Machine Learning Approaches to Ensure Statistical Reproducibility of ESM Simulations

Salil Mahajan, Joe Kennedy, Kate Evans, Min Xu, Matt Norman, Michael Kelleher, Marcia Branstetter, Peter Caldwell, Andy Salinger

ORNL, LLNL, SNL

ORNL is managed by UT-Battelle, LLC for the US Department of Energy
Motivation:

- **E3SM: Software and Algorithms (PI: Andy Salinger, SNL):**
  - Effectively exploit DOE’s leadership class HPC capabilities, improving model trust-worthiness

- **Code Evolution:**
  - Bit-for-bit reproducing changes
    - E.g. Adding a new compset, new output variable
  - Non-b4b changes
    - Different climate (statistics) expected
      - E.g. New parameterizations modules, new tunings
    - Same climate (statistics) expected
      - E.g. code porting, refactoring, GPU kernel, etc.

- **Goal:** Test the null hypothesis that climate simulation is similar for unintended non-b4b changes.
Motivation

- Truncated Floating Point arithmetic:
  - Round-off differences
  - Non-associative:
    - \((-1 + 1) + 2^{-53} \neq -1 + (1 + 2^{-53})\)
  - Optimizations, hybrid architectures

- Climate models:
  - Chaotic, non-linear system

- Round-off differences grow quickly

- Problem: identify systematic bugs in non-BFB reproducible environment.

Lorenz attractor
(Source: en.wikipedia.org/wiki/Chaos_theory)

Root mean squared difference of temperature
for \(~10^6\) grid points from control (Rosinski and Williamson, 1997)

Evolution of Temperature (Courtesy: Matt Norman)
E3SM Testing

- **E3SM Testing Suite (bfb):**
  - * APT (auto promotion test (default length))
  - * CME (compare mct and esmf interfaces (10 days))
  - * ERB (branch/exact restart test)
  - * ERH (hybrid/exact restart test)
  - * ERI (hybrid/branch/exact restart test, default 3+19/10+9/5+4 days)
  - * ERS (exact restart from startup, default 6 days + 5 days)
  - * ERT (exact restart from startup, default 2 month + 1 month (ERS with info dbug = 1))
  - * ICP (cice performance test)
  - * LAR (long term archive test)
  - * NCK (multi-instance validation vs single instance (default length))
  - * NOC (multi-instance validation for single instance ocean (default length))
  - * OCP (open performance test)
  - * P4A (production branch test b40.1850.track1.1deg.006 year 301)
  - * PEA (single pe bfb test (default length))
  - * PEM (pes counts mpi bfb test (seq tests; default length))
  - * PET (openmp bfb test (seq tests; default length))
  - * PFS (performance test setup)
  - * PRS (pes counts hybrid (open-MP/MPI) restart bfb test from startup, default 6 days + 5 days)
  - * SBN (smoke build-namelist test (just run preview_namelist and check_input_data))
  - * SEQ (sequencing bfb test (10 day seq,conc tests))
  - * SMS (smoke startup test (default length))
  - * SSSP (smoke CLM spinup test (only valid for CLM compsets with CLM45 and CN or BGC))

- **Non bit for bit changes:**
  - Convergence test, perturbation growth test and climate reproducibility tests
  - Expert opinion, ad-hoc tests

---

The main thing that distinguishes legacy code from non-legacy code is tests, or rather a lack of tests. —Michael Feathers
Short Independent Simulation Ensemble

\[ T'_j = (1 + x') T_j \]

\( x' \) is uniform random number transformed to range from \((-10^{-14}, 10^{-14})\)
Problem to solve: Multivariate two sample equality of distribution testing for:
  High dimension
  Low sample size
Climate Reproducibility Tests: Ensemble Based Multivariate ML Approach

Accelerate and add rigor to the verification of E3SM for non-BFB changes

- **Approach:**
  - Ensemble vs. ensemble
  - Short (1yr) ensembles

- **Short Ensembles:**
  - Quantify natural variability
  - Computationally efficient (*Mahajan et al. 2017*)

- **Leverage two sample equality of distribution tests from the ML community:**
  - e.g. cross-match test, energy test, kernel test
  - Distribution-free/non-parametric
  - Effective at high dimensions, low sample sizes
  - Used widely in other fields, e.g. genetics, image processing, etc.
Short Independent Simulation Ensembles

