# Fine-tuning evaluation metrics for lossy compression of CESM data

**Allison Baker**
Applications Scalability and Performance Group (ASAP)
Computational Information Systems Laboratory, NCAR

**Dorit Hammerling, Alex Pinard**
Applied Mathematics and Statistics, Colorado School of Mines,

**Haiying Xu**
ASAP, NCAR,

…
and many others!

June 15, 2022

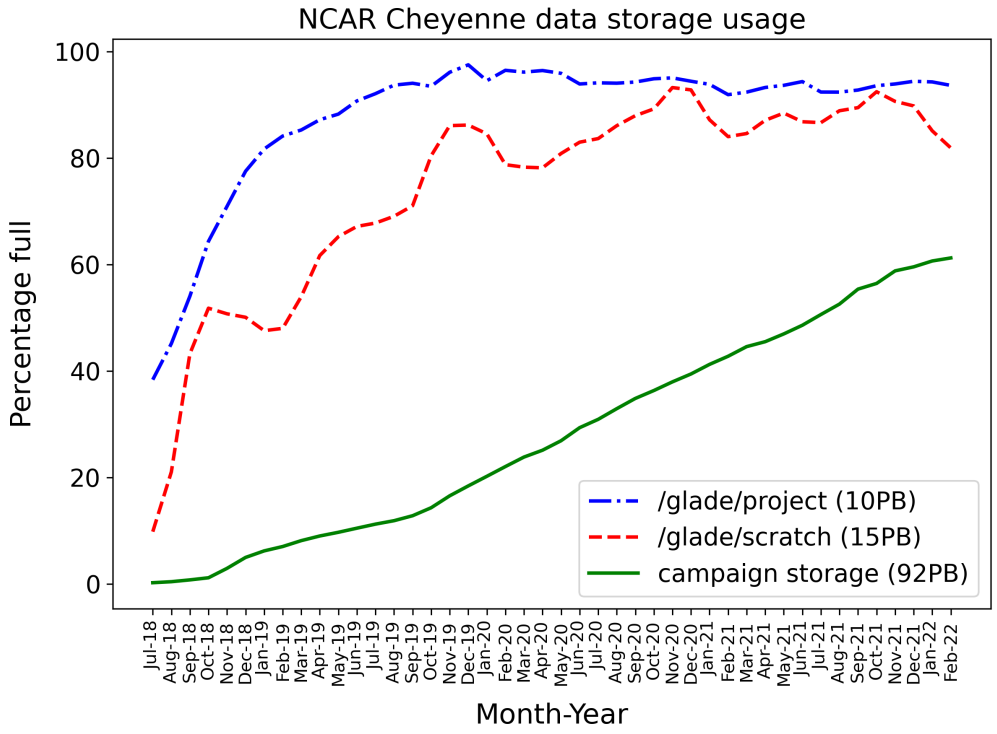
NCAR

NSF

# Why use data compression on CESM data?

INCREASING COMPUTE POWER → MORE AND MORE DATA!

*Data generation trends are unsustainable!*




Storage resources are limiting climate science objectives!

- shorter simulation lengths
- smaller ensembles
- lower resolutions
- less frequent output
- fewer output variables

Idea: Use lossy compression to reduce CESM storage

# Why lossy compression?

Data compression basics

Compression: X → C

Reconstruction: C → R

Data compression types:

Lossless: R = X (e.g., gzip)

Lossy: R ~ X

- Lossless: relatively ineffective on numerical simulation data
  - CESM CAM data: < 2x reduction
  - last bits (digits) are essentially random, e.g.: T = 290.1584967880457
- Lossy: more substantial reduction BUT non-trivial to evaluate info loss


Goal: Use lossy compression to reduce CESM storage
*…without (negatively) impacting science results!*

# Lossy compression and CESM data

Challenges:

- Scientists' reluctance to lose any information
- How best to evaluate the information loss for climate data?
  - often don't know how data will be used/analyzed
  - output variables have very different characteristics
  - spatial and temporal dependencies!
  - simple metrics (like RMSE) are not able to capture all the kinds of compression artifacts that may be of importance

Our focus:

Evaluating the ***effect of lossy compression on CESM data*** via analysis tools that emulate the key aspects of climate data analysis in order to determine ***quantifiable metrics*** that can be used to ***predict optimal compression.***

# Lossy compression and CESM data so far…

1) Establish feasibility:

