Towards Ensuring Statistical Climate Reproducibility of Earth System Models

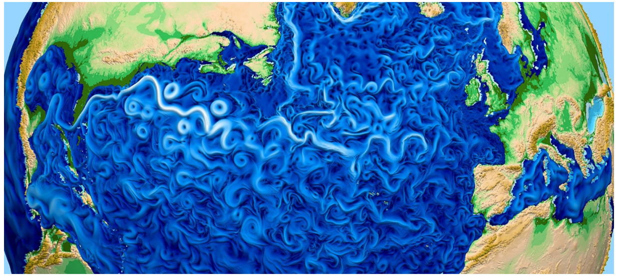
Salil Mahajan[1], Michael Kelleher[1], Joe Kennedy[2], Kate Evans[1]

[1]Oak Ridge National Laboratory, USA, [2]University of Alaska, Fairbanks

Motivation:

E3SM Energy Exascale Earth System Model

- US DOE's Energy Exascale Earth System Model (E3SM) project:
  - Effectively exploit DOE's leadership class HPC capabilities
    - Improving model trust-worthiness

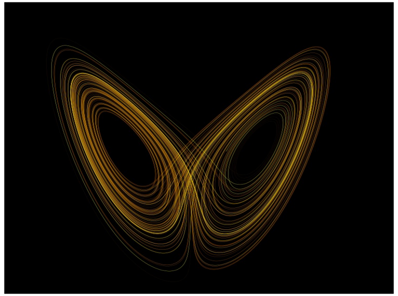


- Code Evolution:
  - Bit-for-bit reproducing changes
    - E.g. Adding a new compset, new output variable, new stealth feature, etc.
  - Non-b4b changes
    - Different climate (statistics) expected
      - E.g. New parameterizations modules, new tunings
    - Same climate (statistics) expected
   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

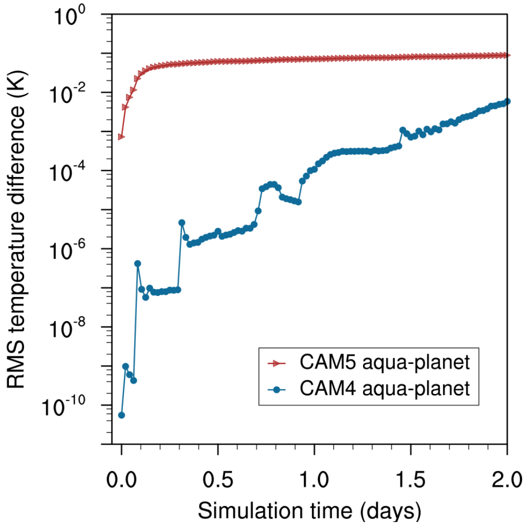
- **Goal:** ***Test the null hypothesis that climate simulation remains statistically equivalent after unintended non-b4b changes.***

Error growth in climate systems

- Truncated Floating Point arithmetic:
  - Round-off errors
  - Non-associative:
    - $(-1 + 1) + 2^{-53} \neq -1 + (1 + 2^{-53})$
  - Optimizations, hybrid architectures, code refactoring, etc. can change the order of operations.
- Climate models:
  - Chaotic, non-linear system
- Round-off differences grow quickly
- Problem: Identify systematic bugs from innocuous error growth in non-BFB reproducible environment.

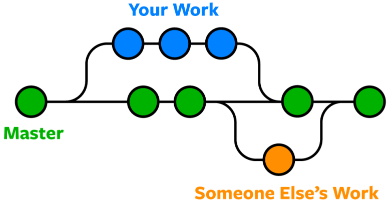


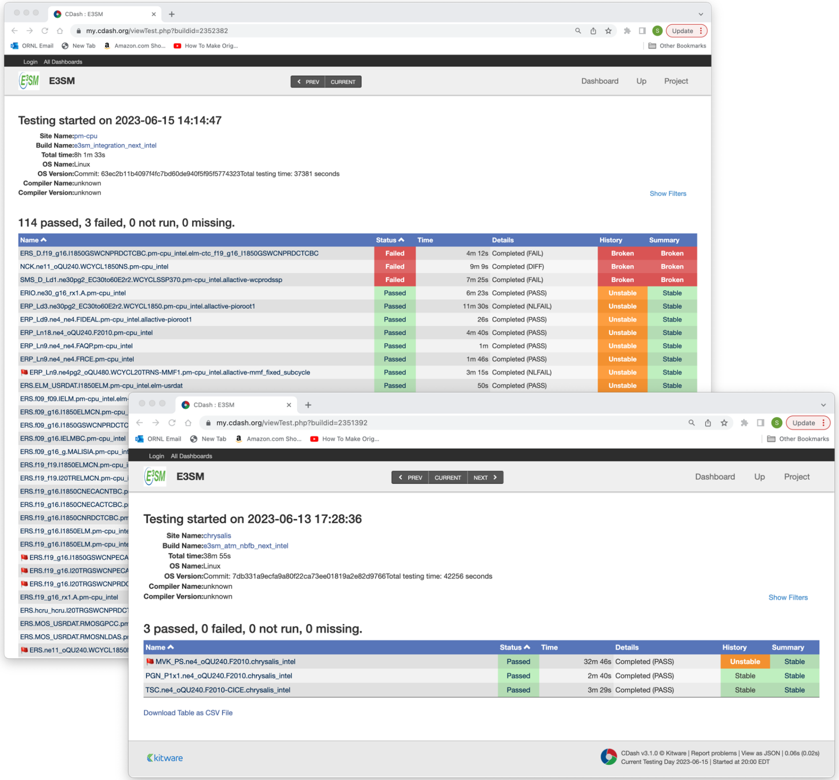
Lorenz attractor
(*Source:en.wikipedia.org/wiki/Chaos_theory*)



Evolution of root mean square temperature difference caused by random perturbations of the order of 10^{-14} K imposed on the temperature initial conditions (Wan et al. 2017)

