Compressing CESM Data… while Preserving Information

Allison H. Baker

Dorit Hammerling
Haiying Xu

Computational Information Systems Laboratory
National Center for Atmospheric Research

…and many other contributors

ICAS 2017
Sept. 12, 2017

Climate models produce lots of data

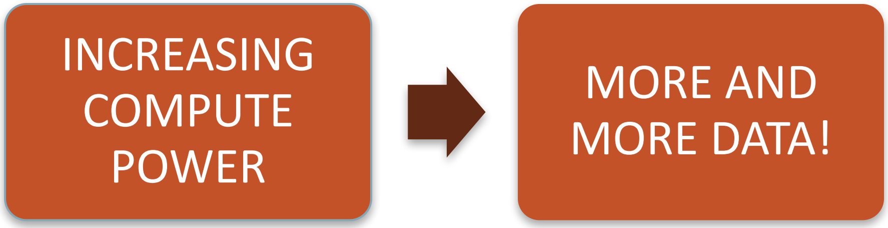
..and it's getting worse!

IPCC Coupled Model Comparison Projects (CMIPs)

- Phase 5 (2013): 2.5 PB of output
- Phase 6 (2018): >20 PB expected (40 PB?)

Storage at NCAR

- More precious than CPU-hours?



INCREASING COMPUTE POWER

MORE AND MORE DATA!

Climate models produce lots of data

..and it's getting worse!

IPCC Coupled Model Comparison Projects (CMIPs)

- Phase 5 (2013): 2.5 PB of output
- Phase 6 (2018): >20 PB expected (40 PB?)

Storage at NCAR

- More precious than CPU-hours?

Storage resources are limiting climate science objectives!

Simulation length, output frequency, ensemble size, output variables, …

Data compression

Compression: $X \Longrightarrow C$

Reconstruction: $C \Longrightarrow \tilde{X}$

Lossless: $X = \tilde{X}$

Lossy: $X \sim \tilde{X}$

- *Lossless* compression is (relatively) ineffective on CESM data
- *Lossy* is much better

Data compression

Compression: $X \Longrightarrow C$

Reconstruction: $C \Longrightarrow \tilde{X}$

Lossless: $X = \tilde{X}$

$$\boxed{\text{Lossy:} \quad X \sim \tilde{X}}$$

- *Lossless* compression is (relatively) ineffective on CESM data
- *Lossy* is much better … *but it makes scientists nervous!*

How to evaluate the effect of lossy compression on climate simulation data?

Lossy data compression

Issue: Quantify the error between X and $\tilde{X}$

Common “simple” compression metrics:

- average error *(peak signal-to-noise ratio, RMSE, …)*
- pointwise error (*max norm)*
- “eye-ball” norm

Lossy data compression

Issue: Quantify the error between X and $\tilde{X}$

Common "simple" compression metrics:

- average error *(peak signal-to-noise ratio, RMSE, …)*
- pointwise error (*max norm)*
- "eye-ball" norm

Not sufficient for evaluating whether compression has (negatively) impacted science results.

What has been done at NCAR so far?

(1) Establish feasibility:

→ evaluate compression in the context of an ensemble.

The compression-introduced differences should not exceed ensemble variability!

- choose appropriate compression with ensemble-based metrics (per-variable)

- impact of compression on solution *is less than* a bit-perturbation to initial conditions

Possible!

What has been done at NCAR so far?

(2) Direct experience:

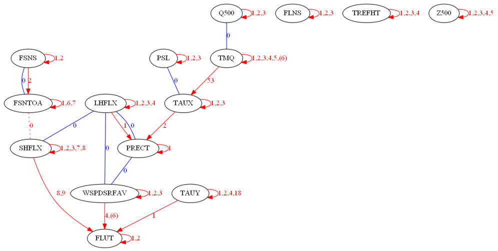
Provide climate scientists with reconstructed data.

