A Collaborative Effort to Improve Lossy Compression Methods for Climate Data

Dorit Hammerling

Department of Applied Mathematics and Statistics
Colorado School of Mines(CSM)
Visiting Appointment: National Center for Atmospheric Research(NCAR)

Joint work with Allison Baker (NCAR), Alexander Pinard (CSM), and Peter Lindstrom (Lawrence Livermore National Laboratory)

. . . and many other contributors

November 17, 2019

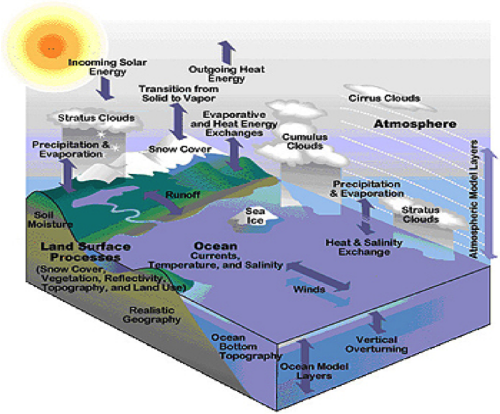
The National Center for Atmospheric Research (NCAR)

- A federally funded research and development center
- Mission: To understand the behavior of the atmosphere and related Earth and geospace systems

The National Center for Atmospheric Research (NCAR)

- A federally funded research and development center
- Mission: To understand the behavior of the atmosphere and related Earth and geospace systems

NCAR's Community Earth System Model



- a "virtual laboratory" to study past, present and future climate states
- describes interactions of the atmosphere, land, river runoff, land-ice, oceans and sea-ice
- complex! Large code base: approx. 1.5 Millions lines of code

Why compress climate data?

Increasing resolution and computational power lead to more and more climate model data. *Flood of data, with no end in sight!*

Storage is costly!
Previous HPC system Yellowstone: $\sim$ 20% of hardware budget for storage
New HPC system Cheyenne starting 2017: $\sim$ 50%

CMIP5 Archive is $\sim$ 3.3 Petabytes of data

CMIP6 Archive $>$ 20 Petabytes! (expected)

Many other examples such as large ensemble projects

Data storage a limiting factor for climate science.

Compression as a tool to store less data with MINIMAL information loss.

Why compress climate data?

Increasing resolution and computational power lead to more and more climate model data. *Flood of data, with no end in sight!*

Storage is costly!
Previous HPC system Yellowstone: $\sim$ 20% of hardware budget for storage
New HPC system Cheyenne starting 2017: $\sim$ 50%

CMIP5 Archive is $\sim$ 3.3 Petabytes of data

CMIP6 Archive $>$ 20 Petabytes! (expected)

Many other examples such as large ensemble projects

Data storage a limiting factor for climate science.

Compression as a tool to store less data with MINIMAL information loss.

Why compress climate data?

Increasing resolution and computational power lead to more and more climate model data. *Flood of data, with no end in sight!*

Storage is costly!
Previous HPC system Yellowstone: $\sim$ 20% of hardware budget for storage
New HPC system Cheyenne starting 2017: $\sim$ 50%

CMIP5 Archive is $\sim$ 3.3 Petabytes of data

CMIP6 Archive $>$ 20 Petabytes! (expected)

Many other examples such as large ensemble projects

Data storage a limiting factor for climate science.

Compression as a tool to store less data with MINIMAL information loss.

Why compress climate data?

Increasing resolution and computational power lead to more and more climate model data. *Flood of data, with no end in sight!*

Storage is costly!
Previous HPC system Yellowstone: $\sim$ 20% of hardware budget for storage
New HPC system Cheyenne starting 2017: $\sim$ 50%

CMIP5 Archive is $\sim$ 3.3 Petabytes of data

CMIP6 Archive $>$ 20 Petabytes! (expected)

Many other examples such as large ensemble projects

Data storage a limiting factor for climate science.