E3SM Testing





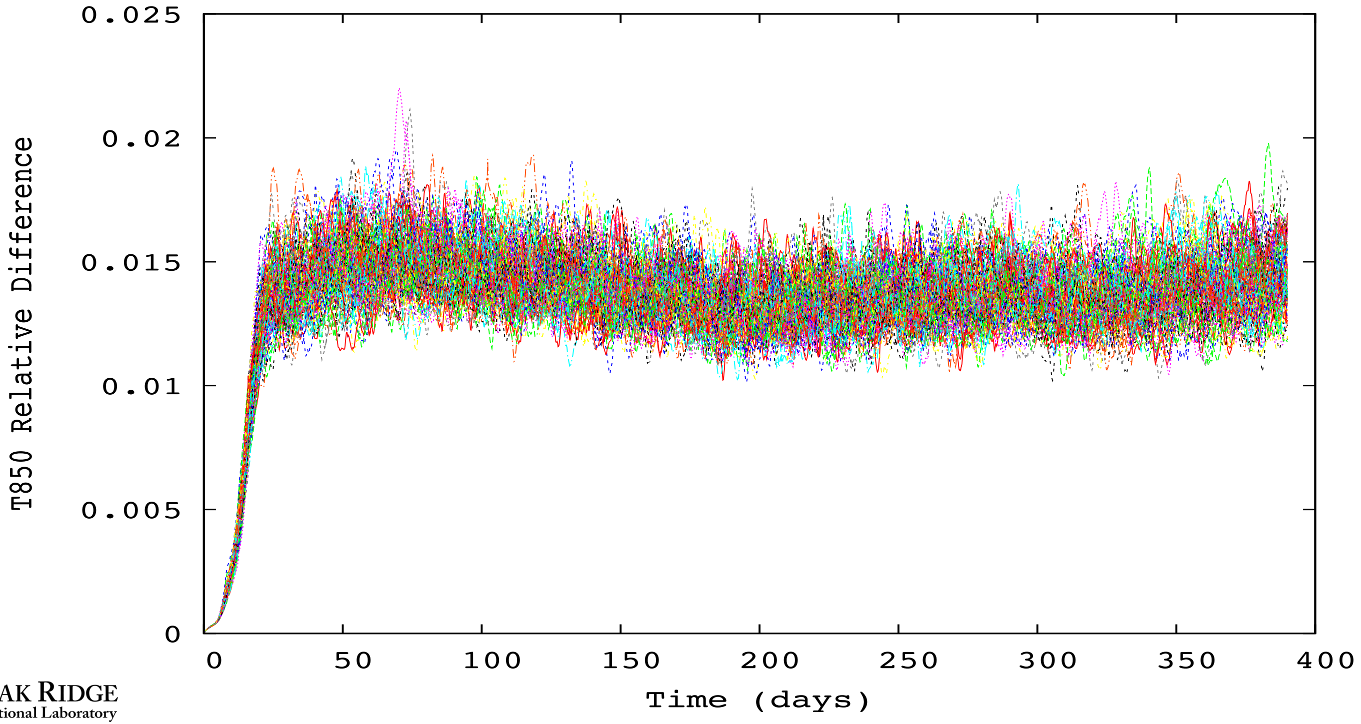
- E3SM Testing Suite (bfb):
  - * APT (auto promotion test (default length))
    * CME (compare mct and esmf interfaces (10 days))
    * ERB (branch/exact restart test)
    * ERH (hybrid/exact restart test)
    * ERI (hybrid/branch/exact restart test, default 3+19/10+9/5+4 days)
    * ERS (exact restart 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))

- 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

Initial Condition Simulation Ensemble

- $T'_j = (1+x')T_j$
  - x' is uniform random number transformed to range from $(-10^{-14}, 10^{-14})$



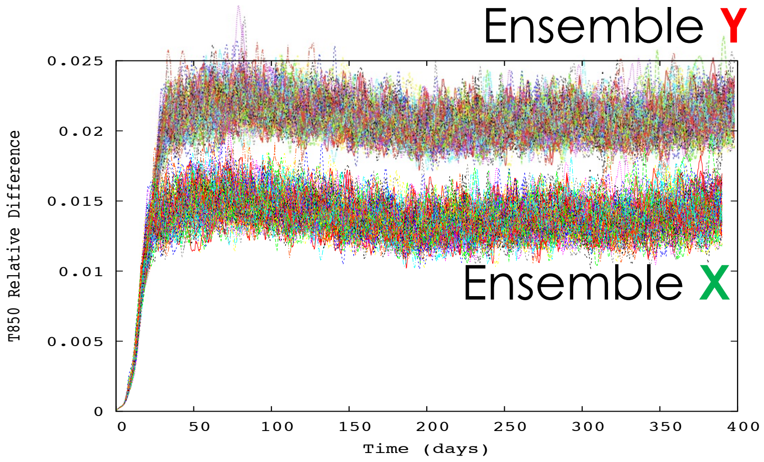
Chaotic nature of the climate system: *L1 Norm of temperature at 850mb as compared to a control run for a 100 EAM runs differing only in initial conditions perturbed by machine precision levels.*

Two Sample Testing Using Ensembles

- **Goal:**
  - Evaluate statistics of the modified ensemble vs. control ensemble after propagation of errors from machine precision differences in initial conditions.
  - Short (1 yr) ensembles

- **Problem statement:**
  - *Multivariate* two sample equality of distribution testing:
    - NULL hypothesis: Statistically Equivalent
    - High dimensions (121 variables)
    - Low sample sizes (~30 ensemble members)

- **Approach:**
  - Use statistical/ML approaches for two sample equality of distribution tests: kernel test, energy test, Kolmogorov-Smirnov (KS) test



OAK RIDGE National Laboratory

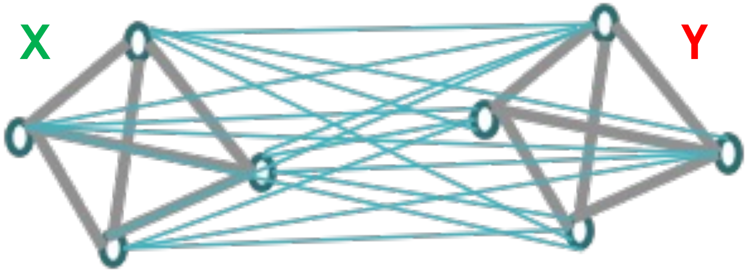
Equality of Distribution Tests

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

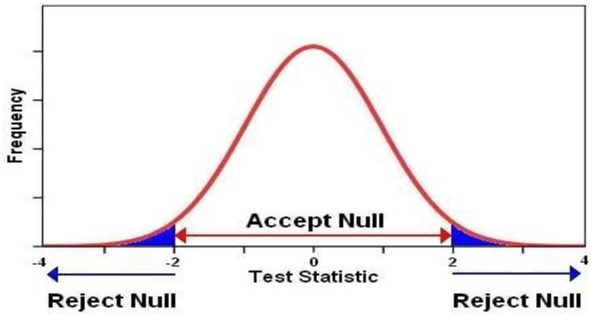
$$e = \frac{nm}{n+m}\left(\overbrace{\frac{2}{nm}\sum_{i=1}^{n}\sum_{k=1}^{m}\|X_i - Y_k\|}^{\text{Sum of Blue lines}} - \overbrace{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|X_i - X_j\| - \frac{1}{m^2}\sum_{l=1}^{m}\sum_{k=1}^{m}\|Y_l - Y_k\|}^{\text{Sum of Grey lines}} \right)$$