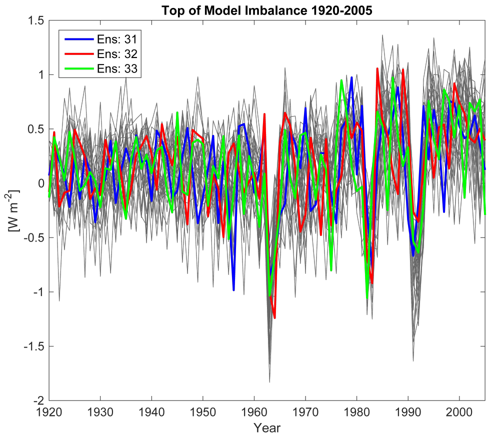
Can climate scientists differentiate between compressed and uncompressed data?

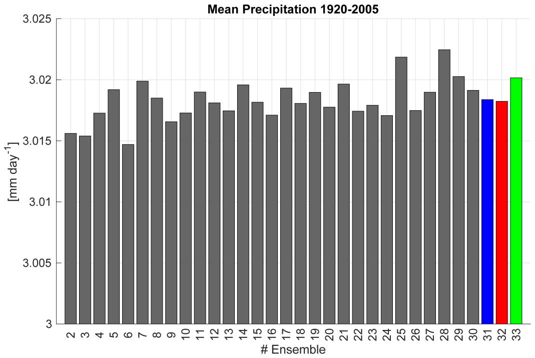
Let your taste decide.
Take the Pepsi Challenge!
Nationwide, more people prefer Pepsi over Coca-Cola!

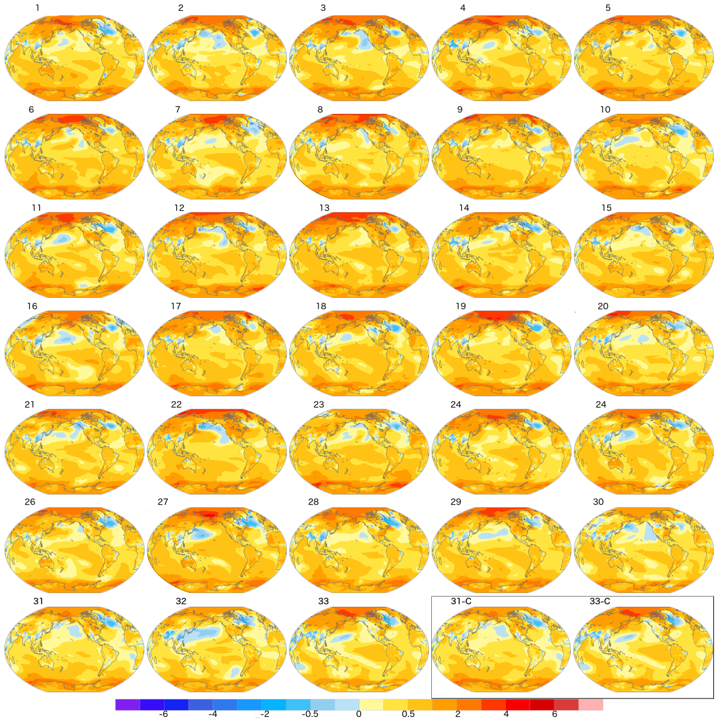
What has been done at NCAR so far?

