Redefining reproducibility for E3SM on multicore systems

Kate Evans, presenter

Plus ORNL project members:
Salil Mahajan (lead),
Matt Norman,
Joe Kennedy,
Min Xu

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

Thanks to the DOE OLCF and the DOE BER ACME-software modernization project for support!

E3SM v1 release, July 2018
Development for v2 is ongoing

- Model Components, each have new features in development:
  - **Atmosphere:** cloud microphysics, aerosols, variable resolution, etc. (EAM)
  - **Land:** biogeochemistry, soil hydrology, land units (ELM)
  - **Ocean:** dycore solvers, coupling to ice (MPAS-O)
  - **Land-ice**: new components (MALI, BISICLES)
  - Etc.
- All components have code updates in anticipation of new computing architectures
  - Code refactoring (Fortran + OpenACC & C++/Kokkos most common)
  - Consideration of new algorithms that favor less local memory, data transfer, efficiency

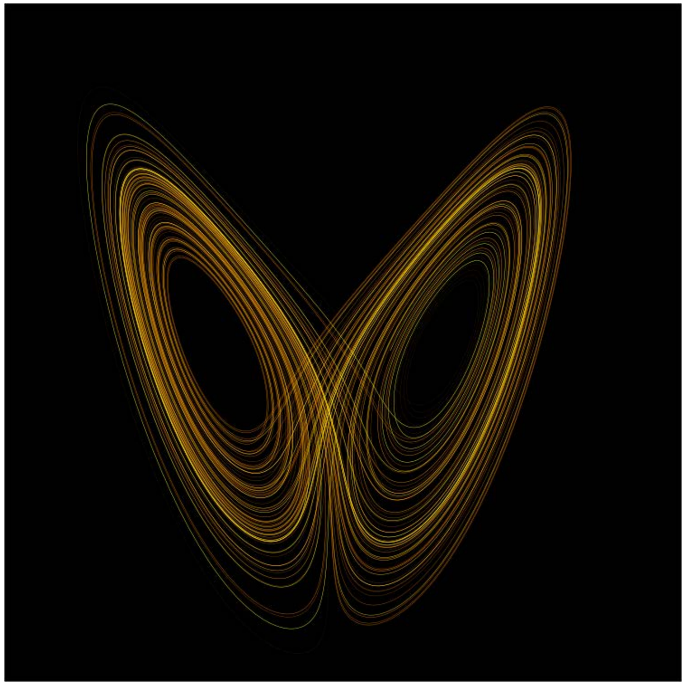
There are categories of code updates:

- Changes that **do no**t affect the climate and **should be** bit-for-bit reproducible
  - E.g. Adding a new compset, inclusion of new output variables
- Changes that **do not** affect the climate and **will not be** bit-for-bit reproducible
  - E.g. code porting, GPU kernel, etc.
  - Climate statistics are the same
- Changes that **do** affect the climate and **will not be** bit-for-bit reproducible
  - E.g. New parameterizations modules, new tunings
  - Climate statistics are not the same

Goal: Test the null hypothesis that climate simulation is "similar".

Motivation: Bit-for-bit is not achievable on target computing systems for E3SM

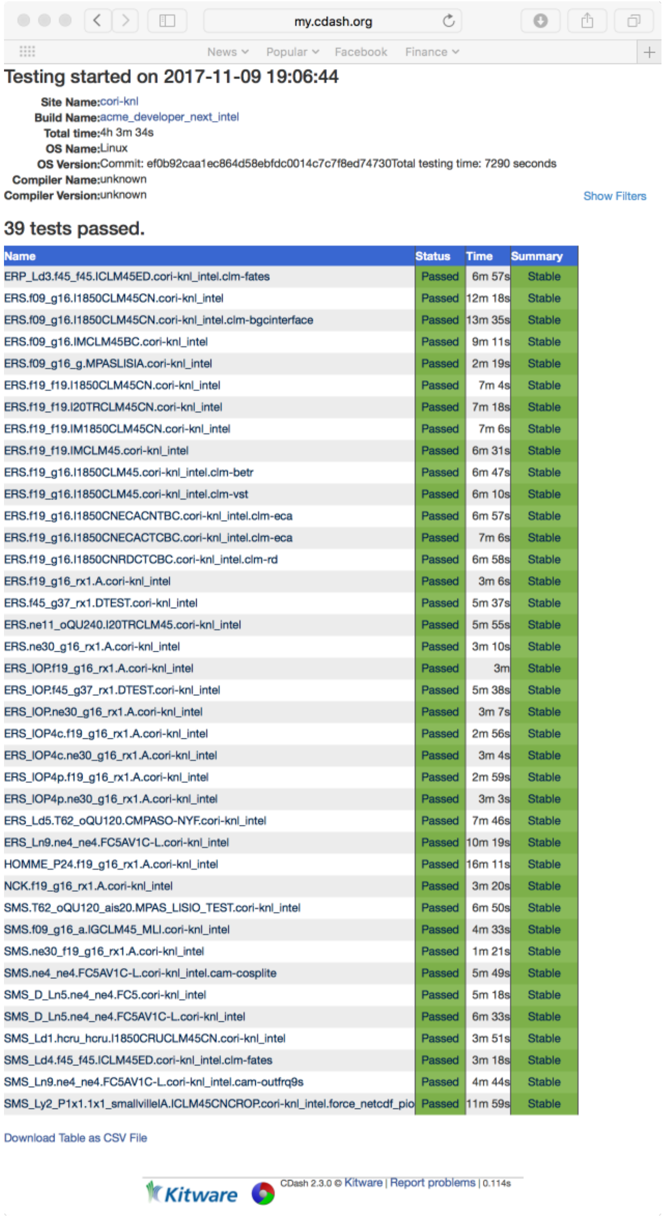
- Truncated Floating Point arithmetic:
  - Round-off differences
  - Non-associative:
    - $(-1 + 1) + 2^{-53} \neq -1 + (1 + 2^{-53})$
  - Optimizations, hybrid architectures, threading
- Climate models are chaotic and non-linear, so round-off differences grow quickly
- Goal: identify systematic bugs in a non-BFB reproducible environment **that allows for a reasonable development cycle**



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

Reproducibility Part 1: BFB

- E3SM Testing Suite:
  - * 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))



Testing started on 2017-11-09 19:06:44

Site Name:cori-knl
Build Name:acme_developer_next_intel
Total time:4h 3m 34s
OS Name:Linux
OS Version:Commit: ef0b92caa1ec864d58ebfdc0014c7c7f8ed74730Total testing time: 7290 seconds
Compiler Name:unknown
Compiler Version:unknown

Show Filters

39 tests passed.