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*



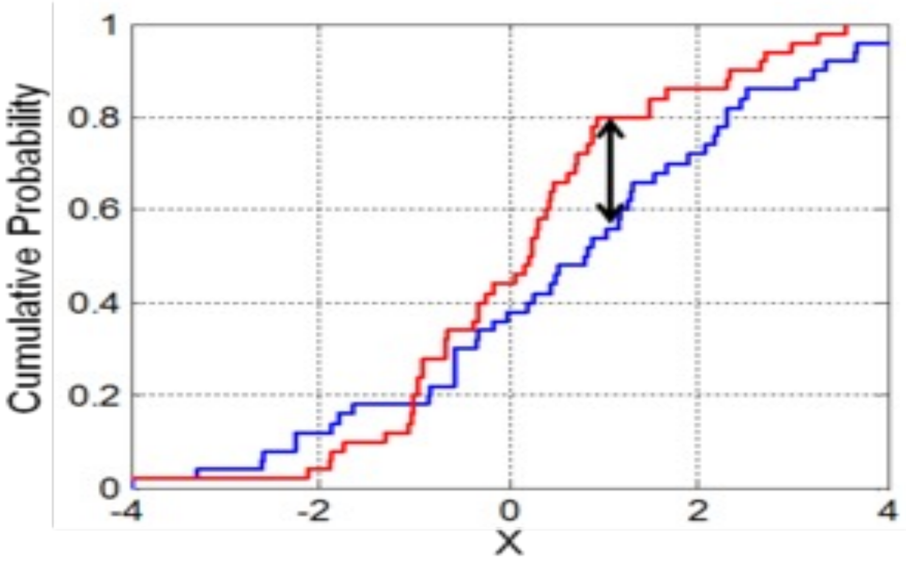
Schematic: Energy Test



Schematic: Null Distribution

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 (**121 variables**).
  - *H_0*: Local Null hypothesis for each variable
  - 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 (*Type I Error Rate*), say 95%.
  - **Null distribution** of *t*: Resampling (150 member ensemble)
  - **Critical value** of *t*: **13**



Schematic Illustration: KS test

OAK RIDGE National Laboratory

Test Case: 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

- **Ensembles:** Machine-precision level random perturbations to the initial 3-D temperature field
  - 30 member 1-yr ensembles
  - $T'_j = (1+x')T_j$, x' is random number transformed to range from $(-10^{-14}, 10^{-14})$

- **Peturbation:** Modify a model tuning parameter:
  - **zm_c0_ocn** (control case: 0.007, modified: 0.045)
  - Deep convection scheme parameter controlling conversion rate of cloud droplets to precipitation

Both KS-test and Energy test reject the null hypothesis

Type II Error Rate (False Negatives)

- What does it mean if the test is a pass?
- What is the false negative rate?
- How small a change than the test detect confidently?

Power Analysis (Type II Error rate)

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

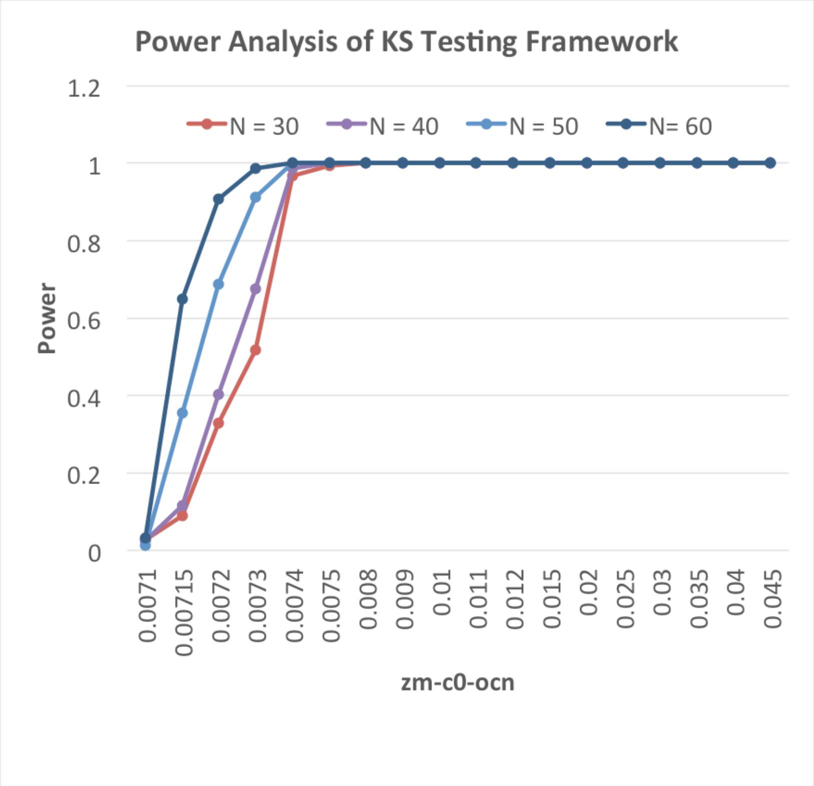
- Turn a tuning parameter knob incrementally: zm_c0_ocn (0.007 to 0.045)

- **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**:
  - Resampling:
    - Randomly pick N=30 (=40, 50, 60) members from the control and modified ensembles
    - Conduct test
    - Repeat (500 times)

Power Analysis: KS Testing Framework

Controlled changes to zm_c0_ocn **(default value = 0.0070)**



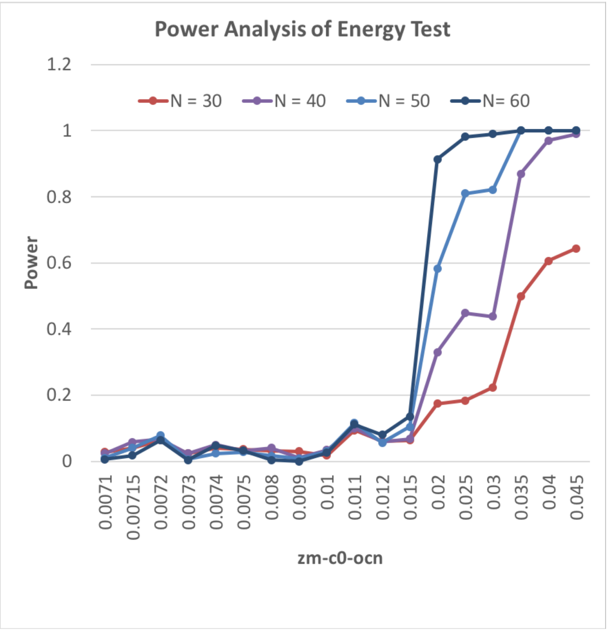
Example of Power Analysis.
Probability of correctly rejecting a false null hypothesis (Power) of the test in detecting changes to a EAM tuning parameter from a control case (*zm_c0_ocn = 0.0070*) for different short simulation (*1yr*) ensemble sizes (*N*).