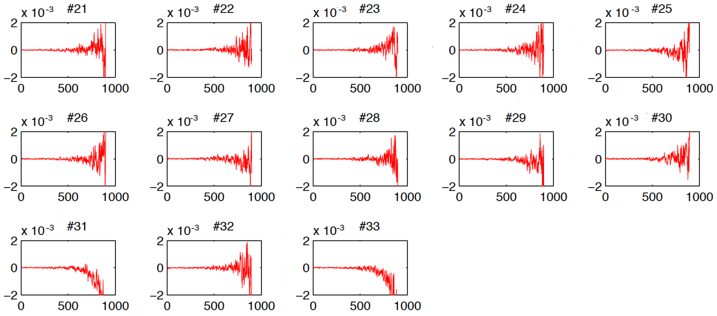
(2) Direct experience …





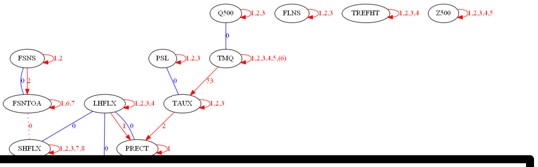


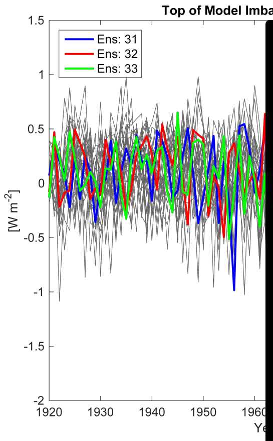




What has been done at NCAR so far?

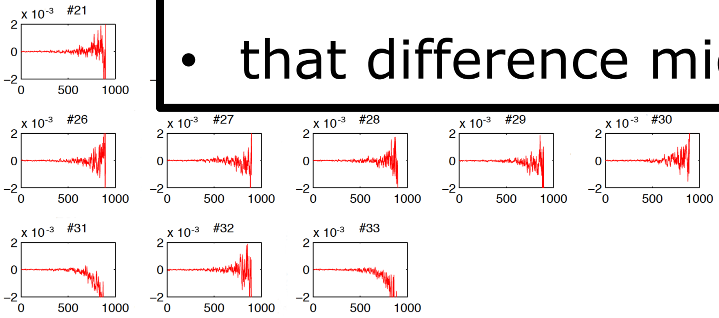
(2) Direct experience …

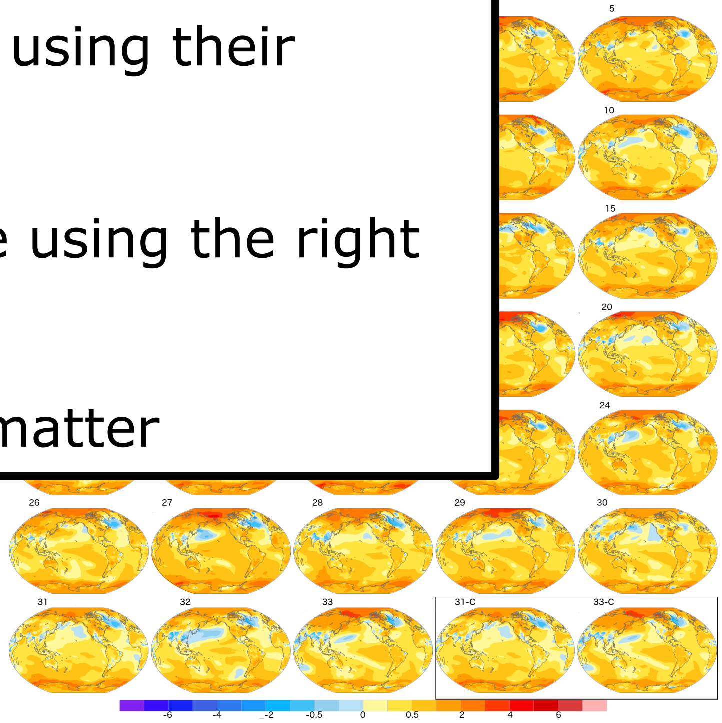




***Main insights*:**

- did not notice a difference using their standard analysis tools
- but can detect a difference using the right (i.e., clever) methods
- that difference might not matter