Compression as a tool to store less data with MINIMAL information loss.

Why compress climate data?

Increasing resolution and computational power lead to more and more climate model data. *Flood of data, with no end in sight!*

Storage is costly!
Previous HPC system Yellowstone: $\sim$ 20% of hardware budget for storage
New HPC system Cheyenne starting 2017: $\sim$ 50%

CMIP5 Archive is $\sim$ 3.3 Petabytes of data

CMIP6 Archive $>$ 20 Petabytes! (expected)

Many other examples such as large ensemble projects

Data storage a limiting factor for climate science.

Compression as a tool to store less data with MINIMAL information loss.

Why are climate scientists reluctant to use compression?

The typical metrics used in the compression community don't have much meaning to climate scientists and are not reassuring to them.

Work with climate scientists to reassure them that compression doesn't change their scientific conclusions, and make sure it indeed doesn't!

Why are climate scientists reluctant to use compression?

The typical metrics used in the compression community don't have much meaning to climate scientists and are not reassuring to them.

Work with climate scientists to reassure them that compression doesn't change their scientific conclusions, and make sure it indeed doesn't!

Spatio-temporal analysis to emulate climate analysis

We have developed spatio-temporal statistical analysis tools that emulate the key aspects of climate data analysis.

- gradients in space and time
- cumulative effects in time
- changes in variability over space or time
- changes in the statistical distribution
  - changes in the extremes
  - changes in skewness
  - ...

Pro-actively work with algorithm developers to use these tools to address any potential issues BEFORE climate scientists use the lossy compressors.

Spatio-temporal analysis to emulate climate analysis

We have developed spatio-temporal statistical analysis tools that emulate the key aspects of climate data analysis.

- gradients in space and time
- cumulative effects in time
- changes in variability over space or time
- changes in the statistical distribution
  - changes in the extremes
  - changes in skewness
  - ...

Pro-actively work with algorithm developers to use these tools to address any potential issues BEFORE climate scientists use the lossy compressors.

Spatio-temporal analysis to emulate climate analysis

We have developed spatio-temporal statistical analysis tools that emulate the key aspects of climate data analysis.

- gradients in space and time
- cumulative effects in time
- changes in variability over space or time
- changes in the statistical distribution
  - changes in the extremes
  - changes in skewness
  - ...

Pro-actively work with algorithm developers to use these tools to address any potential issues BEFORE climate scientists use the lossy compressors.

Temperature compressed with different versions of ZFP

Surface Temperature

- Daily CESM data from 1920–2005
- 31,390 time slices;
- 192 $\times$ 288 grid points
- Smooth in space and time (i.e., strongly correlated)

ZFP

- one of the most effective lossy floating point compressors, transform method
- three version:
  - ZFP-0.5.3
  - ZFP-ROUND
  - ZFP-BETA
- ZFP-0.5.3 and ZFP-ROUND only differ in their rounding
- ZFP-BETA: more compression by encoding fewer bits; same symmetric rounding as ZFP-ROUND

Temperature compressed with different versions of ZFP

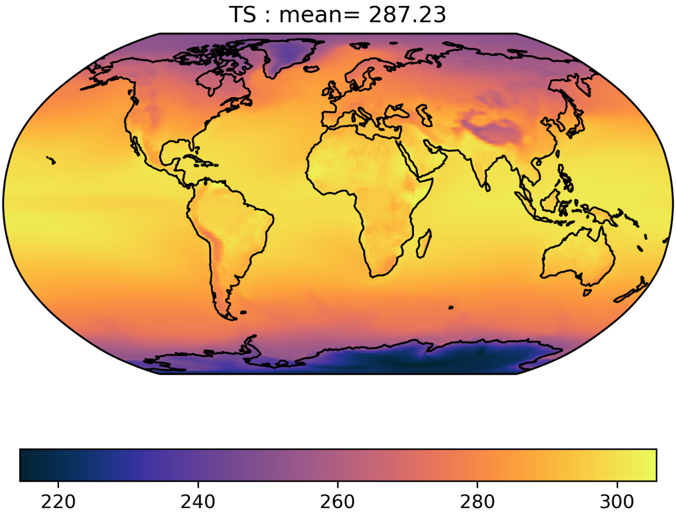
Surface Temperature

- Daily CESM data from 1920–2005
- 31,390 time slices;
- 192 × 288 grid points
- Smooth in space and time (i.e., strongly correlated)