Mahajan et al. 2019

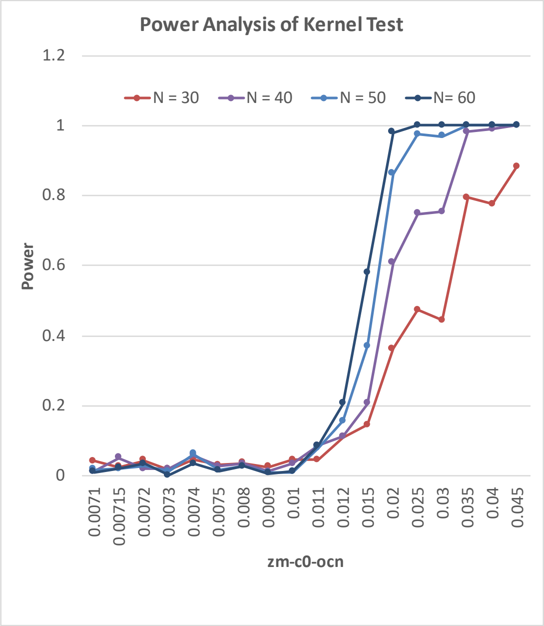
Power Analysis

Controlled changes to zm_c0_ocn **(default value = 0.0070)**

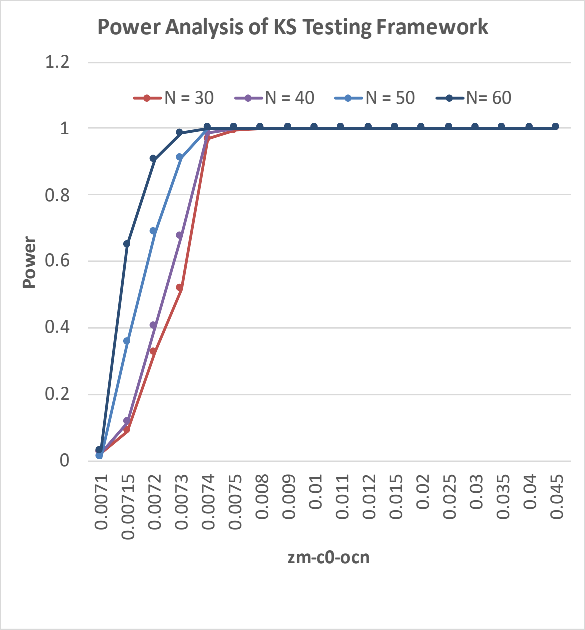
Energy Test



Kernel Test



KS Testing Framework



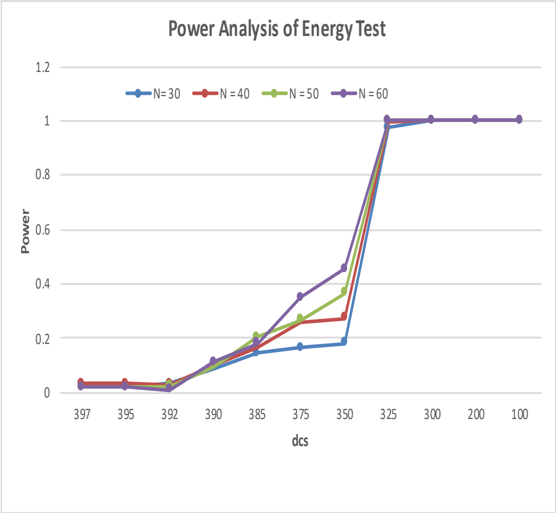
Oak Ridge National Laboratory

Mahajan et al. 2019

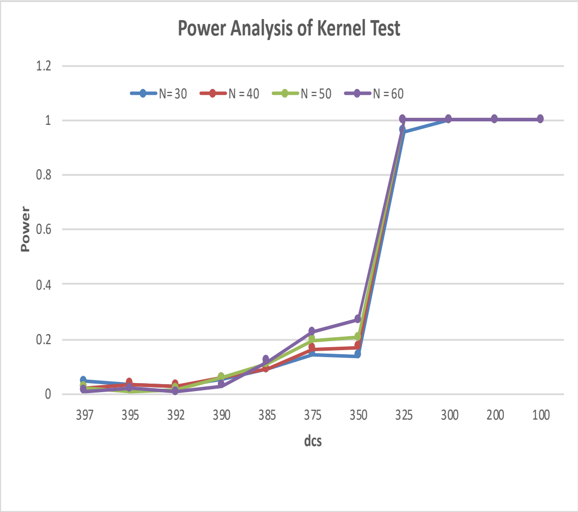
Power Analysis

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

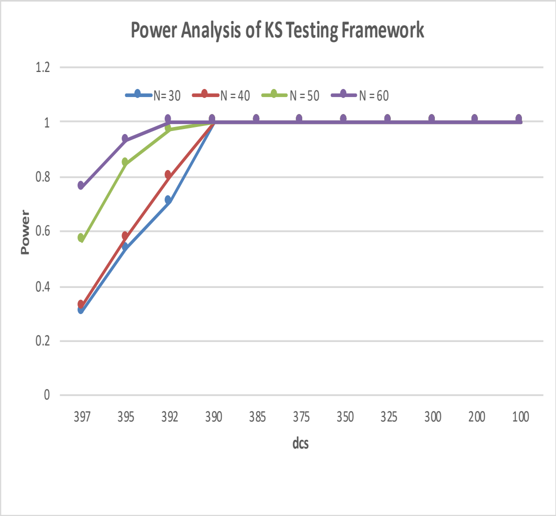
Energy Test



Kernel Test



KS Testing Framework

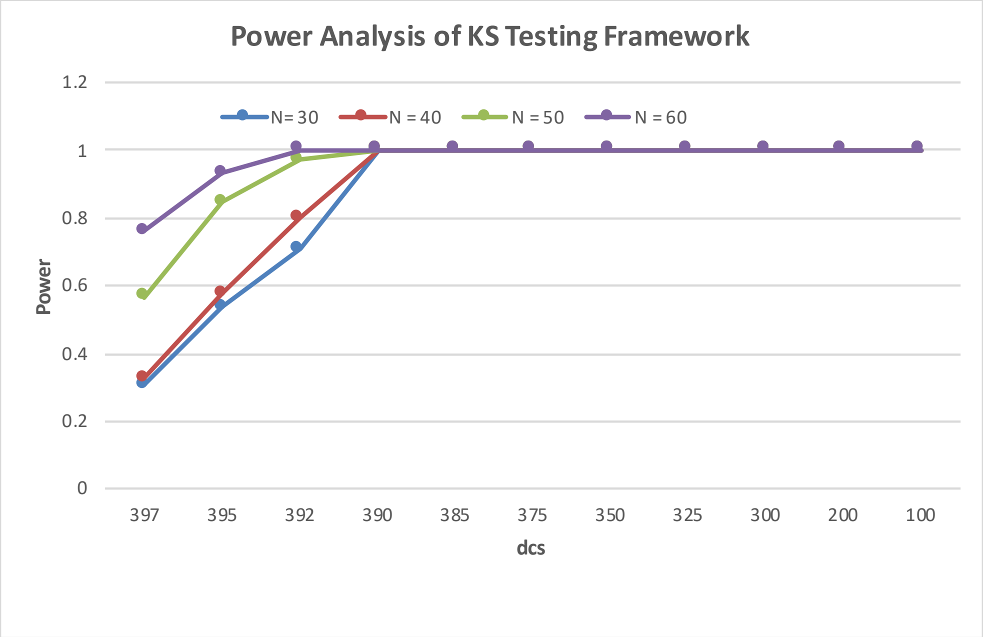


Oak Ridge National Laboratory

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

