Data Compression of Climate Simulation Data

John Dennis (dennis@ucar.edu)

Motivation:

- Historical data generation trends are economically unsustainable
  - online and archive storage will consume CISL hardware budgets
- Storage resources will limit science objectives
  - Not a question of 'if' but 'when'
- Do all experiments need to maintain full 32-byte precision for history files?
- Can we utilize data-compression to reduce online/offline storage needs?

Data-compression basics

- Lossless versus Lossy compression
  - Lossless: No information is lost, full precision is recovered
    - gzip *
  - Lossy: Information is lost as part of the compression algorithm
    - 8-byte → 4-byte
    - Original:

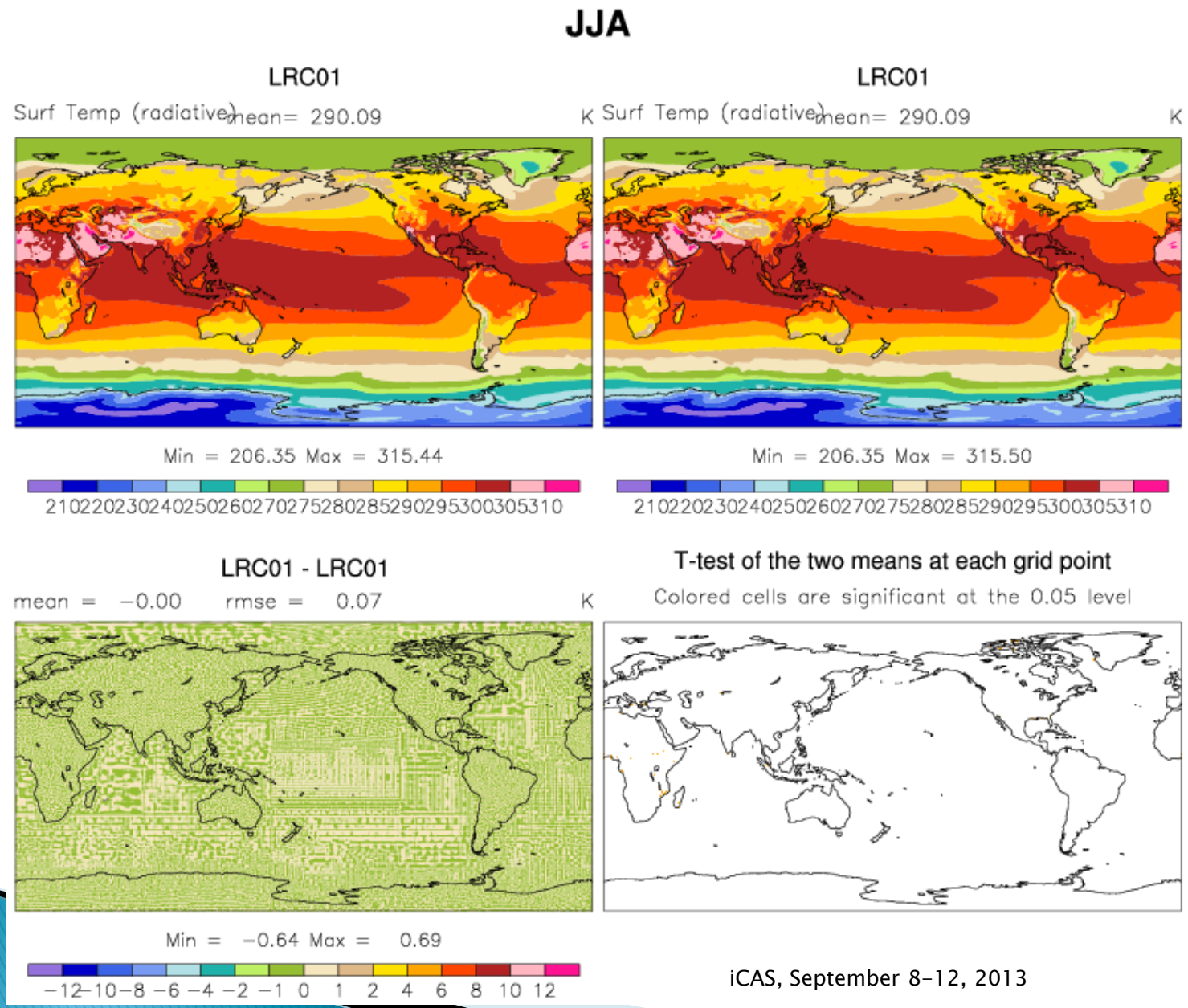
      T = 290.1234567890123
    - Lossy compression:

      T = 290.1234500000000
- Restart files: lossless compression
- History files: lossy or lossless compression

How to evaluate data compression

- Eyeball norm?

TS (Surface Temperature)



New CESM Port-Validation Tool

- M. Levy, J. Dennis, B. Eaton, J. Edwards, A. Mai, D. Nychka, J. Tribbia, M. Vertenstein, D. Williamson, and H. Xu
- Create 101 ensemble
  - Initial perturb: {-5.9,-5.8,…-1,0,1,…5.8,5.9} *10^{-14}
  - 1-year run (annual average output)
- For each ensemble member consider the sub-ensemble containing 100 other members

New CESM Port-Validation Tool (con't)

- For each variable (u) compute 100-member ensemble mean ($\bar{u}$) and standard deviation (σ) at every (i,j,k) point
- Compute root-mean-square (z-score) for the omitted member

- $RMSZ_u = \sqrt{(1/n_x \, \Sigma_{i,j,k} \, ((u_{i,j,k} - \bar{u}_{i,j,k})/\sigma_{i,j,k})^2)}$