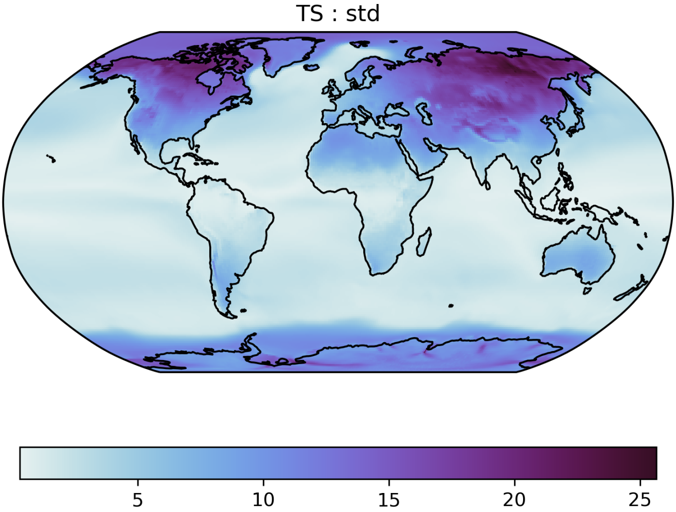
ZFP

- one of the most effective lossy floating point compressors, transform method
- three version:
  - ZFP-0.5.3
  - ZFP-ROUND
  - ZFP-BETA
- ZFP-0.5.3 and ZFP-ROUND only differ in their rounding
- ZFP-BETA: more compression by encoding fewer bits; same symmetric rounding as ZFP-ROUND

Original data: gridcell mean



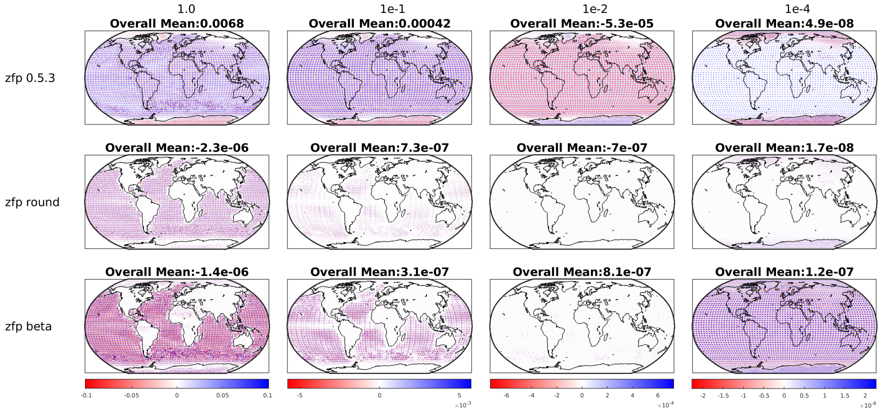
Original data: gridcell standard deviation



TS Mean Error, RMSE and Compression Ratio (CR)