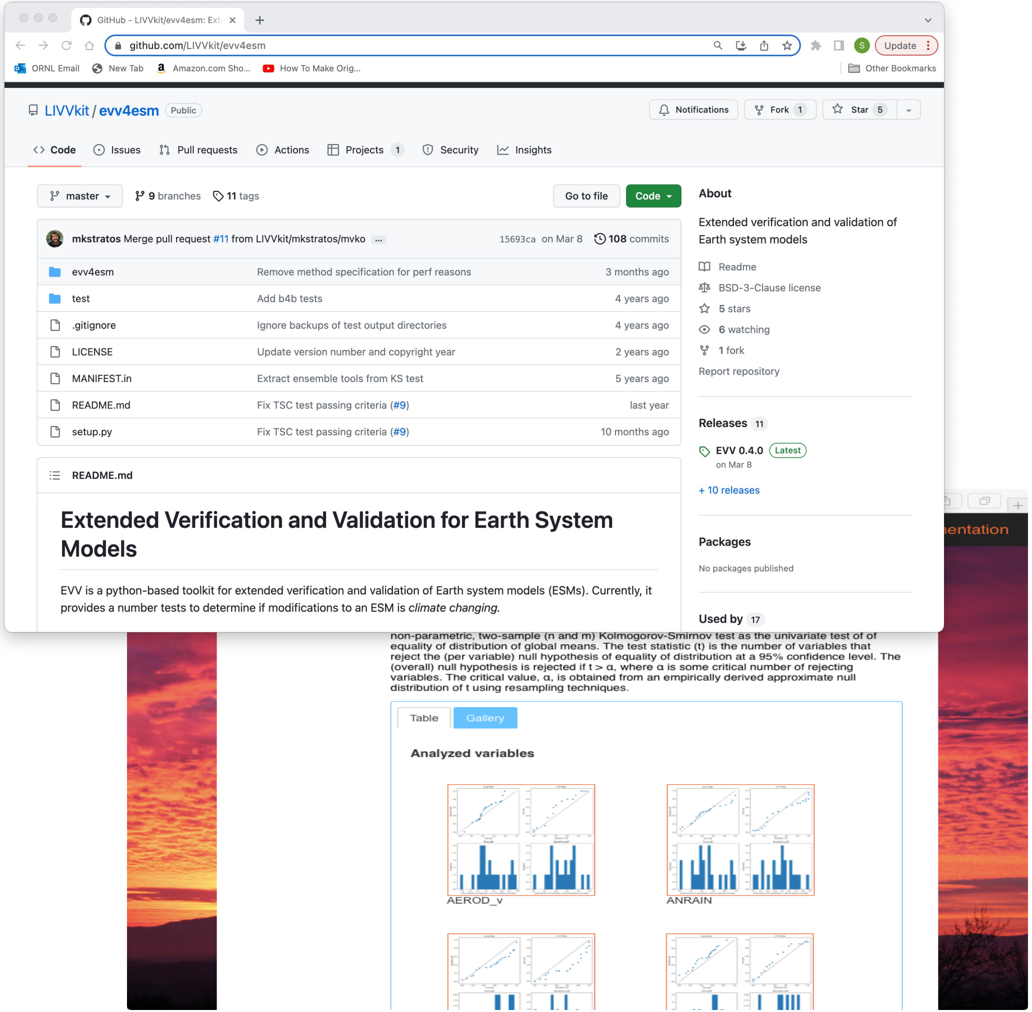


Example of Power Analysis. *Probability of correctly rejecting a false null hypothesis (Power)* of the test in detecting changes to a EAM tuning parameter from a control case (*dcs = 400*) for different short simulation (*1yr*) ensemble sizes (*N*).

Mahajan et al. 2019

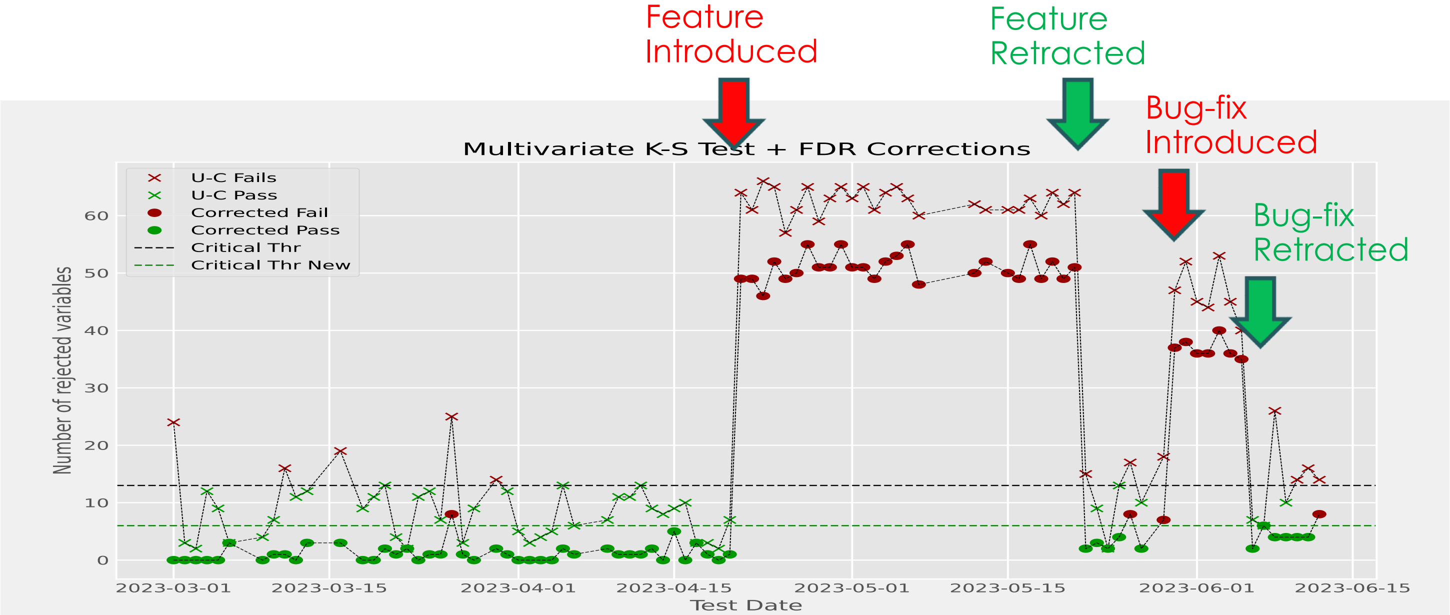
EVV4ESM

- Extended Verification and Validation for Earth System Models (**EVV4ESM**):
    - Python based toolkit
    - Runs control and new ensembles
    - Post-processes model output
    - Conducts reproducibility tests
    - Publishes results and auxiliary plots, tables