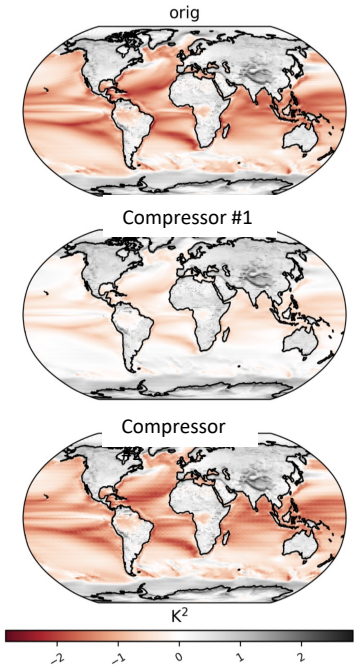
- choose per-variable compression via ensemble-based metrics
  (s.t. compression-induced differences do not exceed variability)

2) Direct experience:

- provide scientists reconstructed CESM-LENS1 data via blind test
- can they differentiate between compressed and uncompressed?
  - did not notice a difference using their standard analysis tools
  - can detect difference using clever methods (might not matter)

3) Spatio-temporal statistical analysis is important

- compression has effects at fine spatial and temporal scales
  that are masked by global statistics….
- need measurements that are *not ensemble-based*
  - ensembles are expensive and inhibit automation
    (properties not known in advance)


N-S contrast variances to measure fine-scale spatial variability (TS)

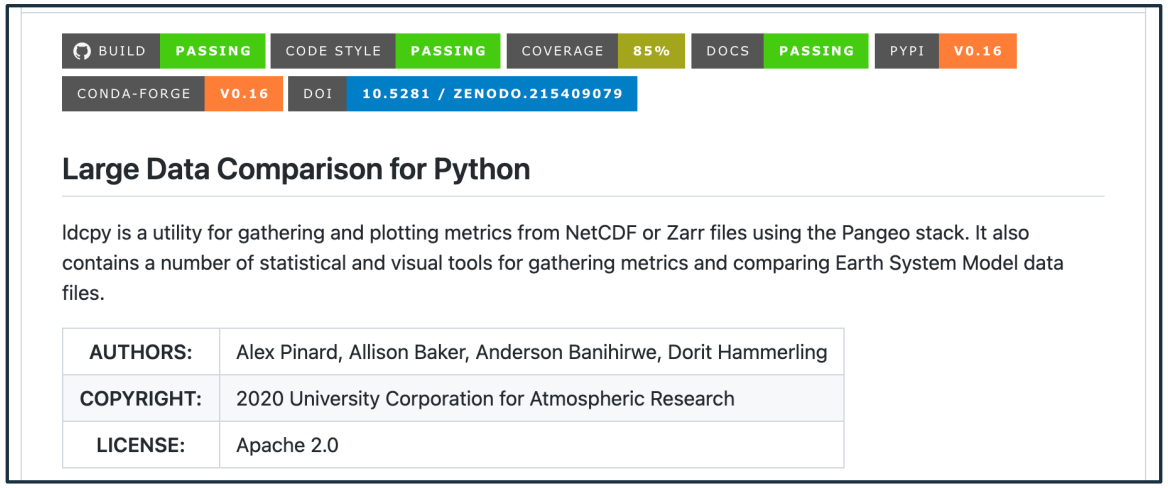
# Spatio-temporal analysis tools

We use analysis tools that emulate the key aspects of climate data analysis:

- gradients in space and time
- cumulative effects in time
- climate-relevant budgets (derived quantities)
- changes in variability over space or time
- changes in the statistical distribution (extremes, skewness)

## ldcpy: Large Data Comparison for Python package

- facilitates data analysis and visual **comparison** of datasets in climate science
- we use it to compare original and compressed data




https://github.com/NCAR/ldcpy

Features:

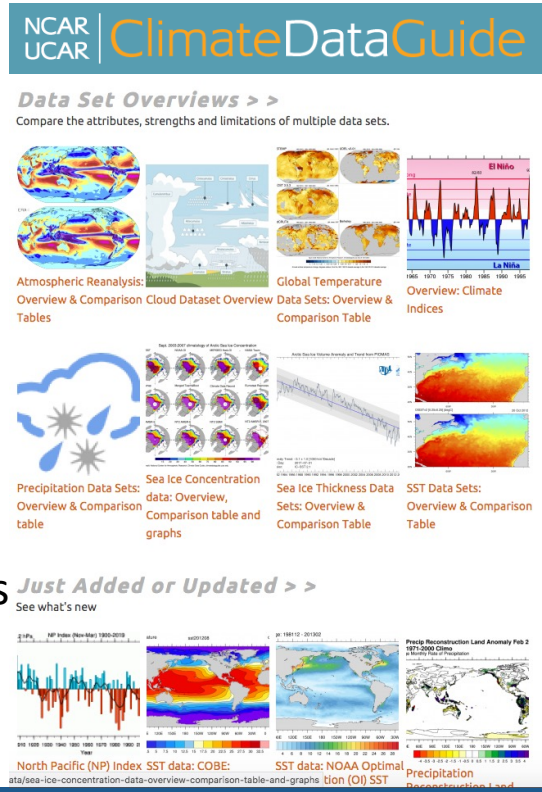
- Interoperability with the **Pangeo** software ecosystem
- Easy interaction through Jupyter Notebooks
- Suitable for wide range of data volumes (single time slice to many years)
- Extensible analysis and plotting capabilities