Name	Status	Time	Summary
ERP_Ld3.f45_f45.ICLM45ED.cori-knl_intel.clm-fates	Passed	6m 57s	Stable
ERS.f09_g16.I1850CLM45CN.cori-knl_intel	Passed	12m 18s	Stable
ERS.f09_g16.I1850CLM45CN.cori-knl_intel.clm-bgcinterface	Passed	13m 35s	Stable
ERS.f09_g16.IMCLM45BC.cori-knl_intel	Passed	9m 11s	Stable
ERS.f09_g16_g.MPASLISIA.cori-knl_intel	Passed	2m 19s	Stable
ERS.f19_f19.I1850CLM45CN.cori-knl_intel	Passed	7m 4s	Stable
ERS.f19_f19.I20TRCLM45CN.cori-knl_intel	Passed	7m 18s	Stable
ERS.f19_f19.IM1850CLM45CN.cori-knl_intel	Passed	7m 6s	Stable
ERS.f19_f19.IMCLM45.cori-knl_intel	Passed	6m 31s	Stable
ERS.f19_g16.I1850CLM45.cori-knl_intel.clm-betr	Passed	6m 47s	Stable
ERS.f19_g16.I1850CLM45.cori-knl_intel.clm-vst	Passed	6m 10s	Stable
ERS.f19_g16.I1850CNECACNTBC.cori-knl_intel.clm-eca	Passed	6m 57s	Stable
ERS.f19_g16.I1850CNECACTCBC.cori-knl_intel.clm-eca	Passed	7m 6s	Stable
ERS.f19_g16.I1850CNRDCTCBC.cori-knl_intel.clm-rd	Passed	6m 58s	Stable
ERS.f19_g16_rx1.A.cori-knl_intel	Passed	3m 6s	Stable
ERS.f45_g37_rx1.DTEST.cori-knl_intel	Passed	5m 37s	Stable
ERS.ne11_oQU240.I20TRCLM45.cori-knl_intel	Passed	5m 55s	Stable
ERS.ne30_g16_rx1.A.cori-knl_intel	Passed	3m 10s	Stable
ERS_IOP.f19_g16_rx1.A.cori-knl_intel	Passed	3m	Stable
ERS_IOP.f45_g37_rx1.DTEST.cori-knl_intel	Passed	5m 38s	Stable
ERS_IOP.ne30_g16_rx1.A.cori-knl_intel	Passed	3m 7s	Stable
ERS_IOP4c.f19_g16_rx1.A.cori-knl_intel	Passed	2m 56s	Stable
ERS_IOP4c.ne30_g16_rx1.A.cori-knl_intel	Passed	3m 4s	Stable
ERS_IOP4p.f19_g16_rx1.A.cori-knl_intel	Passed	2m 59s	Stable
ERS_IOP4p.ne30_g16_rx1.A.cori-knl_intel	Passed	3m 3s	Stable
ERS_Ld5.T62_oQU120.CMPASO-NYF.cori-knl_intel	Passed	7m 46s	Stable
ERS_Ln9.ne4_ne4.FC5AV1C-L.cori-knl_intel	Passed	10m 19s	Stable
HOMME_P24.f19_g16_rx1.A.cori-knl_intel	Passed	16m 11s	Stable
NCK.f19_g16_rx1.A.cori-knl_intel	Passed	3m 20s	Stable
SMS.T62_oQU120_ais20.MPAS_LISIO_TEST.cori-knl_intel	Passed	6m 50s	Stable
SMS.f09_g16_a.IGCLM45_MLI.cori-knl_intel	Passed	4m 33s	Stable
SMS.ne30_f19_g16_rx1.A.cori-knl_intel	Passed	1m 21s	Stable
SMS.ne4_ne4.FC5AV1C-L.cori-knl_intel.cam-cosplite	Passed	5m 49s	Stable
SMS_D_Ln5.ne4_ne4.FC5.cori-knl_intel	Passed	5m 18s	Stable
SMS_D_Ln5.ne4_ne4.FC5AV1C-L.cori-knl_intel	Passed	6m 33s	Stable
SMS_Ld1.hcru_hcru.I1850CRUCLM45CN.cori-knl_intel	Passed	3m 51s	Stable
SMS_Ld4.f45_f45.ICLM45ED.cori-knl_intel.clm-fates	Passed	3m 18s	Stable
SMS_Ln9.ne4_ne4.FC5AV1C-L.cori-knl_intel.cam-outfrq9s	Passed	4m 44s	Stable
SMS_Ly2_P1x1.1x1_smallvilleIA.ICLM45CNCROP.cori-knl_intel.force_netcdf_pio	Passed	11m 59s	Stable

Download Table as CSV File

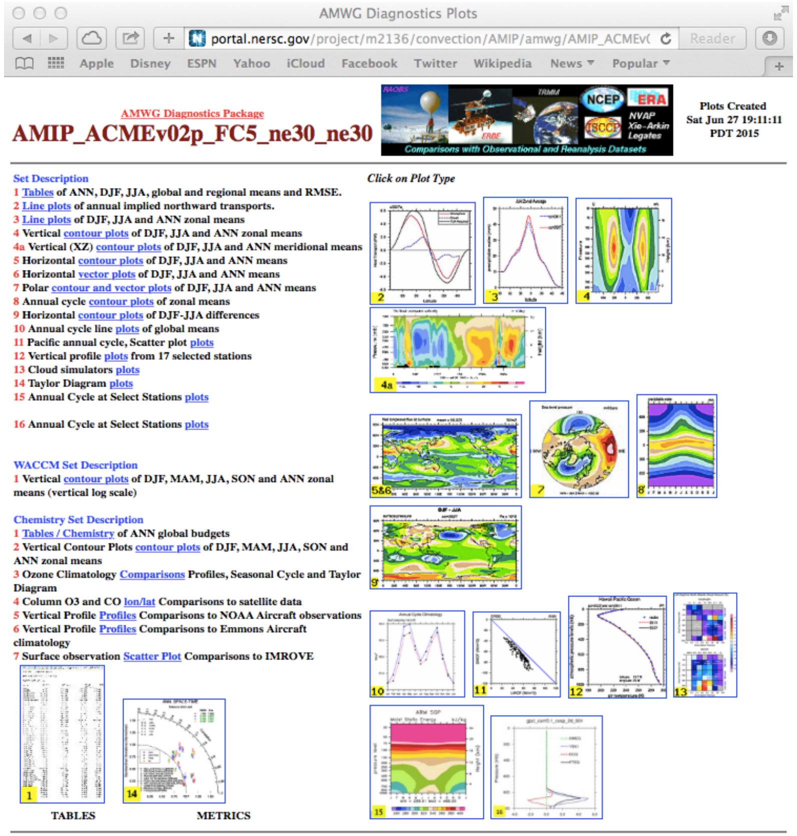
Kitware CDash 2.3.0 © Kitware | Report problems | 0.114s

The main thing that distinguishes legacy code from non-legacy code is tests, or rather a lack of tests. –*Michael Feathers*

Um… what if its not BFB?

Reproducibility Part 2: Expert Opinion

- Some years of a control run
  - scientifically validated on a trusted machine
- Some 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.
- Although: simpler models had less complexity, fewer multiscale features



AMWG Diagnostics Package

AMIP_ACMEv02p_FC5_ne30_ne30

Plots Created
Sat Jun 27 19:11:11
PDT 2015

Set Description
1 Tables of ANN, DJF, JJA, global and regional means and RMSE.
2 Line plots of annual implied northward transports.
3 Line plots of DJF, JJA and ANN zonal means
4 Vertical contour plots of DJF, JJA and ANN zonal means
4a Vertical (XZ) contour plots of DJF, JJA and ANN meridional means
5 Horizontal contour plots of DJF, JJA and ANN means
6 Horizontal vector plots of DJF, JJA and ANN means
7 Polar contour and vector plots of DJF, JJA and ANN means
8 Annual cycle contour plots of zonal means
9 Horizontal contour plots of DJF-JJA differences
10 Annual cycle line plots of global means
11 Pacific annual cycle, Scatter plot plots
12 Vertical profile plots from 17 selected stations
13 Cloud simulators plots
14 Taylor Diagram plots
15 Annual Cycle at Select Stations plots

16 Annual Cycle at Select Stations plots

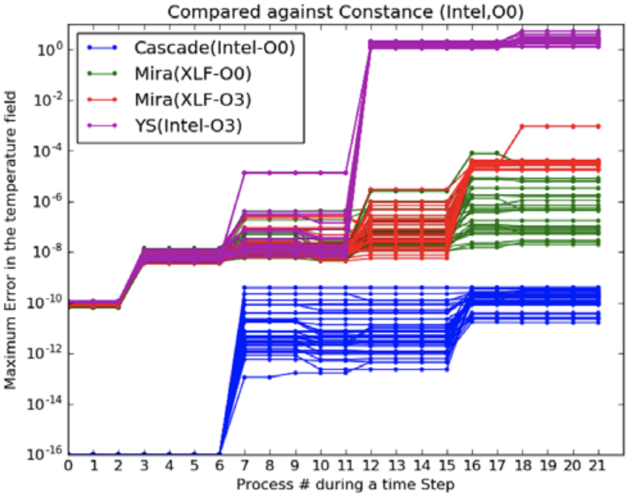
WACCM Set Description
1 Vertical contour plots of DJF, MAM, JJA, SON and ANN zonal means (vertical log scale)

Chemistry Set Description
1 Tables / Chemistry of ANN global budgets
2 Vertical Contour Plots contour plots of DJF, MAM, JJA, SON and ANN zonal means
3 Ozone Climatology Comparisons Profiles, Seasonal Cycle and Taylor Diagram
4 Column O3 and CO lon/lat Comparisons to satellite data
5 Vertical Profile Profiles Comparisons to NOAA Aircraft observations
6 Vertical Profile Profiles Comparisons to Emmons Aircraft climatology
7 Surface observation Scatter Plot Comparisons to IMROVE

Careers “made” on showing the climate is ”good enough” with new numerical dycores, packages, features (e.g. Evans et al 2013, 2014)

Isn’t there a better way?

