OAK RIDGE National Laboratory

Machine Learning Approaches to Ensure Statistical Reproducibility of MPAS-O

Salil Mahajan, Michael Kelleher, Andy Salinger

Oak Ridge National Laboratory

Sandia National Laboratory

ESMD-E3SM PI Meeting 2020

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

U.S. DEPARTMENT OF ENERGY

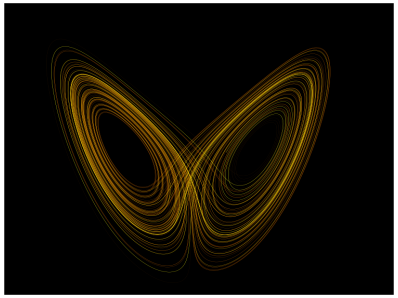
Motivation:

- E3SM Software and Algorithms NGD goals:
  - 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.

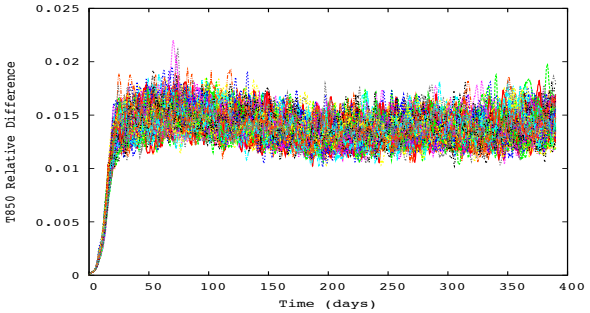
- Goal: Test the null hypothesis that climate simulation remains statistically equivalent after unintended non-b4b changes.

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.



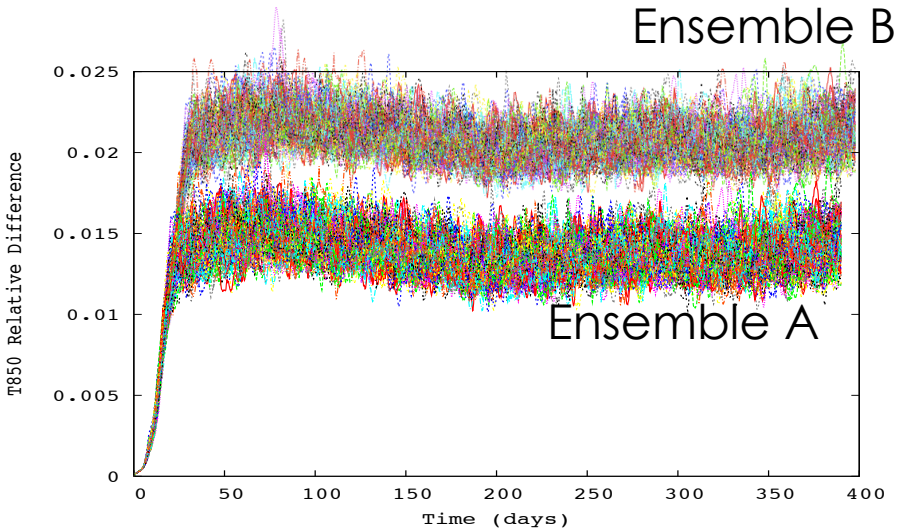
Lorenz attractor
(*Source:en.wikipedia.org/wiki/Chaos_theory*)



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.

ML for Two Sample Testing Using Ensembles

- Approach:
  - Evaluate statistics of the perturbed ensemble vs. control ensemble after propagation of errors from machine precision differences in initial conditions.
  - Short (2yr for MPAS-O) ensembles
- Problem statement: Multivariate two sample equality of distribution testing for:
  - High dimensions
  - Low sample sizes
- Use ML approaches for two sample equality of distribution tests.

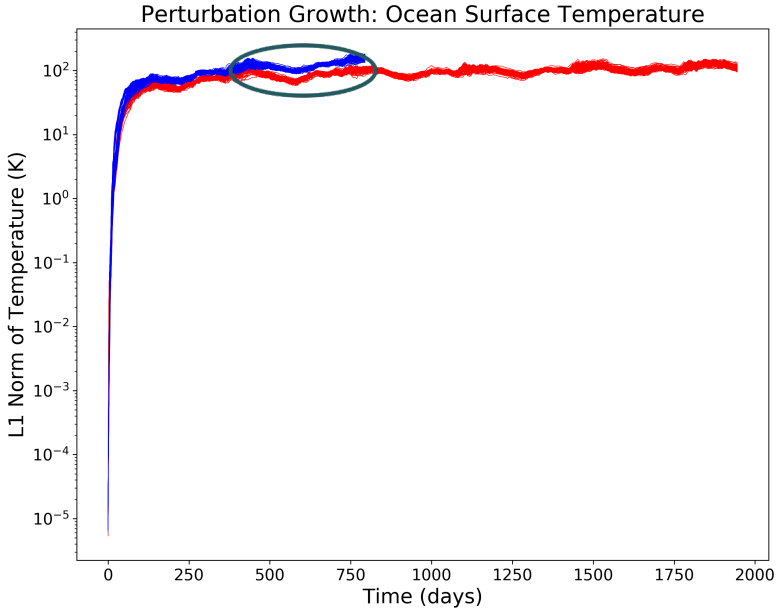


MPAS-O Reproducibility tests: Approach

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

- Generate control and perturbed ensembles at QU240 resolution (7153 cells)
- Evaluate 5 prognostic variables (Baker et al. 2016)
  - SSH, T, U, V, Salinity
  - Annual average of year 2, as error growth converges.
- Ocean variability is spatially very heterogenous (as compared to the atmosphere):
  - So, we 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.
  - Permutation testing

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.