# Identifying artifacts from lossy compression

## 4) quantifiable metrics (and thresholds) for derived quantities and statistics

*We are working on determining a relationship between the suite of derived quantities and statistics and a* ***few quantifiable metrics*** *such that we can use thresholds* ***to determine optimal compression levels****.*

- Pearson correlation coefficient
- Kolmogorov-Smirnov (K-S) test
- Spatial relative error
- Visual similarity

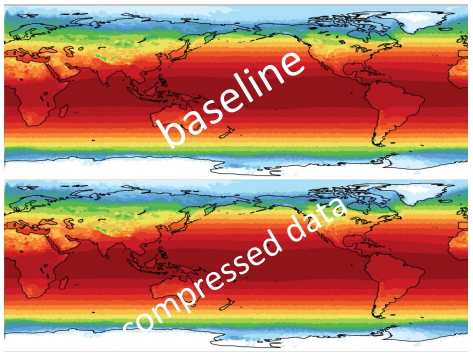



Example:

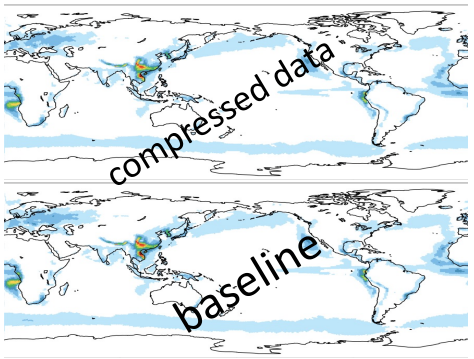
- visualization is critical for post-processing analysis and diagnostics
- it's typically the first interaction with the data

# Visual Similarity Metric

- From a visual evaluation study, we determined a suitable metric, **SSIM**, and **threshold** for CESM data




**IDENTICAL**




**DIFFERENT**

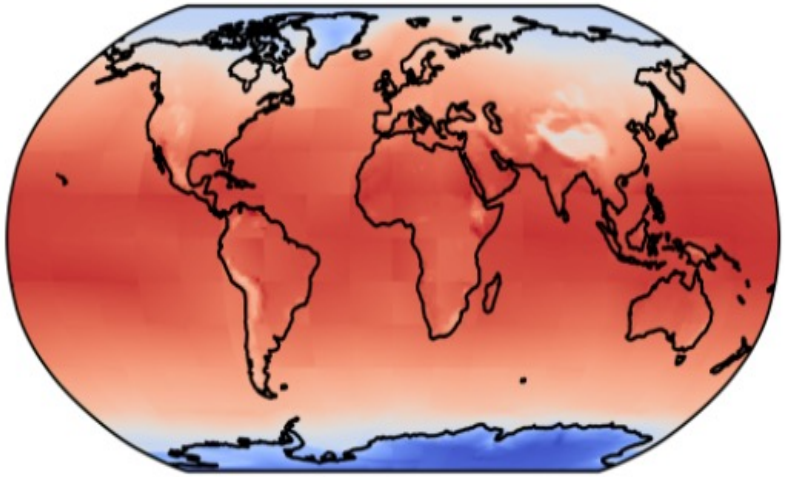
**DSSIM (Data Structural Similarity Index Measure) :**

- newly developed (modified SSIM) to apply *directly to floating-point data*
- no plots = cheaper (key for automation!)
- higher (closer to 1) DSSIM is more conservative
- DSSIM thresholds control quality (e.g., aggressive = .95, moderate =.995, …)
- ***general idea*** of whether images generated from the data are likely to have a difference

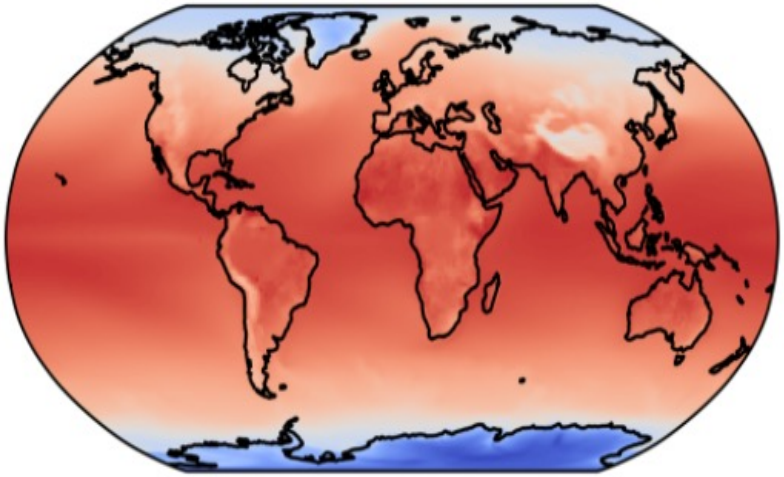
# DSSIM is now our primary metric for selecting compression