- Perturbation growth test (B. Singh, PNNL)
  - a la Rosinski and Williamson (1997)
  - Remove branching/bugs/RNG issues
  - Only one time step, analyze by process
- Time step convergence test (H. Wan, PNNL)
  - Fast; only requires several time steps of data
  - Cannot track errors outside the code where convergence is assessed
- Statistical consistency test (A. Baker, NCAR)
  - Needing only a few time steps for almost all testing
  - Assesses total code output
  - Hard to determine location in code, but being addressed with a code search strategy



Max T (K) difference evolution in various computing environments. Process indices shown on x-axis refer to different physics parameterizations and/or Fortran code modules executed within one model time step. (courtesy, B. Singh, ACME-SM project proposal)

Ensemble Based Multivariate Approach

- Closest to original “expert” method in terms of set up (climate modelers stay in their happy place)
  - This has pros and cons, but means the code is tested just as it runs
- Can also be used for scientific analysis
  - Already being used to analyze long term atmospheric patterns, model sensitivity and UQ for sensor networks
- Suites of statistical tests can be applied.
  - But which ones are best?
- Some tests provide the geographic location of outliers

Ensemble Based Multivariate Approach

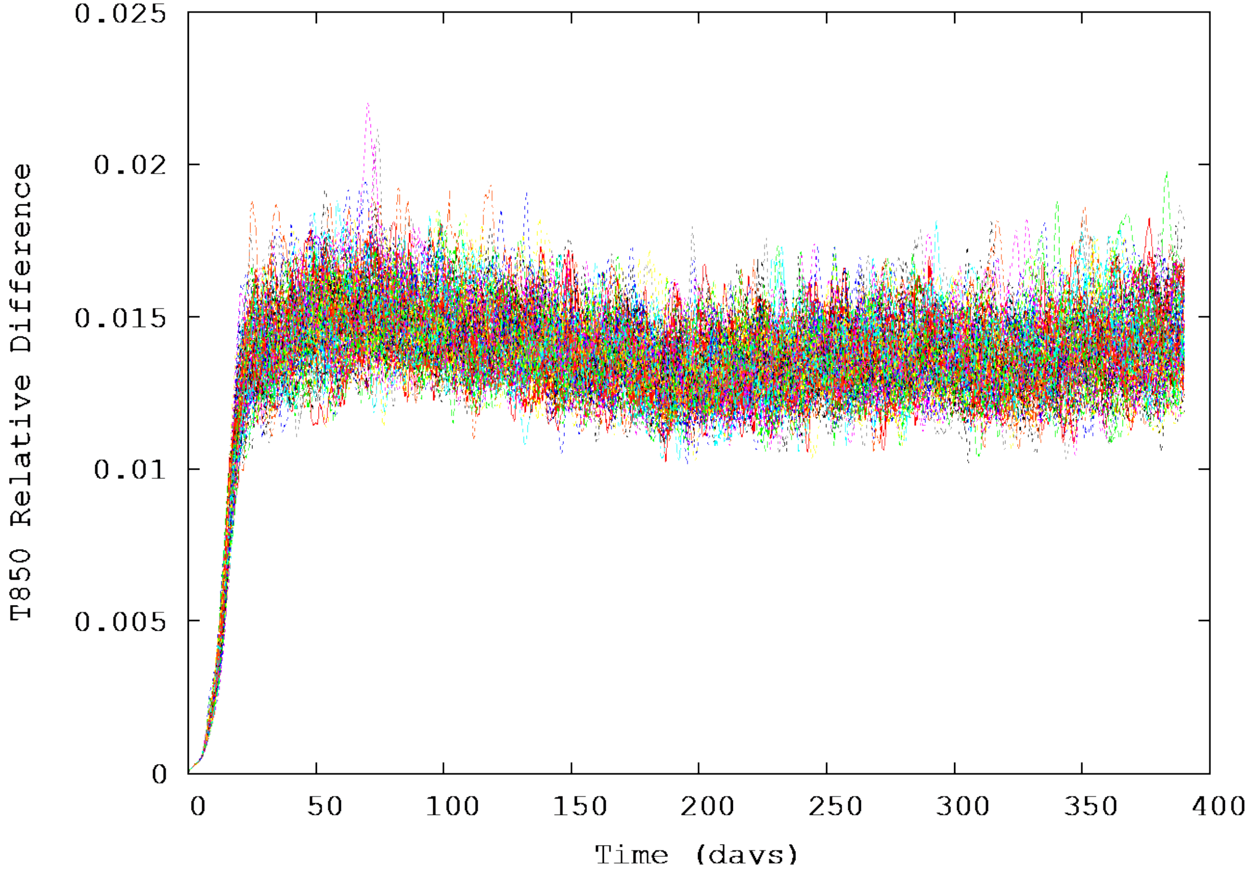
Goal: Accelerate and add rigor to the verification of E3SM for non-BFB changes

- Approach:
  - Ensemble vs. ensemble
  - Short (~1 year) ensembles of control and perturbed runs
- Short Ensembles:
  - Quantify natural variability, span possible climate states
  - Better utilizes multicore machines (*Mahajan et al., 2017)*
- Leverage two sample equality of distribution tests:
  - 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 Ensemble (SISE)

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

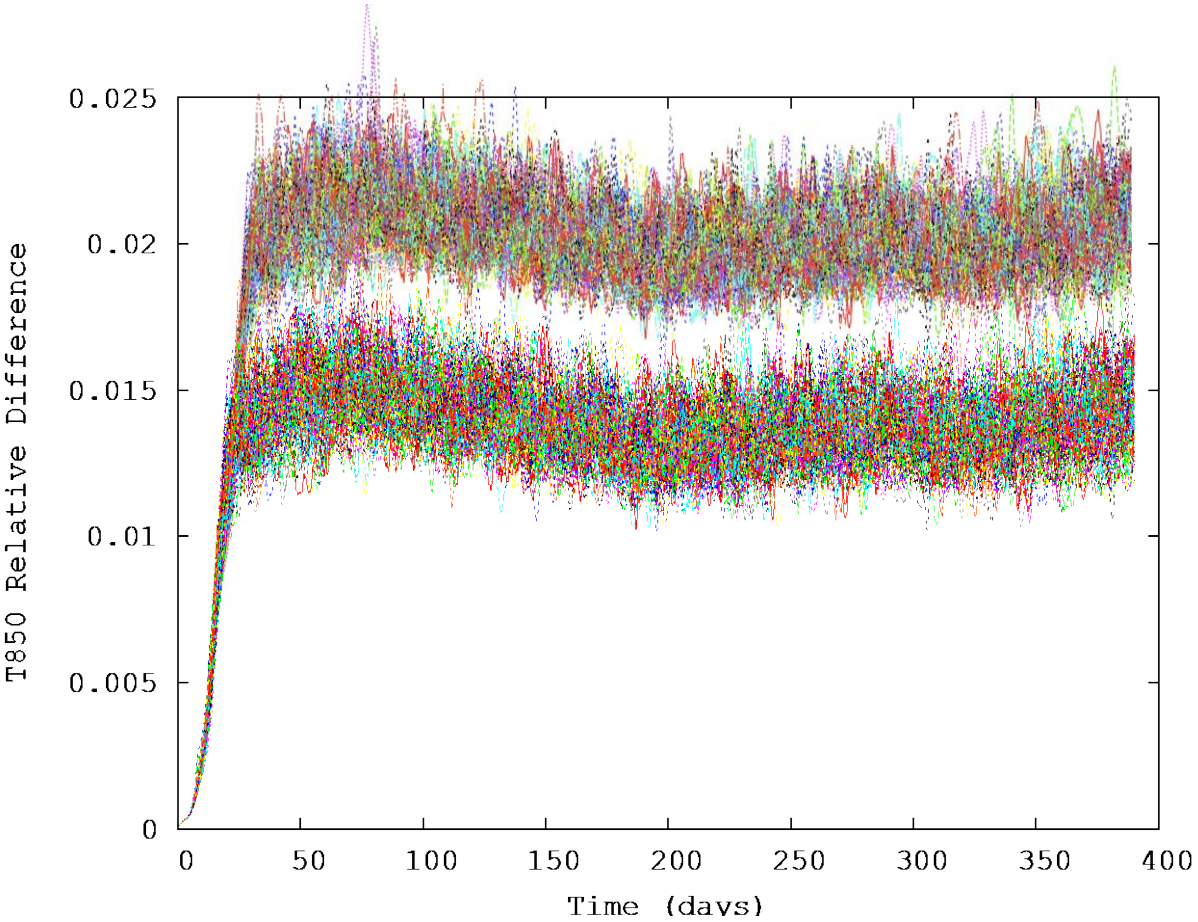
x' is uniform random number transformed to range from (-10^{-14}, 10^{-14})



L1-norm of the absolute differences for hourly 850mb T (Kelvin) *(Courtesy: Matt Norman)*

