Ensemble Consistency Testing for CESM: A new form of Quality Assurance

Dorit Hammerling

Institute for Mathematics Applied to Geosciences
National Center for Atmospheric Research (NCAR)

Joint work with Allison Baker (Application Scalability and Performance Group/NCAR)
and Daniel Milroy (Department of Computer Science/CU Boulder)
. . . and many other contributors

April 5, 2016

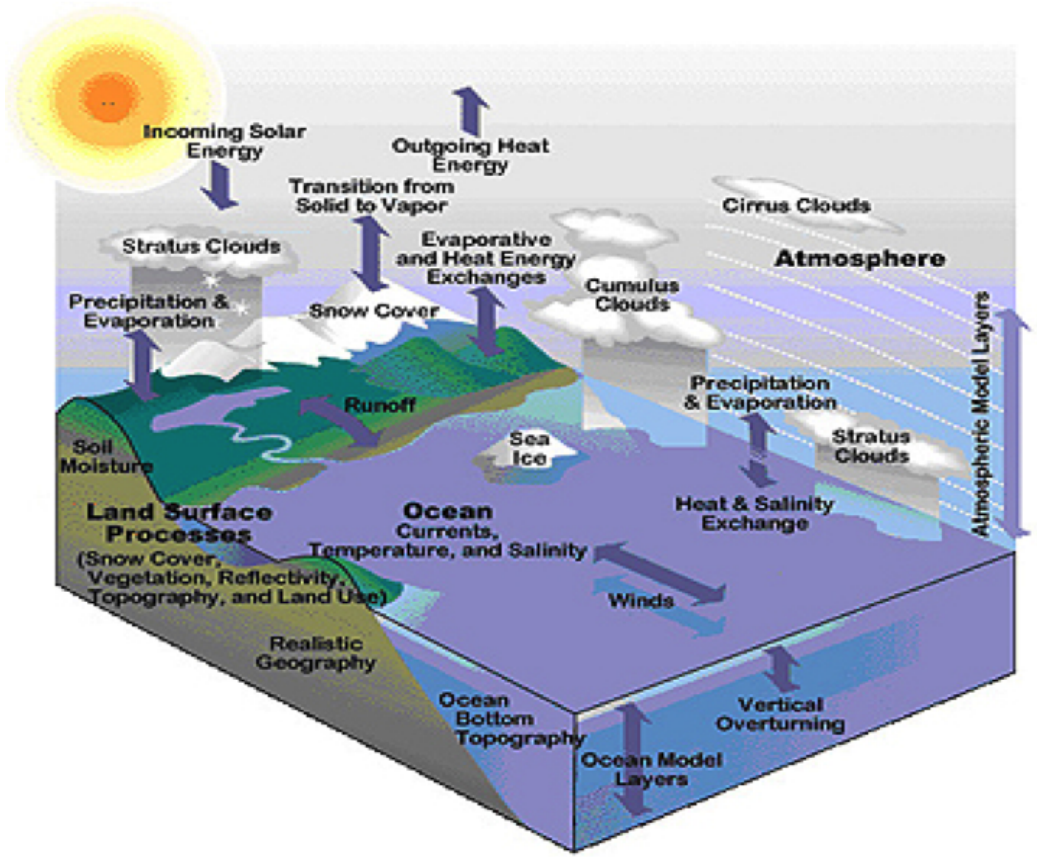
Outline

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Outline

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

NCAR's Community Earth System Model



- a "virtual laboratory" to study past, present and future climate states
- describes interactions of the atmosphere, land, river runoff, land-ice, oceans and sea-ice
- complex! Large code base: approx. 1.5 Millions lines of code

Need for Software Quality Assurance

- Code ported to new environment (different institution)
- To ensure that changes during the CESM development life cycle **do not** adversely effect the results.
  - Code modifications
  - New machine architectures
  - Compiler changes
  - Exascale-computing technologies
  - ...

Bit-for-Bit reproducibility

CESM is deterministic and results are bit-for-bit reproducible

IF

exact same code is run,
with same parameter settings,
same initial conditions,
on same hardware architecture,
using the same compiler,
same implementation of MPI,
same random number generator,
...

$\Rightarrow$ not the case in most applications

Bit-for-Bit reproducibility

CESM is deterministic and results are bit-for-bit reproducible

IF

exact same code is run,
with same parameter settings,
same initial conditions,
on same hardware architecture,
using the same compiler,
same implementation of MPI,
same random number generator,
...

$\Rightarrow$ not the case in most applications

How to assess difference if not bit-for-bit?

X: Original "accepted" data

X_{new} : "New" data

Key question:

If $X \neq X_{new}$ is the code/implementation still "correct"?

$\Rightarrow$ Does the new data still represent the same climate? Or is it "climate-changing"?

How to assess difference if not bit-for-bit?

X: Original "accepted" data

X_{new} : "New" data

Key question:

If $X \neq X_{new}$ is the code/implementation still "correct"?

$\Rightarrow$ Does the new data still represent the same climate? Or is it "climate-changing"?

Evaluating the difference

How to assess whether the difference between X and X_{new} is climate-changing?

Main problem:

There is no clear definition of "climate-changing".

What was done in the past:

Climate scientists compare 400-year simulations.

- computationally intensive
- time-consuming
- subjective
- ...

Evaluating the difference

How to assess whether the difference between X and X_{new} is climate-changing?

Main problem:

There is no clear definition of "climate-changing".

What was done in the past:

Climate scientists compare 400-year simulations.

- computationally intensive
- time-consuming
- subjective
- ...

Evaluating the difference

How to assess whether the difference between X and X_{new} is climate-changing?

Main problem:

There is no clear definition of "climate-changing".

What was done in the past:

Climate scientists compare 400-year simulations.

- computationally intensive
- time-consuming
- subjective
- ...

Outline

New approach: Ensemble Consistency Testing

Leverage climate system's **natural variability** (or rather it's chaotic nature)!

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

- Use "accepted" machine and "accepted" software stack
- One-year CESM simulations with $O(10^{-14})$ perturbations in initial atmospheric temperature
- 1-deg atmosphere model (F-case): 134 variables

$\Rightarrow$ Creates an accepted ensemble distribution that can be used to evaluate new runs.

New approach: Ensemble Consistency Testing

Leverage climate system's **natural variability** (or rather it's chaotic nature)!

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

- Use "accepted" machine and "accepted" software stack
- One-year CESM simulations with $O(10^{-14})$ perturbations in initial atmospheric temperature
- 1-deg atmosphere model (F-case): 134 variables

$\Rightarrow$ Creates an accepted ensemble distribution that can be used to evaluate new runs.