- Packing simulations together is economical as compared to a SLR
- Compare a 100 1-yr ensemble vs. a 100-yr long run
  - Poor Weak and Strong Scaling for 100-yr long run – smaller work load and increased MPI communications with increasing core counts
  - 100x greater workload per node for 100 member 1-yr ensemble on the same no. of nodes
  - Significantly reduced relative MPI and PCI-e overheads for ensembles:
    - Better parallel scaling
  - Faster throughput for ensembles:
    - Large core counts
    - Higher priority (capability scale) on leadership class machines (e.g. OLCF, NERSC, etc.)
  - Example (atmosphere spectral element 2 degree model):
    - Long run (100 years): 1536 elements, 96 nodes, 16 elements per node
    - SISE (100 1yr runs): 48 nodes each, 32 elements per node (total nodes: 4800)

- Usage:
  - Solution reproducibility tests
  - Scientific Applications
Short Ensembles: Scientific Utility

- Control Case (1850S)
- Perturbed Case (2000S)

Fast Response

SST (2000S – 1850S)
Precipitation (2000S – 1850S)

Verma et al. 2019
Equality of Distribution Tests

- **Energy Test** (e.g. Szekely and Rizzo, 2004):
  - e-distance metric

\[
e = \frac{nm}{n+m} \left( \frac{2}{nm} \sum_{i=1}^{n} \sum_{k=1}^{m} \|X_i - Y_k\| - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|X_i - X_j\| - \frac{1}{m^2} \sum_{l=1}^{m} \sum_{k=1}^{m} \|Y_l - Y_k\| \right)
\]

where \(X_1, \ldots, X_n\) and \(Y_1, \ldots, Y_m\) are the multivariate vectors of the baseline and perturbed ensembles.

  - Small values of \(e\) indicate same population
  - Derive null distribution by resampling
Equality of Distribution Tests

• Kernel Test (e.g. Gretton et al. 2006):
  – Maximum mean discrepancy (MMD) metric

\[
MMD = \left( \frac{1}{n^2} \sum_{i,j=1}^{n} k(X_i, X_j) - \frac{2}{nm} \sum_{i,j=1}^{n,m} k(X_i, Y_j) + \frac{1}{m^2} \sum_{i,j=1}^{m} k(Y_i, Y_j) \right)^{\frac{1}{2}}
\]

where \( k \) represents the kernel in its class of functions that maximizes \( MMD \)

  – Small values of MMD indicates same population
  – Derive null distribution by resampling
Equality of Distribution Tests

- **Kolmogorov Smirnov (KS) - Testing Framework:**
  - Null Hypothesis ($H_0$): Two ensembles represent the same climate state.
  - Use global annual means of standard model output variables (158 variables).
  - $H_0$: A variable between the two ensembles belong to the same distribution.
  - Test $H_0$ for each variable using a KS test.
  - Test statistic ($t$): No. of variables that reject $H_0$ at a given confidence level (say 95%).
  - $H_0$ rejected if $t > a$, where $a$ is some critical number for a significance level (Type I error rate).
  - $a$ is empirically from an approximate null distribution of $t$ derived using resampling techniques.
Significance Level (Type I Error Rate): Resampling

- Simulations from the two ensembles of size $n$ and $m$ are pooled together.

- Simulations from the pool are then randomly assigned to one of two groups of sizes $n$ and $m$.

- The $t$-statistic is then computed for the random drawing.

- Repeat

- If all possible random drawings are made, the null distribution of $t$ is exact.
  - We conduct 500 drawings - approximate null distribution.
Model Verification Using Ensembles: Known Climate Changing Perturbation

- **Model:** DOE E3SM v1
- **Configuration:** Active atmosphere land, prescribed cyclical F2000 SSTs and sea-ice distribution (FC5)
- **Spatial Resolution:** ~500km at the equator (5 degrees), 30 vertical layers
- **Machine Configuration:** PGI compiler on Titan
- **Ensembles:** Machine-precision level random perturbations to the initial 3-D temperature field
  - 30 member SISE
  - \( T'_j = (1 + x') T_j \), \( x' \) is random number transformed to range from \((-10^{-14}, 10^{-14})\)
- Turn a tuning parameter knob: zm_c0_ocn (control case: 0.007, modified: 0.045)
KS Testing Framework Results

