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

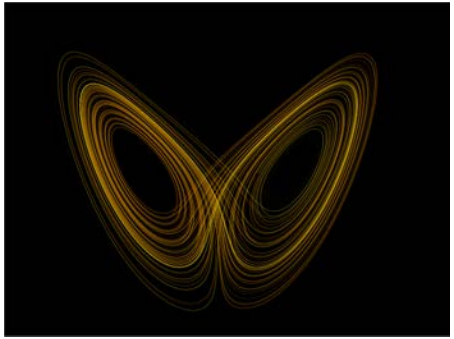
Motivation:

- 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.

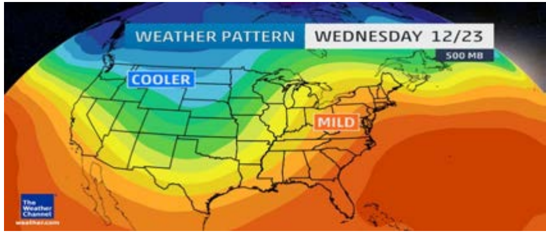
- Goal: Test the null hypothesis that climate simulation is similar for unintended non-b4b changes.

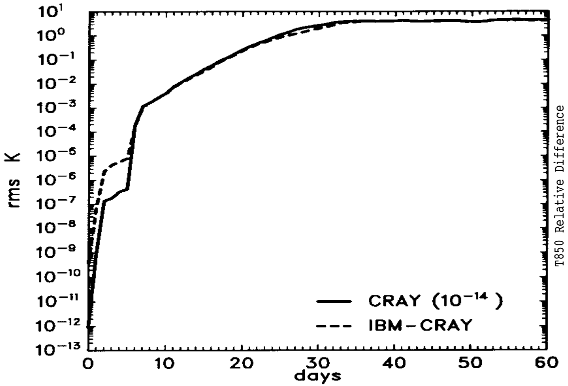
Motivation

- 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.

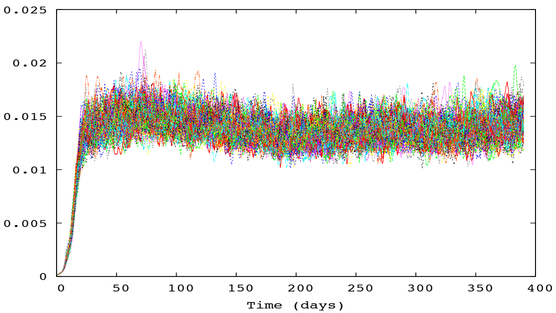


Lorenz attractor
(*Source:en.wikipedia.org/wiki/Chaos_theory*)





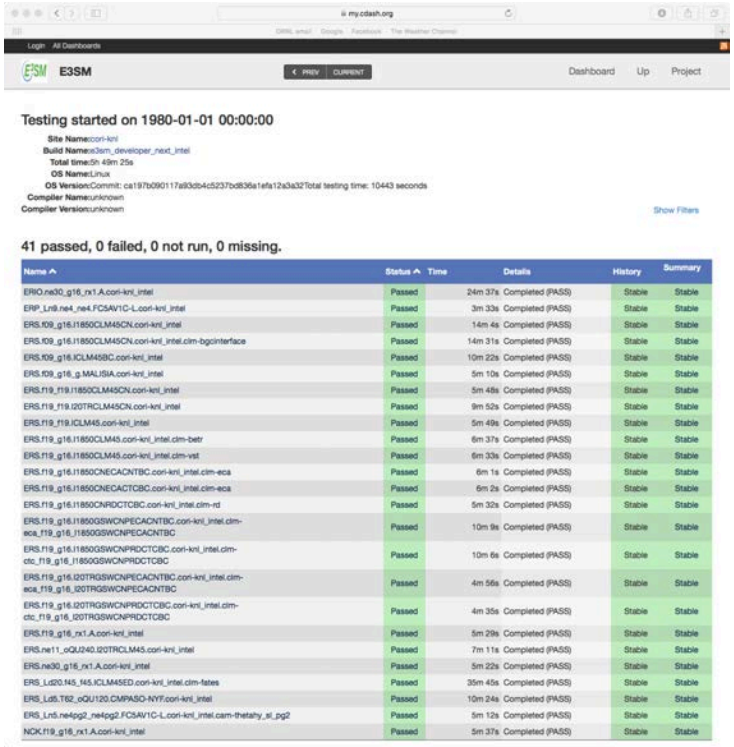
Root mean squared difference of temperature for ~10^6 grid points from control (Rosinski and Williamson, 1997)



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 (pop 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))
    * SSP (smoke CLM spinup test (only valid for CLM compsets with CLM45 and CN or BGC))


E3SM   PREV | CURRENT   Dashboard   Up   Project

Testing started on 1980-01-01 00:00:00

Site Name:cori-knl
Build Name:e3sm_developer_next_intel
Total time:5h 49m 25s
OS Name:Linux
OS Version:Commit: ca197b090117a93db4c5237bd836a1efa12a3a32Total testing time: 10443 seconds
Compiler Name:unknown
Compiler Version:unknown

Show Filters

41 passed, 0 failed, 0 not run, 0 missing.

Name	Status	Time	Details	History	Summary
ERIO.ne30_g16_rx1.A.cori-knl_intel	Passed	24m 37s	Completed (PASS)	Stable	Stable
ERP_Ln9.ne4_ne4.FC5AV1C-L.cori-knl_intel	Passed	3m 33s	Completed (PASS)	Stable	Stable
ERS.f09_g16.I1850CLM45CN.cori-knl_intel	Passed	14m 4s	Completed (PASS)	Stable	Stable
ERS.f09_g16.I1850CLM45CN.cori-knl_intel.clm-bgcinterface	Passed	14m 31s	Completed (PASS)	Stable	Stable
ERS.f09_g16.ICLM45BC.cori-knl_intel	Passed	10m 22s	Completed (PASS)	Stable	Stable
ERS.f09_g16_g.MALISIA.cori-knl_intel	Passed	5m 10s	Completed (PASS)	Stable	Stable
ERS.f19_f19.I1850CLM45CN.cori-knl_intel	Passed	5m 48s	Completed (PASS)	Stable	Stable
ERS.f19_f19.I20TRCLM45CN.cori-knl_intel	Passed	9m 52s	Completed (PASS)	Stable	Stable
ERS.f19_f19.ICLM45.cori-knl_intel	Passed	5m 49s	Completed (PASS)	Stable	Stable
ERS.f19_g16.I1850CLM45.cori-knl_intel.clm-betr	Passed	6m 37s	Completed (PASS)	Stable	Stable
ERS.f19_g16.I1850CLM45.cori-knl_intel.clm-vst	Passed	6m 33s	Completed (PASS)	Stable	Stable
ERS.f19_g16.I1850CNECACNTBC.cori-knl_intel.clm-eca	Passed	6m 1s	Completed (PASS)	Stable	Stable
ERS.f19_g16.I1850CNECACTCBC.cori-knl_intel.clm-eca	Passed	6m 2s	Completed (PASS)	Stable	Stable
ERS.f19_g16.I1850CNRDCTCBC.cori-knl_intel.clm-rd	Passed	5m 32s	Completed (PASS)	Stable	Stable
ERS.f19_g16.I1850GSWCNPECACNTBC.cori-knl_intel.clm-eca_f19_g16_I1850GSWCNPECACNTBC	Passed	10m 9s	Completed (PASS)	Stable	Stable
ERS.f19_g16.I1850GSWCNPRDCTCBC.cori-knl_intel.clm-ctc_f19_g16_I1850GSWCNPRDCTCBC	Passed	10m 6s	Completed (PASS)	Stable	Stable
ERS.f19_g16.I20TRGSWCNPECACNTBC.cori-knl_intel.clm-eca_f19_g16_I20TRGSWCNPECACNTBC	Passed	4m 56s	Completed (PASS)	Stable	Stable
ERS.f19_g16.I20TRGSWCNPRDCTCBC.cori-knl_intel.clm-ctc_f19_g16_I20TRGSWCNPRDCTCBC	Passed	4m 35s	Completed (PASS)	Stable	Stable
ERS.f19_g16_rx1.A.cori-knl_intel	Passed	5m 29s	Completed (PASS)	Stable	Stable
ERS.ne11_oQU240.I20TRCLM45.cori-knl_intel	Passed	7m 11s	Completed (PASS)	Stable	Stable
ERS.ne30_g16_rx1.A.cori-knl_intel	Passed	5m 22s	Completed (PASS)	Stable	Stable
ERS_Ld20.f45_f45.ICLM45ED.cori-knl_intel.clm-fates	Passed	35m 45s	Completed (PASS)	Stable	Stable
ERS_Ld5.T62_oQU120.CMPASO-NYF.cori-knl_intel	Passed	10m 24s	Completed (PASS)	Stable	Stable
ERS_Ln5.ne4pg2_ne4pg2.FC5AV1C-L.cori-knl_intel.cam-thetahy_sl_pg2	Passed	5m 12s	Completed (PASS)	Stable	Stable
NCK.f19_g16_rx1.A.cori-knl_intel	Passed	5m 37s	Completed (PASS)	Stable	Stable