Surface temperature: 2 compressors with the same compression ratio (CR):

DSSIM = .667



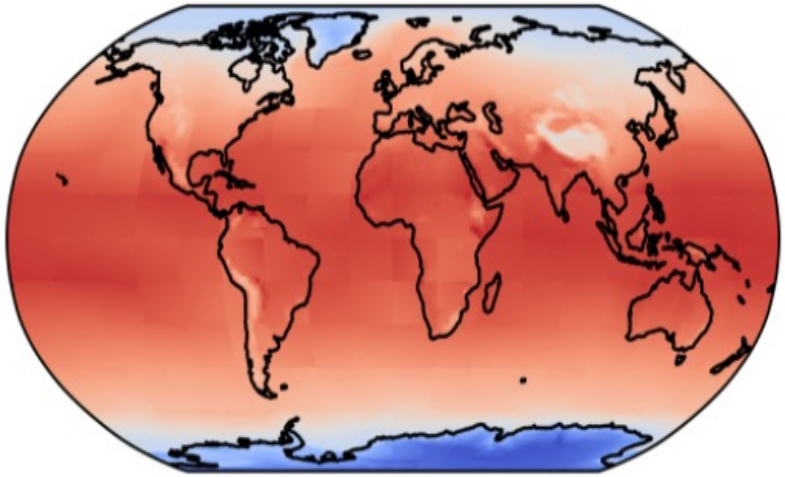
DSSIM = .998



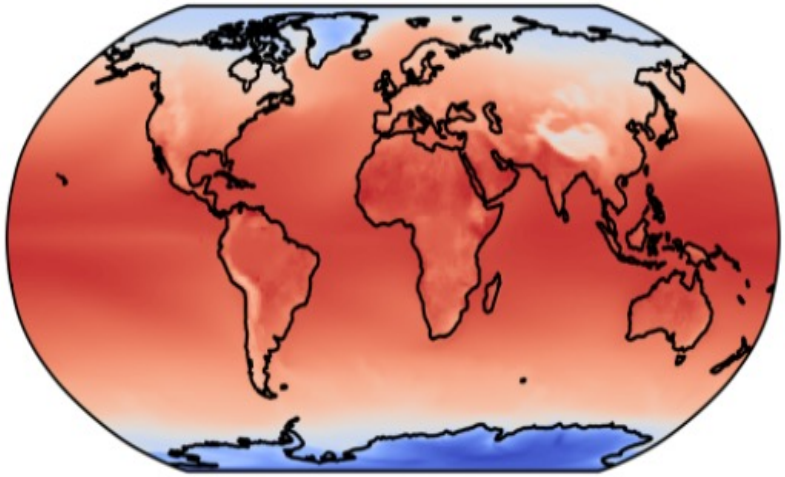
# DSSIM is now our primary metric for selecting compression

Surface temperature: 2 compressors with the same compression ratio (CR):