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Ens. Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default c0_ocn</td>
<td>Default model settings</td>
<td>30</td>
</tr>
<tr>
<td>Perturbed c0_ocn</td>
<td>Perturbed model parameter</td>
<td>30</td>
</tr>
</tbody>
</table>

Comparison

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Test Statistic (t)</th>
<th>Critical No.</th>
<th>H0 Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default vs. perturbed c0_ocn</td>
<td>119</td>
<td>13</td>
<td>Reject</td>
</tr>
</tbody>
</table>
Power Analysis (Type II Error rate)

*Type II error rate: Probability of accepting a false null hypothesis*

- Turn a tuning parameter knob incrementally: zm_c0_ocn (0.007 to 0.045)

- Ensembles:
  - 100 members for each case
  - $T'_j = (1+x')T_j$, $x'$ is random number transformed to range from (-10^{-14}, 10^{-14})

- Power Analysis:
  - Randomly pick N=30 (=40, 50, 60) members from the control and perturbed sets
  - Conduct test
  - Repeat (500 times)
Power Analysis: KS Testing Framework

Controlled changes to zm_c0_ocn tuning parameter in Deep Convection

Example of Power Analysis.
Probability of correctly rejecting a false null hypothesis (Power) of the test in detecting changes to a EAM tuning parameter from a control case (zm_c0_ocn = 0.0070) for different short simulation (1yr) ensemble sizes (N).

Mahajan et al. 2019
Power Analysis

Controlled changes to zm_c0_ocn (= 0.0070, default) tuning parameter in Deep Convection

Energy Test

Kernel Test

KS Testing Framework

Figure 1: Power Analysis of the three tests for the zm_c0_ocn experiment set (ZM_SET): (a) energy test, (b) kernel test and (c) KS testing framework. Empirically (random sampling) estimated statistical Power - probability of rejecting a false null hypothesis - of the test at the 95% confidence level for different ensemble sizes (N = 30, 40, 50, 60). The null hypothesis that two simulation ensembles are statistically identical is tested for each perturbed zm_c0_ocn case against the zm_c0_ocn = 0.007 case. Figure 1c is reproduced from Mahajan et al. [7].
Power Analysis

Controlled changes to dcs (= 400.0, default) tuning parameter in Cloud Microphysics

Energy Test

Kernel Test

KS Testing Framework

Figure 2: Power Analysis of the three tests for the dcs experiment set (MICRO_SET): (a) energy test, (b) kernel test and (c) KS testing framework. Empirically (random sampling) estimated statistical Power (\(\frac{1}{2}\)) - probability of rejecting a false null hypothesis - of the test at the 95% confidence level for different ensemble sizes (N = 30, 40, 50, 60). The null hypothesis that two simulation ensembles are statistically identical is tested for each perturbed dcs case against the dcs = 400.0 case.

Mahajan et al. 2019
Power Analysis: Atmosphere tests

- Expand on Power Analysis:
  - More tuning parameters
    - ice_sed_ai
    - sol_factb_interstitial
    - sol_factic_interstitial
    - cldfrc_dp1
    - zm_conv_lnd
    - dcs
    - zm_conv_ocn
    - zm_conv_dmpdz

- **KS testing framework** most powerful:
  - detects changes of smaller magnitudes confidently
  - compared to **Kernel** and **Energy** test.

Example of Power Analysis. *Probability of correctly rejecting a false null hypothesis (Power) of the test in detecting changes to a EAM tuning parameter from a control case (dcs = 400) for different short simulation (1yr) ensemble sizes (N).*

Mahajan et al. 2019
Test Case: Cori vs. Edison

Evaluate if E3SMv1 DECK simulations on Edison can be reproduced on Cori

- Conducted short simulation (1yr) ensembles on both Edison and Cori:
  - F1850C5-CMIP6 compset
  - ne4 (100 ensemble members)
  - ne30 (30 ensemble members)


- Implications: Cori can be confidently used for remaining DECK simulations
Reproducibility Tests (EAM) on Master

- **Nightly** tests run on Cori (E3SM custom tests)
  - Time step convergence test
  - Perturbation growth test
  - KS testing framework

- On CDASH under E3SM_CustomsTests
  - All runs archived:
  - Large ne4 1yr F1850C5 ensemble available (>1000)
EVV:

- Extended Verification and Validation for Earth System Models (EVV):
  - Python based toolkit:
    - Runs control and perturbed ensembles
    - Post-processes model output
    - Conducts tests
    - Publishes results and auxiliary plots, tables
MPAS-O Reproducibility tests: Ensembles

• Generate ensembles:
  1. Low Res NYF Ocean run:
     • 240 km resolution (7153 cells)
     • Run to quasi-equilibrium – pick base initial condition
     • Perturb initial condition to machine order precision:
       – Add perturbations to 3D temperature field initial condition
       – Save perturbed initial condition files
     • Use create_clone to generate ensembles:
       – each run reading a different perturbed initial condition file

  2. Pertlim capability for MPAS-O (near future):
     • Replicate capability within EAM to MPAS-O
     • Automatically perturb initial conditions
     • Generate ensembles by tweaking a namelist parameter.
     • Replicate multi-instance capability within EAM to MPAS.

Machine Precision Perturbations to \( T_{j} \) at each grid point, \( j \)

\[
x' \] is a uniform random number transformed to range from \((-10^{-14}, 10^{-14})\)
MPAS-O Reproducibility tests: Approach

**Larger Null Hypothesis:** Control and perturbed ensembles belong to the same population

- Generate control and perturbed ensembles at QU240 resolution

- Evaluate 5 prognostic variables (Baker et al. 2016)
  - SSH, T, U, V, Salinity
  - Annual average of year 2.

- Ocean variability is spatially very heterogenous (as compared to the atmosphere):
  - Evaluate at each grid point.

- Conduct fine-grained null hypothesis tests at each grid point:
  - Two sample KS test: Popular non-parametric test
  - Cucconi test: Better power, rank based non-parametric test.

Growth of machine precision differences in oQU240 MPAS-O and ensemble spread: L1 Norm (sum of absolute difference at each grid point, log-scale) of SST of each of the 100 ensemble members with round off differences in initial conditions compared to a reference run for the control (kappa = 1800, red lines) and modified (kappa = 600, blue lines) ensembles.
Cucconi Test

- **Test Statistic:**

\[
CUC = \frac{U^2 + V^2 - 2\rho UV}{2(1 - \rho^2)}.
\]

- **U**: based on squared sum of ranks of samples in Ensemble A in the two sample pool of Ensembles A and B.
- **V**: based on squared sum of contrary-ranks of samples in Ensemble A in the pool.
- **\(\rho\)**: Correlation coefficient between U and V.

- Larger test-statistic indicates that Ensemble A and B come from different populations.
- Popular in other fields like hydrology, quality control, etc. (e.g. Mukherjee and Marozzi et al. 2014)
MPAS-O Reproducibility Tests: Approach

Correct for simultaneous multiple null hypothesis tests (M grid points)

False Discovery Rate (FDR) approach (Wilks et al. 2006, Ventura et al. 2004):

- For single test, null hypothesis is rejected if:
  - Test statistic p-value (p) is less than a critical value, \( \alpha \) (say 0.05): \( p \leq \alpha \)
  - For \( M \) tests, \( \alpha M \) would be rejected for true null hypotheses just by chance

- For multiple tests, FDR constrains critical value (\( \alpha_{FDR} \)) for local hypothesis tests (\( H_0 \)):

  \[
  \alpha_{FDR} = \max_{j=1,2,...,M} \left\{ p_j : p_j \leq \alpha(j/M) \right\}
  \]

  \( p_j \) are sorted p-values of \( M \) tests

- Global Null Hypothesis Test (\( G_0 \)): Reject if \( p_j \leq \alpha_{FDR} \) at any grid point.
- Robust for correlated tests – e.g. spatial correlations (Wilks et a. 2006, Renard et al. 2008).
- Used in testing field significance
FDR Approach: Illustration

\[ \alpha_{FDR} = \max_{j=1,2,\ldots,M} \{p_j : p_j \leq \alpha(j/M)\} \]

Fig. 2. Illustration of the traditional FPR and FDR procedures on a stylized example, with \( q = \alpha = 20\% \). The ordered \( p \)-values, \( p_{(i)} \), are plotted against \( i/n, i = 1, \ldots, n \), and are circled and crossed to indicate that they are rejected by the FPR and FDR procedures, respectively.

