An ensemble-based consistency tool for CESM

Allison H. Baker

Application Scalability and Performance Group

Computational and Information Systems Laboratory

National Center for Atmospheric Research

CESM Software Engineering Working Group

June 16, 2015

CISL Computational & Information Systems Laboratory

NCAR

Many collaborators!

NCAR Earth System Laboratory:

CESM Software Engineering Group

Climate and Global Dynamics Division

NCAR Computational and Information Systems Laboratory:

Application Scalability and Performance Group

Institute for Mathematics Applied to Geosciences

A.H. Baker, D.H. Hammerling, M.N. Levy, H. Xu, J.M. Dennis, B.E. Eaton, J. Edwards, C. Hannay, S. A. Michelson, R. B. Neale, D. Nychka, J. Shollenberger, J. Tribbia, M. Vertenstein, D. Williamson

+ *D. Milroy*, Computer Science, University of Colorado

CISL Computational & Information Systems Laboratory

NCAR

Software Quality Assurance for CESM

Motivation: To insure that changes during the CESM development life cycle ***do not*** adversely effect the code

- » Code modifications
- » New machine architectures
- » Compiler changes

Software Quality Assurance for CESM

Motivation: To insure that changes during the CESM development life cycle ***do not*** adversely effect the code

- » Code modifications
- » New machine architectures
- » Compiler changes

Main issue: Original data = X

"New" data = $\tilde{X}$

If $X \neq \tilde{X}$ *is the code still "correct"*?

CISL Computational & Information Systems Laboratory

NCAR

Software Quality Assurance for CESM

Motivation: To insure that changes during the CESM development life cycle ***do not*** adversely effect the code

- » Code modifications
- » New machine architectures
- » Compiler changes

Main issue: Original data = X

"New" data = $\tilde{X}$

If $X \neq \tilde{X}$ *is the code still "correct"*?

Does the new data still represent the same climate?

CISL Computational & Information Systems Laboratory

NCAR

Bit-for-Bit?

CESM results are bit-for-bit reproducible if:

The exact *same* code is run,

with *same* parameter settings,

and the *same* initial conditions,

on *same* architecture,

using the *same* compiler,

and the *same* MPI, …

not the case in most applications!

Evaluating the differences…

Question: How to assess whether the difference between

X and $\tilde{X}$ is climate changing ?

Main issue: *There is no clear definition of "climate-changing".*

Previous: Climate scientists compare multiple, long simulations:

computationally intensive, time-consuming, subjective

Need an more objective and easy-to-use methodology!

Evaluating the differences…

New methodology: Leverage climate system's ***natural variability***!

Evaluate new data in the context of an *ensemble* of CESM runs

Evaluating the differences…

New methodology: Leverage climate system's ***natural variability*!**

Evaluate new data in the context of an *ensemble* of CESM runs

- Collection of one-year CESM simulations
- $\mathcal{O}(10^{-14})$ perturbations in initial atmospheric temp.
- "accepted" machine and "accepted" software stack