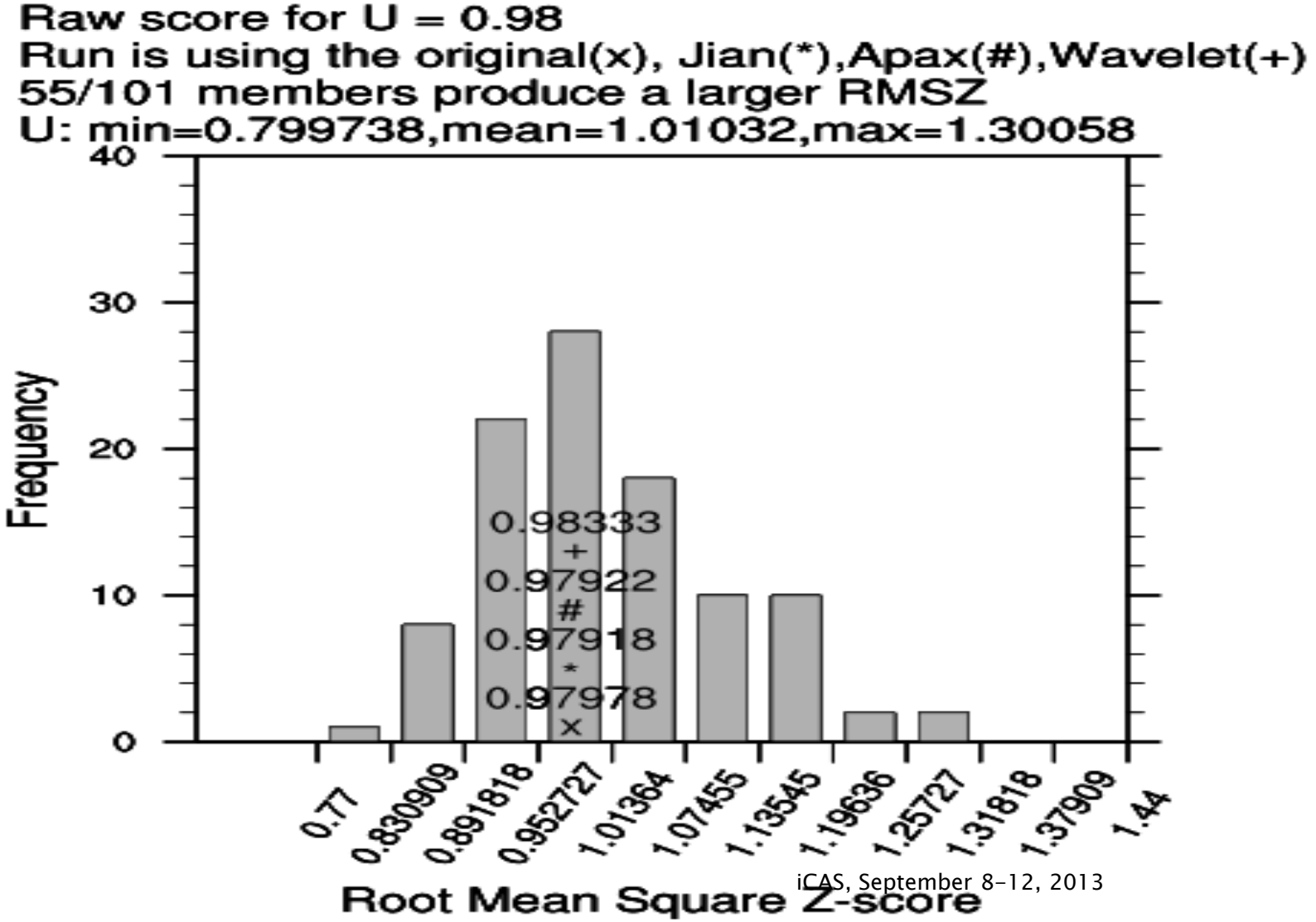
How to evaluate data compression

- ~~Eyeball norm~~
- Leverage CESM Port-Validation Tool
  - RSMZ-ensemble test
    - Choose single ensemble member
    - Compress/decompress member
    - Does decompressed members z-score still belong to ensemble?