New approach: Ensemble Consistency Testing

Leverage climate system's **natural variability** (or rather it's chaotic nature)!

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

- Use "accepted" machine and "accepted" software stack
- One-year CESM simulations with $O(10^{-14})$ perturbations in initial atmospheric temperature
- 1-deg atmosphere model (F-case): 134 variables

$\Rightarrow$ Creates an accepted ensemble distribution that can be used to evaluate new runs.

New approach: Ensemble Consistency Testing

Leverage climate system's **natural variability** (or rather it's chaotic nature)!

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

- Use "accepted" machine and "accepted" software stack
- One-year CESM simulations with $O(10^{-14})$ perturbations in initial atmospheric temperature
- 1-deg atmosphere model (F-case): 134 variables

  $\Rightarrow$ Creates an accepted ensemble distribution that can be used to evaluate new runs.

Ensemble features

Variable dependencies:

Many of the variables are highly correlated ($> .9$)

Difficult to make pass/fail choices based on number of variables because of dependencies.

$\Rightarrow$ Principal Component Analysis (PCA)

- Projection into orthogonal space
- Linear combinations of variables ("scores") instead of individual variables are used as the ensemble distribution and evaluated

Ensemble features

Variable dependencies:

Many of the variables are highly correlated ($> .9$)

Difficult to make pass/fail choices based on number of variables because of dependencies.

$\Rightarrow$ Principal Component Analysis (PCA)

- Projection into orthogonal space
- Linear combinations of variables ("scores") instead of individual variables are used as the ensemble distribution and evaluated

Ensemble features

Variable dependencies:

Many of the variables are highly correlated ($> .9$)

Difficult to make pass/fail choices based on number of variables because of dependencies.

$\Rightarrow$ Principal Component Analysis (PCA)

- Projection into orthogonal space
- Linear combinations of variables ("scores") instead of individual variables are used as the ensemble distribution and evaluated

PCA-based Ensemble Consistency Testing setup

Step 1: Create ensemble

Step 2: Standardize variables and determine transformation matrix ("loadings")

Step 3: Determine distribution of scores

Step 4: Create new runs, apply transformation and determine "new" scores

Step 5: Compare "new" scores to ensemble scores and determine pass or fail

Main idea: compare scores from new runs to scores from ensemble

PCA-based Ensemble Consistency Testing setup

Step 1: Create ensemble

Step 2: Standardize variables and determine transformation matrix ("loadings")

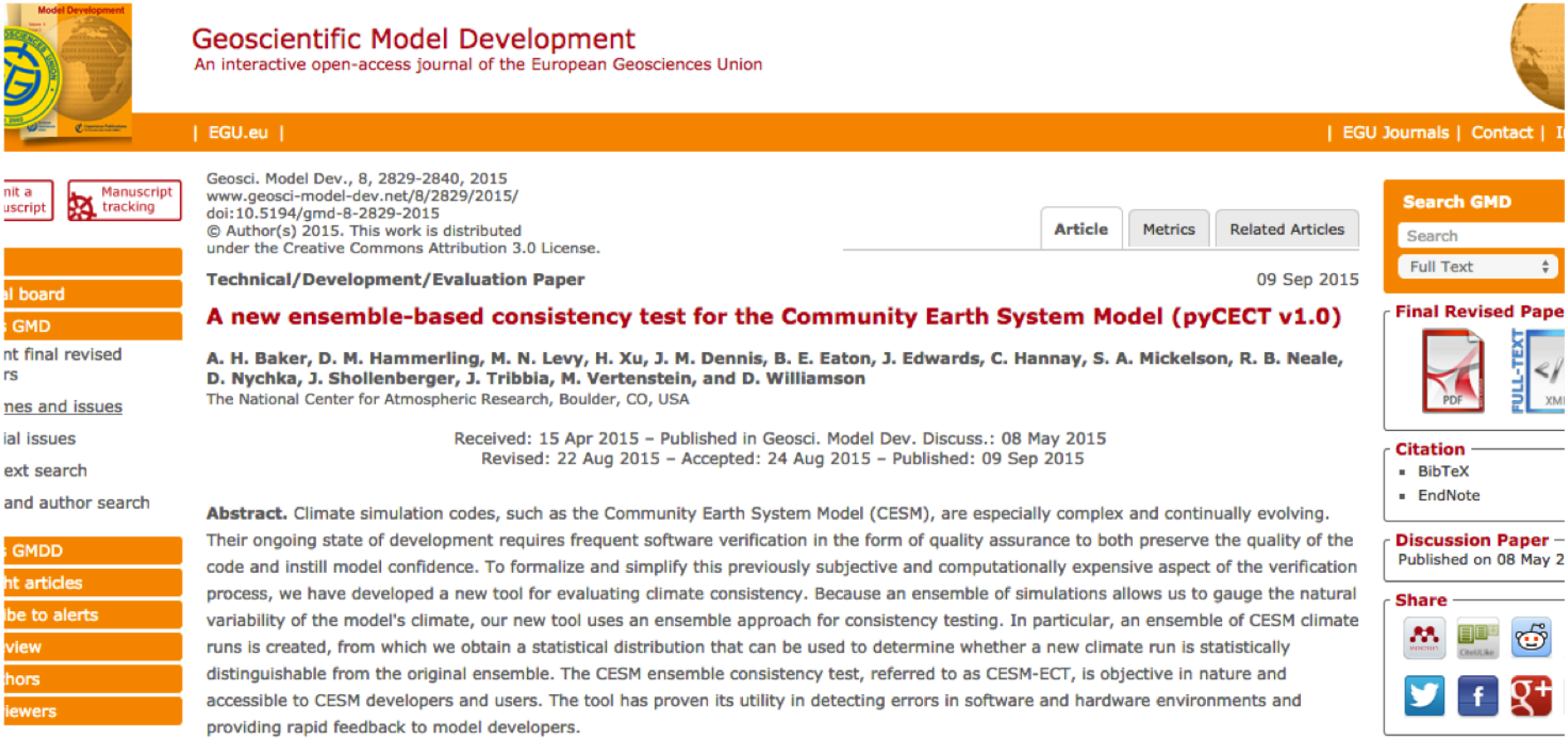
Step 3: Determine distribution of scores

Step 4: Create new runs, apply transformation and determine "new" scores

Step 5: Compare "new" scores to ensemble scores and determine pass or fail

Main idea: compare scores from new runs to scores from ensemble

Manuscript and Code available



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

| EGU.eu | | EGU Journals | Contact |

Manuscript tracking

board
GMD
nt final revised
rs
nes and issues
ial issues
ext search
and author search
GMDD
nt articles
be to alerts
view
hors
iewers

Geosci. Model Dev., 8, 2829-2840, 2015
www.geosci-model-dev.net/8/2829/2015/
doi:10.5194/gmd-8-2829-2015
© Author(s) 2015. This work is distributed
under the Creative Commons Attribution 3.0 License.