Ventura et al. 2004
MPAS-O Reproducibility Tests

Evaluate False Positive Rate:

Bootstrap with Control Ensemble (150 ensemble members):

- Randomly draw two samples with N=M=30 members
- Conduct KS test and Cucconi test for alpha = 0.05
- Repeat 500 times at alpha = 0.05

KS test:
95th percentile of the no. of cells rejecting the local null hypothesis (FDR) = 0
95th percentile of the no. of cells rejecting the local null hypothesis = 426

Cucconi test:
95th percentile of the no. of cells rejecting the local null hypothesis = 15
95th percentile of the no. of cells rejecting the local null hypothesis = 643
MPAS-O Reproducibility Tests: Results

Known Climate Changing Case: GM Kappa = 600 (Default = 1800)
30 member ensembles for test and control case

Both tests reject the null hypothesis that the two ensembles belong to the same population at the 0.05 significance level.
MPAS-O Reproducibility Tests: Power Analysis

Type II error rate: Probability of accepting a false null hypothesis

- Turn a tuning parameter knob incrementally:
  - Gent and McWilliams kappa (600 to 1800):

- Ensembles:
  - 100 members for each case
  - $T'_j = (1+x')T_j$, $x'$ is random number transformed to range from $(-10^{-14}, 10^{-14})$

- Power Analysis:
  - Randomly pick N=30 (=40, 50, 60) members from the control and perturbed sets
  - Conduct test
  - Repeat (500 times)
MPAS-O Reproducibility Tests: Power Analysis

Controlled changes to **GM kappa** tuning parameter in MPAS-O

**Power Analysis.** *Probability of correctly rejecting a false null hypothesis (Power)* of the test in detecting changes to a MPAS-O tuning parameter from a control case (**GM kappa = 1800**) for different ensemble sizes (**N**).
Summary:

- Use short ensembles for model verification as E3SM adapts for Exascale
- Developed a multivariate testing framework for climate reproducibility after perturbation growth:
  - EVV
- Power Analysis of tests to evaluate their detection limits
- Test Cases:
  - Known climate changing perturbations: tuning parameter changes
  - Compiler optimization choices, reproducibility of frozen model after months of software updates
  - Machine port from NERSC’s Edison to Cori of E3SMv1 atmosphere model
- Expanding to include reproducibility testing to MPAS-O
  - Generated control and perturbed GMPAS-NYF ensembles using create_clone
  - KS Test and Cucconi tests with false discovery rates
  - Power Analysis with GM kappa tuning parameter
Next Steps and Challenges

• Future work for MPAS-O tests:
  – Conduct ensembles trajectories from a better quasi-equilibrium initial state
  – Power analysis with other controlled changes
  – Evaluate applicability of low-resolution results at high-resolution
  – Explore other multivariate tests
  – Apply to prior known non-b4b changes and live non-b4b changes

• Integrating tests into EVV/CIME.

• Develop ensemble-based tests for individual software kernels: RRTMGP, MG2, CLUBB, MAM4, etc. (in a SCM framework?)

• Investigate applicability to other model components.

Hack and Pedretti (2000)
Thanks!

• **Acknowledgements:**
  – DOE E3SM Project and CMDV-SM Project
  – Oak Ridge Leadership Computing Facility (OLCF)
  – NERSC
Test for Extremes

- Distribution tests perform **poorly** on distribution with different **tails**
  - Known for univariate tests, unexplored for multivariate tests.
- Use **Generalized Extreme Value (GEV)** theory (e.g. Mahajan et al. 2015, Evans et al. 2014).
  - max./min. of a process belong to GEV distribution.
  - Analogous to **central limit theorem**
  - GEV parameters normally distributed asymptotically

\[
G(z) = \exp \left\{ - \left[ 1 + \xi \left( \frac{z - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}
\]

\[
z : 1 + \xi (z - \mu)/\sigma > 0
\]

where \(\mu\), \(\sigma\) and \(\xi\) represent the location, scale and shape parameter respectively.

For additional information, contact:
Matthew Norman
Computational Climate Scientist
National Center for Computational Sciences
Oak Ridge National Laboratory
normanmr@ornl.gov
climatemodeling.science.energy.gov/acme
Climate Extremes Test

- Null Hypothesis ($G_0$): Simulation of extremes of a variable between two SISE is statistically indistinguishable.

- Annual maxima for each grid point are fit to a GEV distribution.