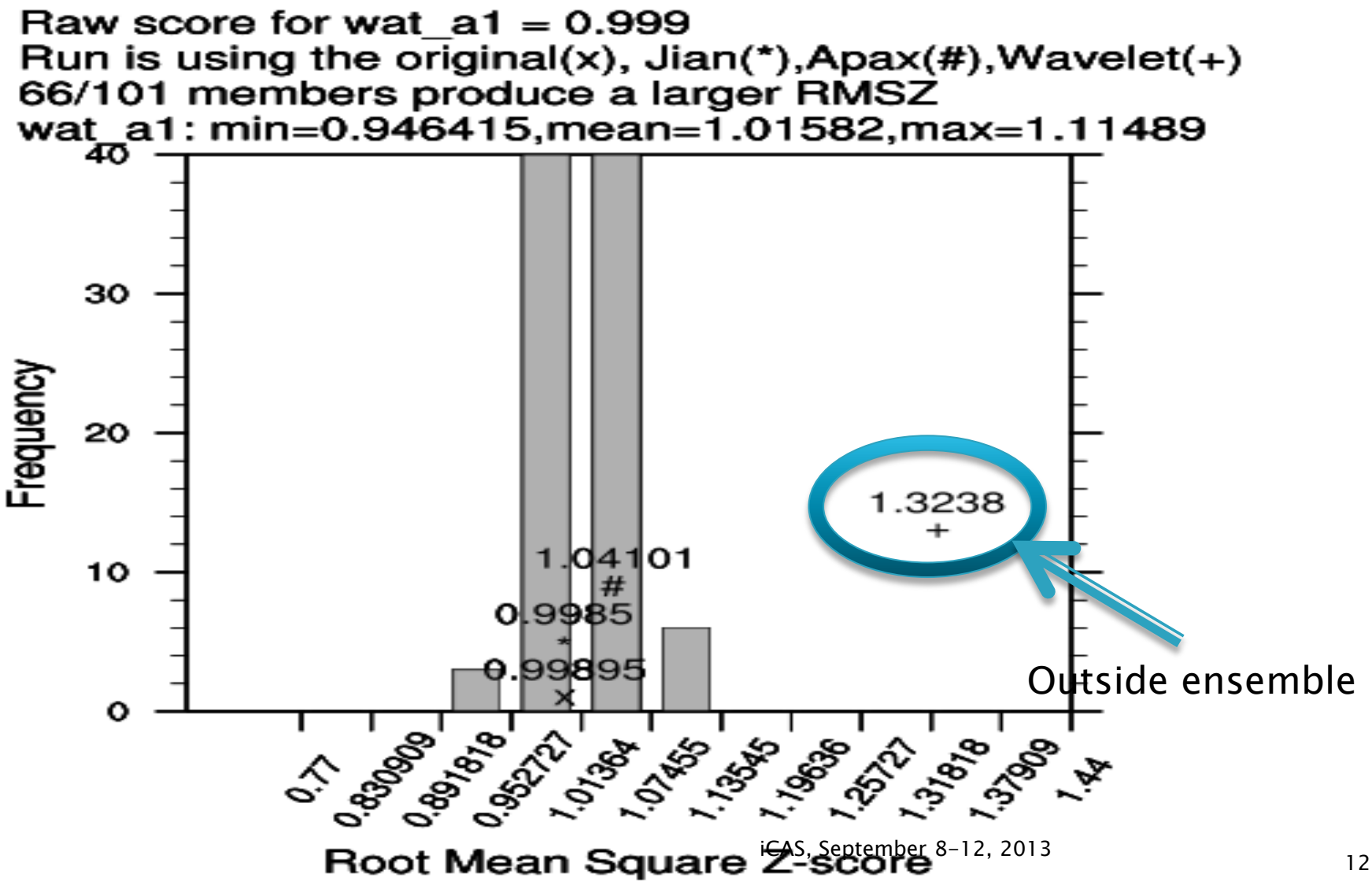
Current Compression Algorithms

- Samplify APAX
  - Fixed rate compression [2:1],[4:1],[5:1],[6:1],[8:1]
  - www.samplify.com
- Climate Compression (CC) [Jian] [5:1]
  - T. Bicer, J. Yin, D. Chiu, G. Agrawal and K. Schuchardt, "Integrated Online Compression to Accelerate Large-Scale Data Analytics Applications", Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS'13), Boston MA, May 2013
- VAPOR wavelet [5:1]
  - J. Clyne, P. Maninni, A. Norton, and M. Rast, "Interactive desktop analysis of high resolution simulations: Application to turbulent plume dynamics: applications to magnetic fields and turbulent flows, *New Journal of Physics*, **10**, 12507 (2008)