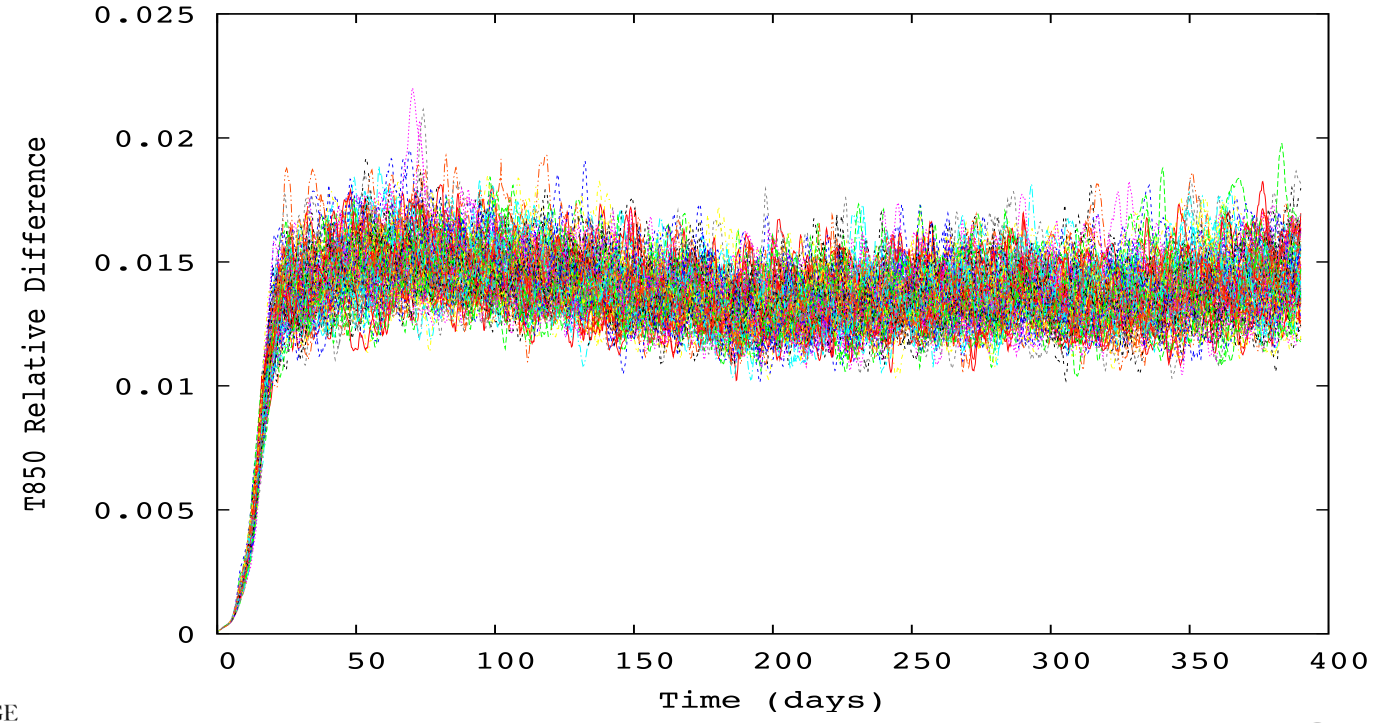
- Non bit for bit changes:
  - Convergence test, perturbation growth test and climate reproducibility tests
  - Expert opinion, ad-hoc tests

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

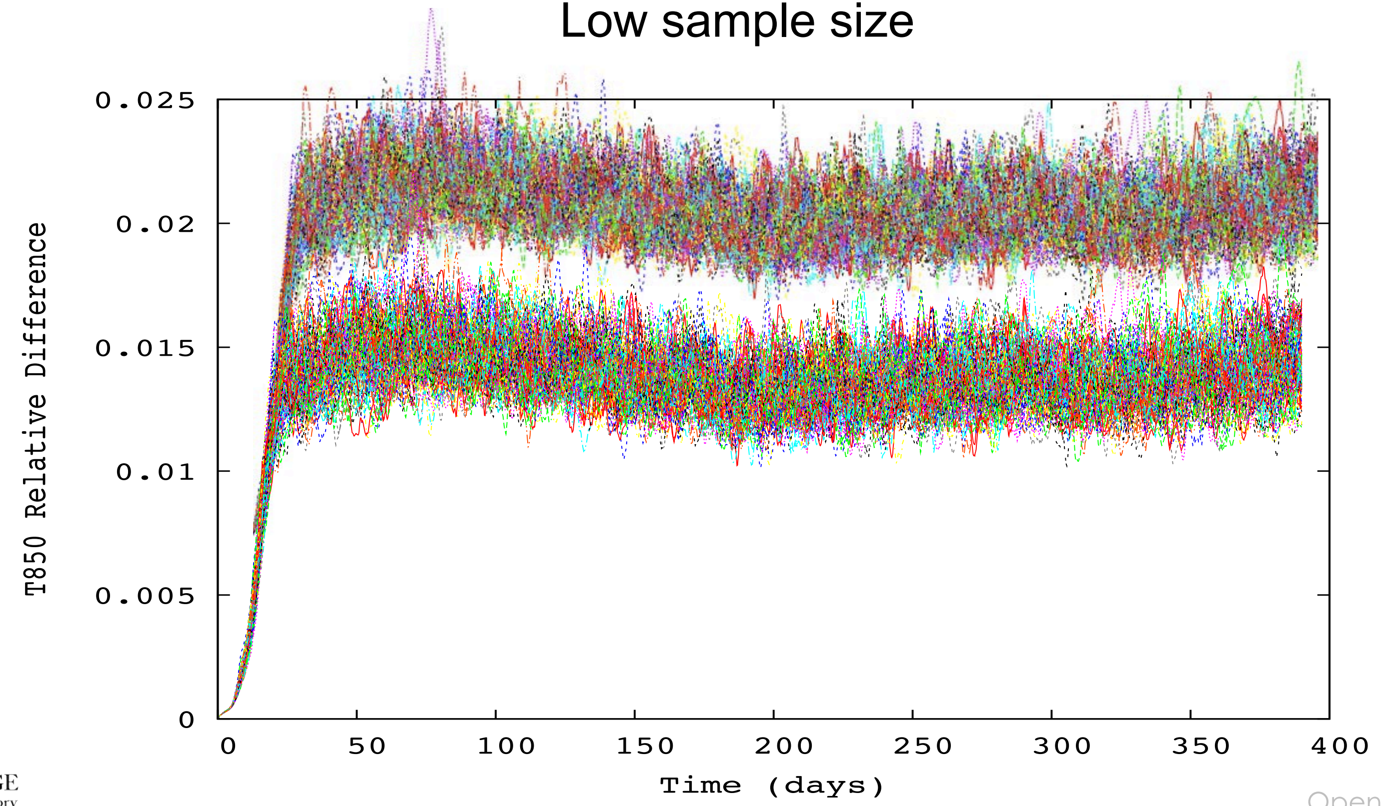
$T'_j = (1+x')T_j$

x' is uniform random number transformed to range from $(-10^{-14}, 10^{-14})$



Short Independent Simulation Ensembles

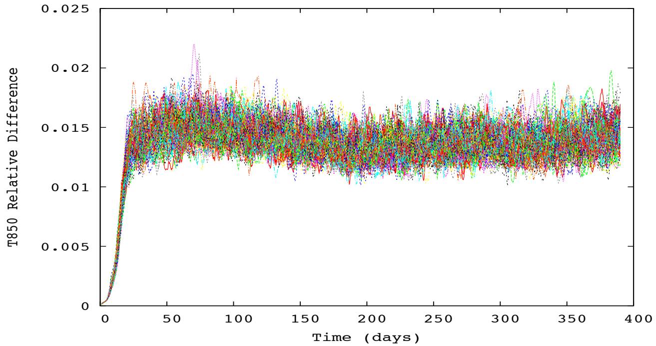
Problem to solve: Multivariate two sample equality of distribution testing for:
High dimension
Low sample size



Climate Reproducibility Tests:
Ensemble Based Multivariate ML Approach

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

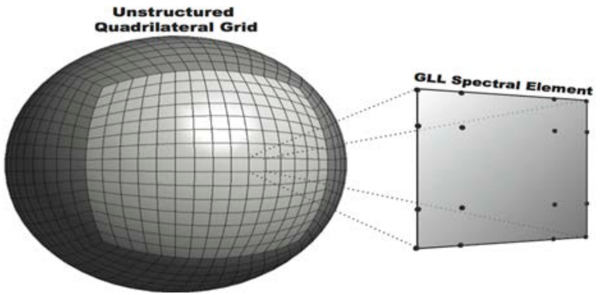
- Approach:
  - Ensemble vs. ensemble
  - Short (1yr) ensembles
- Short Ensembles:
  - Quantify natural variability
  - Computationally efficient (*Mahajan et al. 2017)*



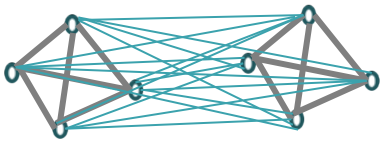
- Leverage two sample equality of distribution tests from the ML community:
  - e.g. cross-match test, energy test, kernel test
  - Distribution-free/non-parametric
  - Effective at high dimensions, low sample sizes
  - Used widely in other fields, e.g. genetics, image processing, etc.

Short Independent Simulation Ensembles

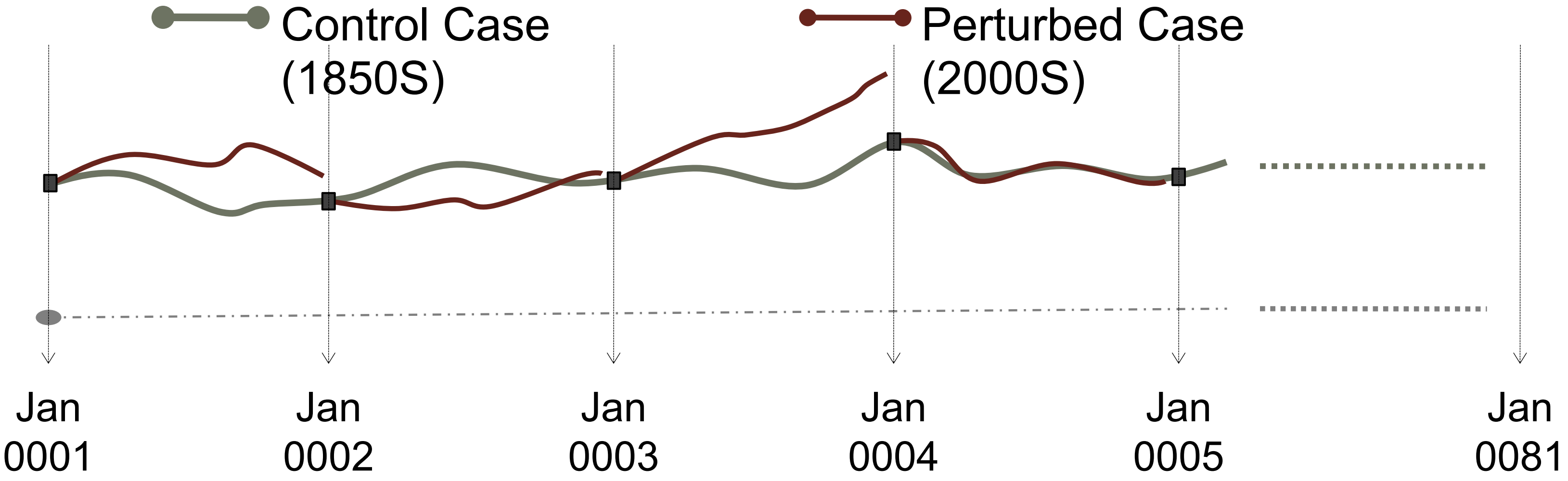
- Packing simulations together is economical as compared to a SLR
- Compare a 100 1-yr ensemble vs. a 100-yr long run
  - Poor Weak and Strong Scaling for 100-yr long run – smaller work load and increased MPI communications with increasing core counts
  - 100x greater workload per node for 100 member 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