Article | Metrics | Related Articles

Technical/Development/Evaluation Paper

09 Sep 2015

A new ensemble-based consistency test for the Community Earth System Model (pyCECT v1.0)

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

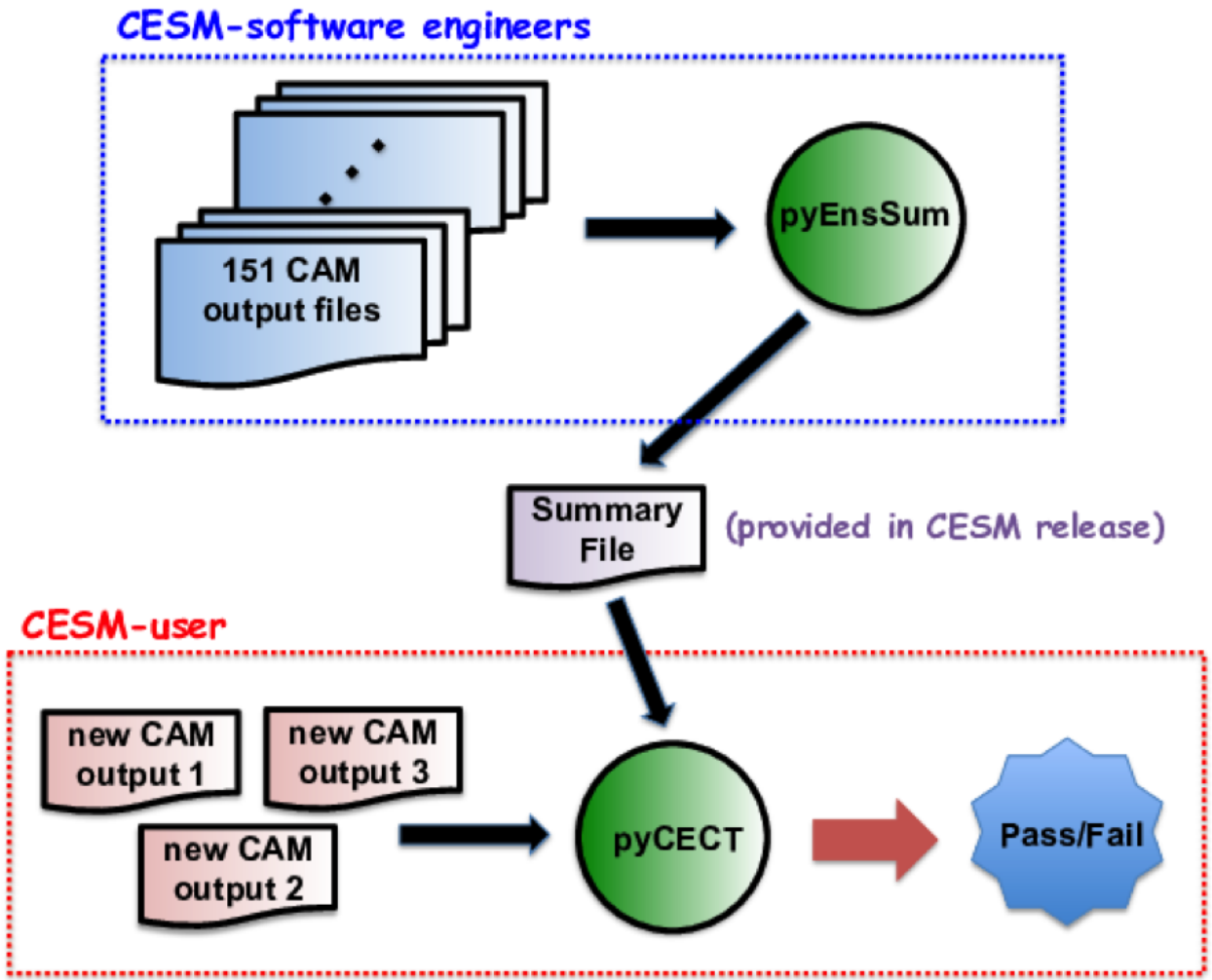
Received: 15 Apr 2015 – Published in Geosci. Model Dev. Discuss.: 08 May 2015
Revised: 22 Aug 2015 – Accepted: 24 Aug 2015 – Published: 09 Sep 2015

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.

Search GMD
Search
Full Text
Final Revised Pape
PDF
FULL-TEXT
Citation
BibTeX
EndNote
Discussion Paper
Published on 08 May 2
Share

https://github.com/NCAR-CISL-ASAP/PyCECT/releases

Schematic of Ensemble Testing framework



Highlights from the manuscript

- Defined testing framework to obtain a 0.5% false-positive rate (conservative!)
- Testing of purposefully constructed cases that should fail
- Testing of other machines
  - Most passed, but some failed
  - mira failed persistently

$\Rightarrow$ Turned out to be an issue with FMA, but led us to scrutinize the approach and ensemble composition

Highlights from the manuscript

- Defined testing framework to obtain a 0.5% false-positive rate (conservative!)
- Testing of purposefully constructed cases that should fail
- Testing of other machines
  - Most passed, but some failed
  - mira failed persistently

$\Rightarrow$ Turned out to be an issue with FMA, but led us to scrutinize the approach and ensemble composition

Highlights from the manuscript

- Defined testing framework to obtain a 0.5% false-positive rate (conservative!)
- Testing of purposefully constructed cases that should fail
- Testing of other machines
  - Most passed, but some failed
  - mira failed persistently

$\Rightarrow$ Turned out to be an issue with FMA, but led us to scrutinize the approach and ensemble composition

Highlights from the manuscript

- Defined testing framework to obtain a 0.5% false-positive rate (conservative!)
- Testing of purposefully constructed cases that should fail
- Testing of other machines
  - Most passed, but some failed
  - mira failed persistently

$\Rightarrow$ Turned out to be an issue with FMA, but led us to scrutinize the approach and ensemble composition

Highlights from the manuscript

- Defined testing framework to obtain a 0.5% false-positive rate (conservative!)
- Testing of purposefully constructed cases that should fail
- Testing of other machines
  - Most passed, but some failed
  - mira failed persistently

$\Rightarrow$ Turned out to be an issue with FMA, but led us to scrutinize the approach and ensemble composition

Outline

How well does CAM-ECT truly work?