DSSIM = .667



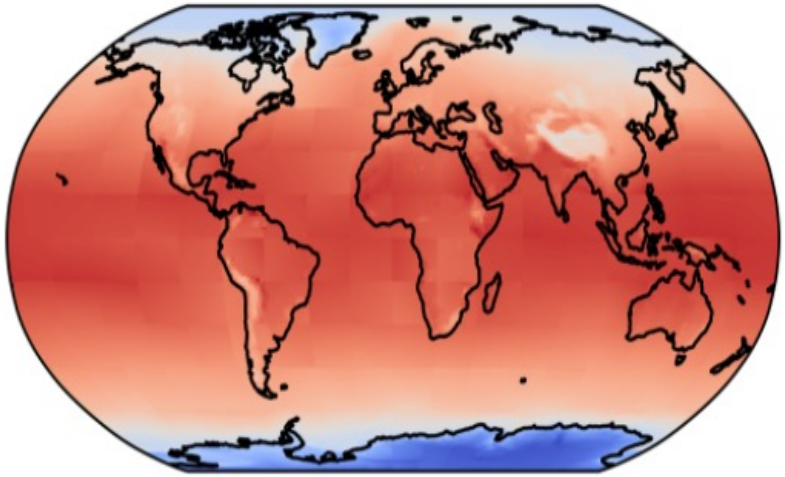
DSSIM = .998



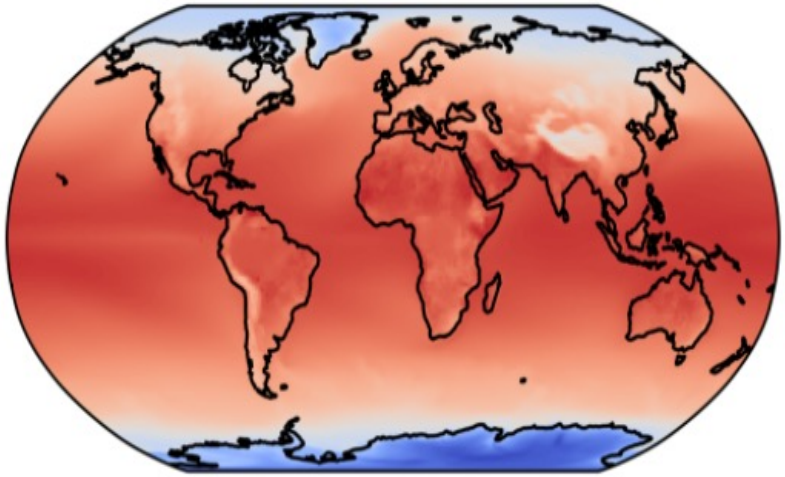
# DSSIM is now our primary metric for selecting compression

Surface temperature: 2 compressors with the same compression ratio (CR):