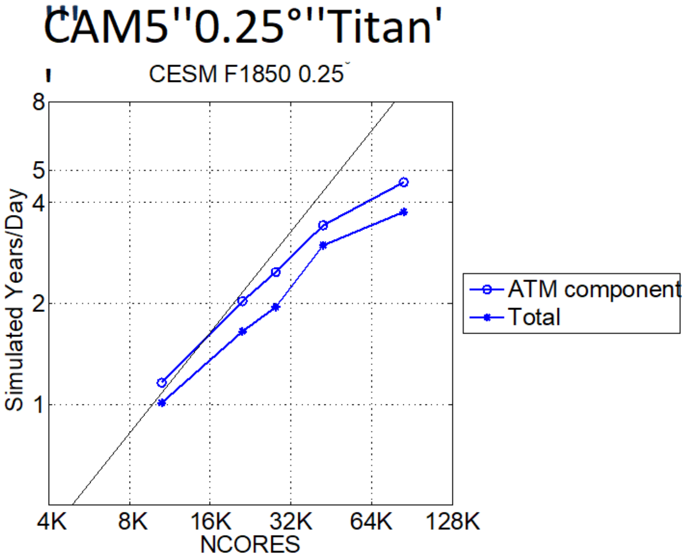
Short Independent Simulation Ensemble (SISE)

Problem to solve: Multivariate two sample equality of distribution testing for high dimension, low sample size



Packing simulations together is **economical** relative to a Single Long Run (SLR)

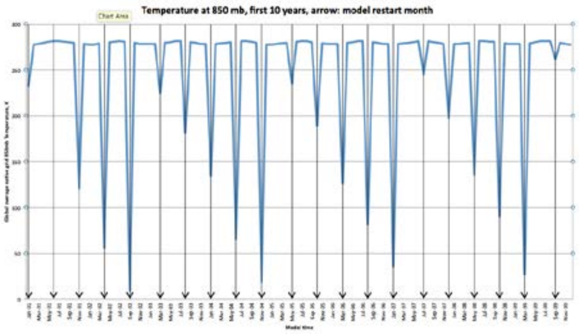
- Single Long run:
  - Less work per core with large core counts
  - Increase in MPI communications
  - Smaller MPI messages -> Large MPI latency
  - MPI cost > 90%
- 100 1-yr SISE vs. 100-yr long run
  - 100x greater workload per node on the same nodes
  - Faster throughput, and easy to use large core counts
    - Significantly reduced relative MPI and PCI-e overheads
    - Higher priority (the cool kids queue) on leadership class machines (e.g. Titan, Cori, etc.)



Strong scaling of a single long run. Courtesy: Mark Taylor and more, circa ~2012

Short Independent Simulation Ensemble (SISE)

- Example: EAM (E3SM atmosphere spectral element) two degree component:
  - SLR (100 years): 1536 elements given 96 nodes, 16 elements per node, takes **weeks** to finish
  - SISE (100 1yr runs): 1536 elements given 48 nodes each, 32 elements per node (total nodes: 4800), takes **less than a day** to finish
  - Took a while to analyze for success, we kept finding new bugs!
    - Random number generator was not so random
    - Restart bug for submonthly configurations for 3D variables



Test: Equality of Distribution

Kolmogorov Smirnov (KS) testing framework:

- Null Hypothesis (H_0): Two simulation ensembles (SISE) represent the same climate state.
- Use **global annual means** of all standard model output variables (**158 variables**)
- H_0: A variable between the two SISEs 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.

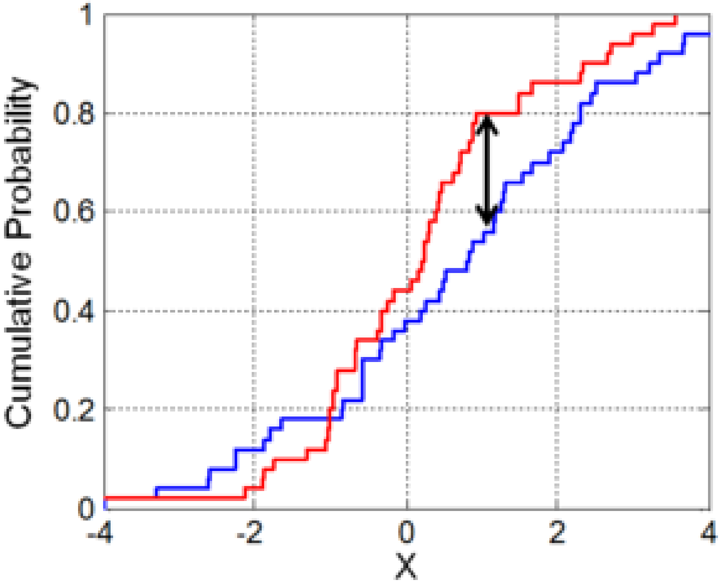
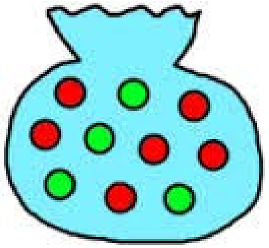
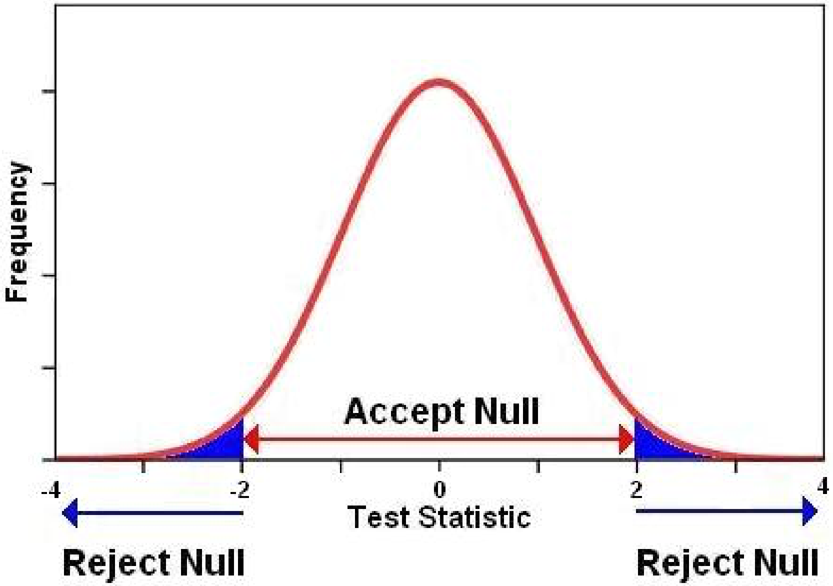


Illustration: KS test

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.





OAK RIDGE National Laboratory

Open slide master to edit

Model Verification Using SISE: E3SM v1

Known Climate Changing Perturbation

- 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 $(-10^{-14}, 10^{-14})$

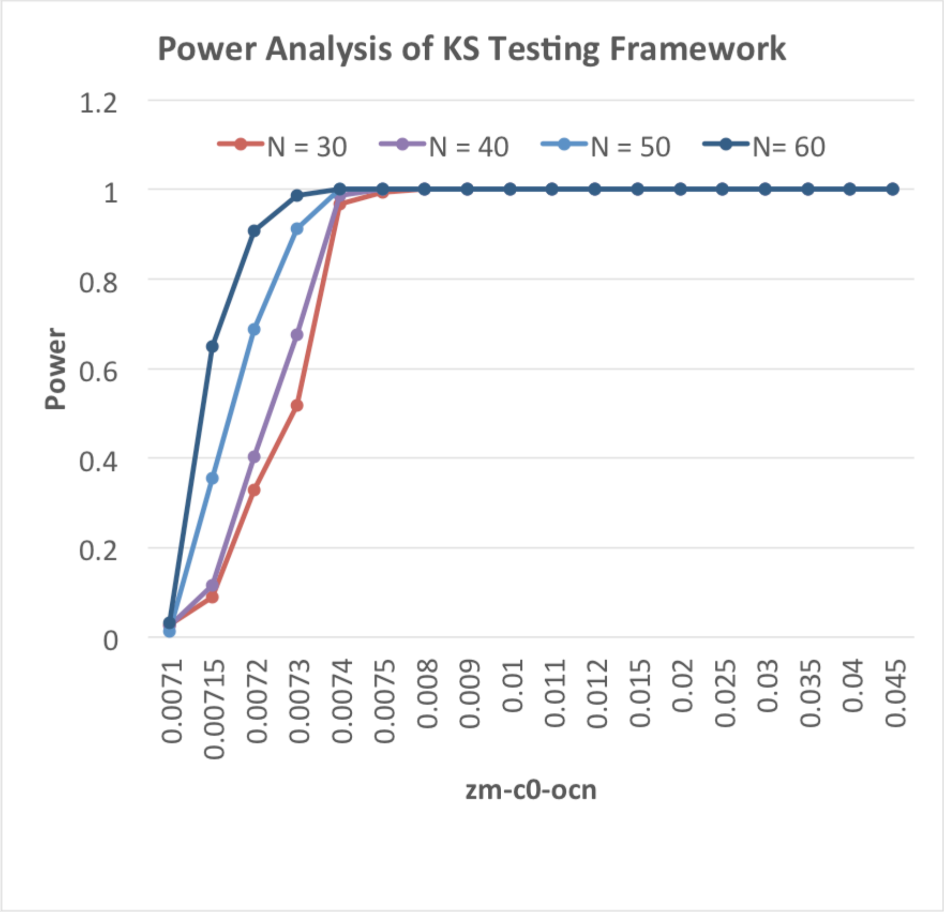
- Turn a tuning parameter knob: zm_c0_ocn (control case: 0.007, modified: 0.045)