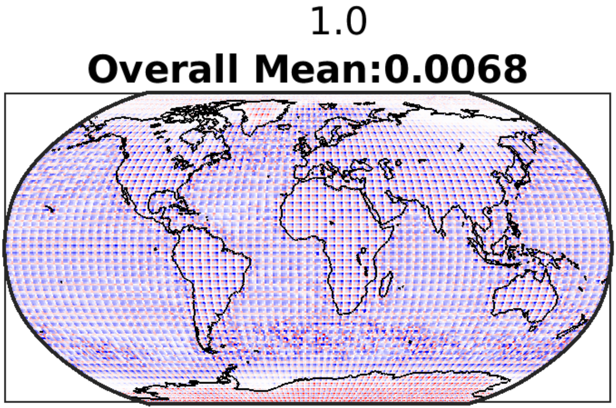
ZFP	Mean Error			RMSE			CR		
tol.	zfp-0.5.3	zfp-round	zfp-beta	zfp-0.5.3	zfp-round	zfp-beta	zfp-0.5.3	zfp-round	zfp-beta
1.0	6.75e-3	-2.25e-6	-1.39e-6	7.39e-2	6.76e-2	1.32e-1	.15	.15	.13
0.5	-3.38e-3	5.20e-7	-2.25e-6	3.88e-2	3.48e-2	6.76e-2	.18	.18	.15
1e-1	4.22e-4	7.34e-7	3.06e-7	5.32e-3	4.56e-3	9.03e-3	.26	.26	.23
1e-2	-5.25e-5	-7.04e-7	8.11e-7	6.71e-4	5.73e-4	1.14e-3	.36	.36	.33
1e-3	6.86e-6	2.70e-7	-1.93e-8	8.44e-5	7.20e-5	1.43e-4	.45	.45	.42
1e-4	4.86e-8	1.72e-8	1.19e-7	3.18e-6	2.11e-6	9.25e-6	.58	.58	.55
1e-5	0.00	0.00	0.00	0.00	0.00	0.00	.67	.67	.64

Selected Results: TS Mean Errors



Sign reversal at the poles: how come?

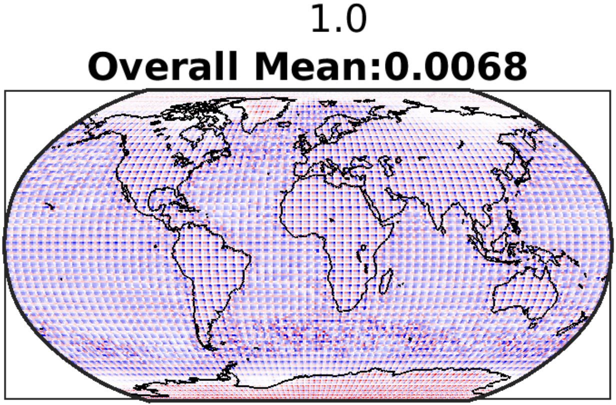
zfp 0.5.3



It is cold at the poles! Binary exponent boundary at 256° K.

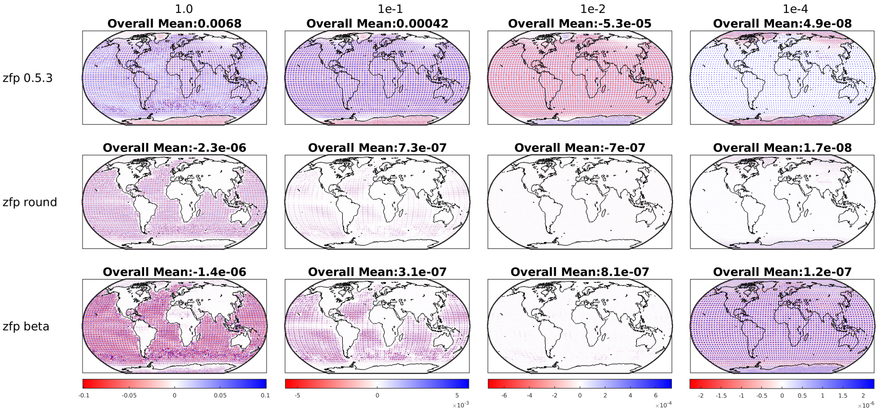
Sign reversal at the poles: how come?

