An ensemble-based consistency test for the Community Earth System Model

Allison Baker

Application Scalability and Performance Group
National Center for Atmospheric Research

International Computing for the Atmospheric Sciences Symposium (ICAS2015)
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


+ D. Milroy, Computer Science, University of Colorado
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”?
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?
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
- $O\left(10^{-14}\right)$ perturbations in initial atmospheric temp.
- “accepted” machine and “accepted” software stack

Creates an “accepted” statistical distribution that can be used to evaluate “new” runs
Composi/on:
• 151 one-year simulations, annual means
• 1-deg active atm. and land (F-case): 120 variables

Compare each new variable value to the ensemble distribution:
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

- Each component is a linear combination of all 120 variables

- 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

**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, …)

**Step 4:** Evaluate new runs
- Determine new scores (apply transformation matrix)
- Compare new scores to ensemble scores: issue pass or fail
CESM Ensemble Consistency Test

**CESM-software engineers**
- 151 CAM output files
- pyEnsSum
- Summary File (provided in CESM release)

**CESM-user**
- new CAM output 1
- new CAM output 3
- new CAM output 2
- pyCECT
- Pass/Fail
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?

Initial Experiments:

- 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

<table>
<thead>
<tr>
<th>Test name</th>
<th>CESM-ECT Results</th>
<th>Number of PCs failing at least 2 runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>CICE4-INTEL</td>
<td>PASS</td>
<td>1</td>
</tr>
<tr>
<td>CICE4-GNU</td>
<td>PASS</td>
<td>0</td>
</tr>
<tr>
<td>CICE4-PGI</td>
<td>PASS</td>
<td>0</td>
</tr>
<tr>
<td>CICE5-INTEL</td>
<td>FAIL</td>
<td>19</td>
</tr>
<tr>
<td>CICE5-GNU</td>
<td>FAIL</td>
<td>20</td>
</tr>
<tr>
<td>CICE5-PGI</td>
<td>FAIL</td>
<td>19</td>
</tr>
</tbody>
</table>
A new ensemble-based consistency test for the Community Earth System Model (pyCECT v1.0)


The National Center for Atmospheric Research, Boulder, CO, USA

Received: 15 Apr 2015 – Published in Geosci. Model Dev. Discuss.: 08 May 2015
Revised: 22 Aug 2015 – Accepted: 24 Aug 2015 – Published: 09 Sep 2015

Abstract. Climate simulation codes, such as the Community Earth System Model (CESM), are especially complex and continually evolving. Their ongoing 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.

https://github.com/NCAR-CISL-ASAP/PyCECT/releases
How to create the ensemble?

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

- size 151, 1-year, Yellowstone machine, Intel compiler
- perturbing the initial condition (IC) for atmospheric temp.

Does the original CESM-ECT ensemble represent the variability of a consistent climate?

- How well do IC perturbation capture “legitimate” differences?
- Is the current distribution sufficient to capture compiler and code changes?

More extensive testing...
Do we need different compilers represented in the ensemble?

More tests:
1. Repeat IC experiments using all 3 CESM-supported Yellowstone compilers (intel, pgi, gnu) – 181 members
2. Same perturbations in initial conditions
3. 120 variables
Ensemble composition

Is the current ensemble distribution sufficient to capture compiler changes?
**Initial Condition effect?**

**Random draws**
- Exclude 30 at random from each 181-member set
- Test the excluded set

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand-Intel-1</td>
<td>8.1</td>
<td>0.0</td>
<td>0.6</td>
<td>9.5</td>
<td>3.5</td>
<td>3.9</td>
<td>4.5</td>
<td>2.3</td>
<td>6.0</td>
</tr>
<tr>
<td>Rand-Intel-2</td>
<td>0.1</td>
<td>4.8</td>
<td>1.9</td>
<td>4.1</td>
<td>11.0</td>
<td>5.8</td>
<td>7.7</td>
<td>2.4</td>
<td>7.8</td>
</tr>
<tr>
<td>Rand-Intel-3</td>
<td>0.3</td>
<td>0.2</td>
<td>8.3</td>
<td>8.9</td>
<td>8.7</td>
<td>8.7</td>
<td>5.2</td>
<td>4.4</td>
<td>12.7</td>
</tr>
</tbody>
</table>