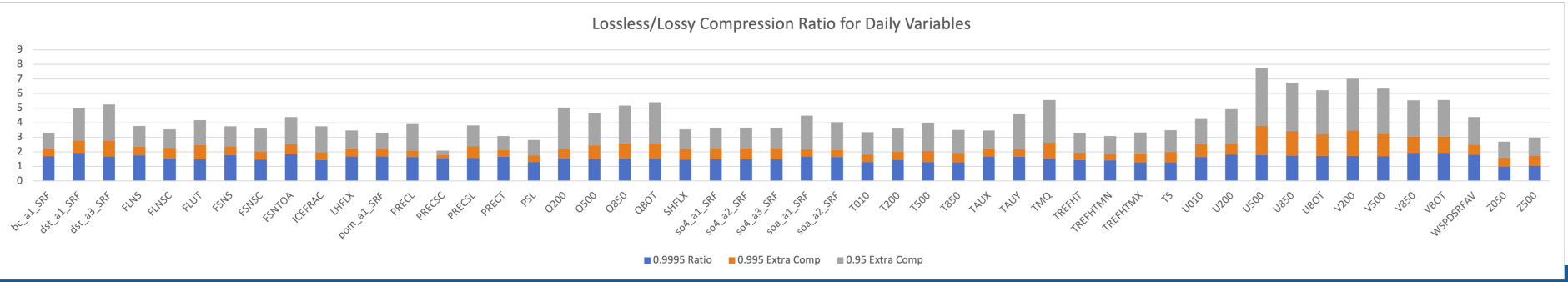
DSSIM = .667



DSSIM = .998

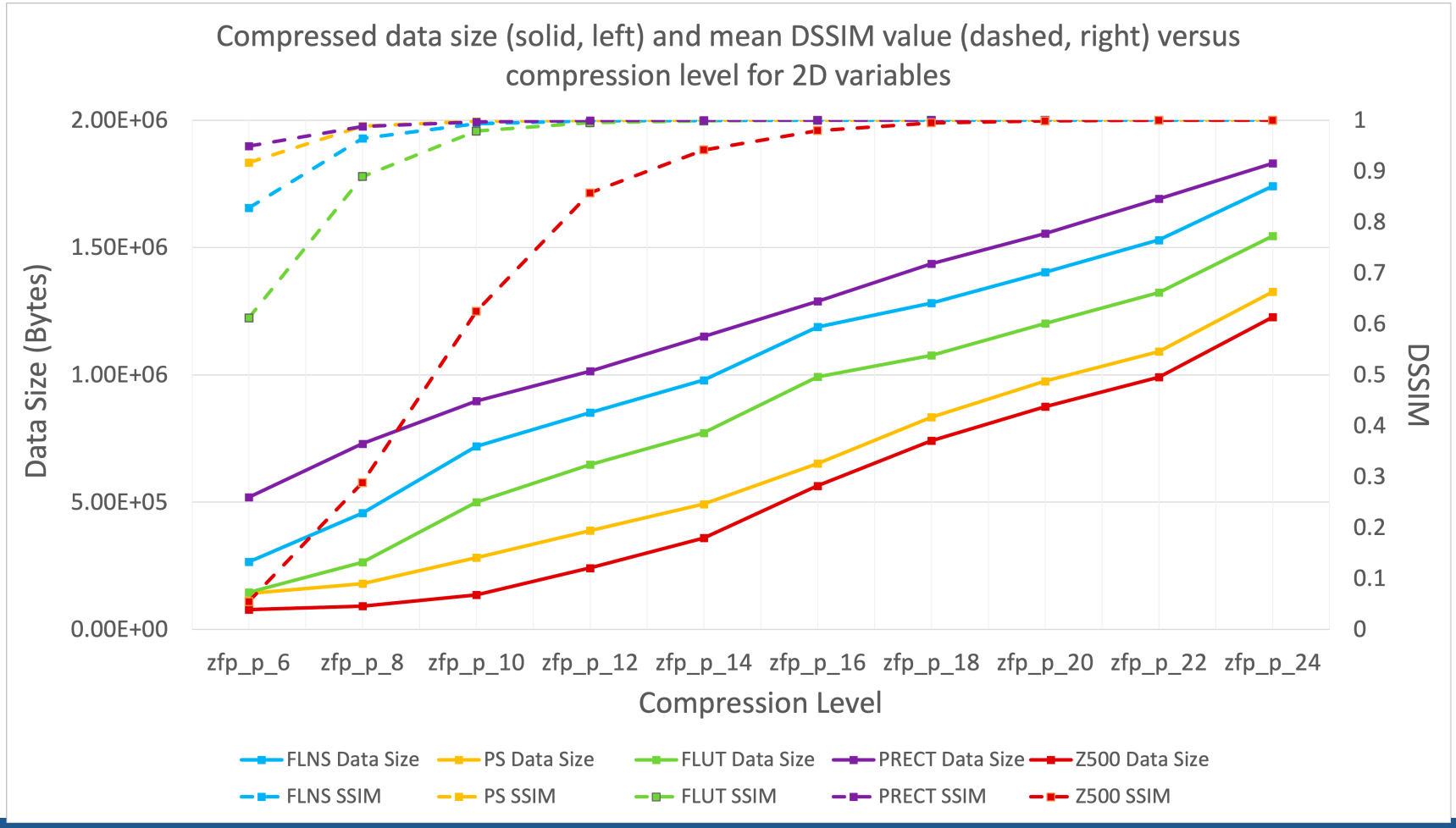


Same DSSIM threshold results in different amounts of compression for each variable…



# Metrics (Identifying artifacts from lossy compression)

- file size reduction is linear with zfp precision
- DSSIM is not (diminishing returns in the data fidelity as we approach lossless compression)

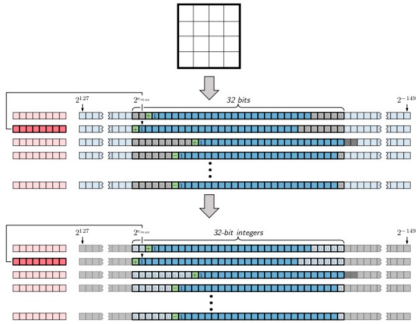


# Current compressor candidates for CESM data

- must work with NetCDF4 (collaboration has been key!)
- recent development: HDF5 filters now exist for two leading DOE compressors

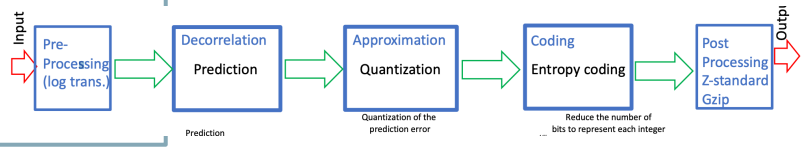
## ZFP (Lindstom, LLNL)

- compresses 4-d blocks (d = dim) via a floating-point representation with a single common exponent per block, an orthogonal block transform, and embedded encoding
- registered ZFP plugin for HDF5




## SZ3 (Capello and Di, ANL)

- Error-bounded lossy compressor framework
- Predictive-based approach with customized encoding
- registered SZ3 plugin for HDF5

