.000003
	1E-4	43.26	1.923554	0	1.939679	0	2.138137	0.000001
SZ3	1E-1	2668.65	1.922057	0.001497	1.936769	0.002910	2.120703	0.017433
	1E-2	618.52	1.923658	0.000104	1.939789	0.000110	2.138293	0.000157
	1E-3	154.48	1.923569	0.000015	1.939695	0.000016	2.138163	0.000027
	1E-4	56.89	1.923554	0	1.939679	0	2.138137	0.000001
ZFP	1E-1	75.40	1.923579	0.000025	1.939711	0.000032	2.138237	0.000101
	1E-2	51.40	1.923551	0.000003	1.939675	0.000004	2.138125	0.000011
	1E-3	35.60	1.923554	0	1.939679	0	2.138138	0.000002
	1E-4	23.30	1.923554	0	1.939679	0	2.138137	0.000001
FPZIP	p=14	39.59	1.703893	0.219661	1.720822	0.218857	1.932323	0.205813
	p=16	37.28	1.868802	0.054752	1.884560	0.055119	2.079579	0.058557
	p=18	36.03	1.910055	0.013499	1.926076	0.013603	2.123459	0.014677
	p=20	34.81	1.920231	0.003323	1.936332	0.003347	2.134537	0.003599

TABLE VII: The Sobolev Norm 2, Hurricane ISABEL

CLOUDf48.log10.bin.f32								
Cmprs & e		CR	Order 0	Diff 0	Order 1	Diff 1	Order2	Diff2
Raw Data		1.00	12.193243	0	12.258431	0	12.639667	0
SZx	1	15.73	12.104116	0.089128	12.160559	0.097872	12.509620	0.130047
	1E-1	13.08	12.187895	0.005348	12.252512	0.005919	12.631488	0.008179
	1E-2	10.75	12.192575	0.000668	12.257692	0.000739	12.638643	0.001024
	1E-3	9.59	12.193160	0.000083	12.258338	0.000092	12.639539	0.000128
SZ2	1	102.83	12.115463	0.077780	12.174132	0.084299	12.522736	0.116931
	1E-1	40.50	12.192696	0.000547	12.257943	0.000487	12.639706	0.000039
	1E-2	26.81	12.193083	0.000160	12.258271	0.000160	12.639513	0.000154
	1E-3	18.56	12.193243	0	12.258431	0	12.639667	0
SZ3	1	111.08	12.243989	0.050746	12.308690	0.050260	12.695751	0.056084
	1E-1	53.44	12.165990	0.027253	12.231398	0.027033	12.613321	0.026346
	1E-2	27.00	12.192990	0.000253	12.258190	0.000240	12.639446	0.000221
	1E-3	18.70	12.193290	0.000047	12.258478	0.000047	12.639713	0.000046
ZFP	1	19.30	12.195410	0.002167	12.260614	0.002184	12.641970	0.002303
	1E-1	12.30	12.193313	0.000070	12.258501	0.000070	12.639736	0.000069
	1E-2	9.62	12.193236	0.000007	12.258424	0.000007	12.639661	0.000006
	1E-3	7.86	12.193246	0.000003	12.258434	0.000003	12.639670	0.000003
FPZIP	p=14	41.22	12.106090	0.087153	12.171206	0.087225	12.552070	0.087597
	p=16	30.83	12.166302	0.026941	12.231404	0.027027	12.612146	0.027521
	p=18	24.45	12.181359	0.011884	12.246458	0.011972	12.627181	0.012486
	p=20	19.94	12.192394	0.000849	12.257587	0.000843	12.638856	0.000811