Evaluating the differences…

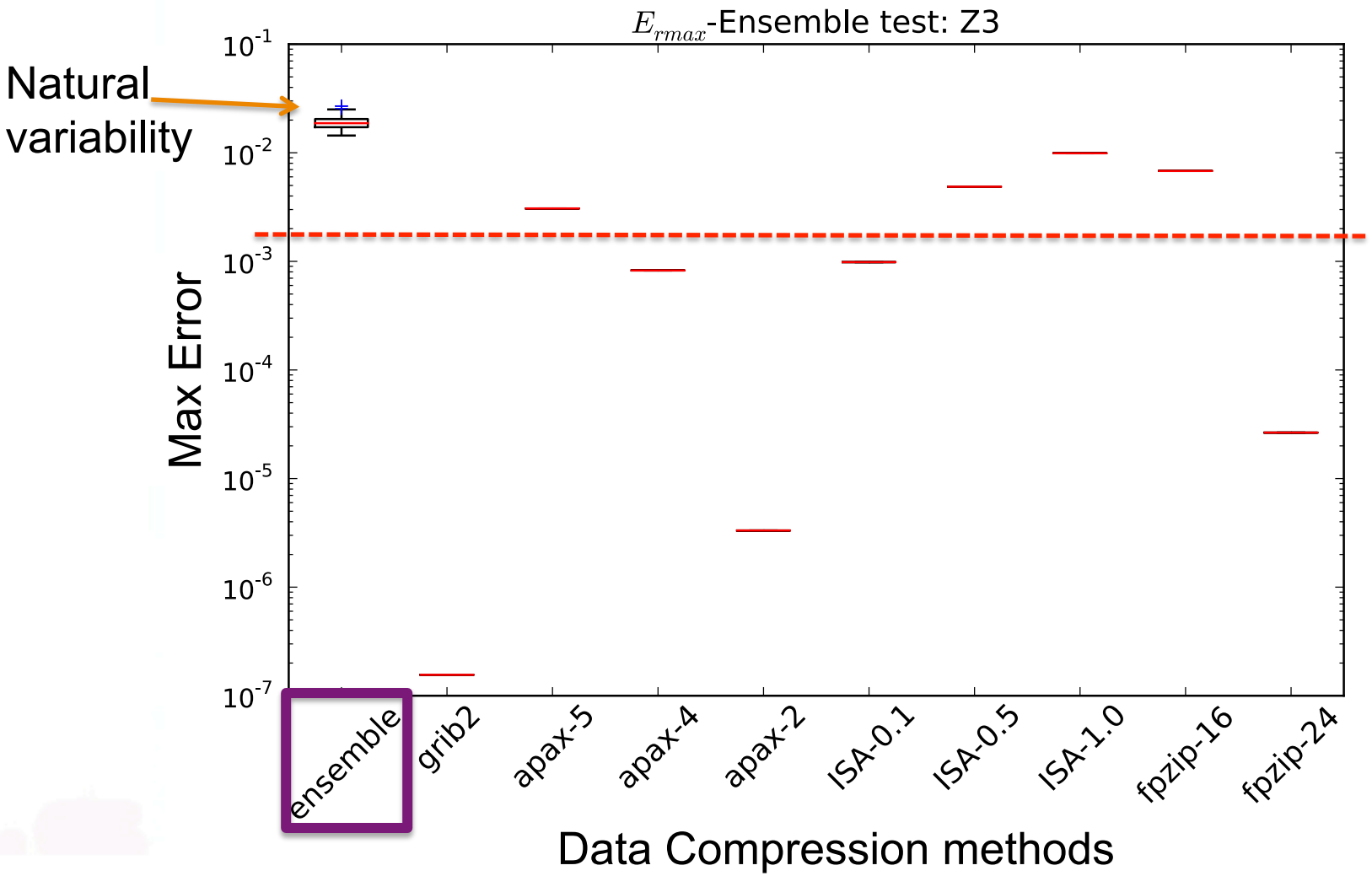
New methodology: Leverage climate system's ***natural variability*!**

Evaluate new data in the context of an *ensemble* of CESM runs

- Collection of one-year CESM simulations
- $\mathcal{O}(10^{-14})$ perturbations in initial atmospheric temp.
- "accepted" machine and "accepted" software stack

Creates an "accepted" statistical distribution that can be used to evaluate "new" runs

Why an Ensemble?