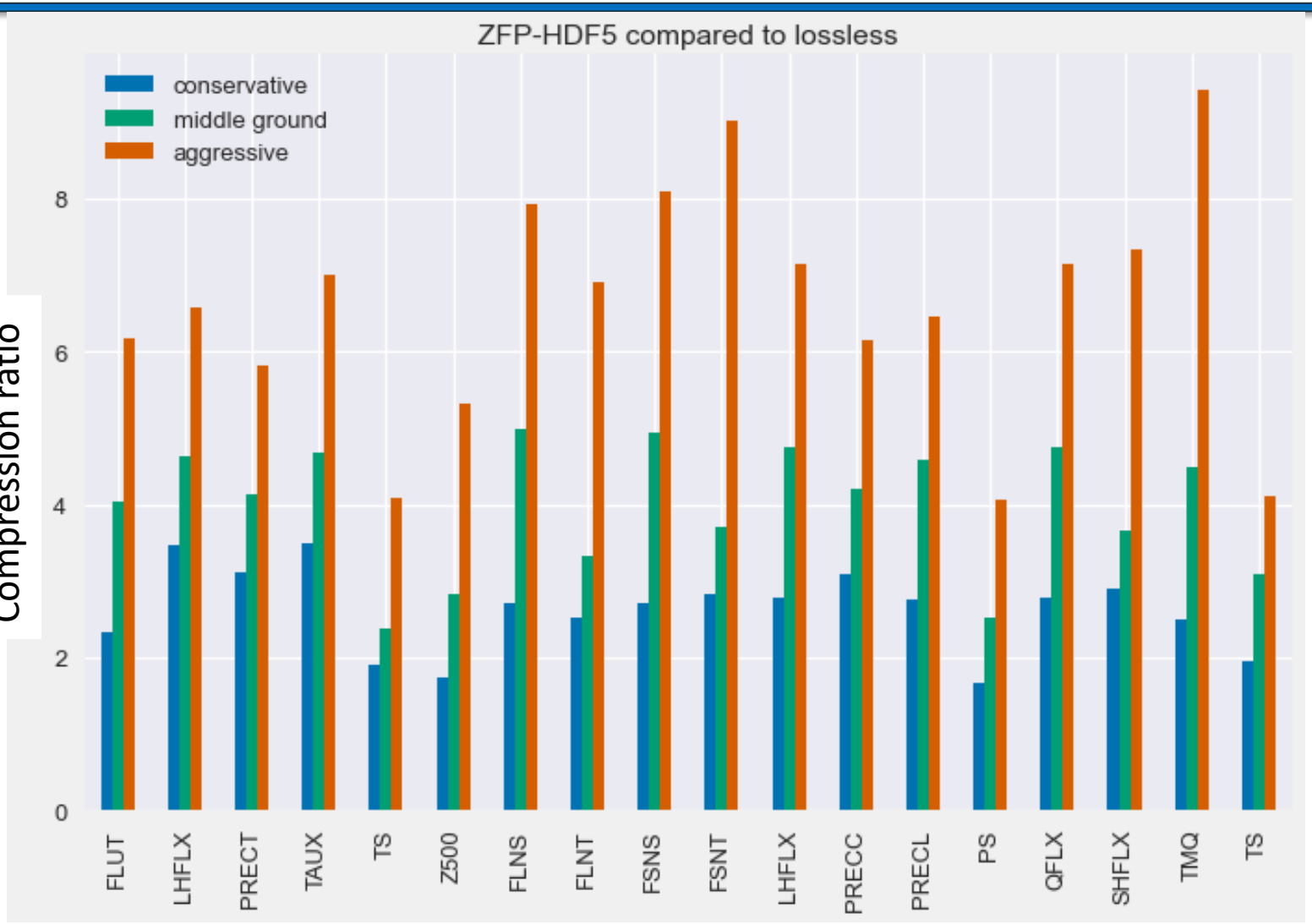



## Bit Grooming / Granular BitRound (Zender, UCI)

- Quantization pre-filter for lossless compression
- Precision-preserving compression (easy to use/understand)
- available now via NCO (and in an upcoming netcdf4 release)


Sperr (Li, NCAR): promising wavelet approach - coming soon!

# Lossy vs. lossless compression on 2D CAM test set




Note: lossless is < 2x reduction from original

# Scientist feedback is critical

CAM test set:

- 21 time-series variables from CESM-LENS1
  - Daily (2yrs): TS, PRECT, TAUX FLUT, Z500, LHFLX
  - Monthly: (5yrs): TS, U, FLNS, CCN3, CLOUD, TMQ, PS, FSNS, FLNT, FSNT, SHFLX, LHFLX, QFLX, PRECC, PRECL
- categorized into three viable sets based on how much lossy compression was applied (e.g., conservative, middle ground, or aggressive)

***We've primarily worked with CAM data thus far due to compressor limitations (missing values, fill values, grids, …), but recent advances (lossy compression via netcdf) have made it easier to look at other component model data***