Modifications expected to be climate-changing fail,
but is what *should* pass actually passing
corresponding to our specified false positive rate?

$\Rightarrow$ Properties of the test rely heavily on the "accepted" ensemble composition.

How well does CAM-ECT truly work?

Modifications expected to be climate-changing fail,
but is what *should* pass actually passing
corresponding to our specified false positive rate?

$\Rightarrow$ Properties of the test rely heavily on the "accepted" ensemble composition.

How well does CAM-ECT truly work?

Modifications expected to be climate-changing fail,
but is what *should* pass actually passing
corresponding to our specified false positive rate?

$\Rightarrow$ Properties of the test rely heavily on the "accepted" ensemble composition.

Revisiting the ensemble composition

- Is using only initial condition perturbations the "right" approach?
- How well do initial condition perturbations capture "legitimate" differences?
- How do runs from different compilers compare to initial perturbation runs?
- How large of an ensemble do we need to capture legitimate changes?

$\Rightarrow$ We created very large ensembles using a combination of initial conditions and different compilers.

Revisiting the ensemble composition

- Is using only initial condition perturbations the "right" approach?
- How well do initial condition perturbations capture "legitimate" differences?
- How do runs from different compilers compare to initial perturbation runs?
- How large of an ensemble do we need to capture legitimate changes?

$\Rightarrow$ We created very large ensembles using a combination of initial conditions and different compilers.

Revisiting the ensemble composition

- Is using only initial condition perturbations the "right" approach?
- How well do initial condition perturbations capture "legitimate" differences?
- How do runs from different compilers compare to initial perturbation runs?
- How large of an ensemble do we need to capture legitimate changes?

$\Rightarrow$ We created very large ensembles using a combination of initial conditions and different compilers.

Revisiting the ensemble composition

- Is using only initial condition perturbations the "right" approach?
- How well do initial condition perturbations capture "legitimate" differences?
- How do runs from different compilers compare to initial perturbation runs?
- How large of an ensemble do we need to capture legitimate changes?

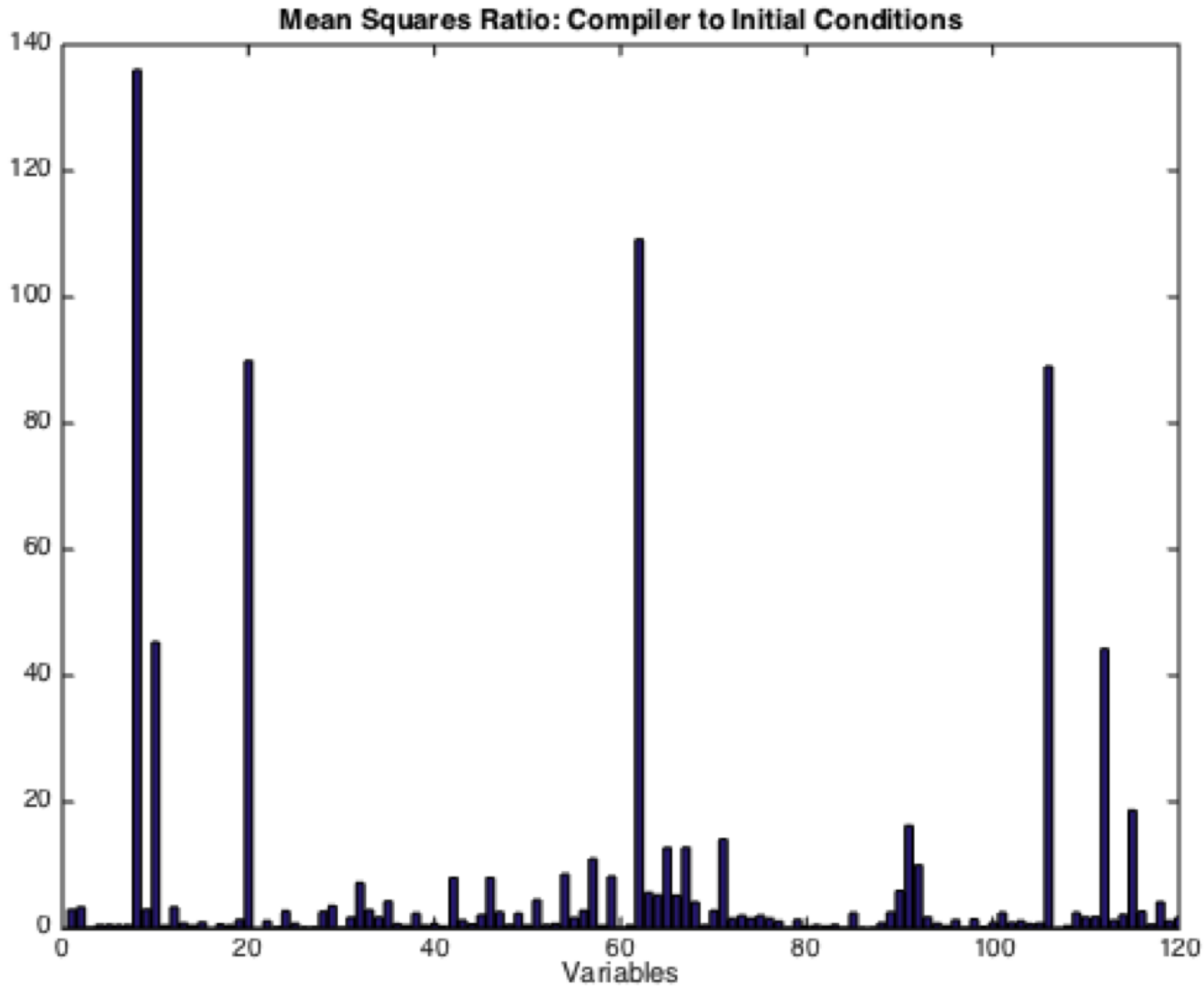
$\Rightarrow$ We created very large ensembles using a combination of initial conditions and different compilers.