Courtesy: David Hall
(https://www.earthsystemcog.org/projects/dcmip-2016/HOMME-NH)

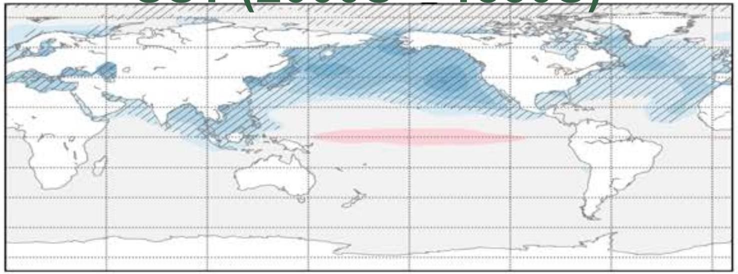


Short Ensembles: Scientific Utility

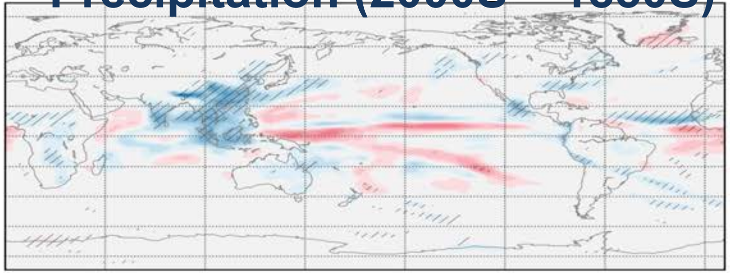


Fast Response

SST (2000S – 1850S)



Precipitation (2000S – 1850S)



Verma et al. 2019

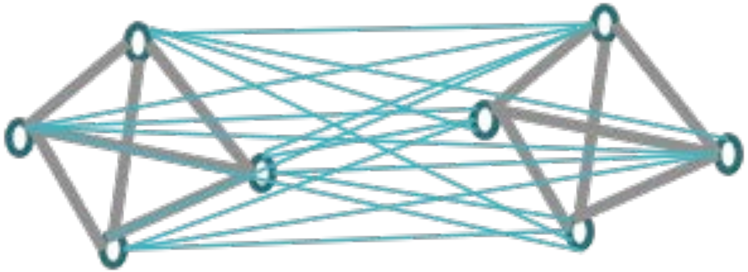
Equality of Distribution Tests

- 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%).

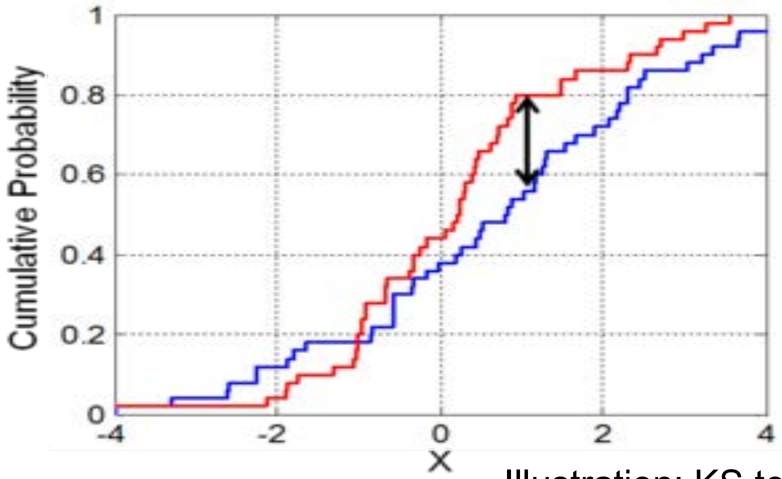
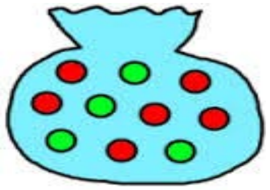


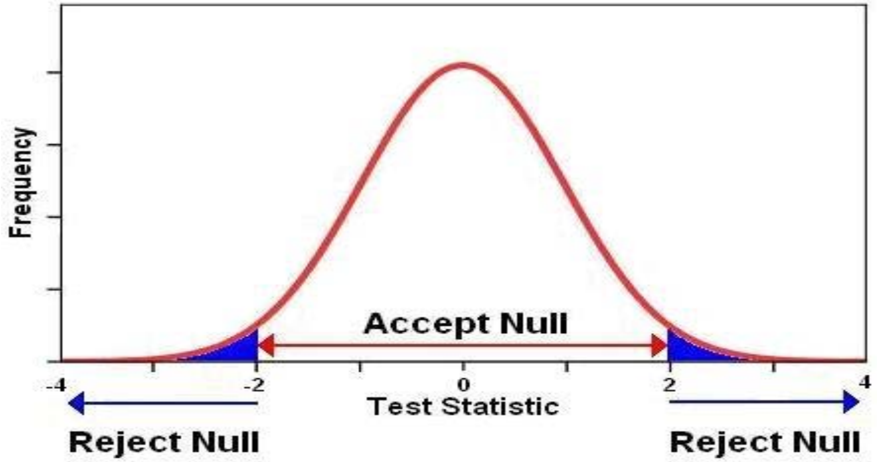
Illustration: KS test

  - 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.



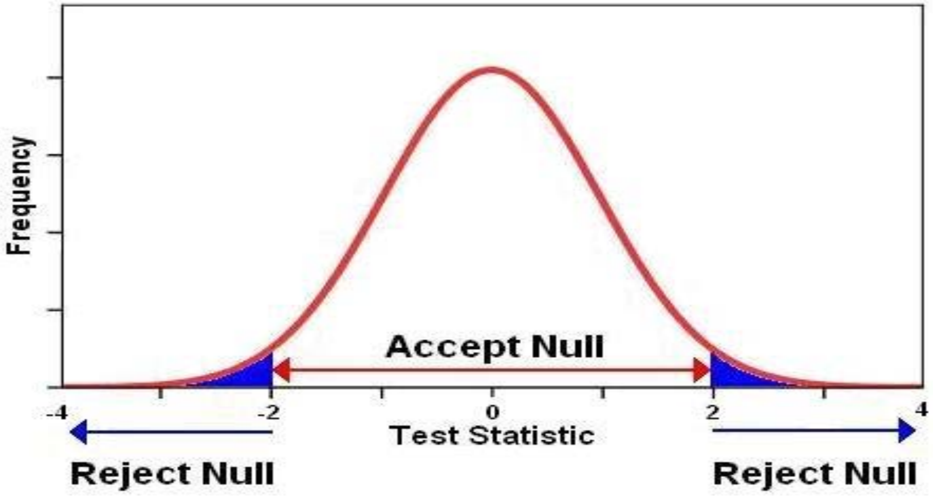


Model Verification Using Ensembles: Known Climate Changing Perturbation

- 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

Name	Description	Ens. Size
Default c0_ocn	Default model settings	30
Perturbed c0_ocn	Perturbed model parameter	30



Comparison	Test Statistic (t)	Critical No.	H0 Test
Default vs. perturbed c0_ocn	119	13	Reject

Power Analysis (Type II Error rate)

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

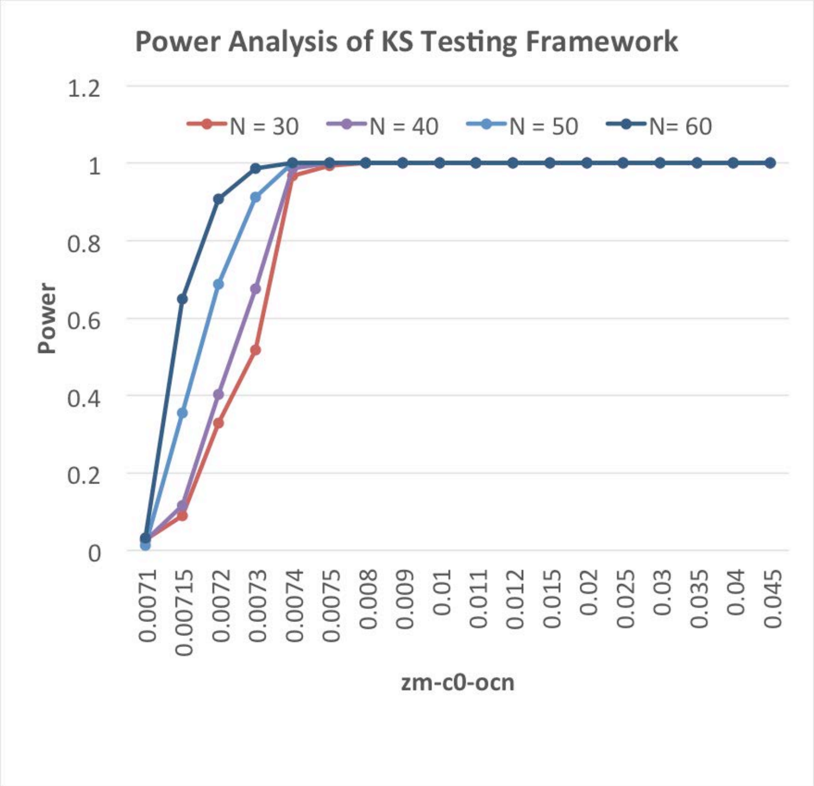
- Turn a tuning parameter knob incrementally: zm_c0_ocn (0.007 to 0.045)