zfp 0.5.3

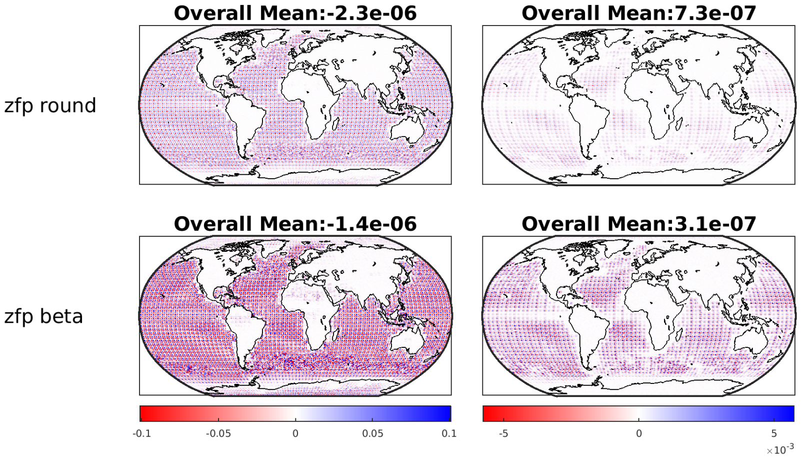


It is cold at the poles! Binary exponent boundary at 256° *K.*

Selected Results: TS Mean Errors

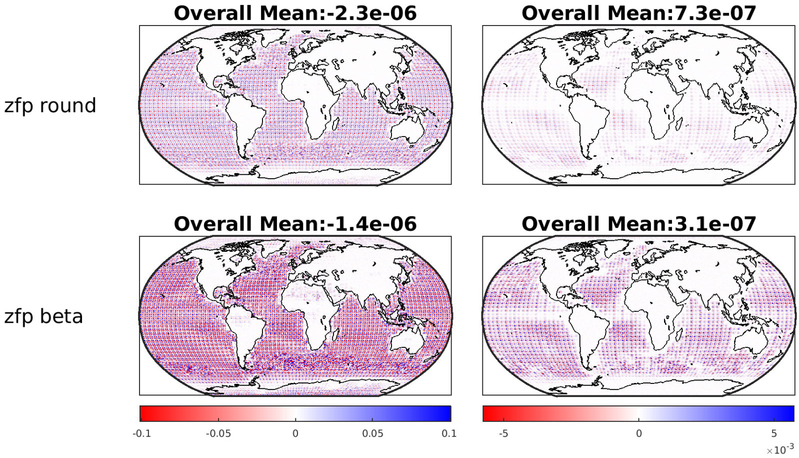


Oceans in turmoil



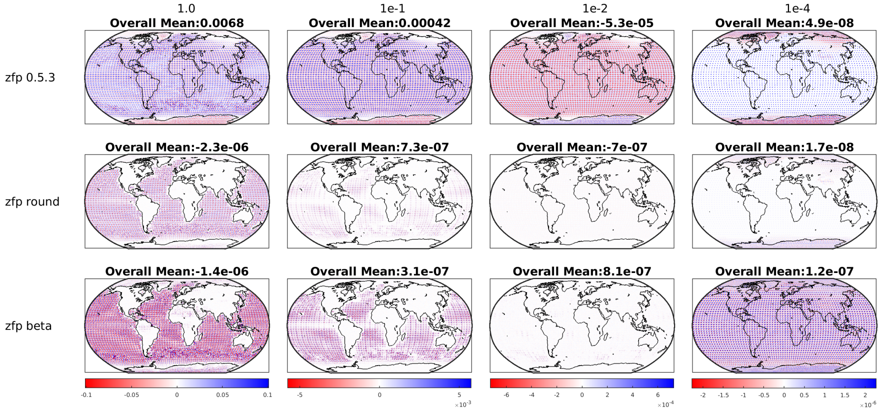
Little variation spatially means ZFP coefficients are small and often quantized to zero where asymmetric rounding works better.