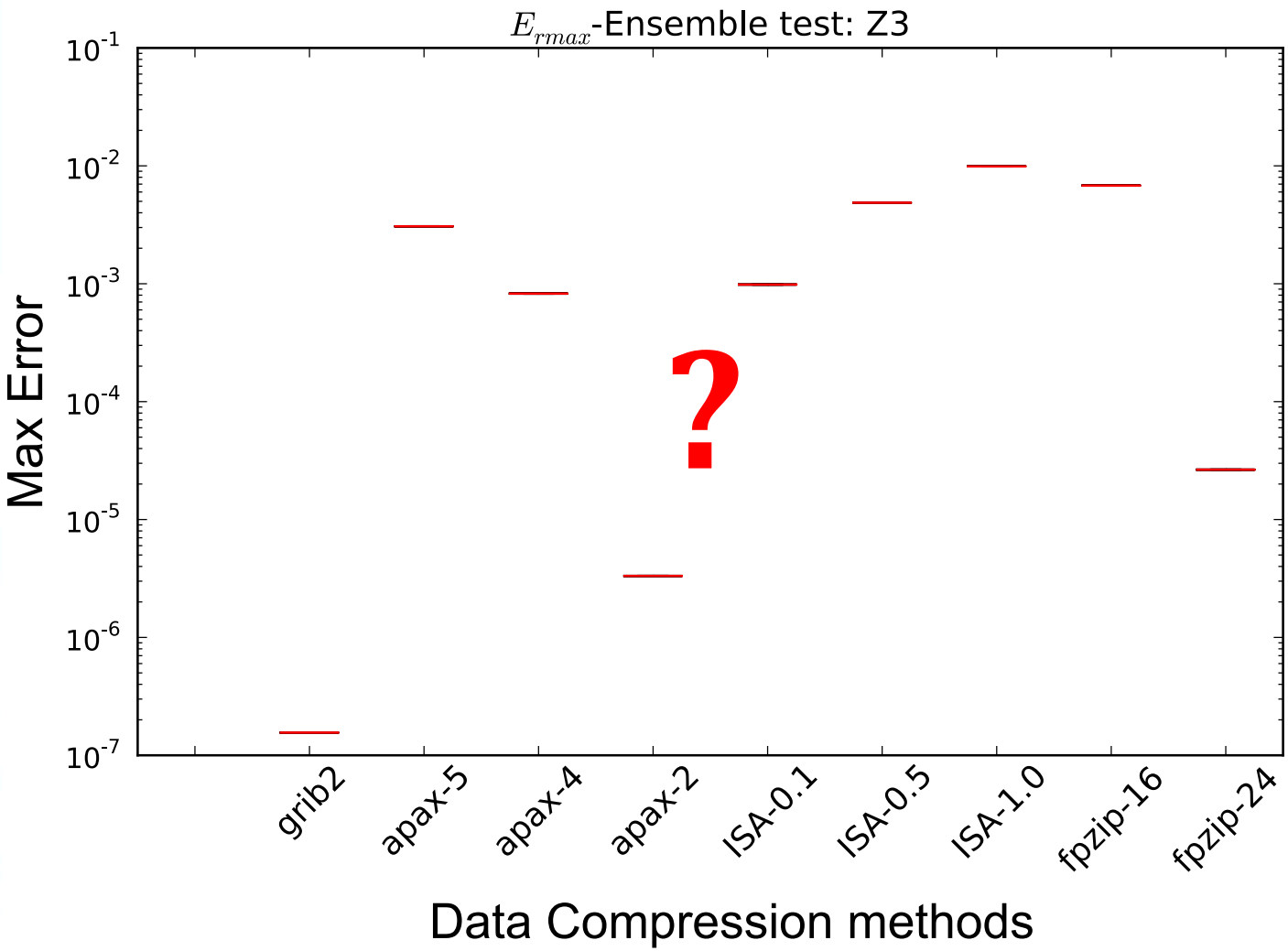


Computational & Information Systems Laboratory

CISL

NCAR

Why an ensemble?



CESM Ensemble

- 151 one-year simulations, annual means
- 1-deg atmosphere model (F-case): **120 variables**

Issue: variable dependencies

many variables are highly correlated!