- 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



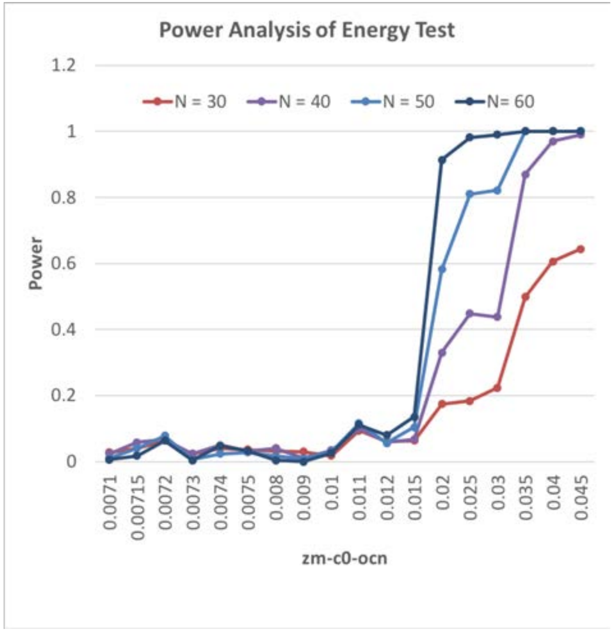
Example of Power Analysis.
Probability of correctly rejecting a false null hypothesis (Power) of the test in detecting changes to a EAM tuning parameter from a control case (*zm_c0_ocn = 0.0070*) for different short simulation (*1yr*) ensemble sizes (*N*).

Mahajan et al. 2019

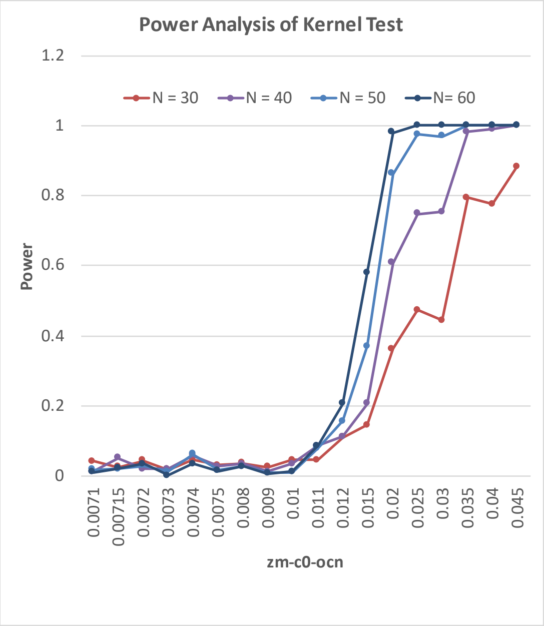
Power Analysis

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

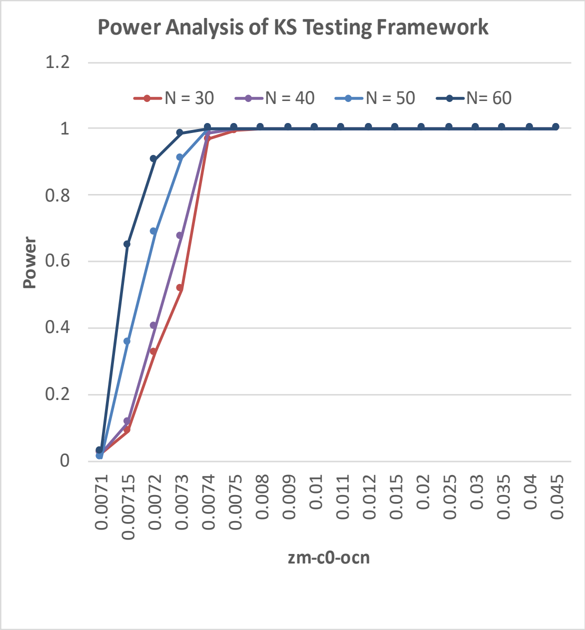
Energy Test



Kernel Test



KS Testing Framework

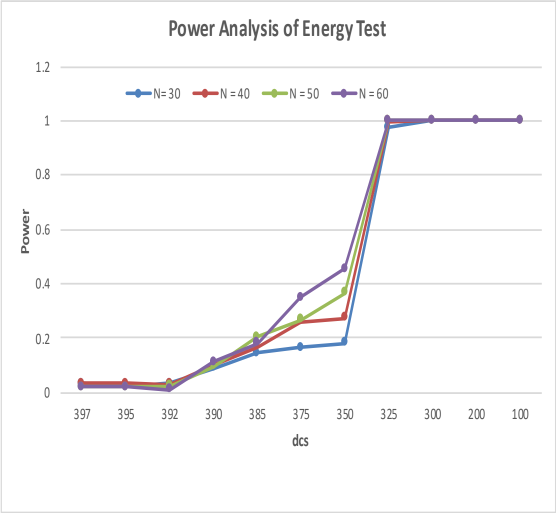


Mahajan et al. 2019

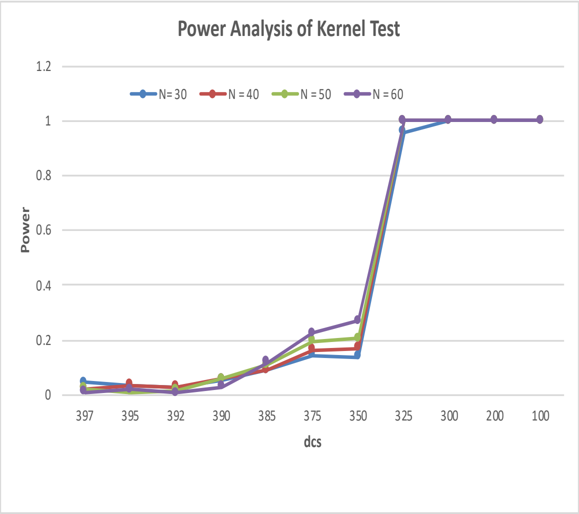
Power Analysis

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

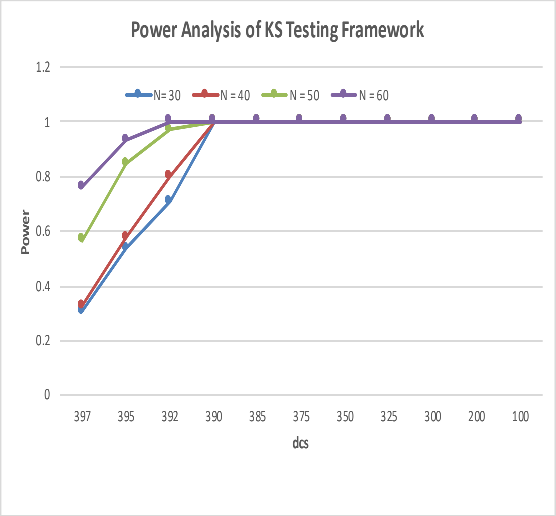
Energy Test



Kernel Test



KS Testing Framework

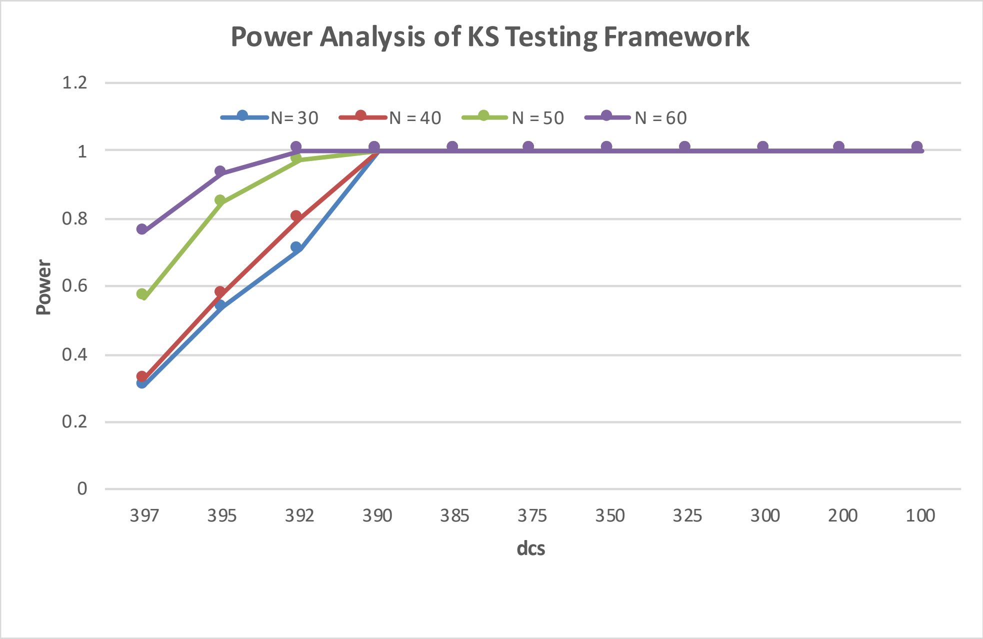


OAK RIDGE National Laboratory
Open slide master to edit
Mahajan et al. 2019

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.



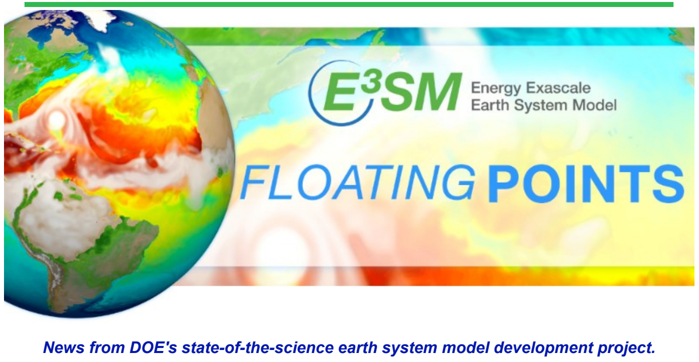
Example of Power Analysis. *Probability of correctly rejecting a false null hypothesis (Power)* of the test in detecting changes to a EAM tuning parameter from a control case (*dcs = 400*) for different short simulation (*1yr*) ensemble sizes (*N*).