Revisiting the ensemble composition

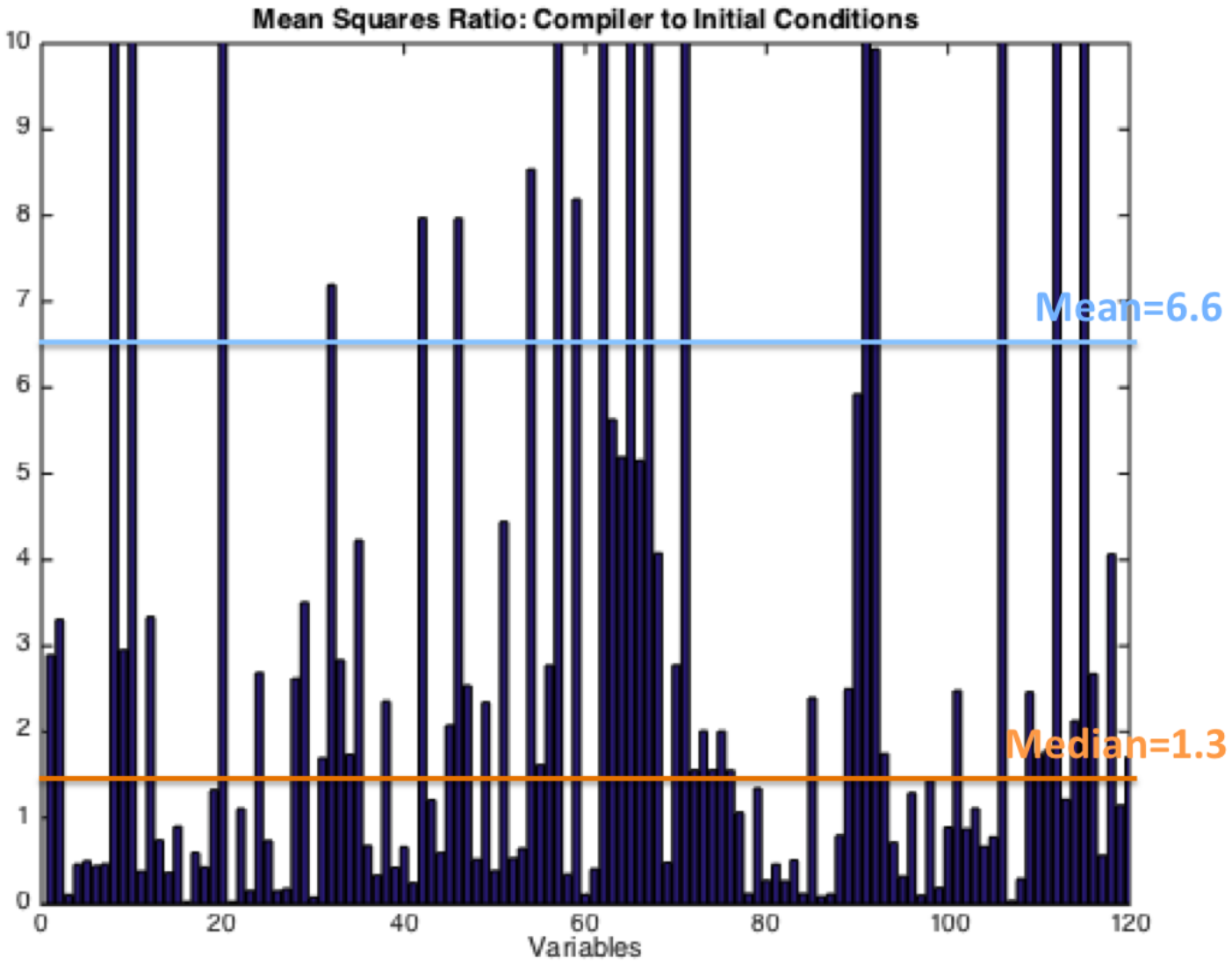
- Is using only initial condition perturbations the "right" approach?
- How well do initial condition perturbations capture "legitimate" differences?
- How do runs from different compilers compare to initial perturbation runs?
- How large of an ensemble do we need to capture legitimate changes?

$\Rightarrow$ We created very large ensembles using a combination of initial conditions and different compilers.