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

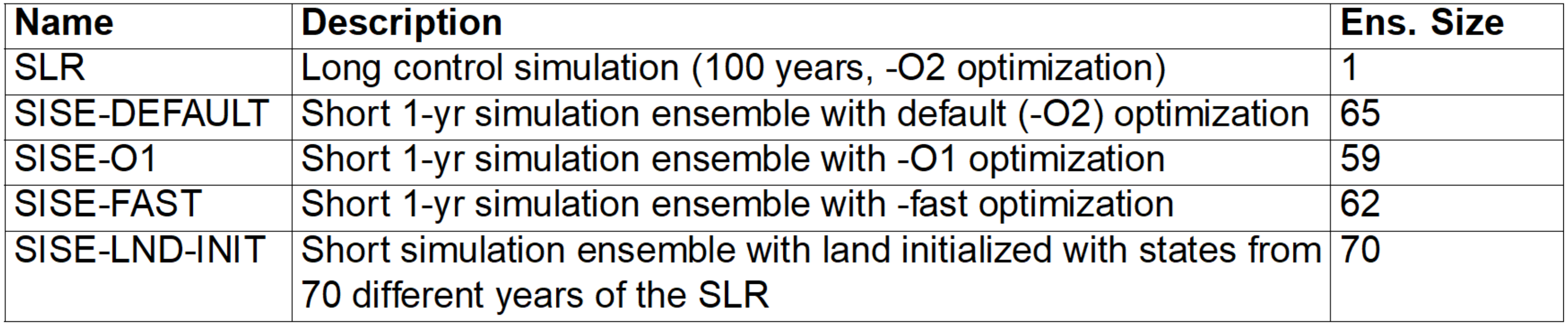
Power Analysis: KS Testing Framework



Fewer ensembles mean less sensitivity. How well do we know how sensitive the world is to changes on forcing?

Model Verification Using SISE: Compiler optimization choices with E3SM v0

- Configuration: Preindustrial, active atmosphere (CAM5) and land (CLM4)
- Spatial Resolution: 208km at the equator (2 degrees), 30 vertical layers
- Machine Configuration: PGI compiler on Titan
- Ensembles: Machine-precision level random perturbations to the initial 3-D temperature field



Name	Description	Ens. Size
SLR	Long control simulation (100 years, -O2 optimization)	1
SISE-DEFAULT	Short 1-yr simulation ensemble with default (-O2) optimization	65
SISE-O1	Short 1-yr simulation ensemble with -O1 optimization	59
SISE-FAST	Short 1-yr simulation ensemble with -fast optimization	62
SISE-LND-INIT	Short simulation ensemble with land initialized with states from 70 different years of the SLR	70

Compiler optimization choices

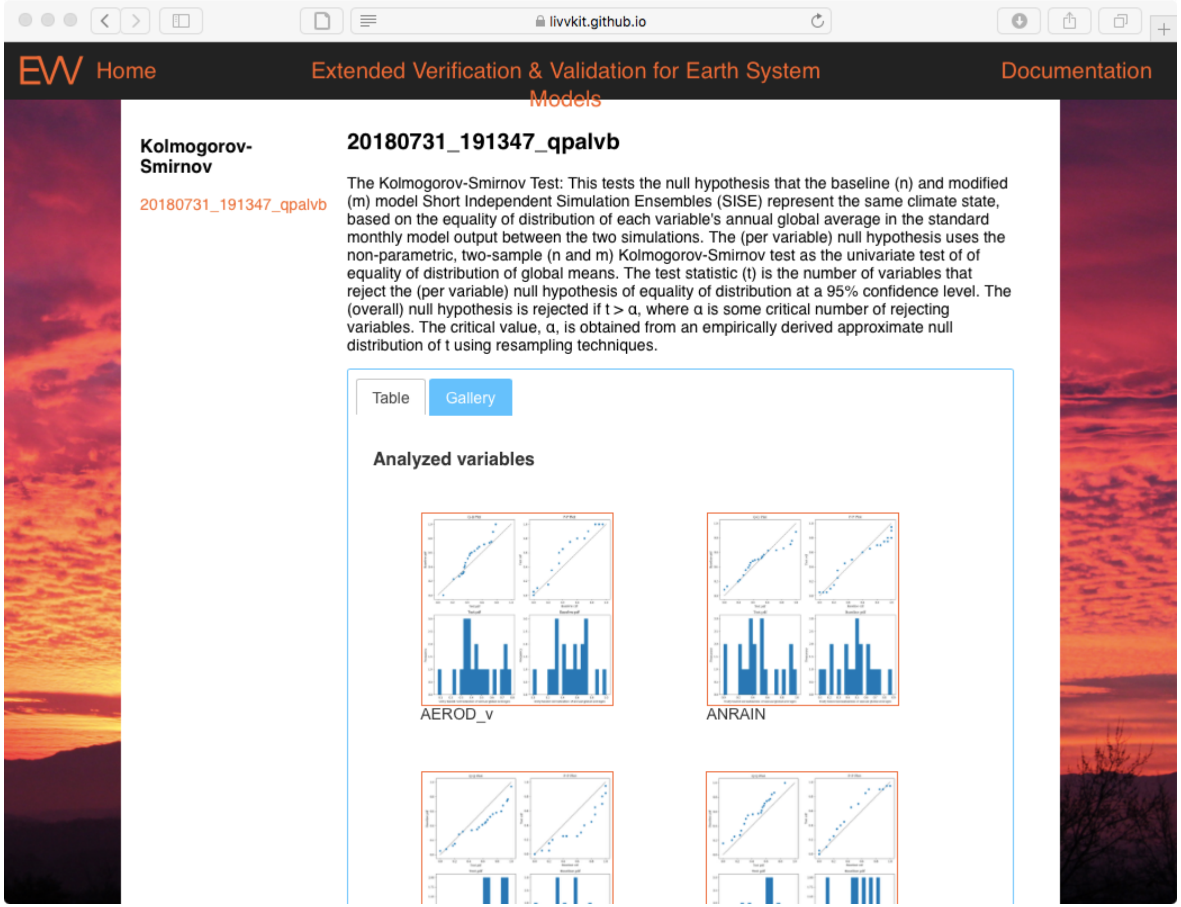
KS Testing Framework Results

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

Aggressive compiler choices (SISE-FAST) with the PGI compiler on Titan **can** result in climate-changing simulations.

Extended Verification and Validation for E3SM:

- Python based toolkit:
  - Runs control and perturbed ensembles
  - Post-processes model output
  - Conducts tests
  - Publishes results and auxiliary plots, tables



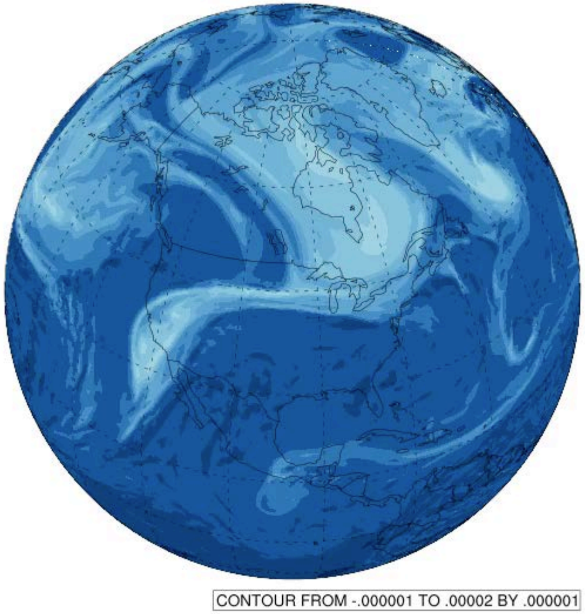
20180731_191347_qpalvb

The Kolmogorov-Smirnov Test: This tests the null hypothesis that the baseline (n) and modified (m) model Short Independent Simulation Ensembles (SISE) represent the same climate state, based on the equality of distribution of each variable's annual global average in the standard monthly model output between the two simulations. The (per variable) null hypothesis uses the non-parametric, two-sample (n and m) Kolmogorov-Smirnov test as the univariate test of of equality of distribution of global means. The test statistic (t) is the number of variables that reject the (per variable) null hypothesis of equality of distribution at a 95% confidence level. The (overall) null hypothesis is rejected if t > α, where α is some critical number of rejecting variables. The critical value, α, is obtained from an empirically derived approximate null distribution of t using resampling techniques.