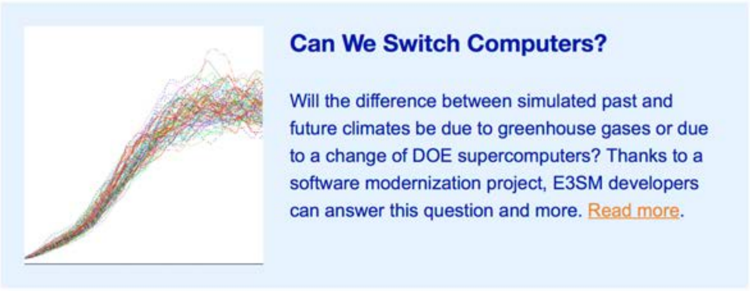
Mahajan et al. 2019

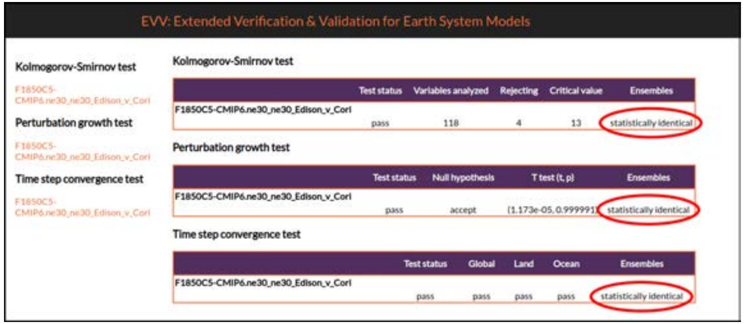
Test Case: Cori vs. Edison

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)
- All three - TSC (Wan, et al.), perturbation growth (Singh, et al.), and KS - climate reproducibility tests passed.
- Implications: Cori can be confidently used for remaining DECK simulations







EVV: Extended Verification & Validation for Earth System Models

Kolmogorov-Smirnov test
F1850C5-CMIP6.ne30_ne30_Edison_v_Cori
Perturbation growth test
F1850C5-CMIP6.ne30_ne30_Edison_v_Cori
Time step convergence test
F1850C5-CMIP6.ne30_ne30_Edison_v_Cori

Kolmogorov-Smirnov test

	Test status	Variables analyzed	Rejecting	Critical value	Ensembles
F1850C5-CMIP6.ne30_ne30_Edison_v_Cori	pass	118	4	13	statistically identical

Perturbation growth test

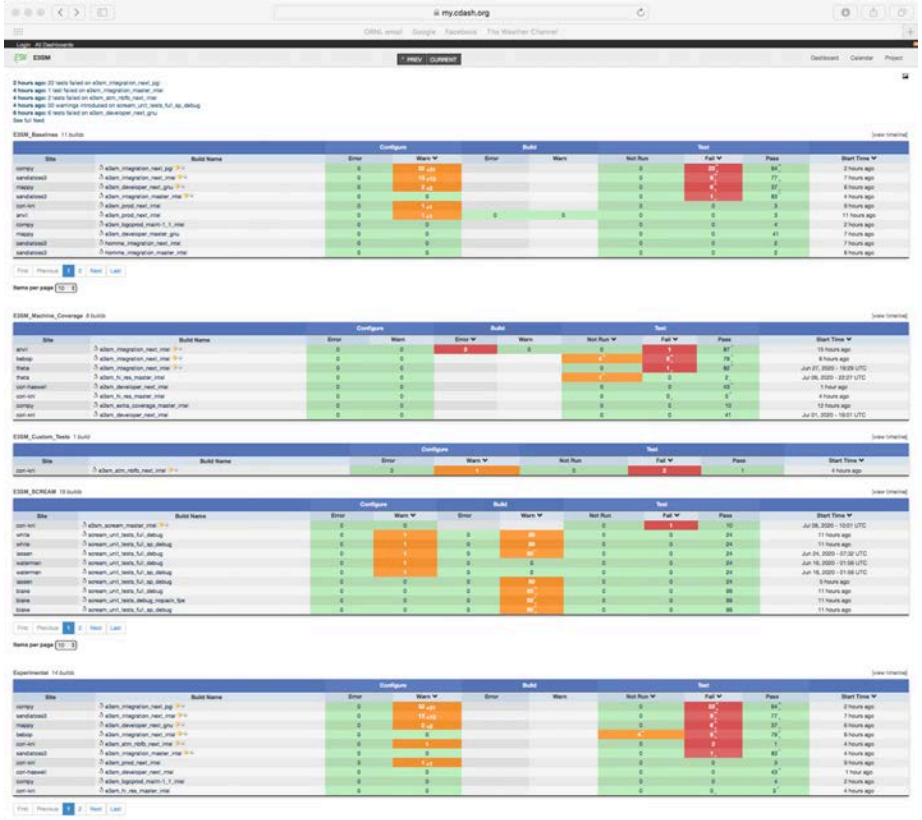
	Test status	Null hypothesis	T test (t, p)	Ensembles
F1850C5-CMIP6.ne30_ne30_Edison_v_Cori	pass	accept	(1.173e-05, 0.999991)	statistically identical

Time step convergence test

	Test status	Global	Land	Ocean	Ensembles
F1850C5-CMIP6.ne30_ne30_Edison_v_Cori	pass	pass	pass	pass	statistically identical

Reproducibility Tests (EAM) on Master

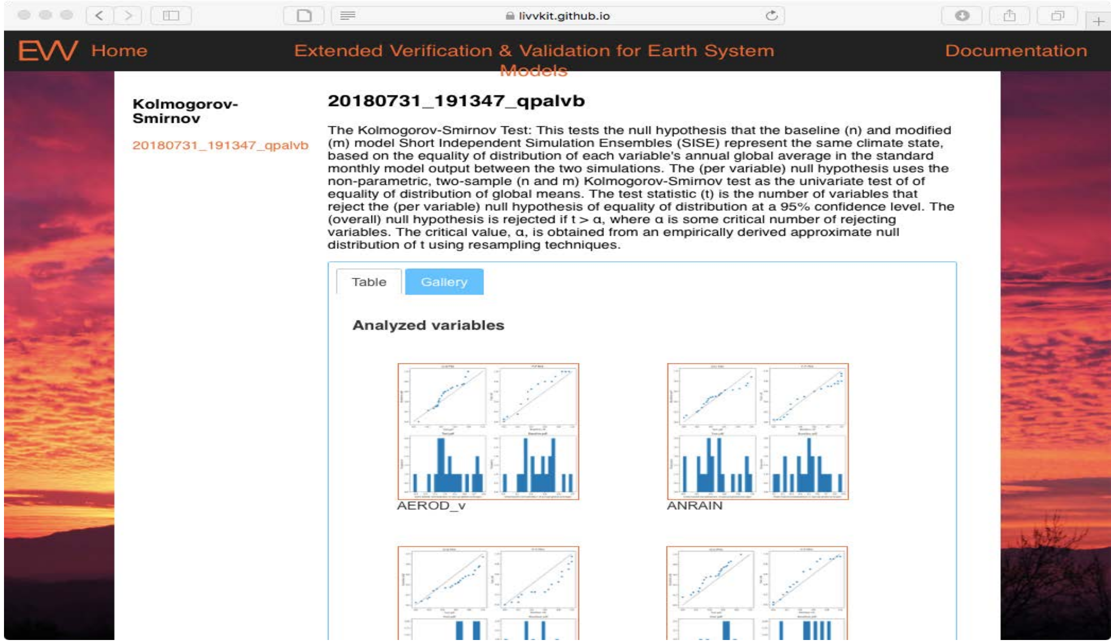
- Nightly tests run on Cori (E3SM custom tests)
  - Time step convergence test
  - Perturbation growth test
  - KS testing framework
- On CDASH under E3SM_Customs_Tests
  - https://my.cdash.org/index.php?project=E3SM
  - All runs archived:
  - Large ne4 1yr F1850C5 ensemble available (>1000)



Oak Ridge National Laboratory

EVV:

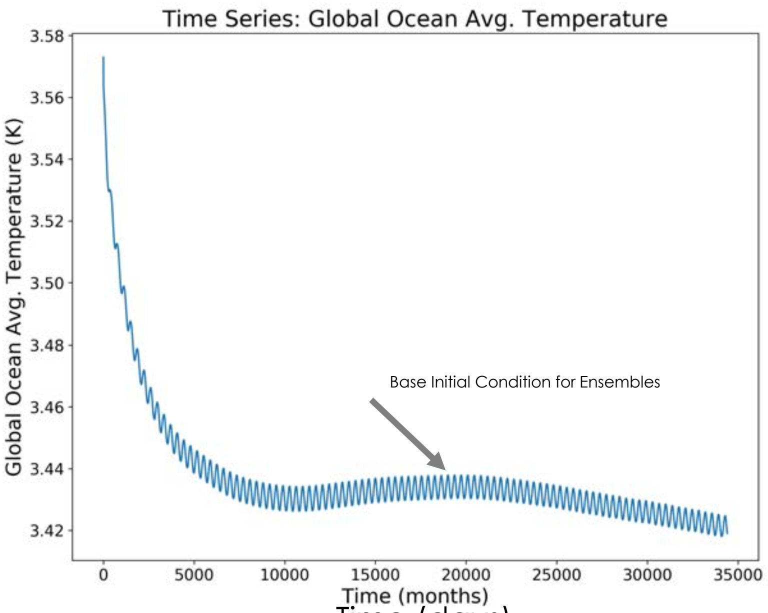
- Extended Verification and Validation for Earth System Models (EVV):
  - Python based toolkit:
    - Runs control and perturbed ensembles
    - Post-processes model output
    - Conducts tests
    - Publishes results and auxiliary plots, tables



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.