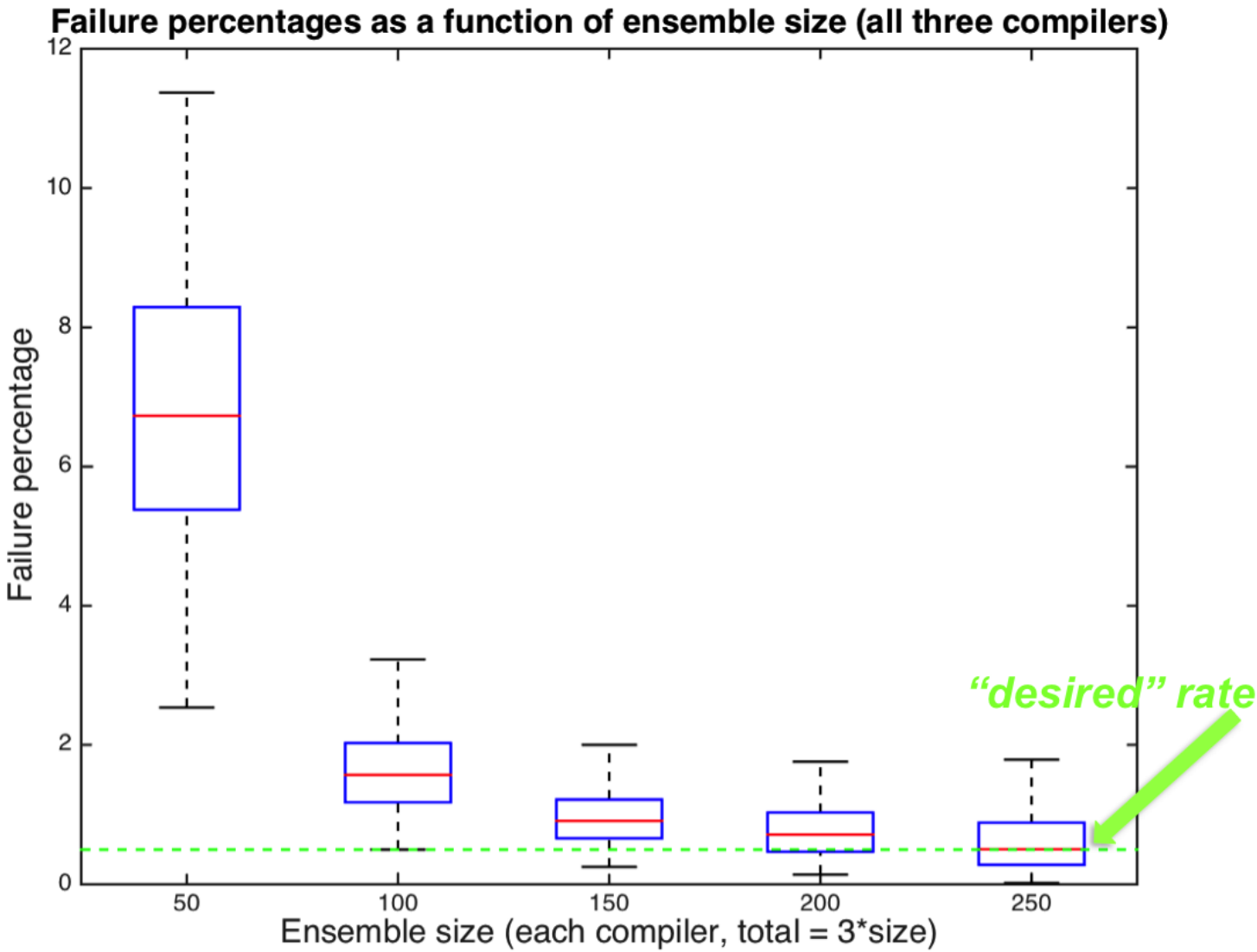
Two-way ANOVA



Two-way ANOVA . . . in more detail



How large of an ensemble?



Purposeful-code-changes experiments

Modification to preq_omega_ps subroutine of prim_si_mod.F90:

Original code:

```
omega_p(i,j,1)=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)*divdp(i,j,1)
```

Modified code:

```
omega_p(i,j,1)=(vgrad_p(i,j,1)-&0.5d0*divdp(i,j,1))/p(i,j,1)
```

Purposeful-code-changes experiments

Modification to preq_omega_ps subroutine of prim_si_mod.F90:

Original code:

```
omega_p(i,j,1)=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)*divdp(i,j,1)
```

Modified code:

```
omega_p(i,j,1)=(vgrad_p(i,j,1)-&0.5d0*divdp(i,j,1))/p(i,j,1)
```

Procedia Computer Science

Volume 80, 2016, Pages 1–12

ICCS 2016. The International Conference on Computational Science

Purposeful code changes . . . indeed have purpose!

Outline

POP-ECT: a version for the ocean component of CESM

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

| EGU.eu | EGU Journals | Contact | Imprint |

Submit a manuscript | Manuscript tracking

doi:10.5194/gmd-2016-3

Discussion papers

Abstract | Discussion | Metrics

About
Editorial board
Articles
Special issues
Highlight articles
Subscribe to alerts
Peer review
For authors
For editors and referees

User ID
Password
New user? | Lost login?

Follow @EGU_GMD

Journal metrics
IF 3.654
IF 5-year

Copernicus Publications
The Innovative Open Access Publisher

Search articles
Search
Title

Download
PDF XML

Short summary
Software quality assurance is critical to detecting errors in large, complex climate simulation codes. We focus on...
Read more

Citation
BibTeX
EndNote

Share

Development and technical paper

28 Jan 2016

Evaluating Statistical Consistency in the Ocean Model Component of the Community Earth System Model (pyCECT v2.0)

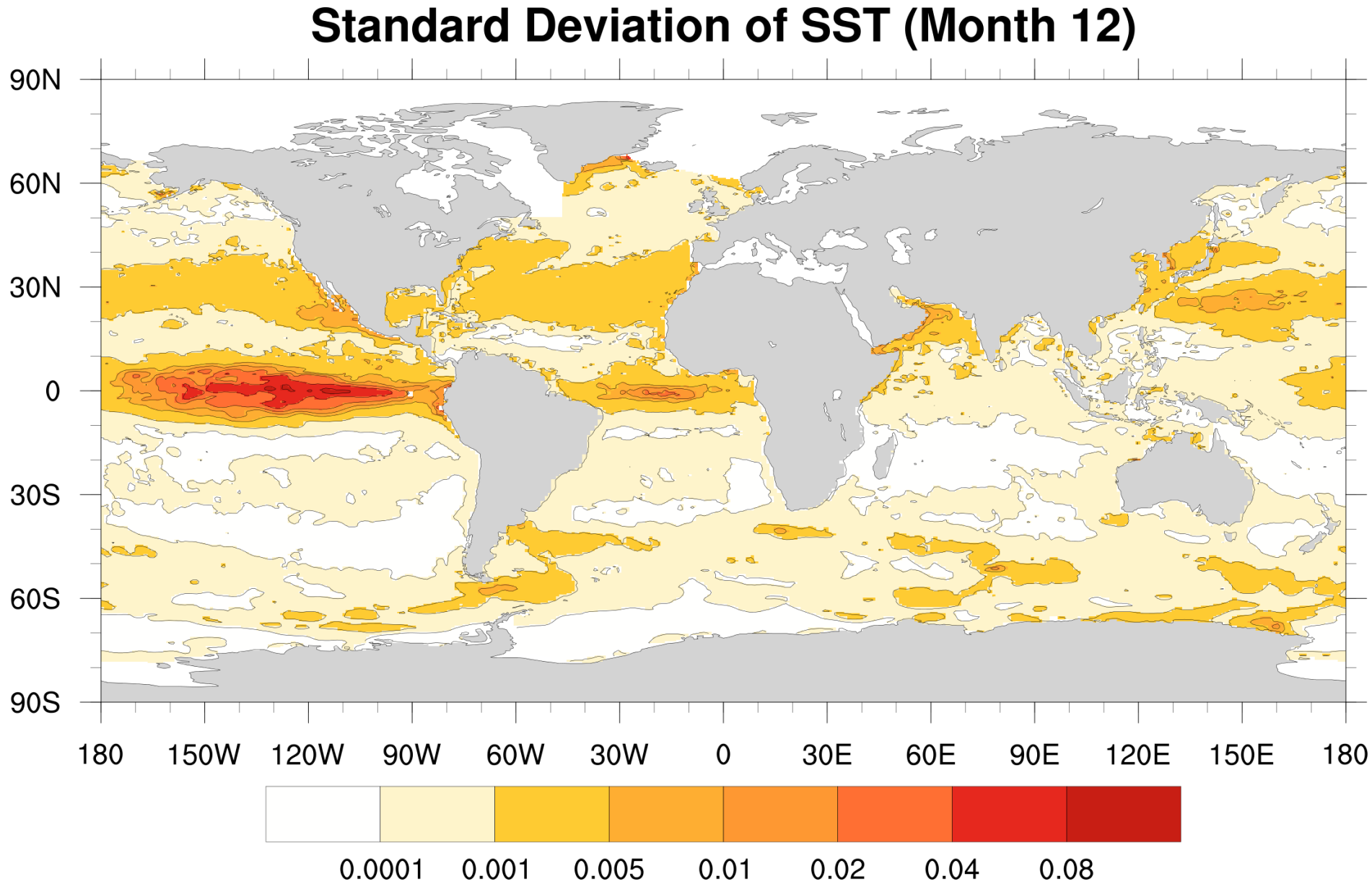
A. H. Baker[1], Y. Hu[2,3], D. M. Hammerling[1], Y. Tseng[1], H. Xu[1], X. Huang[2,3], F. O. Bryan[1], and G. Yang[2,3]
[1]The National Center for Atmospheric Research, Boulder, CO, USA
[2]Center for Earth System Science,Tsinghua University, 100084, China
[3]Joint Center for Global Change Studies, Beijing, 100875, China