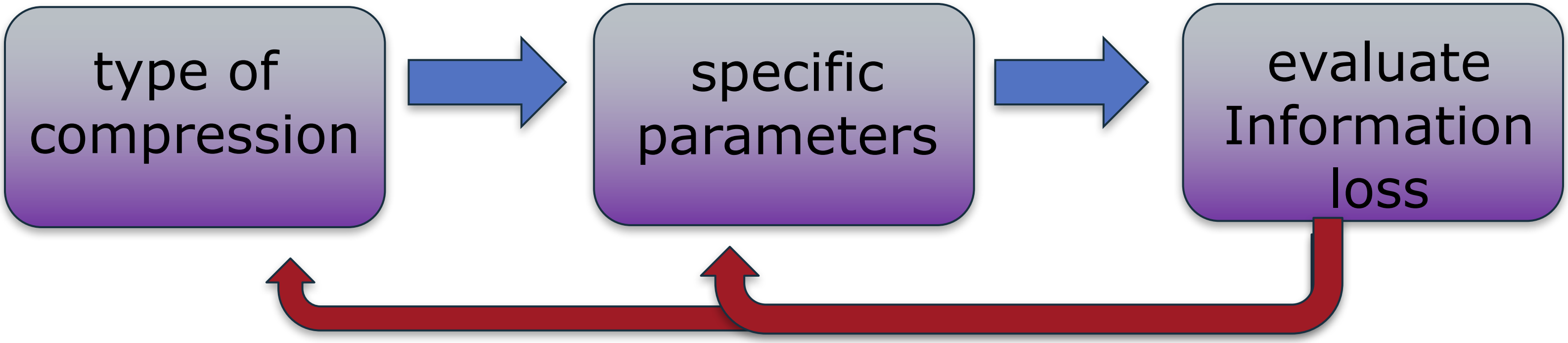
Current: best method for each variable

Determine max compression for each variable that preserves its scientific value

- *Many diverse variables:*
  - constants, abrupt changes, smooth, # of zeros
  - fill values (10^{35}), NANs, missing values

- *Many compression algorithms:*
  - *transform, predictive, statistical, …*

Each variable:



type of compression

specific parameters

evaluate Information loss

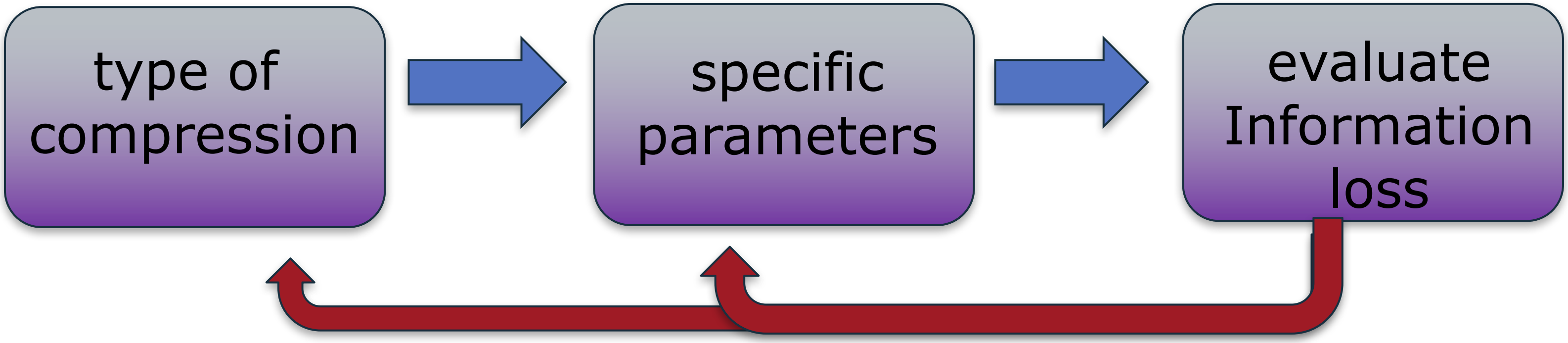
Current: best method for each variable

Determine max compression for each variable that preserves its scientific value

Goal: *automated tool* for CESM workflow

- appropriate metrics
  - reasonable computation cost
- understand compression algorithm properties
- determine important predictive features of data