Time (days)

Machine Precision Perturbations to T at each grid point, j

$$T'_j = (1+x')T_j$$

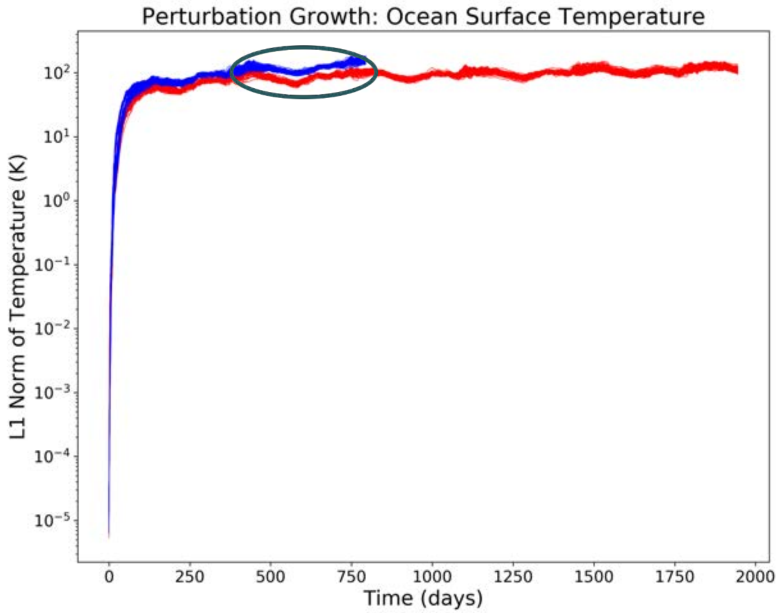
x' is a uniform random number transformed to range from $(-10^{-14}, 10^{-14})$

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 Round-off differences in MPAS-O



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:

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

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.

ρ: 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, α (say 0.05): $p <= \alpha$
  - For M tests, αM would be rejected for true null hypotheses just by chance
- For multiple tests, FDR constrains critical value (α_{FDR}) for local hypothesis tests (H_0):

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

p_j are sorted p-values of M tests

- *Global Null Hypothesis Test* (**G_0**)*: Reject if* $p_j <= \alpha_{FDR}$ 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

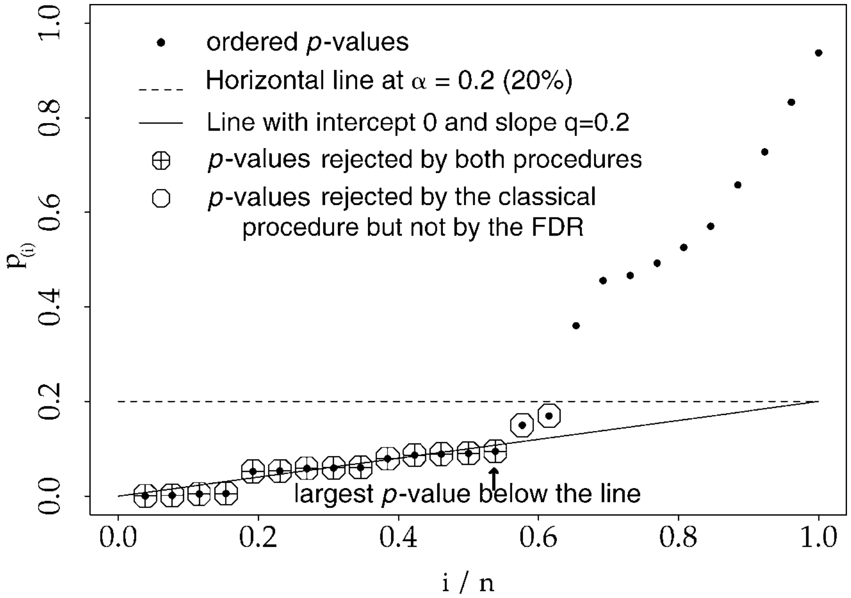


FIG. 2. Illustration of the traditional FPR and FDR procedures on a stylized example, with $q = \alpha = 20\%$. The ordered p-values, $p_{(i)}$, are plotted against i/n, $i = 1, \ldots, n$, and are circled and crossed to indicate that they are rejected by the FPR and FDR procedures, respectively.

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

Ventura et al. 2004

MPAS-O Reproducibility Tests

Evaluate False Positive Rate:

Bootstrap with Control Ensemble (150 ensemble members):

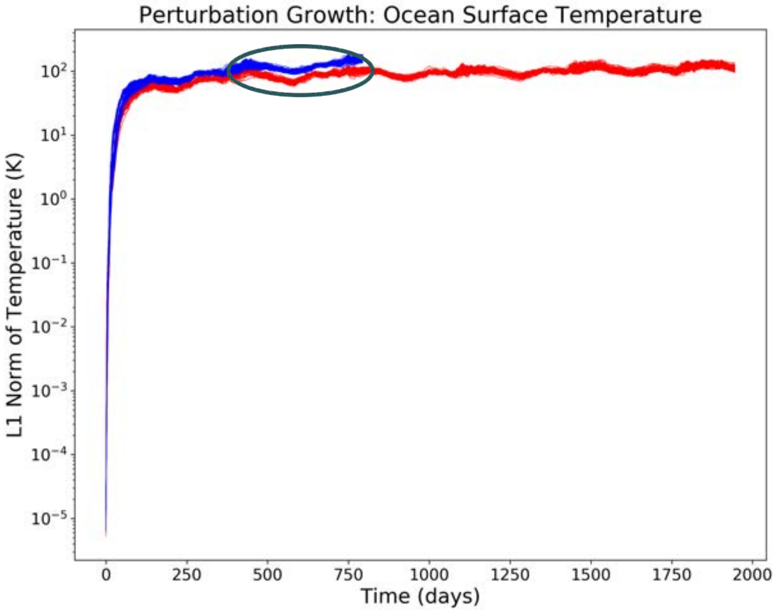
- Randomly draw two samples with N=M=30 members
- Conduct KS test and Cucconi test for alpha = 0.05
- Repeat 500 times at alpha = 0.05

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

Cucconi test:
95^{th} percentile of the no. of cells rejecting the local null hypothesis = 15
95^{th} 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



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.

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

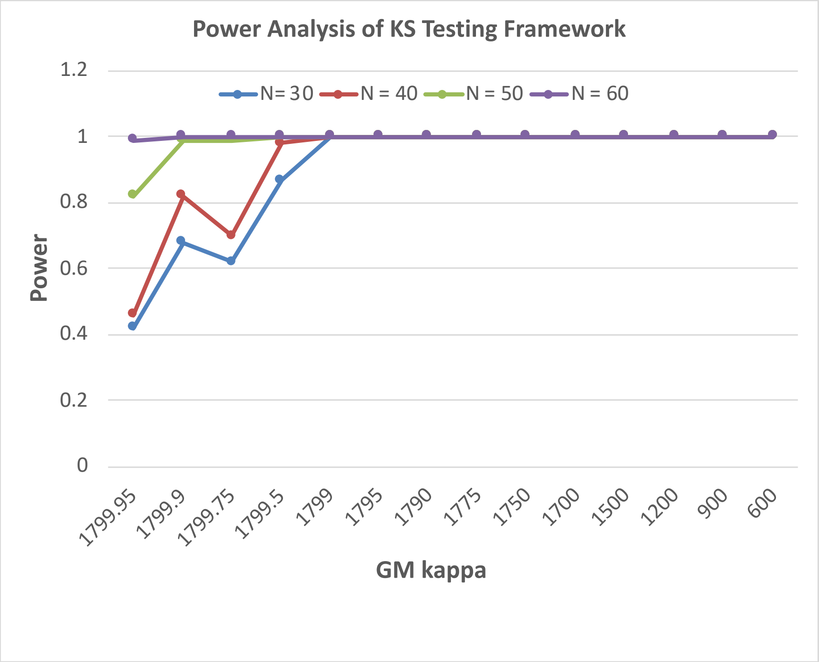
- Turn a tuning parameter knob incrementally:
  - Gent and McWilliams kappa (600 to 1800):

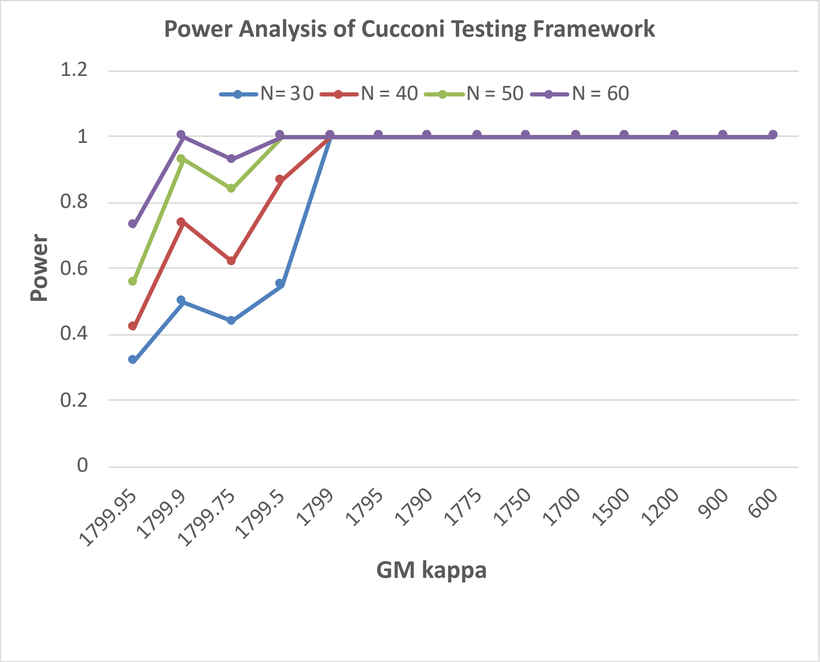
- Ensembles:
  - 100 members for each case
  - $T'_j = (1+x')T_j$, x' is random number transformed to range from $(-10^{-14}, 10^{-14})$

- 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*).