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

Received: 07 Jan 2016 – Accepted: 26 Jan 2016 – Published: 28 Jan 2016

Abstract. The Parallel Ocean Program (POP), the ocean model component of the Community Earth System Model (CESM), is widely used in climate research. Most current work in CESM-POP focuses on improving the model's efficiency or accuracy, such as improving numerical methods, advancing parameterization, porting to new architectures, or increasing parallelism. Because ocean dynamics are chaotic in nature, achieving bit-for-bit (BFB) identical results in ocean solutions cannot be guaranteed for even tiny code modifications, and determining whether model changes are admissible (i.e. statistically consistent with the original results) is non-trivial. In recent work, an ensemble-based statistical approach was shown to work well for statistical consistency testing on atmospheric model data. The general idea of the ensemble-based statistical consistency testing is to use a qualitative measurement of the variability of the ensemble of simulations as a metric with which to compare future simulations and make a determination of statistical distinguishability. Because ocean and atmosphere models have differing characteristics in term of dynamics and time-scales, we present a new statistical method to evaluate ocean model simulation data that requires the evaluation of ensemble means and deviations in a spatial manner. In particular, the statistical distribution from an ensemble of CESM-POP simulations is used to determine the standard score of any new

Ensemble spread in the ocean model component



POP-ECT is conceptually very different

- Only a few variables (compared to over 100 for CAM)
- Variability is spatially heterogeneous, too much lost in using global mean (no sensitivity!)

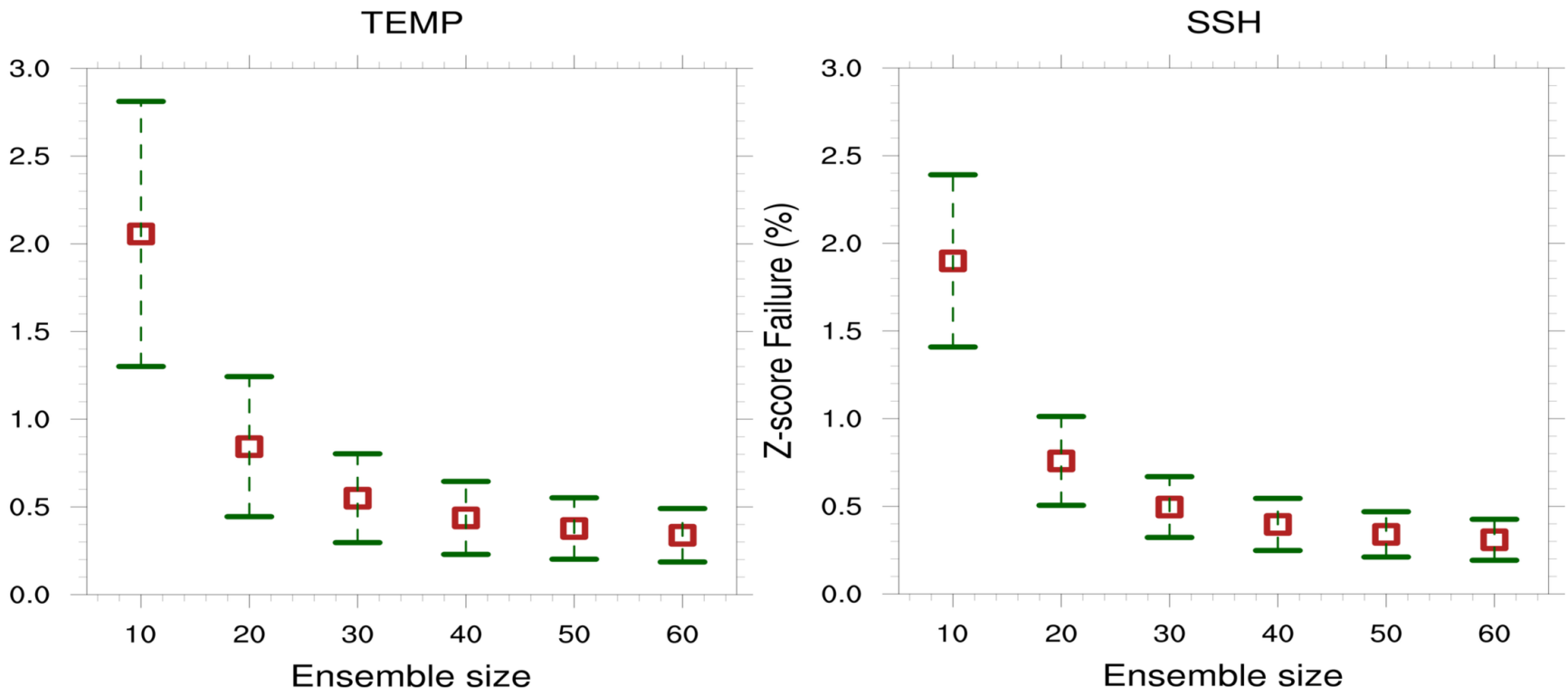
$\Rightarrow$ We evaluate each location as a function of the spatially-varying point-wise ensemble variability and look at their combined exceedances above a threshold.

POP-ECT is conceptually very different

- Only a few variables (compared to over 100 for CAM)
- Variability is spatially heterogeneous, too much lost in using global mean (no sensitivity!)