Each variable:



Metrics: evaluating information loss

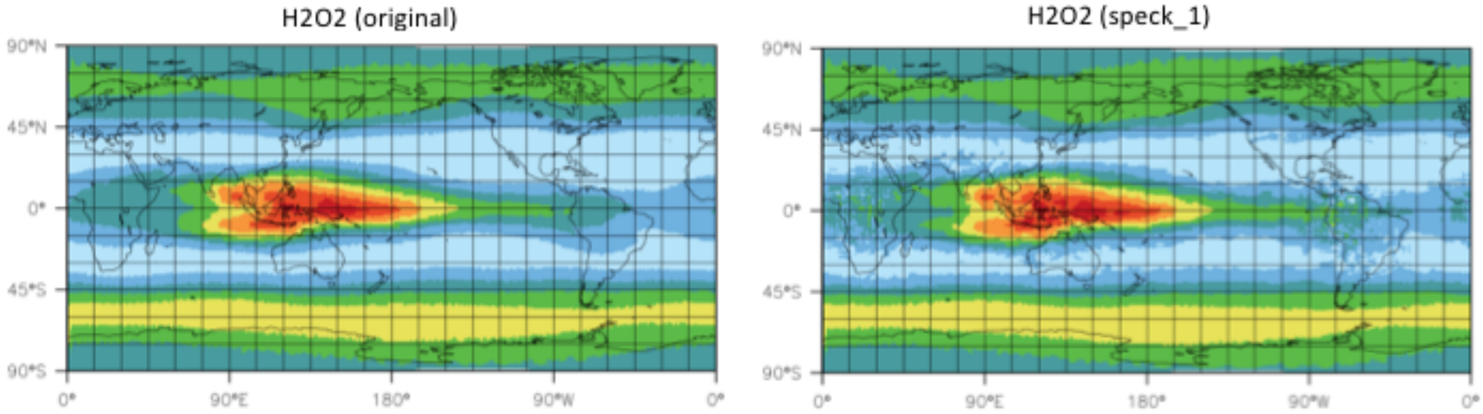
- suite to measure different aspects of data
- not ensemble-based (use only the fields themselves)

(1) Pearson correlation coefficient

(2) Kolmogorov-Smirnov (K-S) test

(3) Spatial relative error

(4) Structural similarity index (SSIM)



Final thoughts

NCAR is suffering from too much data:

- science and $$$
- lossy compression: 4:1 reduction (on average - conservative)

Next:

- determine method-specific parameters (control compression)
  - correlate with features
- further refinement on metrics
  - temporal features (e.g., extremes)
  - derived variables
  - human perception study

Final thoughts

NCAR is suffering from too much data:

- science and $$$
- lossy compression: 4:1 reduction (on average - conservative)

Next:

- determine method-specific parameters (control compression)
  - correlate with features
- further refinement on metrics
  - temporal features (e.g., extremes)
  - derived variables
  - human perception study

Climate scientists compress their simulation data with confidence!

Thanks!

Questions, comments, suggestions:

abaker@ucar.edu

Thanks!

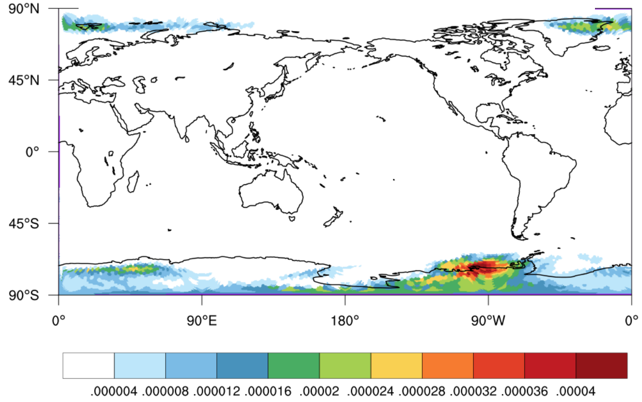
A.H. Baker, D.M. Hammerling, S.A. Mickelson, H. Xu, M.B. Stolpe, P. Naveau, B. Sanderson, i. Ebert-Uphoff, S. Samarasinghe, F. De Simone, F. Carbone, C.N. Gencarelli, J.M. Dennis, J.E. Kay, and P. Lindstrom, "Evaluating Lossy Data Compression on Climate Simulation Data within a Large Ensemble." *Geoscientific Model Development,* 9, 4381-4403, 2016.