Oceans in turmoil

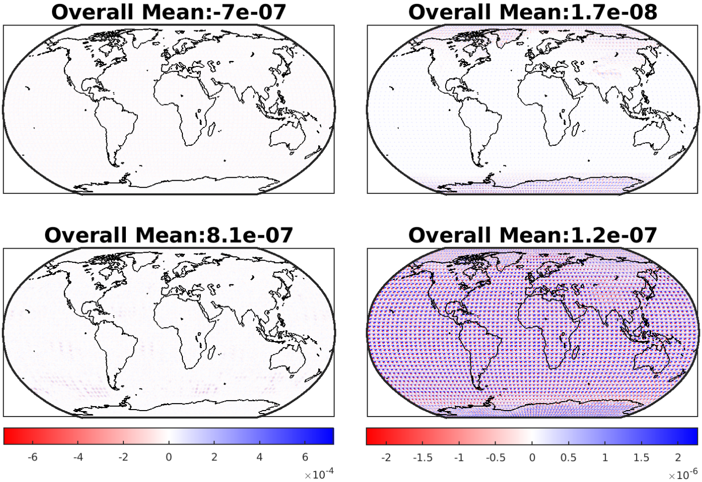


Little variation spatially means ZFP *coefficients are small and often quantized to zero where asymmetric rounding works better.*

Selected Results: TS Mean Errors

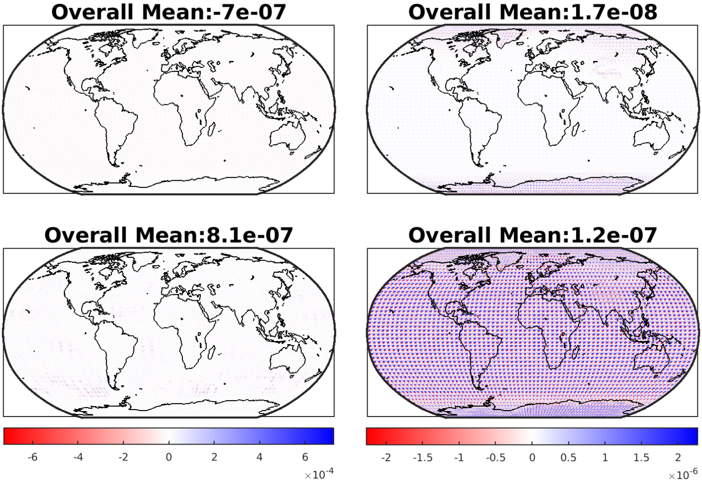


Getting close to machine precision



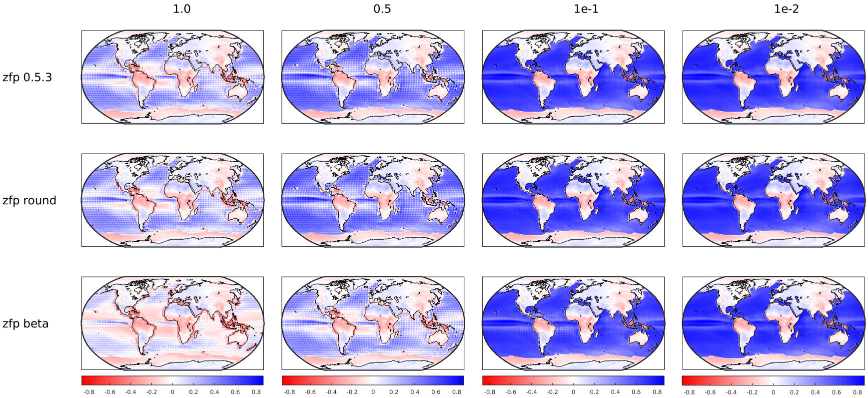
At 1e-4 tolerance, errors are only a few possible discrete values which are poorly approximated by uniform distribution.

Getting close to machine precision



At 1e-4 tolerance, errors are only a few possible discrete values which are poorly approximated by uniform distribution.