Real World Test Cases

- New backwards compatible feature (chemistry) introduced climate changing behavior
  - Reproducibility test flagged climate changing behavior
  - Evaluation, retraction and bug-fix.
- Bug fix (aqueous chemistry) introduced climate changing behavior

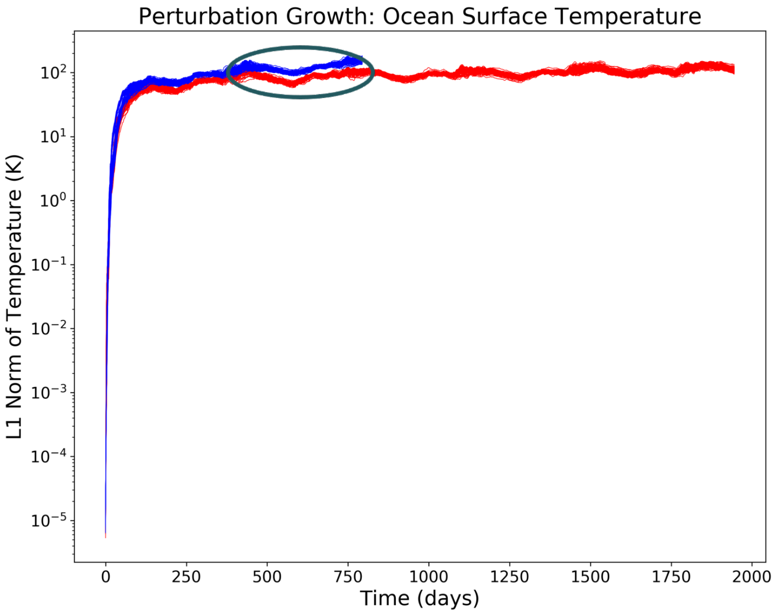


Ocean Model Reproducibility tests: Approach

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

- Generate control and perturbed ensembles at QU240 resolution:
  - 7153 grid points per vertical level
  - 60 vertical levels
- Evaluate 5 prognostic variables (Baker et al. 2016)
  - SSH, T, U, V, Salinity
  - Annual average of year 2.
- Ocean variability is spatially more 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.

Ocean Model 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*):

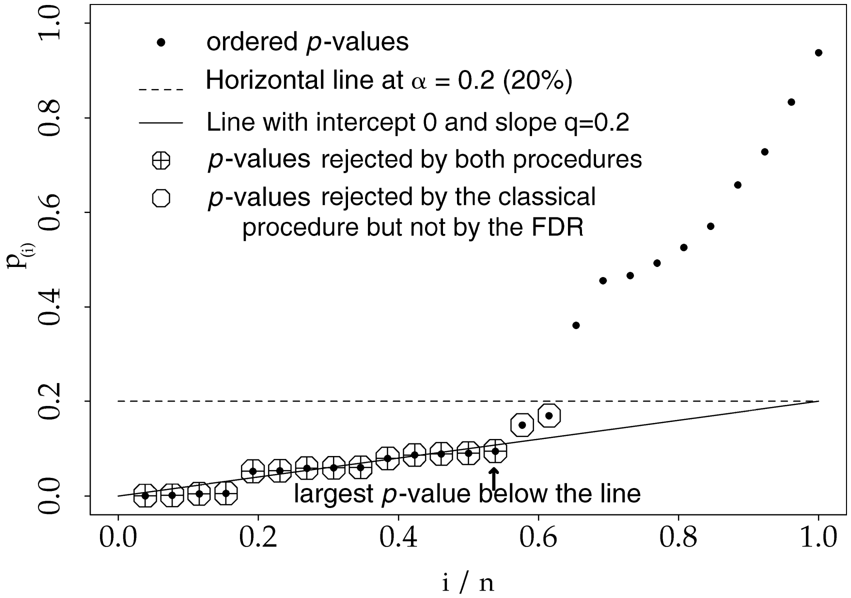
- For single test, null hypothesis is rejected if:
  - Test statistic p-value (p) is less than a critical value, α (say 0.05): $p <= \alpha$
  - For M tests, αM would be rejected for true null hypotheses just by chance
- For multiple tests, FDR constrains critical value (α_{FDR}) for local hypothesis 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* <= α_{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

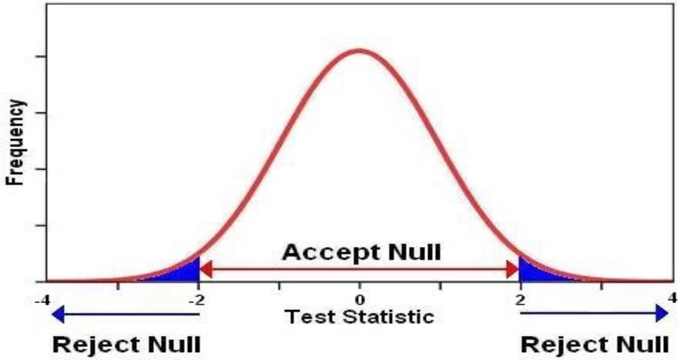


$$\alpha_{FDR} = \max_{j=1,2,\ldots,M} \{p_j : p_j \leq \alpha(j/M)\}$$

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

Ocean Model Reproducibility Tests: Type I Error Rate



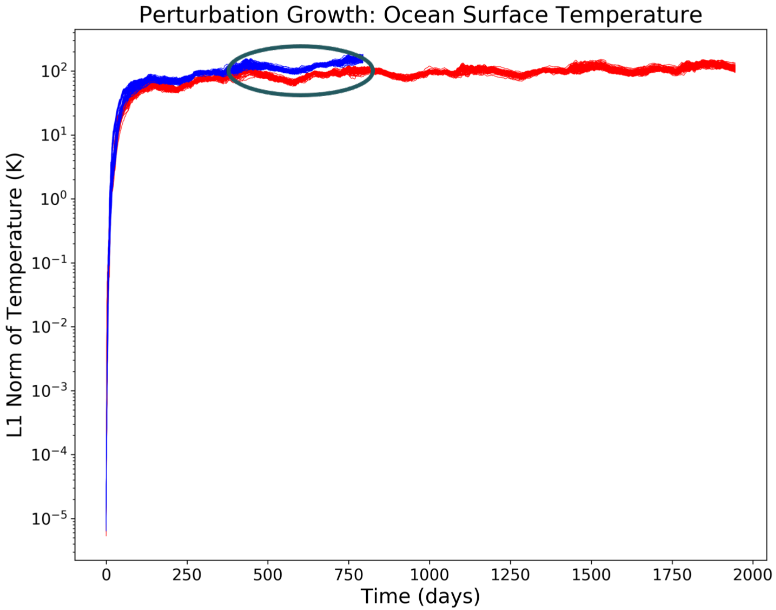
- Bootstrap with Control Ensemble (150 ensemble members):
  - Randomly draw two samples with N=M=30 members
  - Conduct KS test and Cucconi test for alpha = 0.05
  - Repeat 500 times
  - For SSH (7153 ocean cells)

- 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 (FDR) = 15
  - 95th percentile of the no. of cells rejecting the local null hypothesis = 643

Ocean Model Reproducibility Tests: Test case

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.