RMSZ-Ensemble test: U: zonal wind

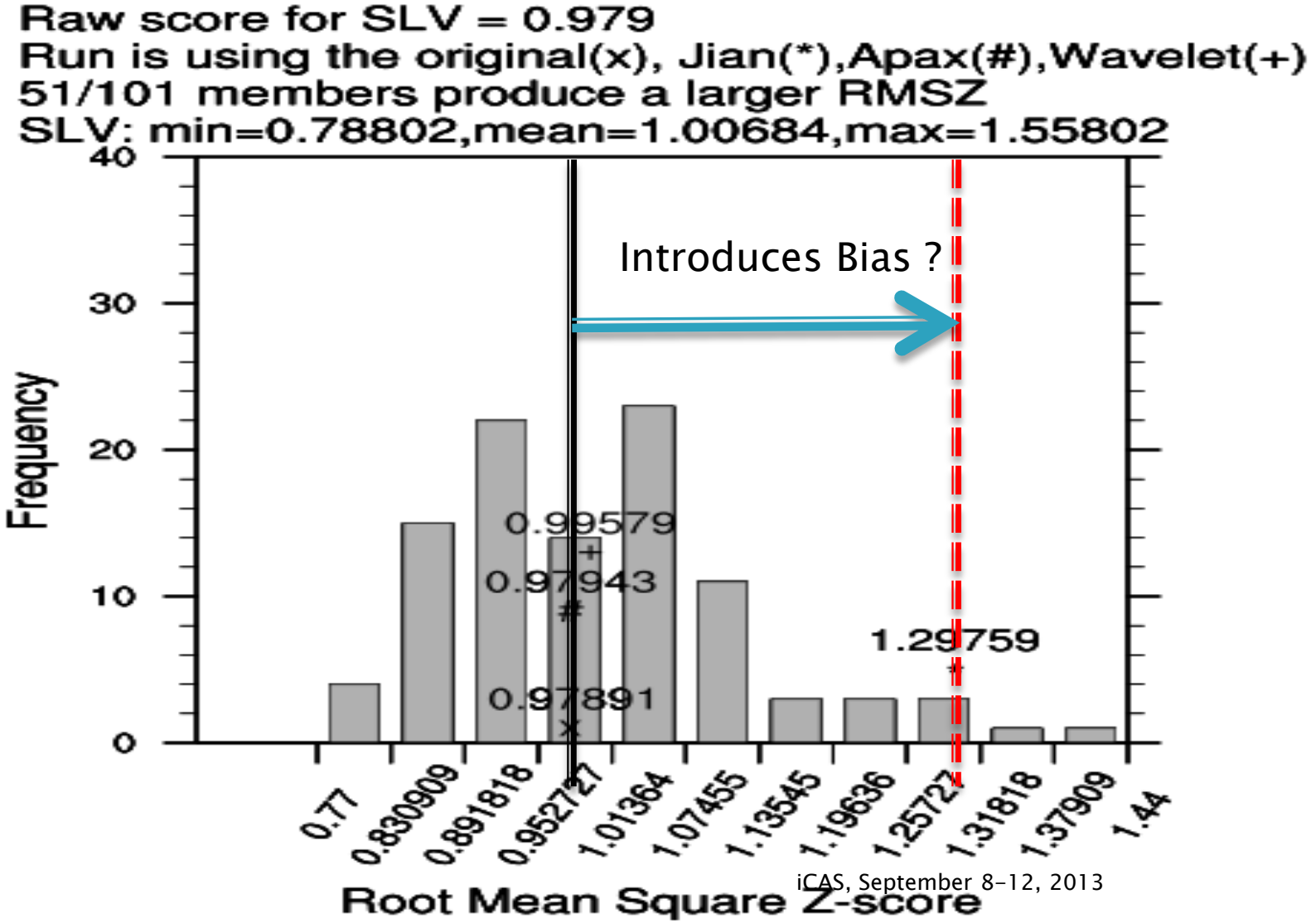


RMSZ-Ensemble test
wat_a1: aerosol water mode 01



Outside ensemble

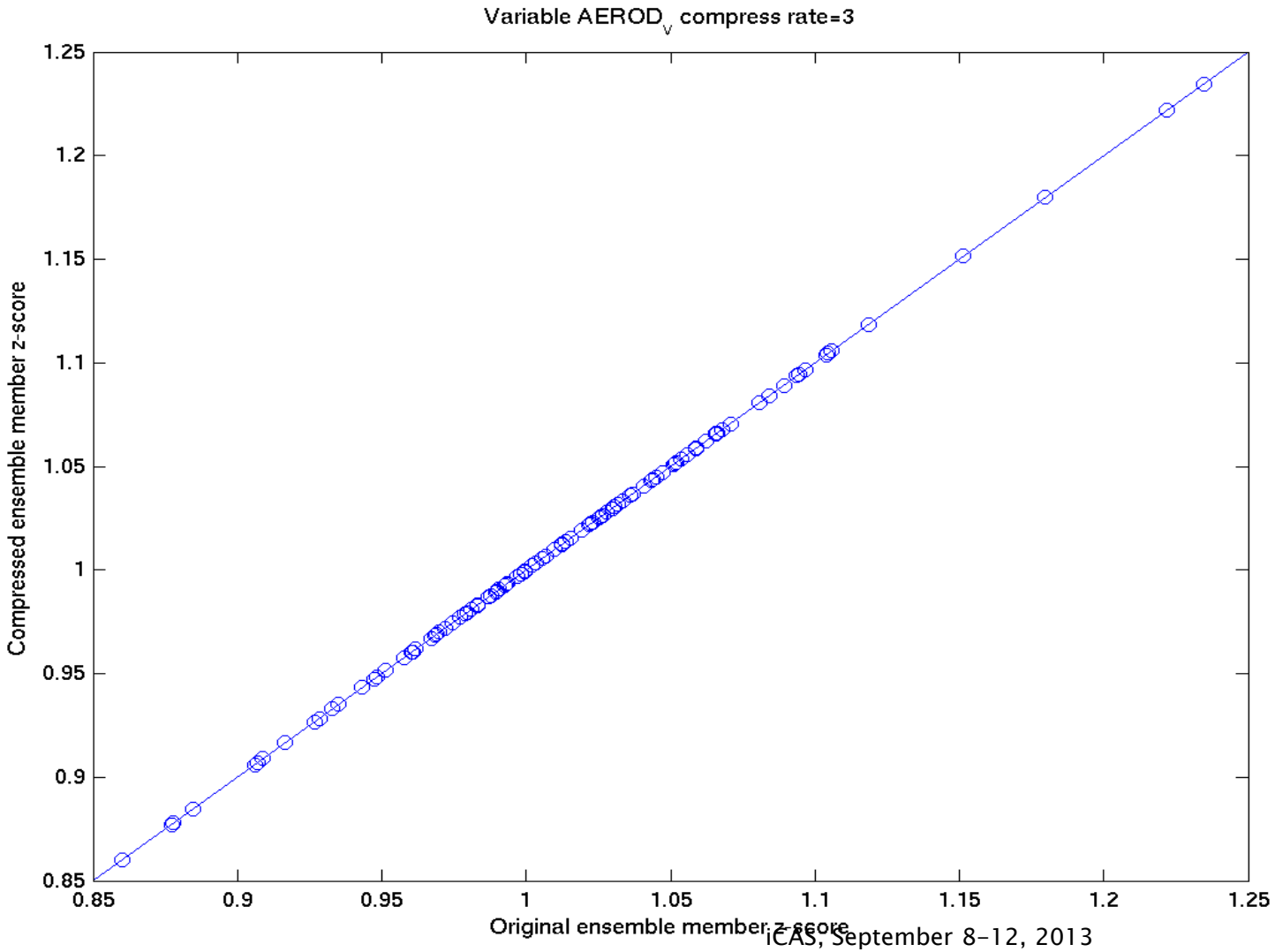
RMSZ-Ensemble test
SLV: Liquid Water virtual static energy