Lag-1 correlations of first differences of deseasonalized TS



- 1e-2 visually identical to original for all three versions
- dampening and gridding artifacts at looser tolerances

Summary of evaluation work

- Compression has effects at fine spatial and temporal scales that are masked by global statistics
- Useful insights come from investigating metrics which vary at lower magnitudes than the data itself
- Collaboration is key to address issues that are highlighted in these analyses, for example, adaptive rounding schemes

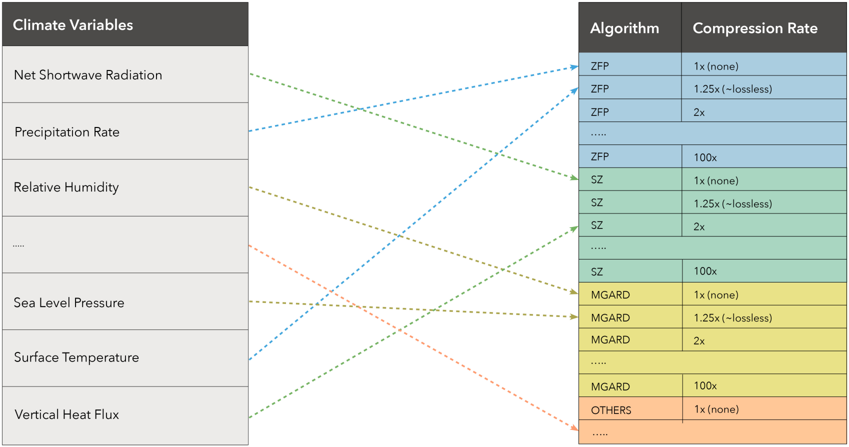
Next goal: develop a Python library (integrated with Pangeo) so climate scientist and compression algorithm developer can see effects for themselves

Summary of evaluation work

- Compression has effects at fine spatial and temporal scales that are masked by global statistics
- Useful insights come from investigating metrics which vary at lower magnitudes than the data itself
- Collaboration is key to address issues that are highlighted in these analyses, for example, adaptive rounding schemes

Next goal: develop a Python library (integrated with Pangeo) so climate scientist and compression algorithm developer can see effects for themselves

Where we would like to be some years from now . . .



References

Baker, A. H., Hammerling, D. M., Mickelson, S. A., Xu, H., Stolpe, M. B., Naveau, P., Sanderson, B., Ebert-Uphoff, I., Samarasinghe, S., De Simone, F., Carbone, F., Gencarelli, C. N., Dennis, J. M., Kay, J. E., and Lindstrom, P. (2016). Evaluating lossy data compression on climate simulation data within a large ensemble. *Geoscientific Model Development*, 9(12):4381–4403.

Baker, A. H., Hammerling, D. M., and Turton, T. L. (2019). Evaluating image quality measures to assess the impact of lossy data compression applied to climate simulation data. *Computer Graphics Forum*, 38(3):517–528.

Baker, A. H., Xu, H., Hammerling, D. M., Li, S., and Clyne, J. P. (2017). *Toward a Multi-method Approach: Lossy Data Compression for Climate Simulation Data*, pages 30–42. Springer International Publishing.

Guinness, J. and Hammerling, D. M. (2018). Compression and conditional emulation of climate model output. *Journal of the American Statistical Association*, 113(521):56–67.

Hammerling, D. M., Baker, A., Pinard, A., and Lindstrom, P. (2019). A collaborative effort to improve lossy compression for climate data. *In The 5th International Workshop on Data Analysis and Reduction for Big Scientific Data (DRBSD-5) in Conjunction with SC19*.

Nardi, J., Feldman, N., Poppick, A., Baker, A., and Hammerling, D. M. (2018). Statistical analysis of compressed climate data. Technical report, NCAR Technical Note NCAR/TN-547+STR.

Thanks! Any questions: *hammerling@mines.edu*