⇒ ***Difficult to make pass/fail choices based on number of variables because of variable dependencies***

CESM Ensemble

Composition:

- 151 one-year simulations, annual means
- 1-deg atmosphere model (F-case): **120 variables**

Compare each variable to the ensemble:

Issue: variable dependencies

many variables are highly correlated!

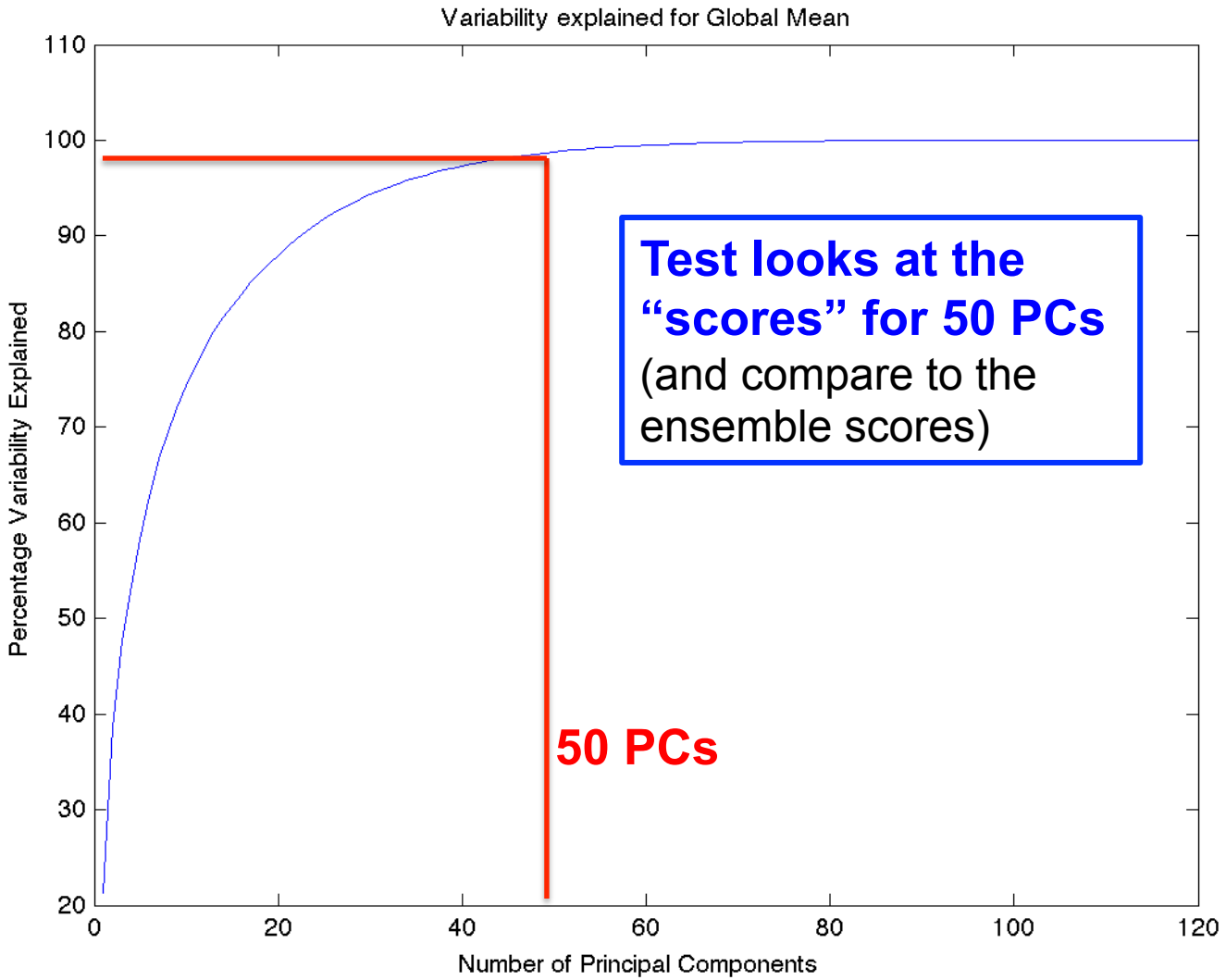
⇒ ***Difficult to make pass/fail choices based on number of variables because of variable dependencies***

