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 $\rightarrow$ 4-byte
    - Original: $T = 290.1234567890123$
    - Lossy compression:
      $T = 290.1234500000000$

- Restart files: lossless compression
- History files: lossy or lossless compression

iCAS, September 8–12, 2013
How to evaluate data compression

- Eyeball norm?
TS (Surface Temperature)

**JJA**

**LRC01**

Surf Temp (radiative) mean = 290.09

Min = 206.35 Max = 315.44

LRC01 - LRC01

mean = -0.00 rmse = 0.07

Min = -0.64 Max = 0.69

**T-test of the two means at each grid point**

Colored cells are significant at the 0.05 level

iCAS, September 8–12, 2013
New CESM Port-Validation Tool

- Create 101 ensemble
  - Initial perturb: \{-5.9,-5.8,...,-1,0,1,...,5.8,5.9\} \times 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 (\(\sigma\)) at every (i,j,k) point
- Compute root-mean-square (z-score) for the omitted member

\[
\text{RMSZ}_u = \sqrt{\frac{1}{n_x} \sum_{i,j,k} \left( \frac{(u_{i,j,k} - \bar{u}_{i,j,k})}{\sigma_{i,j,k}} \right)^2}
\]
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?

iCAS, September 8–12, 2013
Current Compression Algorithms

- **Samplify APAX**
  - Fixed rate compression [2:1, 4:1, 5:1, 6:1, 8:1]
  - [www.samplify.com](http://www.samplify.com)

- **Climate Compression (CC) [Jian] [5:1]**

- **VAPOR wavelet [5:1]**
RMSZ–Ensemble test: U: zonal wind

Raw score for U = 0.98
Run is using the original(x), Jian(*), Apax(#), Wavelet(+) 55/101 members produce a larger RMSZ
U: min=0.799738, mean=1.01032, max=1.30058
RMSZ–Ensemble test
wat_a1: aerosol water mode 01

Raw score for wat_a1 = 0.999
Run is using the original (x), Jian (*), Apax (#), Wavelet (+)
66/101 members produce a larger RMSZ
wat_a1: min=0.946415, mean=1.01582, max=1.11489

Outside ensemble
RMSZ–Ensemble test
SLV: Liquid Water virtual static energy

Raw score for SLV = 0.979
Run is using the original(x), Jian(*), Apax(#), Wavelet(+) 51/101 members produce a larger RMSZ
SLV: min=0.78802, mean=1.00684, max=1.55802

Introduces Bias?
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

U 53: min = 0.799738, mean = 1.01032, max = 1.30058

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 \( \geq \)
  - RMSZ–bias test \( \geq \)
- 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
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

- N. Huebbe, A. Wegener, J. Kunkel, Y. Ling, and T. Ludwig, “Evaluating Lossy Compression on Climate Data”, ISC13