Oak Ridge National Laboratory

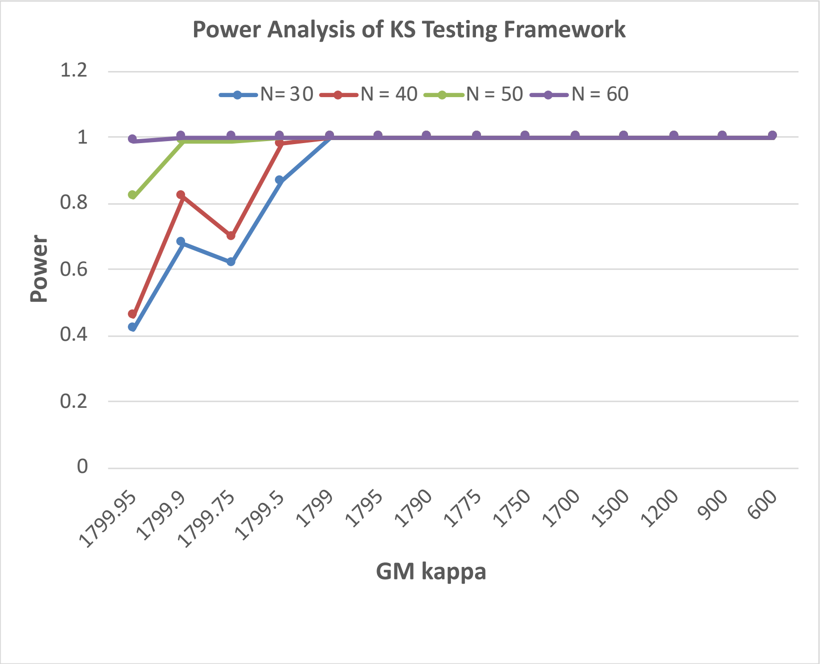
Ocean Model Reproducibility Tests: Power Analysis

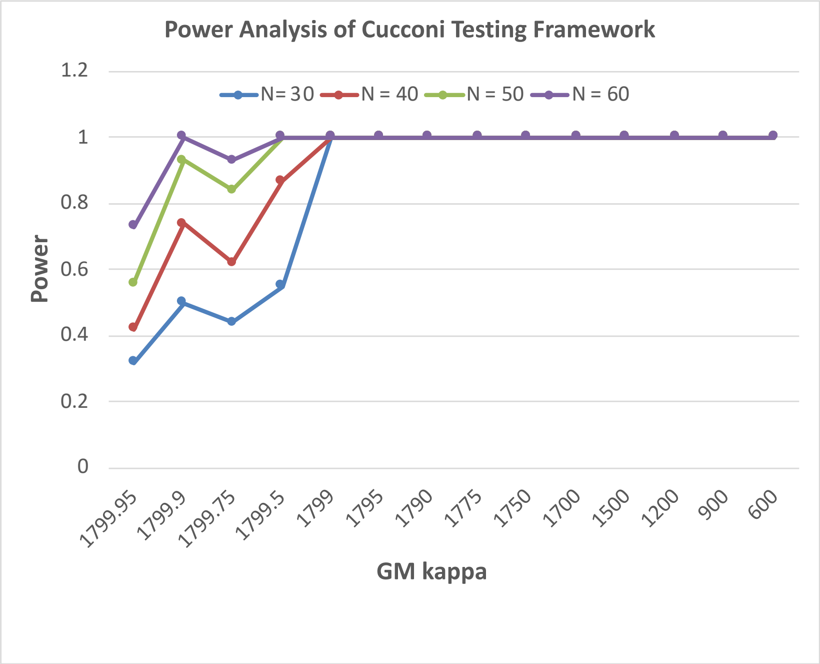
Type II error rate: Probability of accepting a false null hypothesis

- 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)

Ocean Model Reproducibility Tests: Power Analysis

Controlled changes to GM kappa **(default value = 1800)**





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 ESMs adapts for Exascale: GPU ports, AI/ML based kernels, etc.
- Developed a **multivariate testing framework** for climate reproducibility after perturbation growth in atmosphere and ocean models :
  - **EVV4ESM** toolkit
- **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
  - **Real world scenarios/success stories**: Machine ports, climate changing bug-fixes, climate changing stealth features, etc.
- **Future work:**
  - Apply to other known test cases with non-b4b changes
  - Evaluate applicability of low-resolution results at high-resolution
  - Apply FDR correction to the atmosphere KS testing framework
  - Evaluate other ML based tests
  - Build tests for individual software kernels: e.g. individual physics packages like RRTMGP, MG2, CLUBB, MAM4, etc.
  - Build tests for other modeling components – sea-ice, land

Thanks!

- Acknowledgements:
  - DOE E3SM Project and CMDV-SM Project
  - Oak Ridge Leadership Computing Facility (OLCF)
  - Argonne Leadership Computing Facility (ALCF)
  - National Energy Research Scientific Computing (NERSC)

E3SM Energy Exascale Earth System Model

U.S. DEPARTMENT OF ENERGY | Office of Science

BIOLOGICAL AND ENVIRONMENTAL RESEARCH
Climate and Environmental Sciences Division

- References:

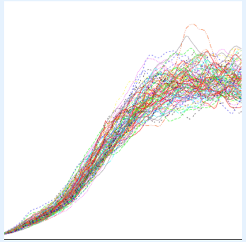
  - Mahajan S., A. L. Gaddis, K. J. Evans and M. R. Norman, 2017: Exploring an ensemble-based approach to atmospheric climate modeling and testing at scale, Procedia Computer Science, 108, 735-744, doi: 10.1016/j.procs.2017.05.259
  - Mahajan, S., K. J. Evans, Joe Kennedy, M. L. Branstetter, M. Xu, M. Norman (2019): "A multivariate approach to ensure statistical reproducibility of climate model simulations", Proceedings of the Platform for Advanced Scientific Computing (PASC) 2019
  - Mahajan, S., K. J. Evans, Joe Kennedy, M. L. Branstetter, M. Xu, M. Norman (2019): "Ongoing solution reproducibility of earth system models as they progress toward exascale computing", Special Issue for Computational Reproducibility at Exascale Workshop, 2017, Super Computing 2017 in International Journal of High Performance Computing Applications
  - Mahajan, S. (2021): Ensuring Statistical Reproducibility of Ocean Model Simulations in the Age of Hybrid Computing, Platform for Advanced Scientific Computing, Association for Computing Machinery, New York, NY, USA, Article 1, 19, https://doi.org/10.1145/3468267.3470572