=> Principal Component Analysis

Principal Component Analysis (PCA)-based testing

- Rotate (project) data into an *orthogonal* subspace that better represents the variance in the data
- Look only at components that represent the most variance (dimension reduction)
- Can determine a false positive rate

Principal Component Analysis (PCA)-based testing



Test looks at the "scores" for 50 PCs
(and compare to the ensemble scores)

CESM Ensemble Consistency Test

Step 1: *Create an ensemble of CESM runs*

- Use “accepted” machine and “accepted” software stack

CESM Ensemble Consistency Test

Step 1: *Create an ensemble of CESM runs*

- Use “accepted” machine and “accepted” software stack

Step 2: *Create ensemble summary file*

- Standardize variables
- Determine transformation matrix
- Determine distribution of scores for ensemble

CESM Ensemble Consistency Test

Step 1: *Create an ensemble of CESM runs*

- Use “accepted” machine and “accepted” software stack

Step 2: *Create ensemble summary file*

- Standardize variables
- Determine transformation matrix
- Determine distribution of scores for ensemble

Step 3: *Create “new” runs (new platform, code base, …)*

CESM Ensemble Consistency Test

Step 1: *Create an ensemble of CESM runs*

- Use "accepted" machine and "accepted" software stack

Step 2: *Create ensemble summary file*

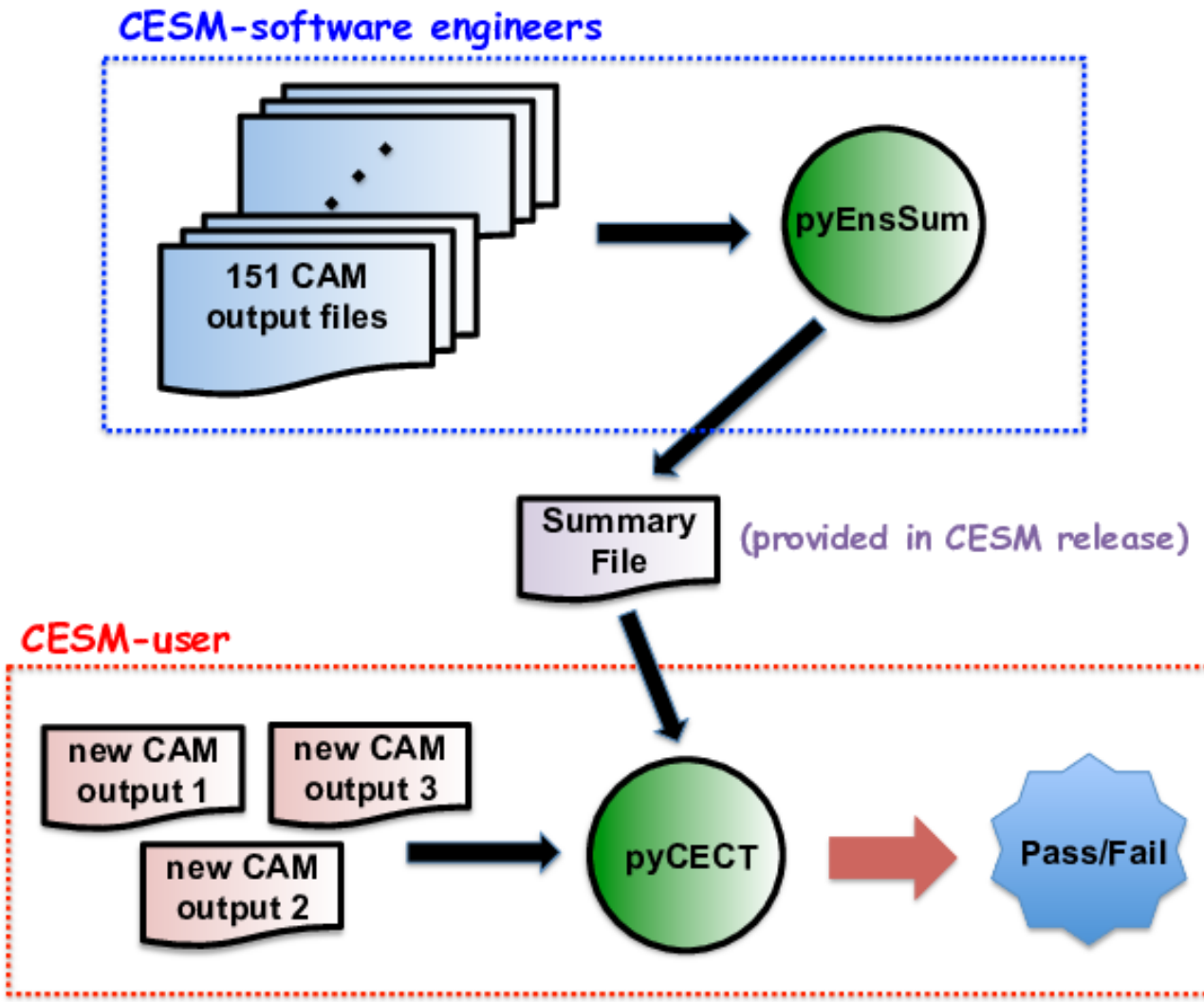
- Standardize variables
- Determine transformation matrix
- Determine distribution of scores for ensemble