*Don't want pass/fail dependent on which random sample from ensemble ...*
Compiler vs. IC effects

Two-way analysis of variance (ANOVA):

![Graph showing mean squares ratio: Compiler to Initial Conditions](image-url)
Compiler vs. IC effects

Two-way analysis of variance (ANOVA):

Mean=6.6

Median=1.3

Compiler has more influence....
Refine Ensemble

- Compiler aggregate ensembles (size 453)

<table>
<thead>
<tr>
<th>Experiments</th>
<th>Intel-Gnu-PGI Rand-1</th>
<th>Intel-Gnu-PGI Rand-2</th>
<th>Intel-Gnu-PGI Rand-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand-Intel-1</td>
<td>0.8</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>Rand-Intel-2</td>
<td>0.2</td>
<td>1.8</td>
<td>0.4</td>
</tr>
<tr>
<td>Rand-Intel-3</td>
<td>0.8</td>
<td>0.5</td>
<td>1.9</td>
</tr>
</tbody>
</table>

- Low failure rates
Simple code changes?

Is the current ensemble distribution sufficient to capture reasonable code changes to CAM?
(e.g., mathematically identical and “small”)

Example 1 (Combine)

Original:
\[
\omega_p(i,j,1) = \frac{\text{vgrad}_p(i,j,1)}{p(i,j,1)}
\]
\[
\omega_p(i,j,1) = \omega_p(i,j,1) - 0.5d0/p(i,j,1) \ast \text{divdp}(i,j,1)
\]

Modified:
\[
\omega_p(i,j,1) = \frac{(\text{vgrad}_p(i,j,1) - 0.5d0 \ast \text{divdp}(i,j,1))}{p(i,j,1)}
\]
Example 2 (Expand)

Original:
\[
\text{phii}(i,j,nlev) = R_{gas} \cdot T_v(i,j,nlev) \cdot hkl \\
\text{phi}(i,j,nlev) = \text{phis}(i,j) + R_{gas} \cdot T_v(i,j,nlev) \cdot hkk
\]

Modified:
\[
\text{phii}(i,j,nlev) = T_v(i,j,nlev) \cdot hkl \\
\text{phii}(i,j,nlev) = R_{gas} \cdot \text{phii}(i,j,nlev) \\
\text{tt_real} = T_v(i,j,nlev) \cdot hkk \\
\text{phi}(i,j,nlev) = \text{phis}(i,j) + R_{gas} \cdot \text{tt_real}
\]
Simple code changes?

"original"

<table>
<thead>
<tr>
<th>Category</th>
<th>Intel (151)</th>
<th>Intel-Gnu-Pgi (150)</th>
<th>Intel-Gnu-Pgi (300)</th>
<th>Intel-Gnu-Pgi (453)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Failure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combine</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expand</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Div-to-Mult.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edge Order</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rand. Num.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CISL
Computational & Information Systems Laboratory
Next Steps

1) Investigate ensemble size
   (stability of PC calculations)

2) Length of ensemble runs
   (shorter?)
Next Steps

1) Investigate ensemble size  
   (stability of PC calculations)

2) Length of ensemble runs  
   (shorter?)

3) Fine-grained testing capability for failures  
   (to identify groups of variables that cause failure)
Next Steps

1) Investigate ensemble size
   (stability of PC calculations)

2) Length of ensemble runs
   (shorter?)

3) Fine-grained testing capability for failures
   (to identify groups of variables that cause failure)

4) Evaluate spatial patterns in addition to global
   (e.g. regional features, boundaries ocean/land, spatial structure)

5) Evaluate spatial relationships between variables
   (cross-covariance studies)