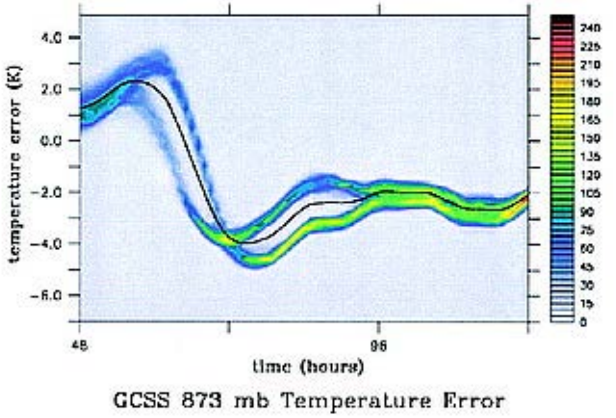
Summary:

- 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.

Ensemble spread in SCM



Hack and Pedretti (2000)

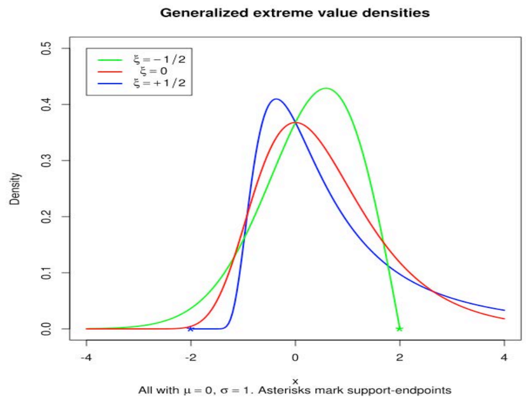
OAK RIDGE National Laboratory

Thanks!

- Acknowledgements:
  - DOE E3SM Project and CMDV-SM Project
  - Oak Ridge Leadership Computing Facility (OLCF)
  - NERSC

Test for Extremes

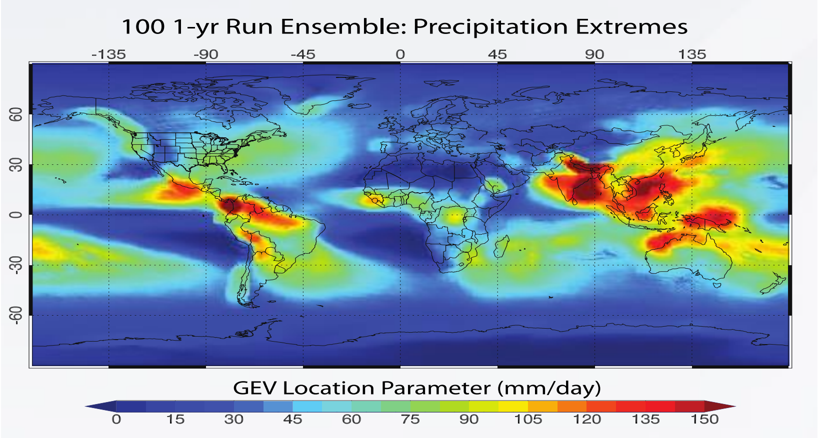
- Distribution tests perform poorly on distribution with different tails
  - Known for univariate tests, unexplored for multivariate tests.
- Use Generalized Extreme Value (GEV) theory (*e.g. Mahajan et al. 2015, Evans et al. 2014*).
  - max./min. of a process belong to GEV distribution.
  - Analogous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



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

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

where μ, σ and ξ represent the location, scale and shape parameter respectively.

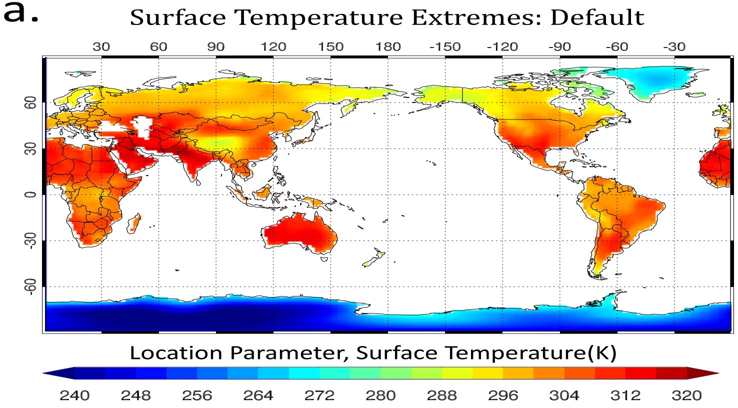


Climate Extremes Test

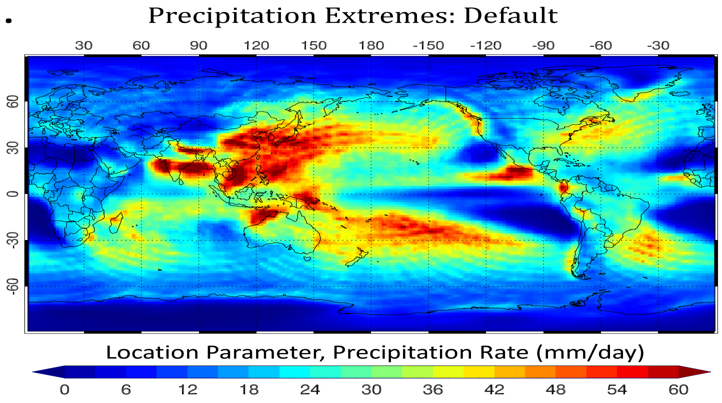
- Null Hypothesis (***G_0***): Simulation of extremes of a variable between two SISE is statistically indistinguishable.
- Annual maxima for each grid point are fit to a GEV distribution.
- *G_0:* Extremes at each grid point are statistically indistinguishable
- Test statistic (*g*): No. of grid points that reject *G_0*
- ***G_0*** rejected if $t > b$, where *b* is some critical number, obtained using resampling techniques.

Climate Extremes

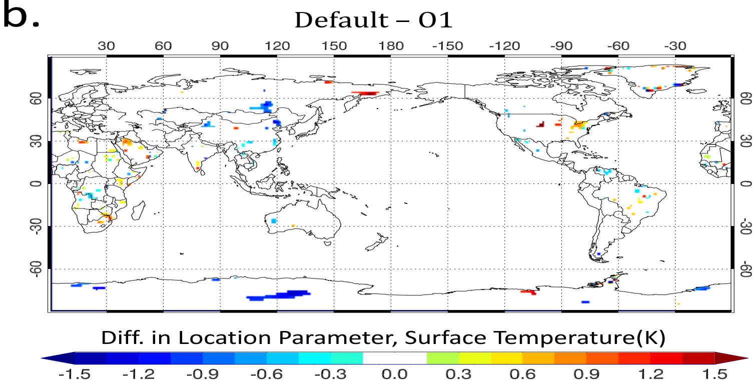
a.



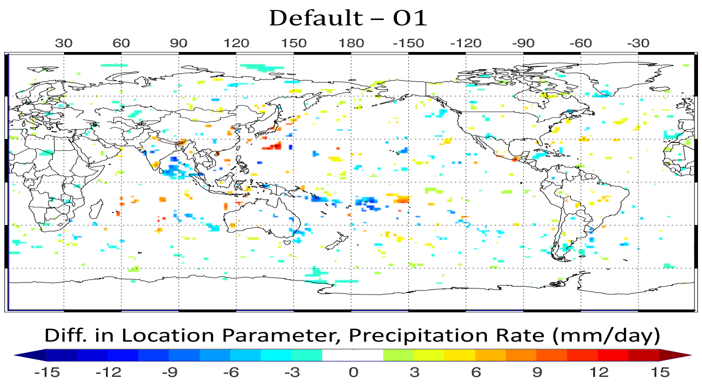
c.



b.



d.



o edit

Climate Extremes

Comparison	Variable	Test statistic (g)	Critical value (β)	G_0 Test
SISE-DEFAULT vs. SISE-O1	Precipitation Rate	5.1%	6.5%	Accept G_0
	Surface Temperature	5.0%	9.6%	Accept G_0
SISE-DEFAULT vs. SISE-FAST	Precipitation Rate	4.7%	6.3%	Accept G_0
	Surface Temperature	3.6%	9.6 %	Accept G_0
SISE-O1 vs. SISE-FAST	Precipitation Rate	5.2%	6.5%	Accept G_0
	Surface Temperature	10.3%	9.8%	Reject G_0

- 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

Single Long Run (SLR) vs. SISE

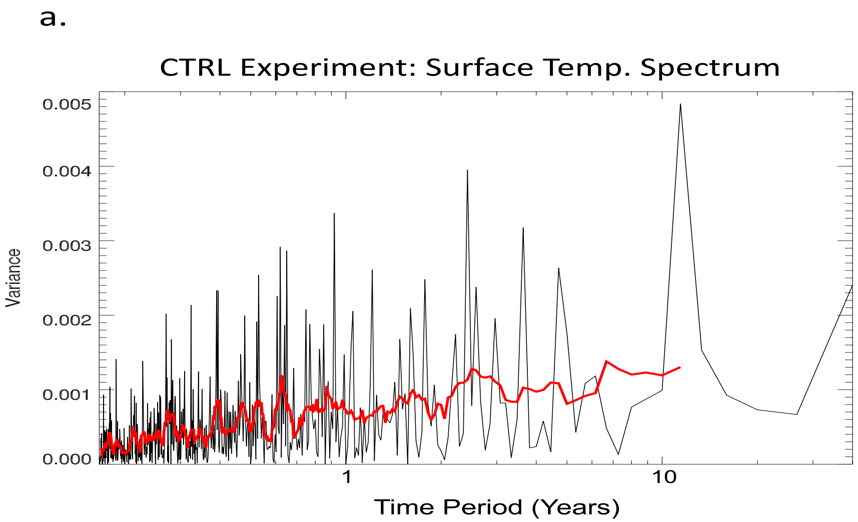
- SLR is clearly distinct from the SISE-DEFAULT

