

Robert Underwood, Sheng Di, Julie Bessac, Franck Cappello
Argonne National Laboratory

Understanding the Effects of Modern Lossless and Lossy Compressors on the Community Earth Science Model

Why Lossy Compression?

Climate and Weather Science Has Big Data

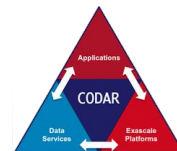
Compressors Considered

How do modern compressors work?

Compressor	Description
SZ	Predicts values using regression and a lorenzo (neighbor-based) method + Zstandard
ZFP	Transforms values using a near orthogonal transform
MGARD	Represents data using multi-level nodal coefficients + Zstandard
Digit Rounding	Rounds to a specified precision then lossless (Zlib/lz77) compression
Bit Grooming	Exploits floating point properties to compress then lossless (Zlib/lz77) compression
TThresh	Represents the data as a higher order Singular Value Decomposition (SVD)
SZ3	Variant of SZ using Interpolation and 2-level lorenzo prediction
FPZIP	Lorenzo Prediction + Custom Encoding
ZStandard	LZ77 with a large window + a finite state machine-based encoding method
And more	Created a benchmark that for any LibPressio-based compressor

Interfacing with Compressors

LibPressio






- Every compressor has a different interface - we use LibPressio
 - Supports 30+ lossless and lossy compression methods
 - LibPressio abstracts between many of these differences
 - Provides HDF5 Filter, CLI, R, and Python NumCodecs-like interfaces
 - Provides an interface to provide metrics capabilities
- LibPressio's OptZConfig automatically finds settings for quality requirements
 - Uses non-linear black box optimization techniques
 - Works with any LibPressio Compressor

R. Underwood, J. C. Calhoun, S. Di, A. Apon and F. Cappello, "OptZConfig: Efficient Parallel Optimization of Lossy Compression Configuration," in IEEE Transactions on Parallel and Distributed Systems, doi: 10.1109/TPDS.2022.3154096.

Compressors Considered

What Classes of Compressors Do We Consider?

-  Lossy Compressor
-  Lossless and Lossy Compressor
-  Lossless Compressor

Compressor	Description
SZ	Predicts values using regression and a lorenzo (neighbor-based) method
ZFP	Transforms values using a near orthogonal transform
MGARD	Represents data using multi-level nodal coefficients
Digit Rounding	Rounds to a specified precision then lossless (Zlib/lz77) compression
Bit Grooming	Exploits floating point properties to compress then lossless (Zlib/lz77) compression
TThresh	Represents the data as a higher order Singular Value Decomposition (SVD)
SZ3	Variant of SZ using Interpolation and 2-level lorenzo prediction
FPZIP	Lorenzo Prediction + Custom Encoding
ZStandard	LZ77 with a large window + a finite state machine-based encoding method
And more	Created a benchmark that for any LibPressio-based compressor

- Latest Version Evaluated Previously – Included for Comparison
- Older Version Evaluated Previously; Changes Improve Results
- Not Previously Evaluated to Our Knowledge

Compressors Considered

What have prior works considered that we consider?

Compressor	Description
SZ	Predicts values using regression and a lorenzo (neighbor-based) method
ZFP	Transforms values using a near orthogonal transform
MGARD	Represents data using multi-level nodal coefficients
Digit Rounding	Rounds to a specified precision then lossless (Zlib/lz77) compression
Bit Grooming	Exploits floating point properties to compress then lossless (Zlib/lz77) compression
TThresh	Represents the data as a higher order Singular Value Decomposition (SVD)
SZ3	Variant of SZ using Interpolation and 2-level lorenzo prediction
FPZIP	Lorenzo Prediction + Custom Encoding
ZStandard	LZ77 with a large window + a finite state machine-based encoding method
And more	Created a benchmark that for any LibPressio-based compressor

Quantities of Interest

How Similar is Good Enough?

These values are from
previous speaker

Name	Description	Values	Threshold
SSIM/d-SSIM	Structural Image Similarity Metric	[0,1]	$\geq 0.995^*$
KS-test	P value of for the Kolmogorov-Smirnov	[0,1]	≥ 0.05
R^2	Pearson's Coefficient of Determination	[0,1]	≤ 0.99999
Spatial Relative Error	% of elements that exceed a value range relative error δ	[0,1]	$\leq 5\%$ at $\delta \leq 1e-4$

* several different values have been proposed: 0.98 [Baker 2017], 0.99 [Klöwer 2021], and .99995 [Baker 2019]. The listed value comes from [Pinard 2020]

Quantities of Interest

How Similar is Good Enough?

In practice, one of these is the limiting factor

Name	Description	Values	Threshold
SSIM/d-SSIM	Structural Image Similarity Metric	[0,1]	$\geq 0.995^*$
KS-test	P value of for the Kolmogorov-Smirnov	[0,1]	≥ 0.05
R^2	Pearson's Coefficient of Determination	[0,1]	≤ 0.99999
Spatial Relative Error	% of elements that exceed a value range relative error δ	[0,1]	$\leq 5\%$ at $\delta \leq 1e-4$

* several different values have been proposed: 0.98 [Baker 2017], 0.99 [Klöwer 2021], and .99995 [Baker 2019]. The listed value comes from [Pinard 2020]

The CESM datasets

Model	Datatype	Total Size	Buffer Size
Atmos	float32	1.5TB	642MB
Ocean	float32	235GB	1.35GB
Land	float32	41GB	216KB
Ice	float32	33GB	480KB

Which Compressors do best?

It Depends



- Right now
 - lossy compressors (SZ2/SZ3, ZFP) do better on Atmosphere and Ocean
 - Specialized encoding methods (FPZip) do best on Land and Ice
- Why?
 - The metadata (i.e. settings used) and entropy data (i.e. huffman trees) cost too much overhead for small buffers
 - Z-standard's features may point a way forward for lossy improvements
 - “Common Dictionary” and “External Metadata”
 - Would require improvements to HDF5 to fully leverage
 - However, Z-standard's principle not well suited for floating point
 - Larger Chunk sizes improve the situation
 - more data per netcdf variable, set larger chunks in netcdf
 - The latency differences may be smaller than you'd think

Which Compressors do best?

Which metrics are the bottleneck?

- For SZ3, the KS-test tends to be limiting
 - To pass the everything, SZ needs a REL error bound $\ll 1e-15$
 - KS, dSSIM, SRE, R^2 : CR~1.3 (slightly over Z-standard)
 - dSSIM, SRE, R^2 : CR~59.81
 - SRE and R^2 : CR~93
- For ZFP, the d-SSIM tends to be limiting
 - To pass everything, ZFP needs a REL error bound of $<6e-5$
 - KS, dSSIM, SRE, R^2 : CR~3.8
 - KS, SRE, R^2 : CR~3.8
 - SRE and R^2 : CR=13.27
- Spatial Relative Error and R^2 tended to be met
- The best lossless compressor: FPZIP CR=2.2
- Other lossy compressors have either worse CR or can't pass these tests!

Questions

runderwood@anl.gov

Understanding the Effects of Modern Lossless and Lossy Compressors on the Community Earth Science Model

Robert Underwood, Sheng Di, Julie Bessac, Franck Cappello
Argonne National Laboratory