- $G_0$: Extremes at each grid point are statistically indistinguishable

- Test statistic ($g$): No. of grid points that reject $G_0$

- $G_0$ rejected if $t > b$, where $b$ is some critical number, obtained using resampling techniques.
Climate Extremes

a. Surface Temperature Extremes: Default

b. Default – O1

c. Precipitation Extremes: Default

d. Default – O1
## Climate Extremes

### Table 1: Comparison of SISE Simulations

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Variable</th>
<th>Test statistic ((g))</th>
<th>Critical value ((\beta))</th>
<th>(G_0) Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>SISE-DEFAULT vs. SISE-O1</td>
<td>Precipitation Rate</td>
<td>5.1%</td>
<td>6.5%</td>
<td>Accept (G_0)</td>
</tr>
<tr>
<td></td>
<td>Surface Temperature</td>
<td>5.0%</td>
<td>9.6%</td>
<td>Accept (G_0)</td>
</tr>
<tr>
<td>SISE-DEFAULT vs. SISE-FAST</td>
<td>Precipitation Rate</td>
<td>4.7%</td>
<td>6.3%</td>
<td>Accept (G_0)</td>
</tr>
<tr>
<td></td>
<td>Surface Temperature</td>
<td>3.6%</td>
<td>9.6%</td>
<td>Accept (G_0)</td>
</tr>
<tr>
<td>SISE-O1 vs. SISE-FAST</td>
<td>Precipitation Rate</td>
<td>5.2%</td>
<td>6.5%</td>
<td>Accept (G_0)</td>
</tr>
<tr>
<td></td>
<td>Surface Temperature</td>
<td>10.3%</td>
<td>9.8%</td>
<td>Reject (G_0)</td>
</tr>
</tbody>
</table>

- All SISE simulations are identical to each other in terms of their simulation of climate extremes.

- The result is in contrast to the result of the KS-testing framework.

- It suggests that either optimization choices do not effect climate extremes, or climate extremes are not a good metric to evaluate answer changes that

---

Climate variability. Normalized temporal variance spectrum (red: smoothed with a moving average window of 11) of monthly global-average surface temperature after the seasonal cycle is removed, for (a) the SLR long simulation of 80 years and (b) SISE-DEFAULT one year simulation ensemble of 80 years. The SLR simulation, broken into an ensemble of 1-yr segments, is clearly distinct from the SISE-default ensemble set. Individual simulations in the SISE become independent of each other in a few days (Fig. 2). Aggressive compiler optimizations can significantly change model climate statistics. Although, unforced low-frequency atmospheric intrinsic variability implies that SISE, initialized with atmospheric variability comparable to that of the actual atmosphere, should be able to simulate a climate that looks like the real one. The SLR simulation, broken into an ensemble of 1-yr segments, is clearly distinct from the SISE-default ensemble set. Individual simulations in the SISE become independent of each other in a few days (Fig. 2).
Single Long Run (SLR) vs. SISE

• **SLR** is clearly distinct from the **SISE-DEFAULT**

---

**KS Testing Framework Results**

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Test Statistic ($t$)</th>
<th>Critical Value ($\alpha$)</th>
<th>$H_0$ Test Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLR vs. SISE-DEFAULT</td>
<td>80 (50.6 %)</td>
<td>15</td>
<td>Reject $H_0$</td>
</tr>
<tr>
<td>SLR vs. SISE-LND-INIT</td>
<td>74 (48 %)</td>
<td>13</td>
<td>Reject $H_0$</td>
</tr>
</tbody>
</table>
SLR vs. SISE

- Atmospheric models show that free atmospheric-only internal variability can include variability on longer time-scales (e.g. James and James, 1989, Lorenz, 1990, Held, 1993, Marshall and Molteni, 1993).