Analyzed variables

Summary:

- Short runs and ensembles are the only viable path for model verification as model expense grows
- A multivariate testing framework (EVE) is presented for climate reproducibility:
- We demonstrated this with known climate changing perturbations (and provided detection limits), choice of compiler optimization, and verifying how frozen the model was after months of software updates
- Future work:
  - Evaluate applicability of low-resolution results at high-resolution
  - Apply to shorter runs (monthly and daily vs. yearly)
  - Optimize multivariate tests, e.g. use different kernel functions, distance metrics

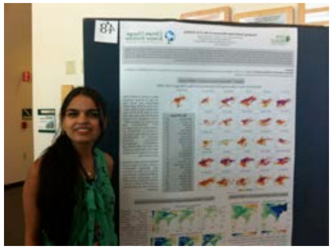


Potential vorticity at 300mb, E3SM present day test run

We are hiring! If you have expertise* in one or more of the following I would like to talk to you

- A passion for coding for >petaflop systems
- Understanding of modeling the atmosphere
  - Dynamics
  - clouds
- Understanding of modeling the ocean
- Understanding software ecosystems and good habits
- Diversity in every dimension

* i.e. ninjas



Model Verification Using SISE: 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
- Ensembles: Machine-precision level random perturbations to the initial 3-D temperature field

- 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

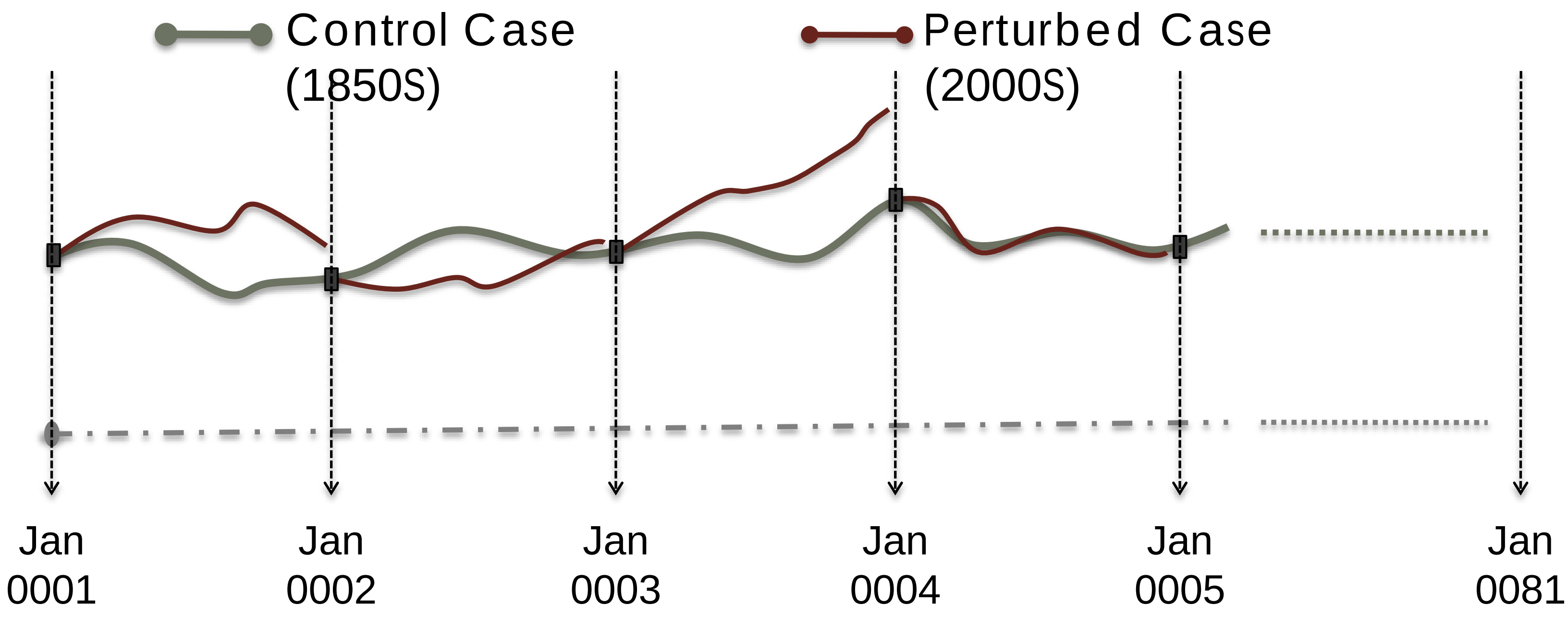
Name	Ens. Size	CIME	PGI	p-netcdf
v0.4-2015	30	4.0	15.3	1.5.0
master	30	5.0	17.5	1.7.0
v0.4	27	4.0	17.5	1.7.0

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!

Short Ensembles: Scientific Utility



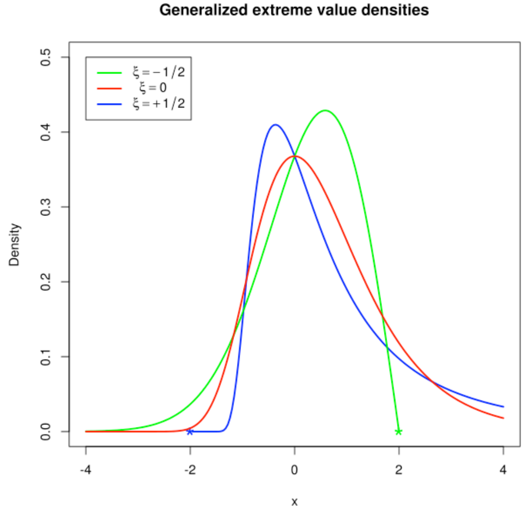
Fast Response

SST (2000S – 1850S)

Precipitation (2000S – 1850S)