A. H. Baker, H. Xu, D. M. Hammerling, S. Li, and J. Clyne, "Toward a Multi-method Approach: Lossy Data Compression for Climate Simulation Data", *International Workshop on Data Reduction for Big Scientific Data (*DRBSD-*1), ISC'17*, 2017.

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." *Proc. of the 23rd International ACM Symposium on High Performance Parallel and Distributed Computing* (HPDC14), pp. 203-214, 2014.

Comparing two types of compression

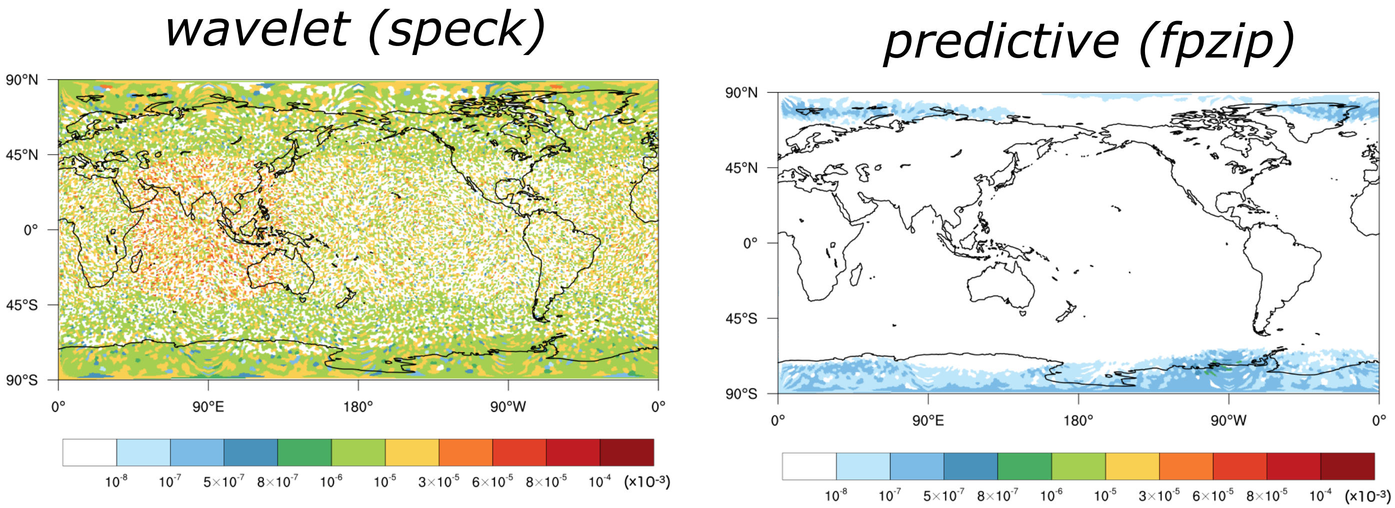
CLOUD



CLOUD

- large range (8 orders magnitude)
- 22% zeros
- small numbers

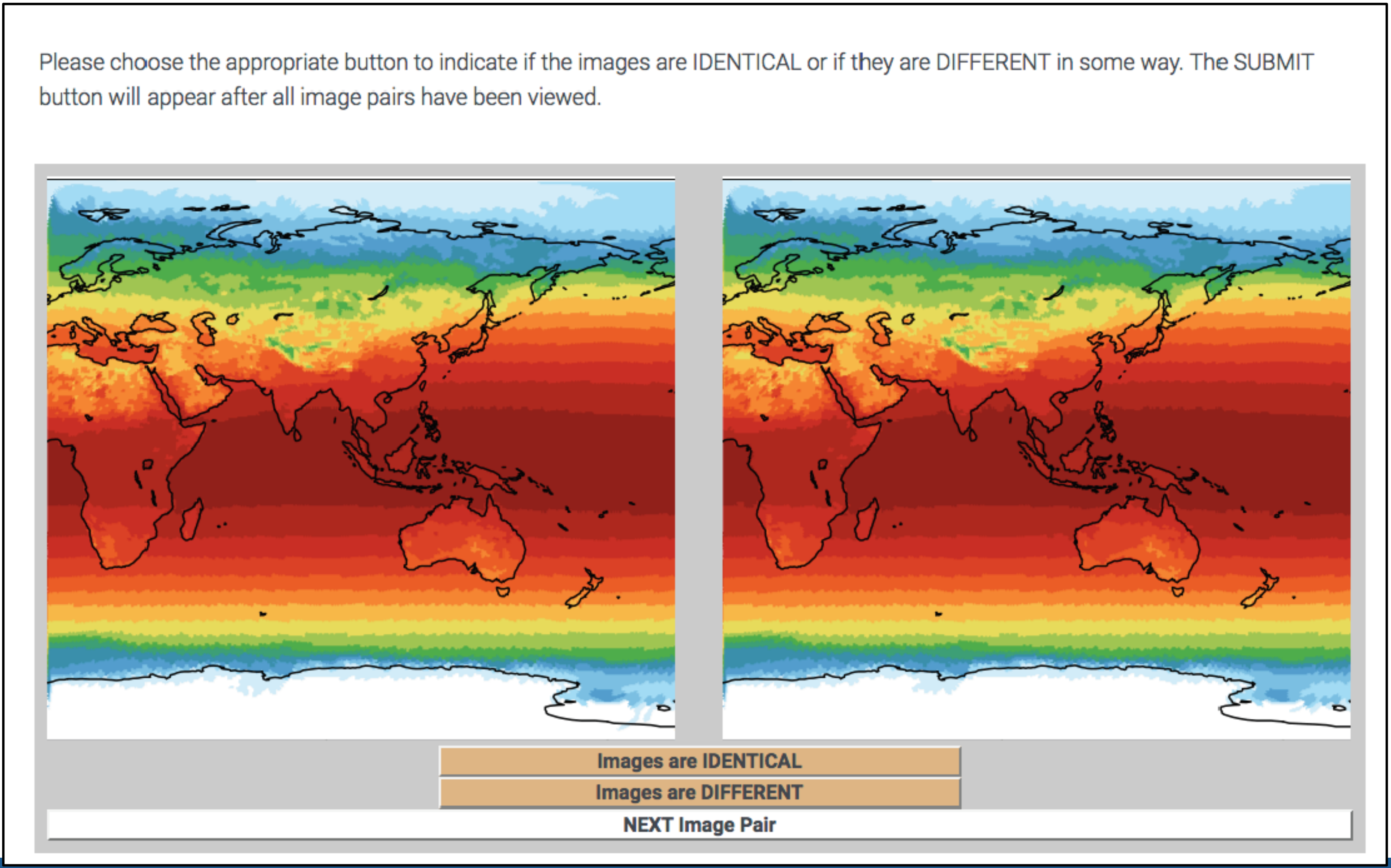
Absolute Error


wavelet (speck)

predictive (fpzip)

SSIM: human perception pilot study

-Visualization is essential!



Please choose the appropriate button to indicate if the images are IDENTICAL or if they are DIFFERENT in some way. The SUBMIT button will appear after all image pairs have been viewed.