Step 3: *Create "new" runs (new platform, code base, …)*

Step 4: *Evaluate new runs*

- Determine new scores (apply transformation matrix)
- Compare "new" scores to ensemble scores: issue **pass or fail**

Provided with CESM release



CESM Ensemble Consistency Test

Advantages:

- User-friendly (climate-modeling expertise is *not* required)
- Better feedback for model developers
- Flexible accept/reject criteria

Many uses:

- Port-verification (new CESM-supported architectures)
- Heterogeneous computing platforms
- Exploration of new algorithms, solvers, compiler options, …
- Evaluation of data compression on CESM data

Does it work?

Experimental Studies:

- **Modifications not expected to be climate-changing**
  - 5 of 5 compiler and threading modifications *pass*
- **Modifications expected to be climate-changing**
  - 10 of 11 CAM parameter modifications *fail*
- **CESM-supported machines as modifications**
  - Some borderline failures - *Currently investigating*

Practical applications

- **Error in cloud generator code only manifested on big endian machine**
  - ❖ Decisive failures on big endian machine
- **Errors in new version of Community Ice Code**
  - ❖ Not detected in standalone component testing

Test name	CESM-ECT Results	Number of PCs failing at least 2 runs
CICE4-INTEL	PASS	1
CICE4-GNU	PASS	0
CICE4-PGI	PASS	0
CICE5-INTEL	FAIL	19
CICE5-GNU	FAIL	20
CICE5-PGI	FAIL	19

Manuscript

Geoscientific Model Development
An interactive open-access journal of the European Geosciences Union

Geosci. Model Dev. Discuss., 8, 3823-3859, 2015
www.geosci-model-dev-discuss.net/8/3823/2015/
doi:10.5194/gmdd-8-3823-2015

Technical/Development/Evaluation Paper

08 May 2015

A new ensemble-based consistency test for the Community Earth System Model

A. H. Baker, D. M. Hammerling, M. N. Levy, H. Xu, J. M. Dennis, B. E. Eaton, J. Edwards, C. Hannay, S. A. Mickelson, R. B. Neale, D. Nychka, J. Shollenberger, J. Tribbia, M. Vertenstein, and D. Williamson
The National Center for Atmospheric Research, Boulder, CO, USA

Review Status
This discussion paper is under review for the journal Geoscientific Model Development (GMD).