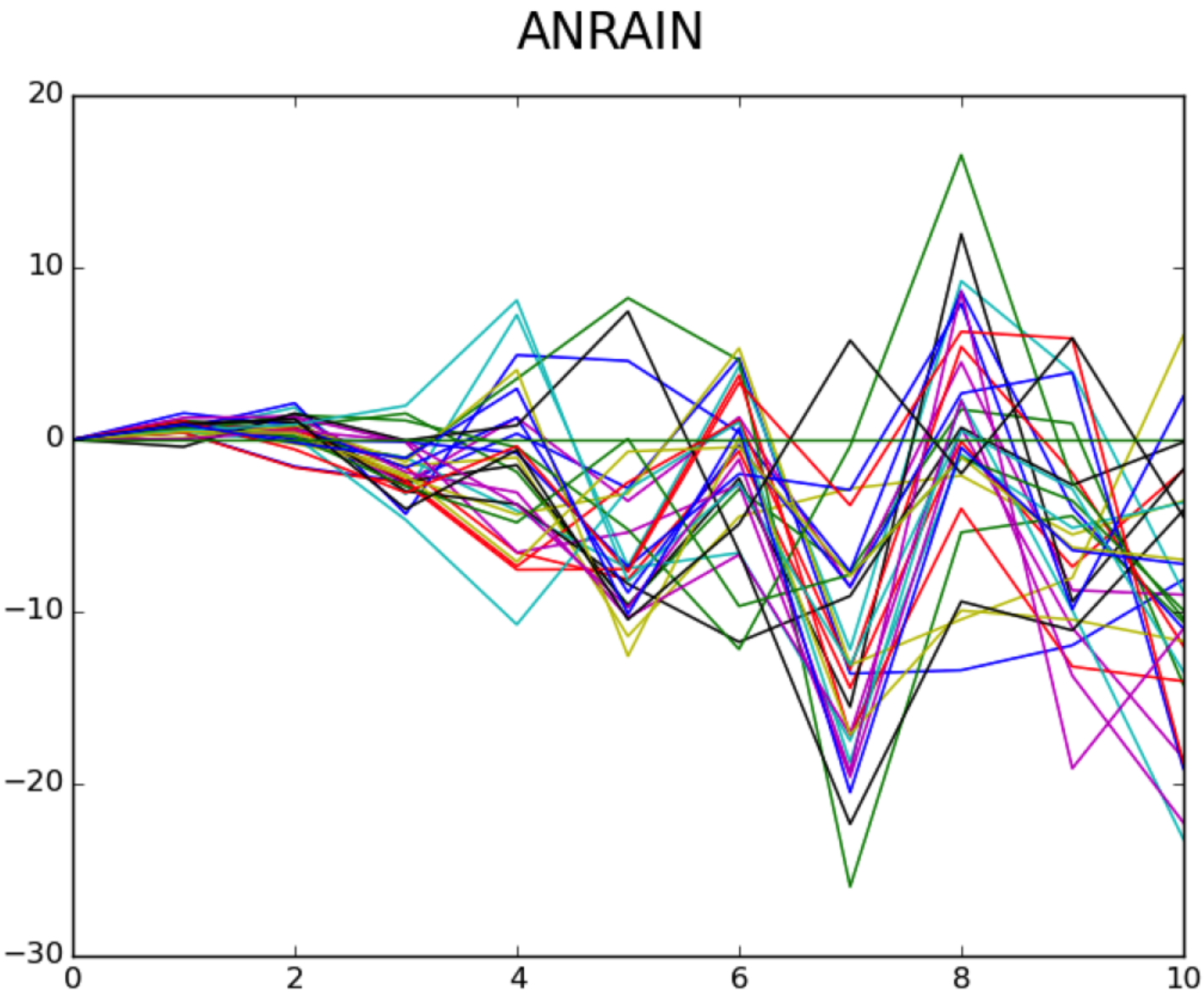
$\Rightarrow$ We evaluate each location as a function of the spatially-varying point-wise ensemble variability and look at their combined exceedances above a threshold.

Empirical evaluation of ensemble size for POP-ECT



Outline

How fast does the ensemble spread?



Ultra-fast Ensemble Consistency Testing

- We only use the first 9 time steps (corresponds to about 5 hours model time)
- Early results very promising: we catch errors and runs that should pass do
- Early results also indicate that we might need a large ensemble size
  - Runs a very cheap
  - And we don't need to store the ensemble long term, just the (very small!) summary file
- Idea is to use Ultra-fast ECT in combination with yearly ECT as a first (necessary, but non-sufficient) test

$\Rightarrow$ Might be ideal for quick check after code changes.

Ultra-fast Ensemble Consistency Testing

- We only use the first 9 time steps (corresponds to about 5 hours model time)
- Early results very promising: we catch errors and runs that should pass do
- Early results also indicate that we might need a large ensemble size
  - Runs a very cheap
  - And we don't need to store the ensemble long term, just the (very small!) summary file
- Idea is to use Ultra-fast ECT in combination with yearly ECT as a first (necessary, but non-sufficient) test

$\Rightarrow$ Might be ideal for quick check after code changes.

Ultra-fast Ensemble Consistency Testing

- We only use the first 9 time steps (corresponds to about 5 hours model time)
- Early results very promising: we catch errors and runs that should pass do
- Early results also indicate that we might need a large ensemble size
  - Runs a very cheap
  - And we don't need to store the ensemble long term, just the (very small!) summary file
- Idea is to use Ultra-fast ECT in combination with yearly ECT as a first (necessary, but non-sufficient) test

  $\Rightarrow$ Might be ideal for quick check after code changes.

Ultra-fast Ensemble Consistency Testing

- We only use the first 9 time steps (corresponds to about 5 hours model time)
- Early results very promising: we catch errors and runs that should pass do
- Early results also indicate that we might need a large ensemble size
  - Runs a very cheap
  - And we don't need to store the ensemble long term, just the (very small!) summary file
- Idea is to use Ultra-fast ECT in combination with yearly ECT as a first (necessary, but non-sufficient) test

$\Rightarrow$ Might be ideal for quick check after code changes.

Ultra-fast Ensemble Consistency Testing

- We only use the first 9 time steps (corresponds to about 5 hours model time)
- Early results very promising: we catch errors and runs that should pass do
- Early results also indicate that we might need a large ensemble size
  - Runs a very cheap
  - And we don't need to store the ensemble long term, just the (very small!) summary file
- Idea is to use Ultra-fast ECT in combination with yearly ECT as a first (necessary, but non-sufficient) test

$\Rightarrow$ Might be ideal for quick check after code changes.

Outline

Summary

- Ensemble Consistency testing is a subjective and user-friendly way for quality assurance
- Estimating a high dimensional covariance matrix well demands a large ensemble
- Ultra-fast testing very promising so far
- Ocean testing demands a conceptually different approach than atmosphere

Future work:

- Fine-grained testing: what is the source of the inconsistency
- More thorough study of ensemble composition and size