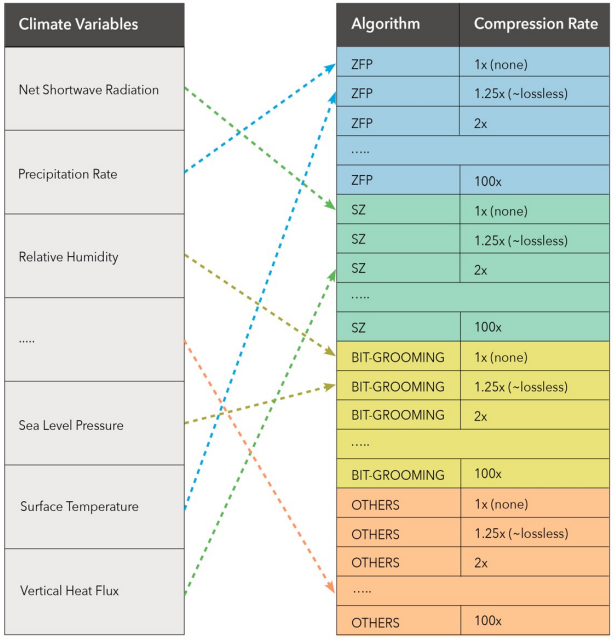
CICE and CLM test sets:

- coming soon! (working groups have recently provided variable lists)

Plan to compress entire CESM-LENS1 data set and redo some analyses!

# Current focus: optimal compression settings

*For lossy compression to become a practical option for CESM scientists, we need a tool for auto-selection of the ideal compression algorithm and parameters for each variable.*




## Overview:

- we determined metrics that relate the similarity of a compressed dataset to the original (e.g., DSSIM + complementary measures)
- if a compressed dataset passes these metrics, then the dataset is acceptable
- *optimally compressed:* a compressed dataset is smallest in storage size among all datasets that pass these metrics
- *our focus is on prediction of compression algorithm and settings (bases on features of the data) that will result in an optimally compressed dataset (NEXT TALK!)*

# Lessons learned / Concluding thoughts…

- work closely with application scientists
  - *adoption of lossy compression requires addressing CESM-user concerns*
  - *ldcpy package facilitates feedback transparency*
- treat variables individually
  - *no "one-size-fits-all"*

- preserve spatio-temporal features/properties
  - *compression has effects at fine spatial and temporal scales*
- consider derived variables
  - *e.g., surface energy balance, global precipitation, top of model radiation budget*
- identify detectable vs. consequential differences
  - *a skilled researcher will be able to detect compression – but does it matter?*
- collaborate with compression algorithm developers
  - *understand methods and effects, provide feedback*

# Thanks!

***We are working to make lossy compression a reality for climate model data. The goal is that applying compression is not something suspicious, but is rather analogous to carefully choosing grid resolutions, output frequency, and computation precisions.***

# Thanks!

Some of our papers …


A.H. Baker, A. Pinard, D.M. Hammerling. DSSIM: a structural similarity index for floating-point data, *submitted to Vis 2022.*

A. Poppick, J. Nardi, N. Feldman, A.H. Baker, A. Pinard, D.M. Hammerling. A Statistical Analysis of Lossily Compressed Climate Model Data, *Computers and Geosciences*, vol. 145, 2020.

A. Pinard, D.M. Hammerling, A.H. Baker. Assessing Differences in Large Spatio-temporal Climate Datasets with a New Python package, *Proceedings of the International Workshop on Big Data Reduction*, BigData2020, 2020.

A.H. Baker, D.M. Hammerling, T.L. Turton, Evaluating image quality measures to assess the impact of lossy data compression applied to climate simulation data, *Computer Graphics Forum, 2019.*

D.M. Hammerling, A.H. Baker, A. Pinard, P. Lindstrom, A collaborative effort to improve lossy compression for climate data, 5th Int'l Workshop on Data Analysis and Reduction for Big Scientific Data, SC19, 2019.

A.H. Baker, H. Xu, D. M. Hammerling, S. Li, J. Clyne, Toward a Multi-method Approach: Lossy Data Compression for Climate Simulation Data, LNCS (ISC 17), 2017.

A.H. Baker, D.M. Hammerling, S.A. Mickelson, H. Xu, M. B. Stolpe, P. Naveau, B. Sanderson, I. Ebert-Uphoff, S. Samarasinghe, F. De Simone, F. Carbone, C.N. Gencarelli, J.M. Dennis, J.E. Kay, P. Lindstrom, Evaluating Lossy Data Compression on Climate Simulation Data within a Large Ensemble. *GMD,* 2016.

A.H. Baker, H. Xu, J.M. Dennis, M.N. Levy, D. Nychka, S.A. Mickelson, J. Edwards, M. Vertenstein, A. Wegener, A Methodology for Evaluating the Impact of Data Compression on Climate Simulation Data, HPDC 2014.