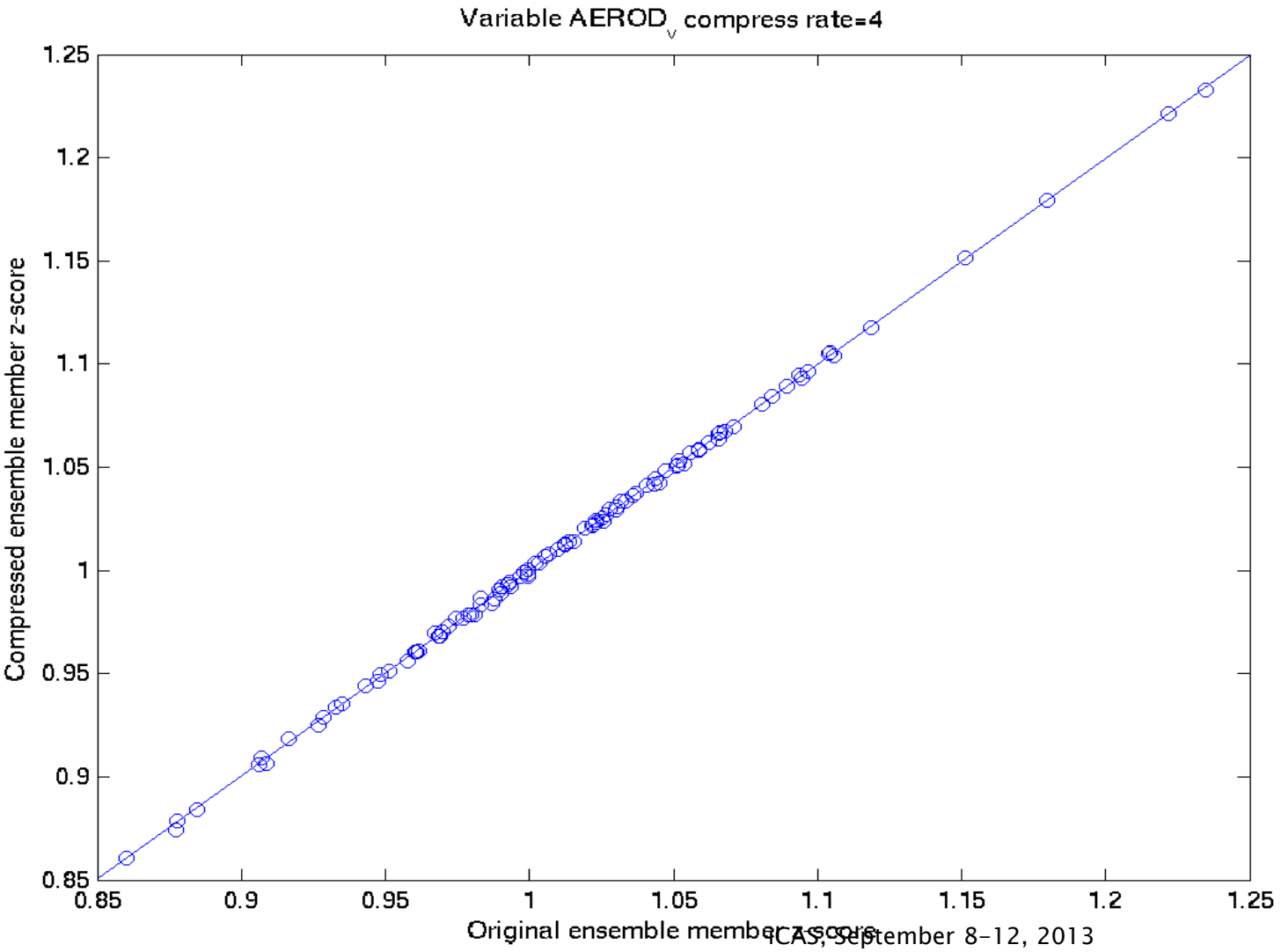
How to evaluate data compression

- ~~Eyeball norm~~
- Leverage CESM Port-Validation Tool
  - RSMZ-ensemble test
    - Choose single ensemble member
    - Compress/decompress member
    - Does decompressed members z-score still belong to ensemble?
  - RMSZ-bias test
    - Compress/decompress all members
    - Calculate z-score versus uncompressed ensemble
    - Compare z-score of compressed versus original
    - Does compression/decompression introduce bias?