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

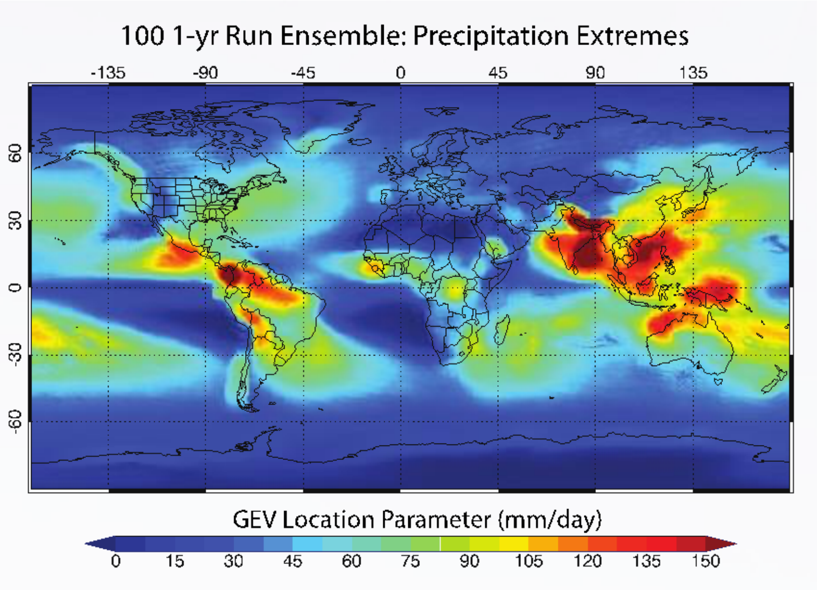


All with μ = 0, σ = 1. Asterisks mark support-endpoints

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

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

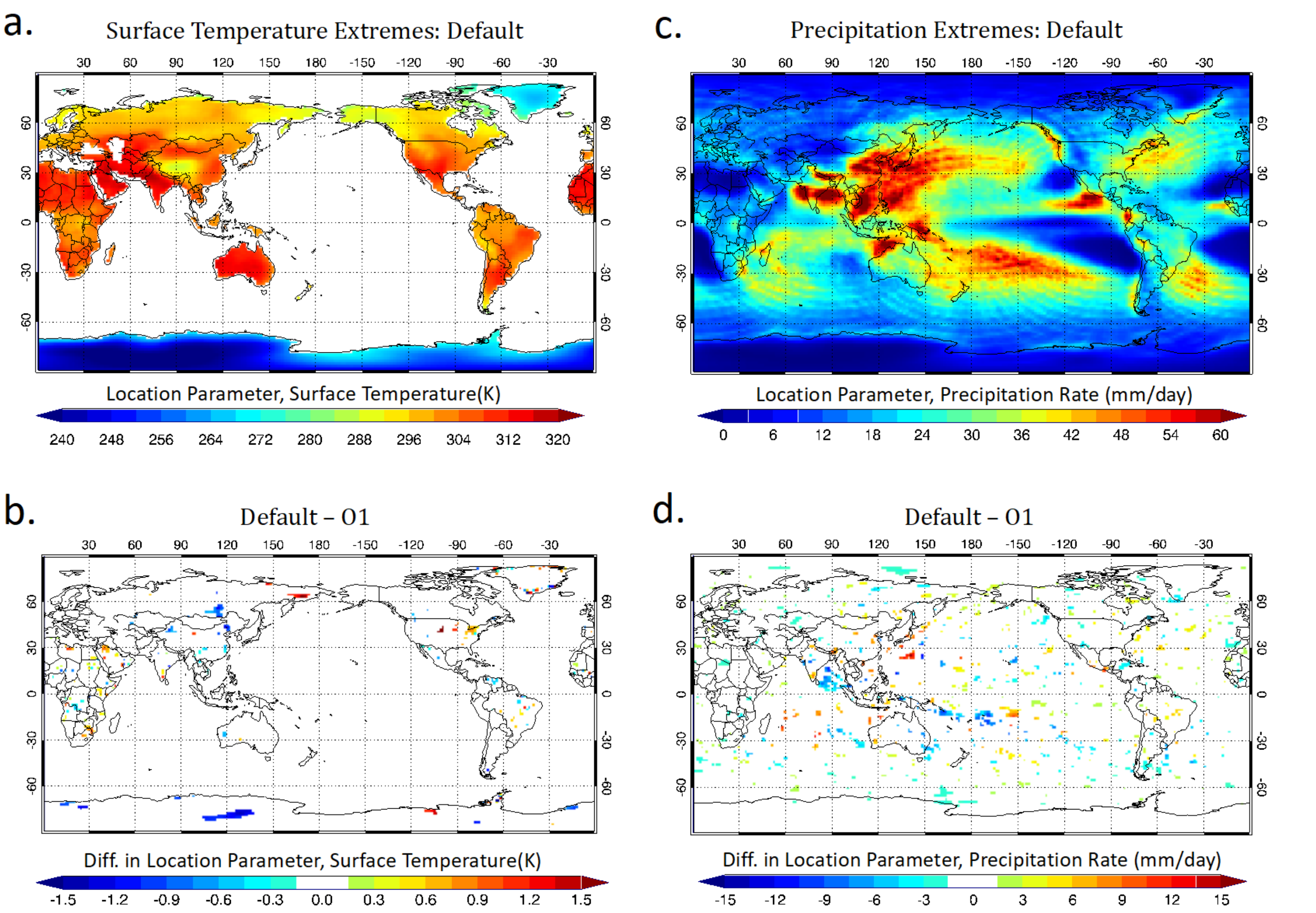
where μ, σ and ξ represent the location, scale and shape parameter respectively.


100 1-yr Run Ensemble: Precipitation Extremes
-135
-90
-45
0
45
90
135
60
30
0
-30
-60
GEV Location Parameter (mm/day)
0
15
30
45
60
75
90
105
120
135
150

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.

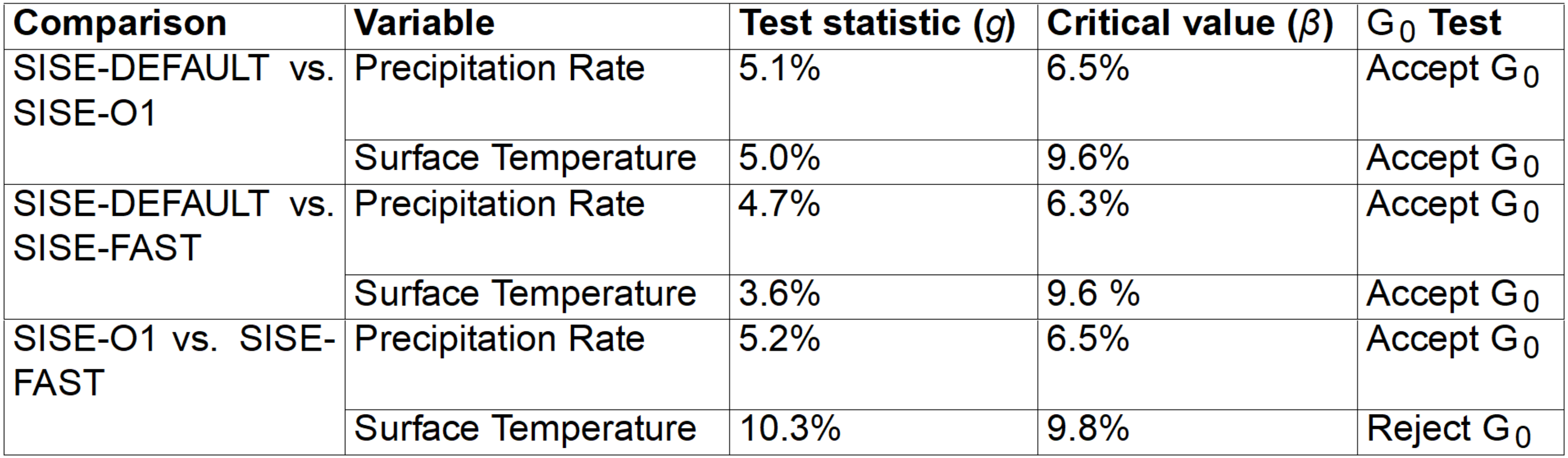
Climate Extremes



O
Nat

to edit

Climate Extremes



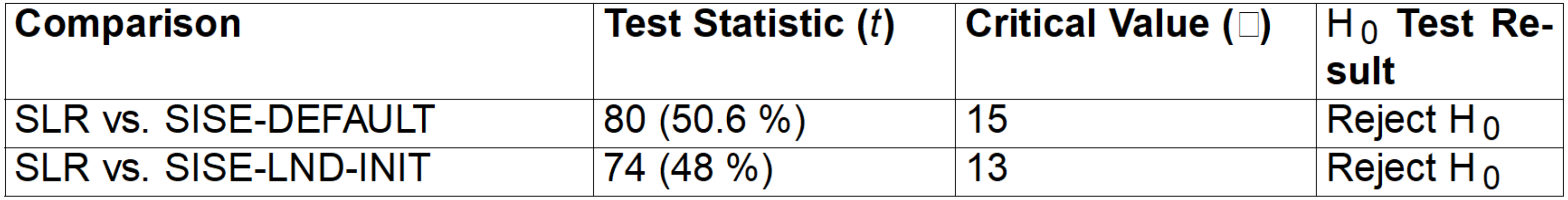
Comparison	Variable	Test statistic (*g*)	Critical value (*β*)	G_0 Test
SISE-DEFAULT vs. SISE-O1	Precipitation Rate	5.1%	6.5%	Accept G_0
	Surface Temperature	5.0%	9.6%	Accept G_0
SISE-DEFAULT vs. SISE-FAST	Precipitation Rate	4.7%	6.3%	Accept G_0
	Surface Temperature	3.6%	9.6 %	Accept G_0
SISE-O1 vs. SISE-FAST	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
- Climate extremes are not a good metric to evaluate answer changes that might effect the simulation of the climate, with 60 ensemble members.