Received: 15 April 2015 – Accepted: 16 April 2015 – Published: 08 May 2015

Abstract. Climate simulations codes, such as the Community Earth System Model (CESM), are especially complex and continually evolving. Their on-going state of development requires frequent software verification in the form of quality assurance to both preserve the quality of the code and instill model confidence. To formalize and simplify this previously subjective and computationally-expensive aspect of the verification process, we have developed a new tool for evaluating climate consistency. Because an ensemble of simulations allows us to gauge the natural variability of the model's climate, our new tool uses an ensemble approach for consistency testing. In particular, an ensemble of CESM climate runs is created, from which we obtain a statistical distribution that can be used to determine whether a new climate run is statistically distinguishable from the original ensemble. The CESM Ensemble Consistency Test, referred to as CESM-ECT, is objective in nature and accessible to CESM developers and users. The tool has proven its utility in detecting errors in software and hardware environments and providing rapid feedback to model developers.

Citation: Baker, A. H., Hammerling, D. M., Levy, M. N., Xu, H., Dennis, J. M., Eaton, B. E., Edwards, J., Hannay, C., Mickelson, S. A., Neale, R. B., Nychka, D., Shollenberger, J., Tribbia, J., Vertenstein, M., and Williamson, D.: A new ensemble-based consistency test for the Community Earth System Model, Geosci. Model Dev. Discuss., 8, 3823-3859, doi:10.5194/gmdd-8-3823-2015, 2015.

NCAR | NSF

Ensemble Composition

Effectiveness of CESM-ECT method relies heavily on the "accepted" ensemble composition

- *size 151, Yellowstone machine, Intel compiler*
- *perturbing the initial condition (IC) for atmospheric temp.*

Does the original ensemble represent the variability of a consistent climate?

- *Is using IC perturbation the "right" approach?*
- *How well do IC perturbation capture "legitimate" differences?*
- *Is the current ensemble distribution sufficient to capture compiler changes**?***

CISL Computational & Information Systems Laboratory

NCAR

Ensemble Composition

Do we need different compilers represented in the ensemble?
(compiler variability vs. IC variability)

More experiments:

1. Repeat IC experiments using other compilers (intel, pgi, gnu)

2. Include multiple compilers in ensemble (random draws)

3. Investigate ensemble size/ stability of PC

More detailed info on Dorit Hammerling's poster

Don't want pass/fail dependent on which random samble from ensemble ….

CISL Computational & Information Systems Laboratory

NCAR

Failure percentages

Test runs	453 3-compiler ensemble rand1	453 3-compiler ensemble rand2	453 3-compiler ensemble rand3
intel-extra30-rand1	0.8	0.0	0.2
intel-extra30-rand2	0.1	0.4	0.2
intel-extra30-rand3	0.5	0.3	0.7
gnu-extra30-rand1	0.4	0.4	0.0
gnu-extra30-rand2	0.2	0.1	0.3
gnu-extra30-rand3	0.1	0.2	1.8
pgi-extra30-rand1	0.5	0.4	0.1
pgi-extra30-rand2	0.1	1.4	0.7
pgi-extra30-rand3	0.3	0.5	1.1

Agrees with specified 0.5% false positive rate!

Next Steps

- *Investigate borderline machine failures*
- *Introduce simple, legitimate code changes and test*
  (Do minimal code changes pass the CESM-ECT?)
- *Fine-grained testing capability for failures*
  (to identify groups of variables that cause failure)
- *Evaluate spatial patterns in addition to global*
  (e.g. regional features, boundaries ocean/land, spatial structure)
- *Evaluate spatial relationships between variables*
  (cross-covariance studies)

Thanks!