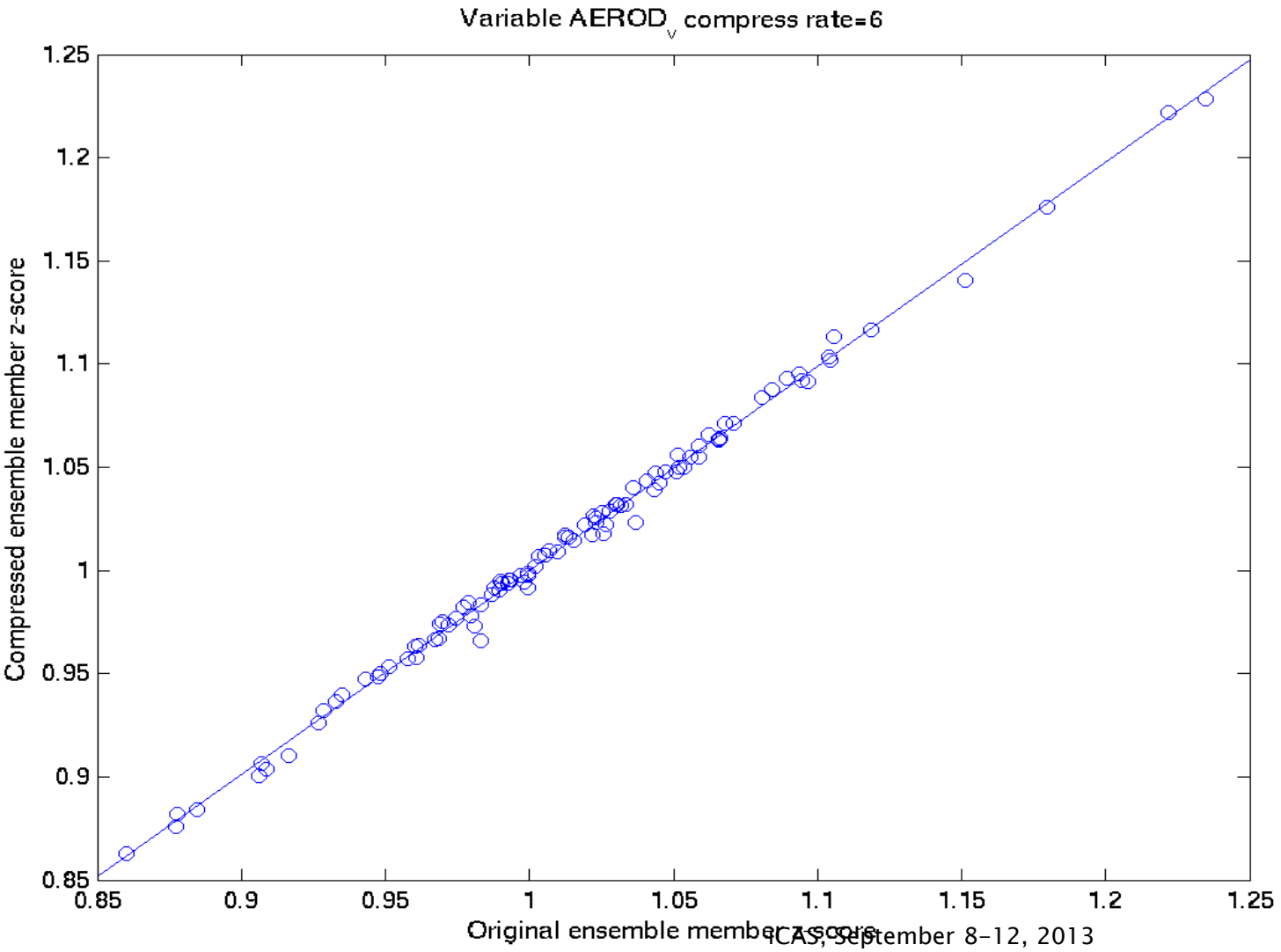
RMSZ-bias test
AEROD_V: Total Aerosol Optical Depth [3:1]



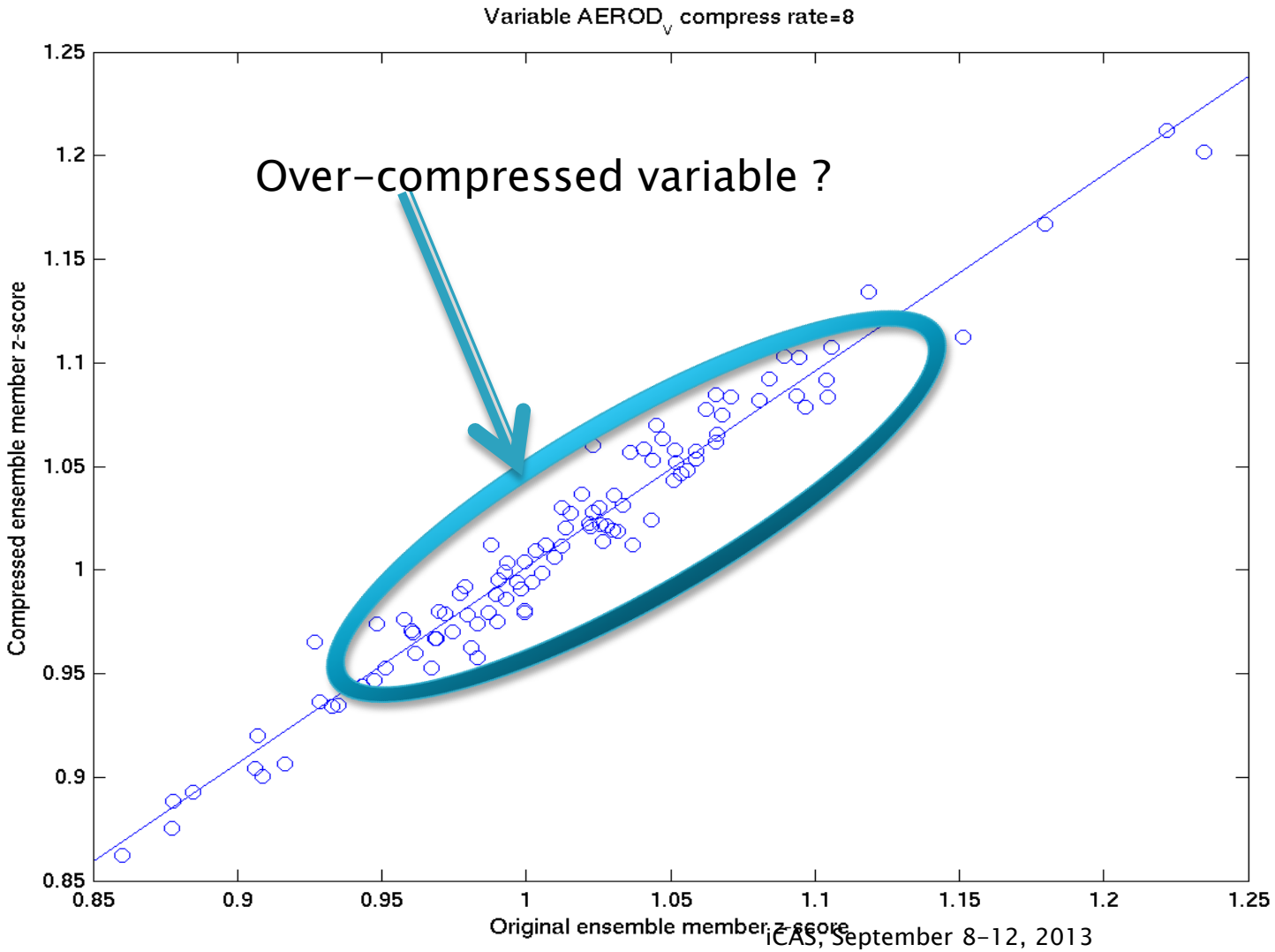
RMSZ-bias test
AEROD_V: Total Aerosol Optical Depth [4:1]