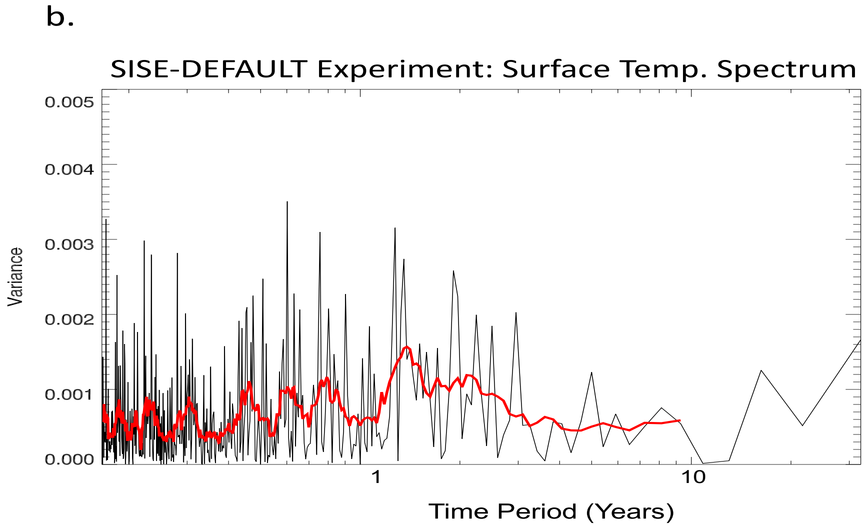
KS Testing Framework Results

Comparison	Test Statistic (t)	Critical Value (α)	$\mathbf{H_0}$ Test Result
SLR vs. SISE-DEFAULT	80 (50.6 %)	15	Reject $\mathbf{H_0}$
SLR vs. SISE-LND-INIT	74 (48 %)	13	Reject $\mathbf{H_0}$

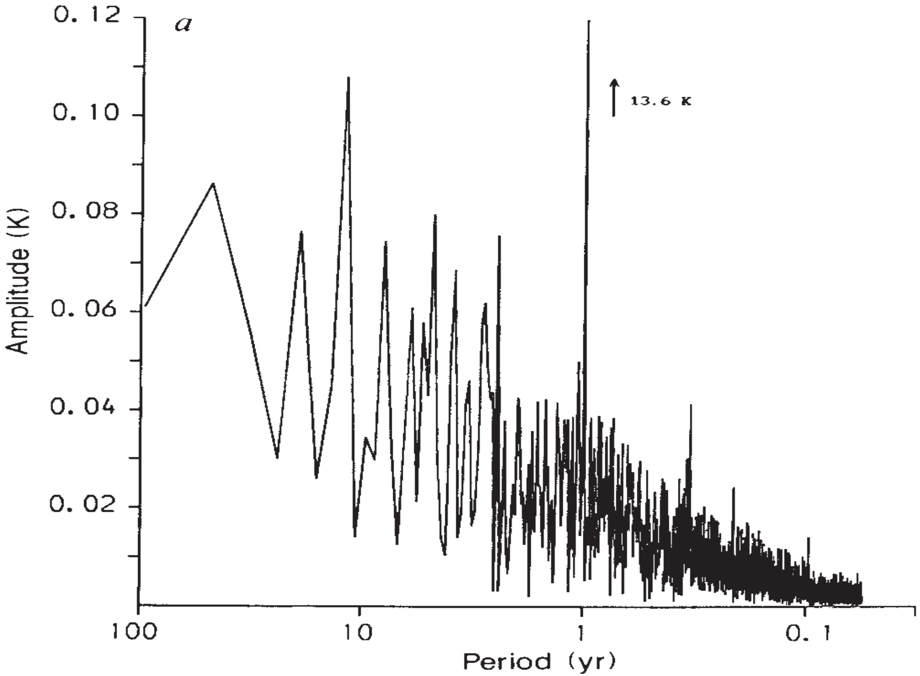
SLR vs. SISE

- 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*).





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):

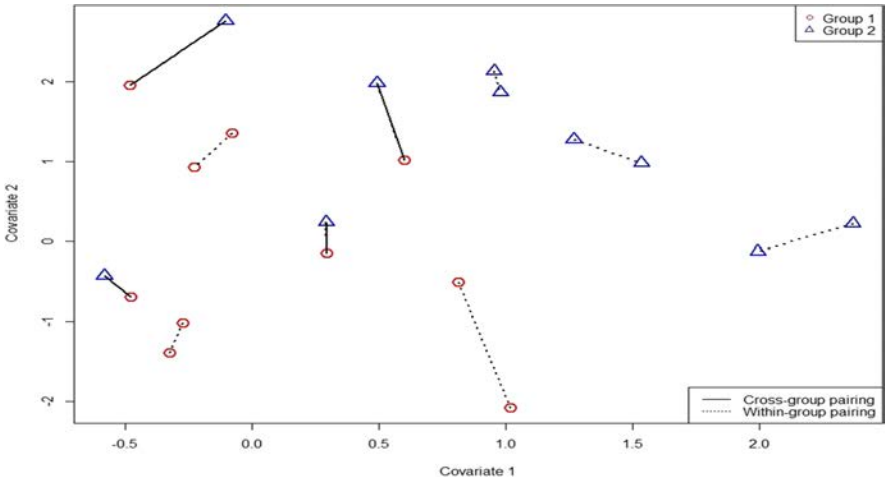
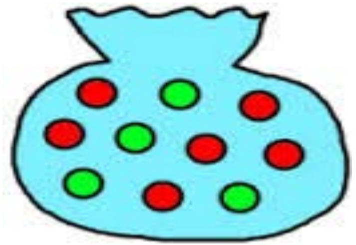


Illustration of cross matching for a bivariate case with $n = m = 10$. *(Ruth, 2014)*

Cross-Match Test



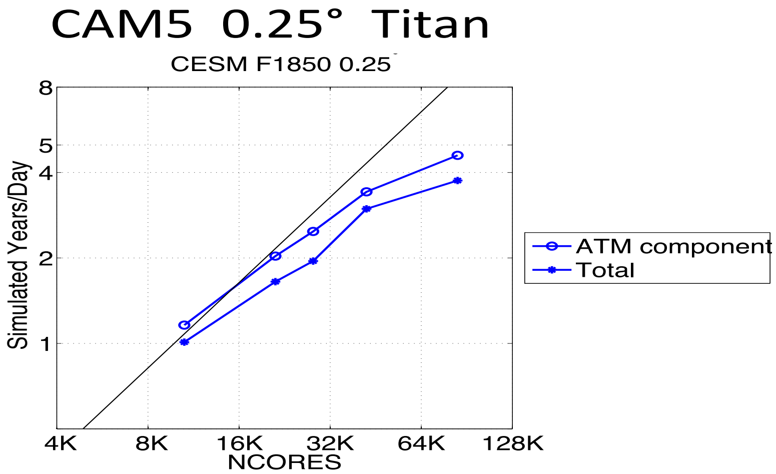
- Null distribution of T-statistic:

$$P(T = a_1) = \frac{2^{a_1} (N/2)!}{\binom{N}{n} \left(\frac{n-a_1}{2}\right)! \, a_1! \left(\frac{m-a_1}{2}\right)!}$$

  - 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

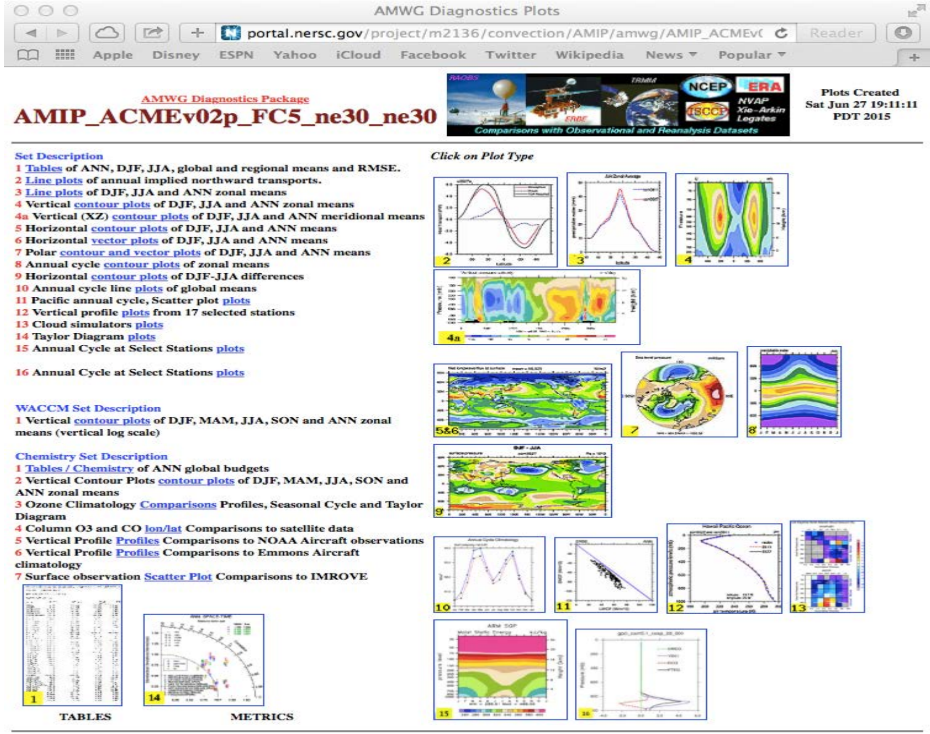
- To enhance throughput, use more cores:
  - 5 simulated years per day (required)
- But, scaling (weak or strong) is not perfect:
  - Less work per core with large core counts
  - Increase in MPI communications
  - Smaller MPI messages
  - Large MPI latency
- MPI cost: 90%



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

Comparison	**Test Statistic (t)**	**Critical Value (α)**	**$\mathrm{H_0}$ Test**
SISE-DEFAULT vs. SISE-O1	1 (0.6%)	17	Accept $\mathrm{H_0}$
SISE-DEFAULT vs. SISE-FAST	24 (15.2%)	14	Reject $\mathrm{H_0}$
SISE-O1 vs. SISE-FAST	23 (14.6%)	16	Reject $\mathrm{H_0}$

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

Name	Ens. Size	CIME	PGI	p-netcdf
v0.4-2015	30	4.0	15.3	1.5.0
master	30	5.0	17.5	1.7.0
v0.4	27	4.0	17.5	1.7.0

Frozen model configuration v0 vs. v1

Comparison	Test Statistic (t)	Critical no. (α)	H0 Test
v0.4-2015 vs. master	6 (3.6%)	13	Accept H0
v0.4 vs. master	8 (4.2%)	13	Accept H0
v0.4-2015 vs. v0.4	5 (3%)	13	Accept H0

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