Test Case: Cori vs. Edison

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

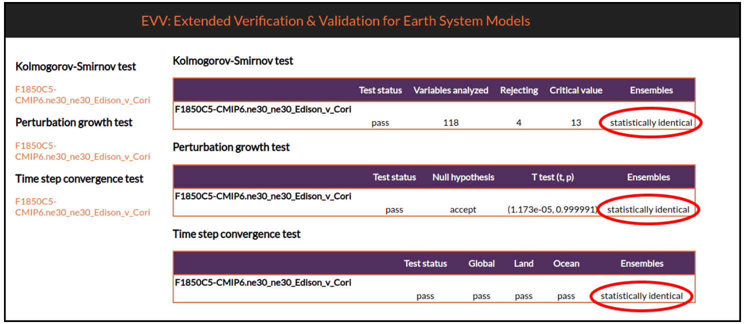
- Conducted short simulation (1yr) ensembles on both Edison and Cori:
  - F1850C5-CMIP6 compset
  - ne4 (100 ensemble members)
  - ne30 (30 ensemble members)
- All three - TSC (Wan, et al.), perturbation growth (Singh, et al.), and KS - climate reproducibility tests passed.
- Implications: Cori can be confidently used for remaining DECK simulations

News from DOE's state-of-the-science earth system model development project.



Can We Switch Computers?

Will the difference between simulated past and future climates be due to greenhouse gases or due to a change of DOE supercomputers? Thanks to a software modernization project, E3SM developers can answer this question and more. Read more.



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

Significance Level (Type I Error Rate): Resampling

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

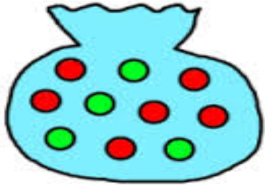
- Simulations from the pool are then randomly assigned to one of two groups of sizes *n* and *m*.

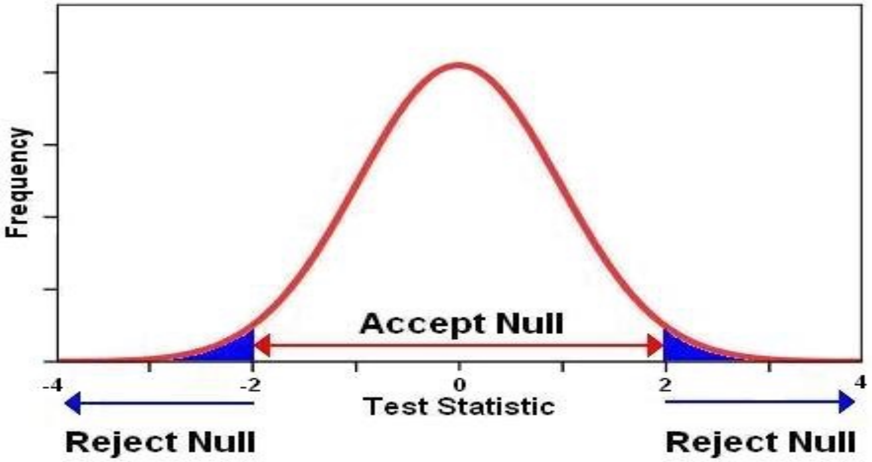
- The *t-statistic* is then computed for the random drawing.

- Repeat

- If all possible random drawings are made, the null distribution of *t* is exact.

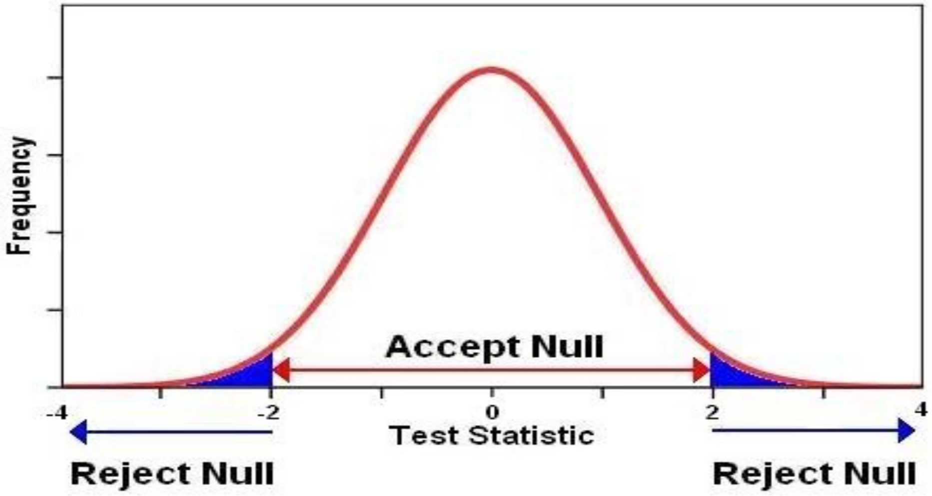
  – We conduct 500 drawings - approximate null distribution.





KS Testing Framework Results

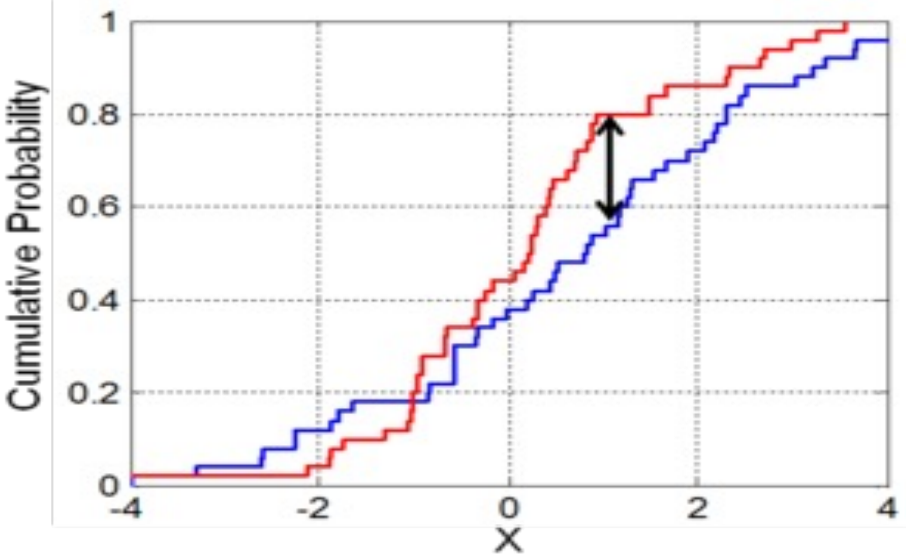
Name	Description	Ens. Size
Default c0_ocn	Default model settings	30
Perturbed c0_ocn	Perturbed model parameter	30



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

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 (121 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%).
  - Null distribution: Resampling



Schematic Illustration: KS test

- ***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.

OAK RIDGE National Laboratory

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

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)

Reproducibility Tests (EAM) on Master

- Nightly tests run on Chrysalis (E3SM machine)
  - 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