MPAS-O Reproducibility Tests: Approach

Type I error rate (False Positive Rate): Probability of falsely rejecting a null hypothesis

- *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

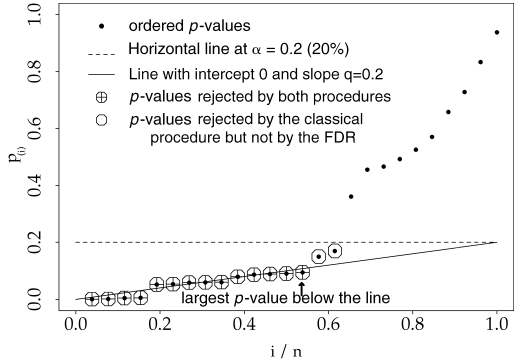


FIG. 2. Illustration of the traditional FPR and FDR procedures on a stylized example, with $q = \alpha = 20\%$. The ordered *p*-values, $p_{(i)}$, are plotted against i/n, $i = 1, \ldots, n$, and are circled and crossed to indicate that they are rejected by the FPR and FDR procedures, respectively.

Ventura et al. 2004

OAK RIDGE National Laboratory

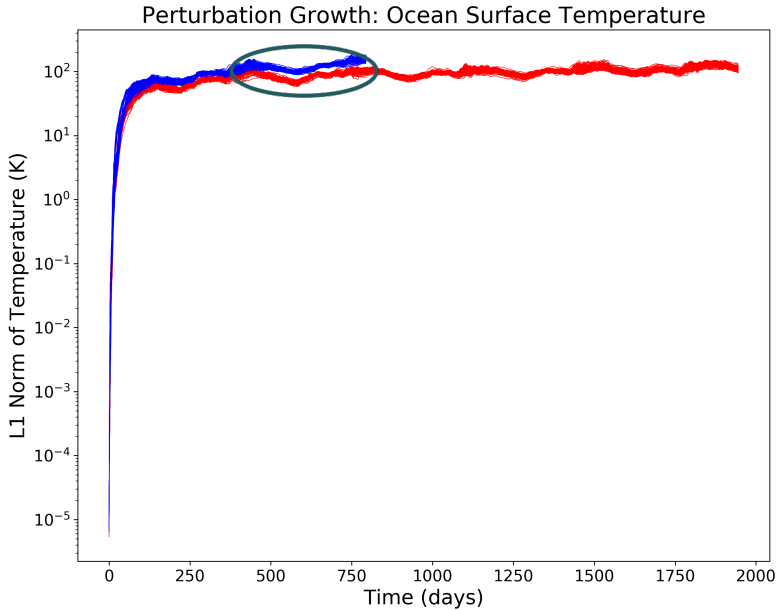
MPAS-O Reproducibility Tests

Evaluation of 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
- 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

MPAS-O Reproducibility Tests: Test Case

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



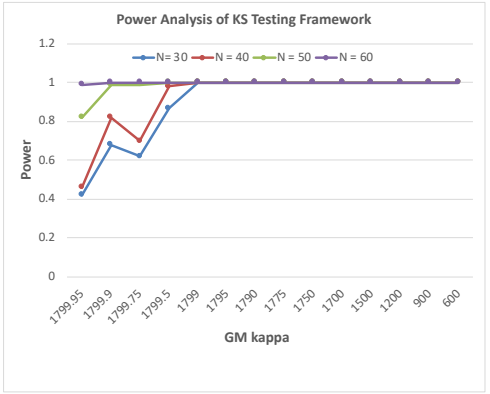
Growth of machine precision differences in oQU240 MPAS-O and ensemble spread: L1 Norm (sum of absolute difference at each grid point, log-scale) of SST of each of the 100 ensemble members with round off differences in initial conditions compared to a reference run for the control (kappa = 1800, red lines)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

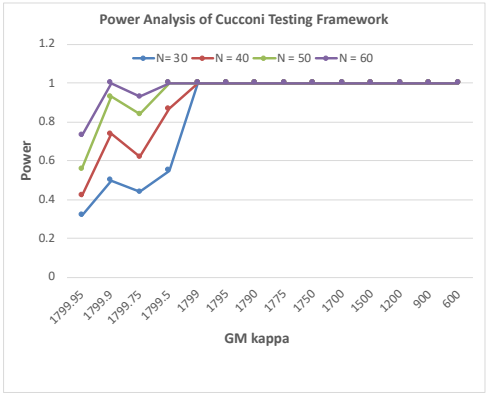
Both KS and Cucconi tests reject the null hypothesis that the two ensembles belong to the same population at the 0.05 significance level.

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
  - Initial condition at each grid box, j, of each ensemble member perturbed as:
    - $T'_j = (1+x')T_j$, 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)
- Result:
  - Both tests can catch small differences in GM Kappa with high confidence. For example, the tests with 30 member ensembles can detect changes in GM Kappa from 1800 to 1799.





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 using ML techniques as E3SM adapts for Exascale
- Developed a ML based multivariate testing framework for climate reproducibility for MPAS-O using short ensembles:
  - To be ported to EVV – the reproducibility testing framework for EAM.
- Test Cases:
  - Correctly detects known climate changing perturbations by tuning parameter changes
  - Working with developers of a new implementation of the barotropic solver in MPAS-O to ensure climate reproducibility.
- Power Analysis:
  - Both the KS test and the Cucconi testing frameworks can catch small changes to tuning parameters with high confidence, with increasing power with increasing number of ensemble members
  - Provides a framework to developers to evaluate impact of non-b4b changes.

Acknowledgements:
E3SM, NERSC, OLCF