RMSZ-bias test
AEROD_V: Total Aerosol Optical Depth [6:1]

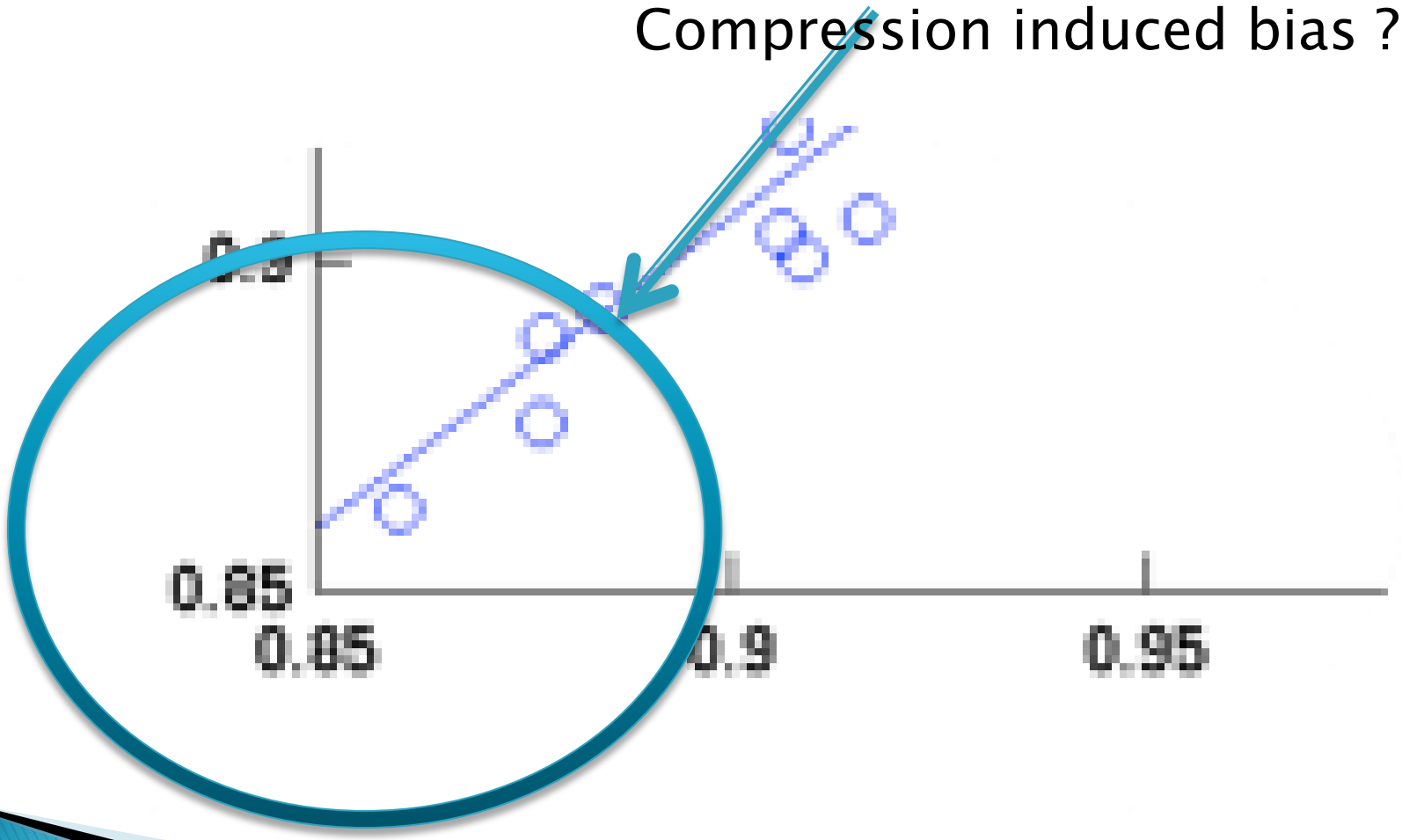


RMSZ-bias test
AEROD_V: Total Aerosol Optical Depth [8:1]