Single Long Run (SLR) vs. SISE

- SLR is clearly distinct from the SISE-DEFAULT

KS Testing Framework Results

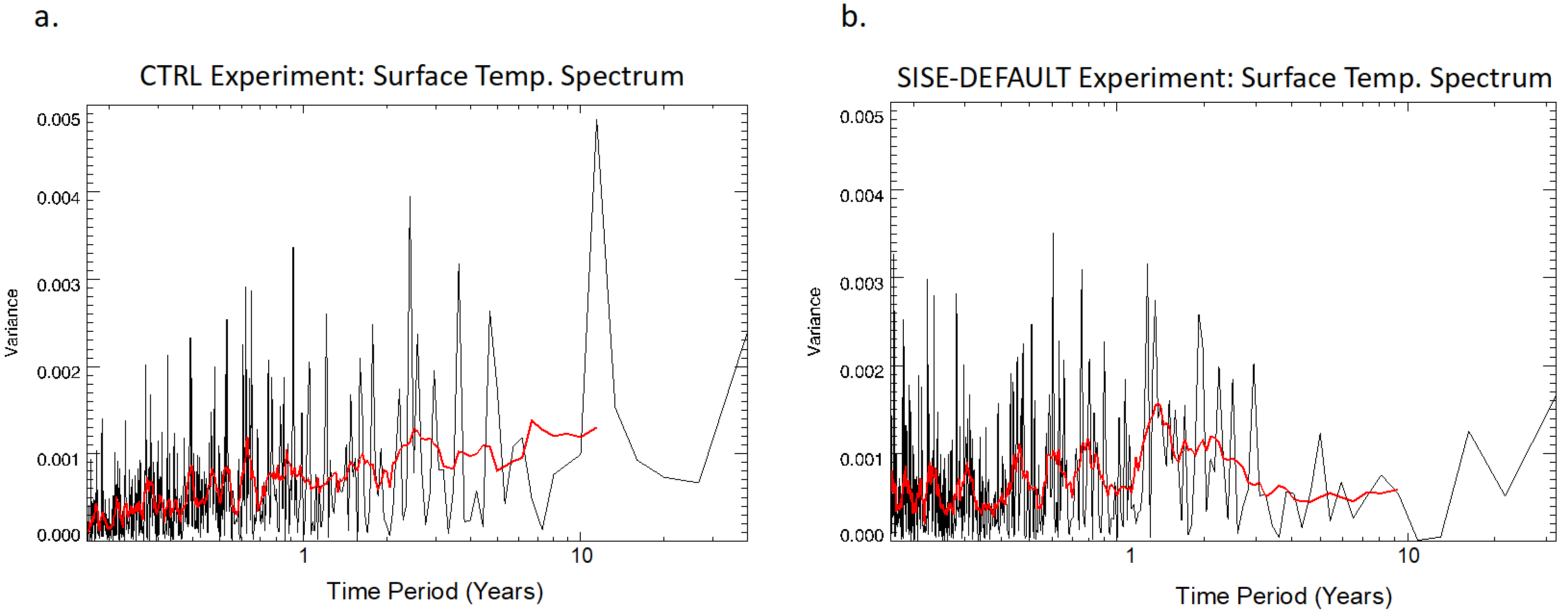


Comparison	**Test Statistic (*t*)**	**Critical Value (□)**	H_0 **Test Result**
SLR vs. SISE-DEFAULT	80 (50.6 %)	15	Reject H_0
SLR vs. SISE-LND-INIT	74 (48 %)	13	Reject H_0

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*).

- This low frequency variability is not captured by SISE.



Atmospheric Low-frequency Variability

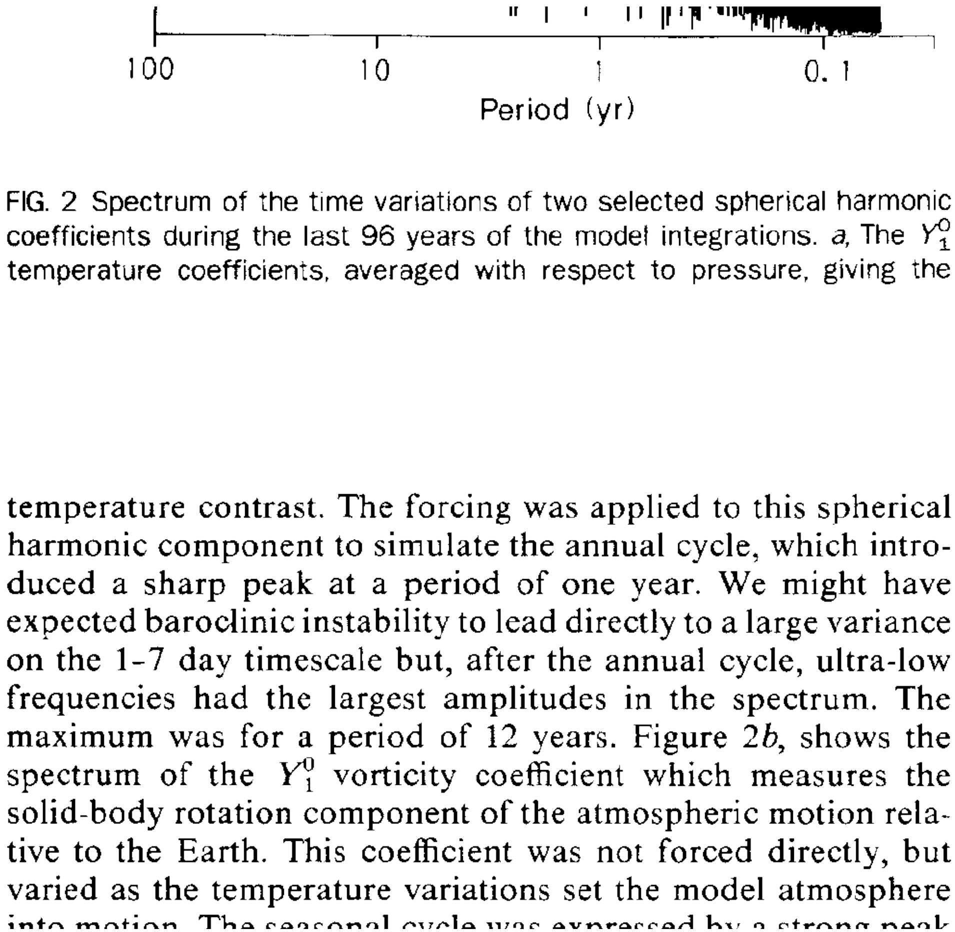

100
10
1
0. 1
Period (yr)

FIG. 2 Spectrum of the time variations of two selected spherical harmonic coefficients during the last 96 years of the model integrations. *a*, The Y_1^0 temperature coefficients, averaged with respect to pressure, giving the

temperature contrast. The forcing was applied to this spherical harmonic component to simulate the annual cycle, which introduced a sharp peak at a period of one year. We might have expected baroclinic instability to lead directly to a large variance on the 1-7 day timescale but, after the annual cycle, ultra-low frequencies had the largest amplitudes in the spectrum. The maximum was for a period of 12 years. Figure 2*b*, shows the spectrum of the Y_1^0 vorticity coefficient which measures the solid-body rotation component of the atmospheric motion relative to the Earth. This coefficient was not forced directly, but varied as the temperature variations set the model atmosphere into motion. The seasonal cycle was expressed by a strong peak

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):
  - No. of pairs with one control run and one perturbed run
- Test the null hypothesis using the exact null distribution

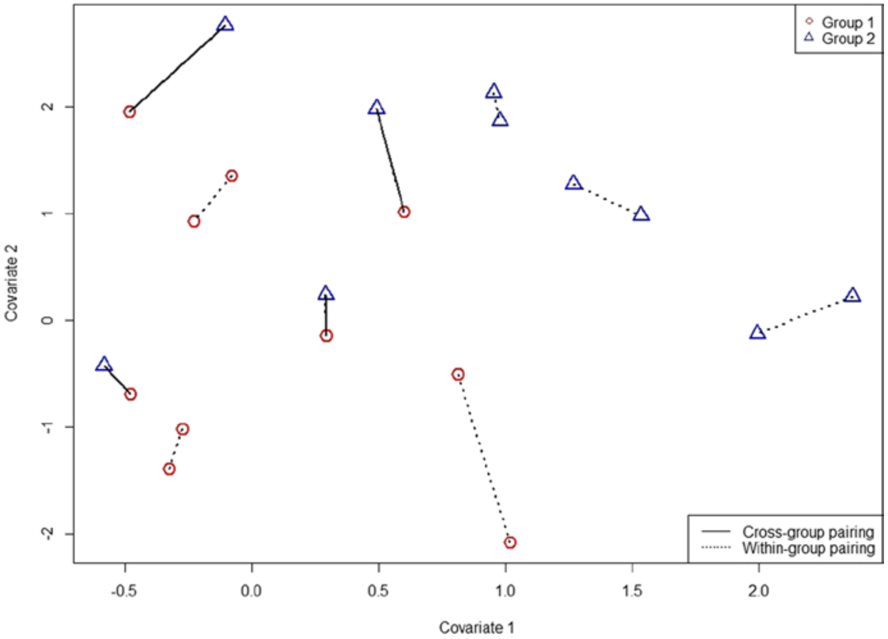
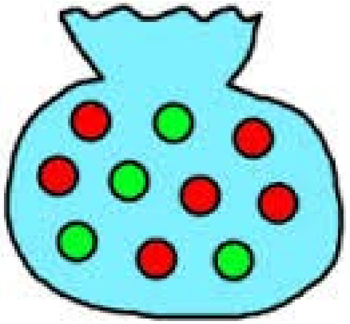


Illustration of cross matching for a bivariate case with $n = m = 10$. *(Ruth, 2014)*

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
  - For e.g. for $n = m = 9$, $P(a_1 \leq 1) = 0.0259$