---

a. 
CTRL Experiment: Surface Temp. Spectrum

b. 
SISE-DEFAULT Experiment: Surface Temp. Spectrum
Atmospheric Low-frequency Variability

James and James, Nature, 1989
Multivariate Cross-Match Test

- $n$ 1-yr control runs (~C)
- $m$ 1-yr modified runs (~M)
- Coarse grained: global annual means
- **Multivariate** vector for each run (size ~130)
- Pool vectors, $N = n + m$
- Pair vectors based on min. Mahalanobis distance
- $H_0$: $C = M$
- Test-statistic ($T$):
  - $T$: number of pairs with one control run

Illustration of cross matching for a bivariate case with $n = m = 10$. (Ruth, 2014)
Cross-Match Test

- Null distribution of T-statistic:

\[ P(T = a_1) = \frac{2^{a_1} (N/2)!}{\binom{N}{n} \binom{n-a_1}{2}! a_1! \binom{m-a_1}{2}!} \]

- i.e. when both samples belong to the same population

- where \( a_1 \) is the no. of pairs with one control and one perturbed vector

- Based on simple combinatorial arguments, thus exact

- Analogous to the probability of drawing one red and one green ball
Single Long Runs: Scalability

- To enhance throughput, use more cores:
  - 5 simulated years per day (required)

- But, scaling (weak or strong) is not perfect:
  - Less work per core with large core counts
  - Increase in MPI communications
  - Smaller MPI messages
  - Large MPI latency

- MPI cost: 90%

Courtesy: Mark Taylor, AMWG meeting
Climate State Approach

- **Several years of a control run**
  - scientifically validated on a trusted machine
- **Several years of the perturbed run**
- **Expert opinion** from a subjective evaluation of plots, tables, etc.
- **Expensive, slow and subjective, no quantitative standardized metric or cost function analysis.**
- **Need for tests** for the multivariate problem of climate model verification.
Test Case: Optimization Choices

- **Model:** DOE E3SM v0.4
- **Configuration:** F1850C5
- **Spatial Resolution:** 208km at the equator (2 degrees), 30 vertical layers
- **Machine Configuration:** PGI compiler on Titan

### KS Testing Framework Results

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Test Statistic ($t$)</th>
<th>Critical Value ($\alpha$)</th>
<th>$H_0$ Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>SISE-DEFAULT vs. SISE-O1</td>
<td>1 (0.6%)</td>
<td>17</td>
<td>Accept $H_0$</td>
</tr>
<tr>
<td>SISE-DEFAULT vs. SISE-FAST</td>
<td>24 (15.2%)</td>
<td>14</td>
<td>Reject $H_0$</td>
</tr>
<tr>
<td>SISE-O1 vs. SISE-FAST</td>
<td>23 (14.6%)</td>
<td>16</td>
<td>Reject $H_0$</td>
</tr>
</tbody>
</table>

Aggressive compiler choices (SISE-FAST) with the PGI compiler on Titan can result in climate-changing simulations.
Test Case: Model Verification Using Ensembles: Frozen model configuration v0 vs. v1

- **Configuration**: F1850C5 compset (frozen after v0 bug-fixes, v0.4)
- **Spatial Resolution**: 208km at the equator (2 degrees), 30 vertical layers

- **Goal**: Evaluate if efforts towards exascale computing impact climate reproducibility:
  - New scientific features, code refactoring
  - CIME (Common Infrastructure for Modeling the Earth System) update
  - Compiler and Software library updates

<table>
<thead>
<tr>
<th>Name</th>
<th>Ens. Size</th>
<th>CIME</th>
<th>PGI</th>
<th>p-netcdf</th>
</tr>
</thead>
<tbody>
<tr>
<td>v0.4-2015</td>
<td>30</td>
<td>4.0</td>
<td>15.3</td>
<td>1.5.0</td>
</tr>
<tr>
<td>master</td>
<td>30</td>
<td>5.0</td>
<td>17.5</td>
<td>1.7.0</td>
</tr>
<tr>
<td>v0.4</td>
<td>27</td>
<td>4.0</td>
<td>17.5</td>
<td>1.7.0</td>
</tr>
</tbody>
</table>
Frozen model configuration v0 vs. v1

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Test Statistic (t)</th>
<th>Critical no. (α)</th>
<th>H0 Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>v0.4-2015 vs. master</td>
<td>6 (3.6%)</td>
<td>13</td>
<td>Accept H0</td>
</tr>
<tr>
<td>v0.4 vs. master</td>
<td>8 (4.2%)</td>
<td>13</td>
<td>Accept H0</td>
</tr>
<tr>
<td>v0.4-2015 vs. v0.4</td>
<td>5 (3%)</td>
<td>13</td>
<td>Accept H0</td>
</tr>
</tbody>
</table>

Software infrastructure updates are not climate changing. Frozen model configuration reproducible!