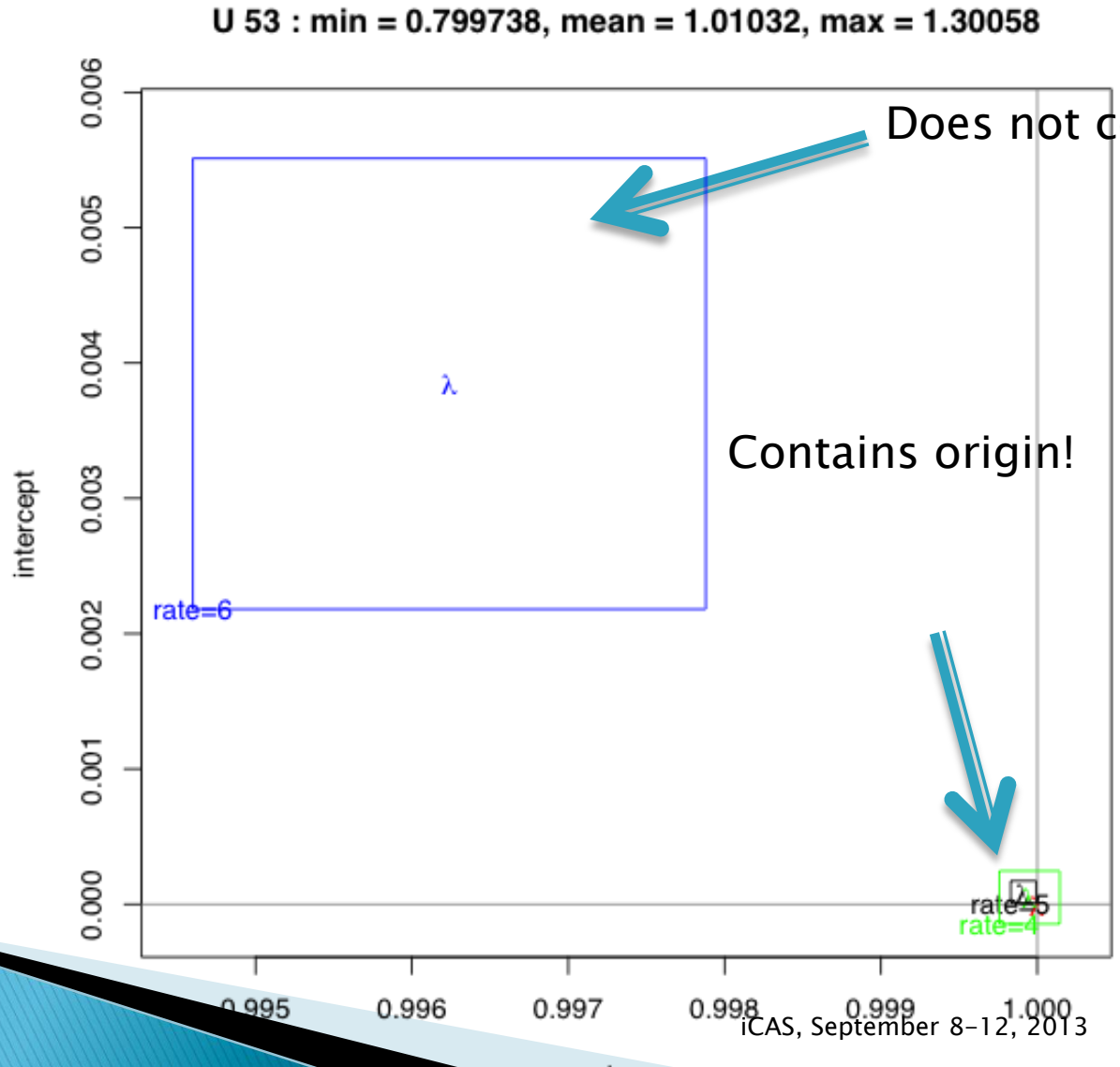


RMSZ-bias test:
AEROD_V: Total Aerosol Optical Depth [8:1]

Does not pass through origin.
Compression induced bias ?



RMSZ-bias test confidence intervals
U: velocity



Does not contain origin!

Contains origin!

Evaluating Compression

- Using Samplify APAX (2:1,4:1,6:1)
- For a variable (u) choose highest compression rate such that
  - RMSZ-ensemble test
  - RMSZ-bias test
- 139 variables
  - 2 variables: [1:1] no compression
  - 42 variables: [2:1]
  - 51 variables: [4:1]
  - 44 variables: [6:1]
- Overall 30.4% of original file size

Next Steps I

- Evaluate other lossy compression algorithms
  - Need to be Open Source
    - APAX is not a long term option
  - Potential options
    - Grib2
    - Grib2 w/JPEG 2000
    - ISABELA
    - fpzip
    - sengcom

Next Steps II: The Pepsi Challenge

Climate Scientist

Conclusions

- We can not ignore the increasing cost of output data manipulation and storage
- Statistical approach to evaluating compression algorithms
- Impact of data-compression on solution is less then bit-perturbation to initial conditions
- Potential 3x reduction in online/offline storage

It is not about the loss of information, it is about doing more science!

Questions

John Dennis
dennis@ucar.edu

References

- T. Bicer, J. Yin, D. Chiu, G. Agrawal and K. Schuchardt, “Integrated Online Compression to Accelerate Large-Scale Data Analytics Applictions”, Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS’13), Boston MA, May 2013
- N. Huebbe, A. Wegener, J. Kunkel, Y. Ling, and T. Ludwig, “Evaluating Lossy Compression on Climate Data”, ISC13
- S. Lakshminarasimhan, N. Shah, S. Ethier, S. Klasky, R. Latham, R. Ross, and N. F. Samatova. Compressing the incompressible with ISABELA: In-situ reduction of spatio-temporal data. In E. Jeannot, R. Namyst, and J. Roman, editors, Euro-Par (1), volume 6852 of Lecture Notes in Computer Science, pages 366-379. Springer, 2011.
- Peter Lindstrom and Martin Isenburg
"Fast and Efficient Compression of Floating-Point Data"
IEEE Transactions on Visualization and Computer Graphics, 12(5):1245-1250, September-October 2006