Improvements in Reproducibility Testing Through False Discovery Rate Correction

Michael E. Kelleher, Salil Mahajan

10 November 2023

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

Problem Introduction

- Evaluate impact of code changes on simulated climate with E3SM
- Nightly testing suite: hundreds of individual tests across multiple machines
- Most evaluate bit-for-bit reproducibility
- Non-bit-for-bit tests evaluate if a change has modified the simulated climate
- This includes the multivariate Kolmogorov-Smirnov (MVK) test (Mahajan et al., 2019)

Methods Introduction

- The multivariate Kolmogorov-Smirnov (MVK) test compares two "short" independent ensembles
- Each is a 30-member ensemble of 14-month low resolution simulations.
- A baseline is generated after each approved "climate changing" code modification
- A test ensemble is performed each night, then a comparison is done
- Software packages evv4esm and LIVVkit perform the data analysis and create a user friendly web page of the results

Methods Introduction

- The test performed is the Kolmogorov-Smirnov test: comparison of two CDFs
- This test is used on 120 variables output by the E3SM Atmosphere Model (EAM)
- 150 member ensembles were conducted (with E3SM v1)
- Power analysis used to determine a threshold: number of statistically significant different variables to determine a "changed climate"

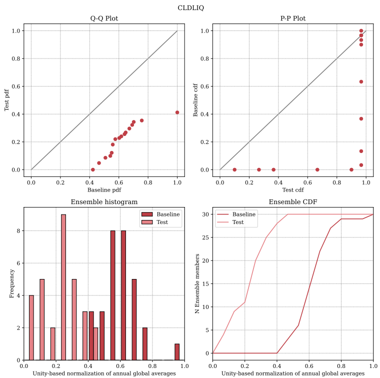

Figure 1: Rejected variable: CLDLIQ, Grid box averaged cloud liquid amount [kg/kg]

Methods Introduction

- The test performed is the Kolmogorov-Smirnov test: comparison of two CDFs
- This test is used on 120 variables output by the E3SM Atmosphere Model (EAM)
- Power analysis used to determine a threshold: number of statistically significant different variables to determine a "changed climate"
- 150 member ensembles were conducted (with E3SM v1)

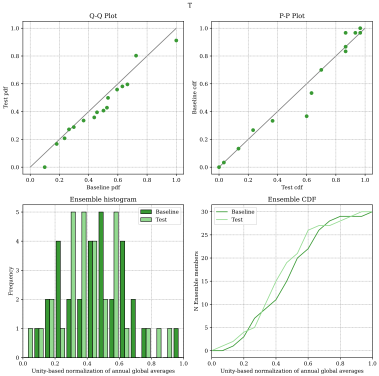

Figure 2: Accepted variable: T, Temperature [K]

Operational results

- Problem: several nightly tests which are bit-for-bit are above the failure threshold, and thus are incorrectly identified as climate changing
- Solution 1: Make each nightly ensemble use same set of seeds
- Solution 2: Use FDR correction to account for multiple tests

Operational results

- Problem: several nightly tests which are bit-for-bit are above the failure threshold, and thus are incorrectly identified as climate changing
- ~~Solution 1: Make each nightly ensemble use same set of seeds~~
- Solution 2: Use FDR correction to account for multiple tests

Operational results

Solution 2: Use FDR correction to account for multiple tests

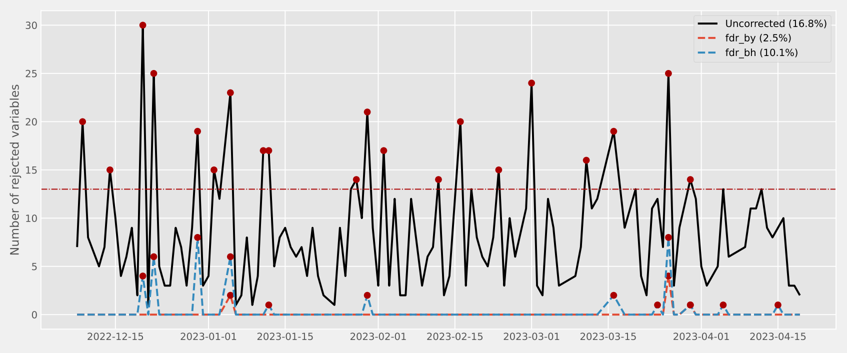

Figure 3: Number of tests with a global rejection by date

Ensemble Setup

- Can we do away with bootstrapping a large ensemble to find a threshold?
- Start by…generating a large ensemble
  - Using the same setup and simulation duration as operational tests
  - "Ultra-low" resolution: 7.5° atmosphere / 240 km ocean, 14 month simulation
  - Each variation has a 120 member ensemble

Ensemble Setup

- Two parameters (so far) tested to determine how small of a change can be detected
- Highly sensitive: `clubb_c1`, less sensitive: `effgw_oro` in E3SMv1 (Qian et al., 2018).
- Comparisons are made using 500 bootstrap iterations of random draws from each ensemble

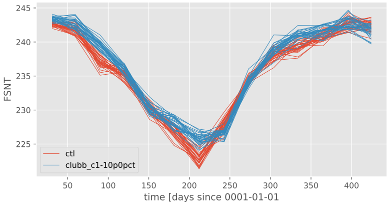

Figure 4: Ensemble plot of FSNT (Net solar flux at top of model $[W\,m^{-2}]$)

Bootstrap comparison method

- Each ensemble has 120 members, select 30 at random from each
- Compare the distributions using K-S test which generates a p value for each variable
- Use the false discovery rate correction to correct p-values (Wilks, 2016)

$$p^*_{(i)} = p_{(i)} * (i/N) \qquad (1)$$

- That is, after sorting, i^{th} p-value is corrected by i/N the null hypothesis $H_{(i)}$ is rejected if $p^*_{(i)} \leq \alpha$
- Global null hypothesis (do these simulations have the same climate) is rejected if any $H_{(i)}$ is rejected

Bootstrap comparison method

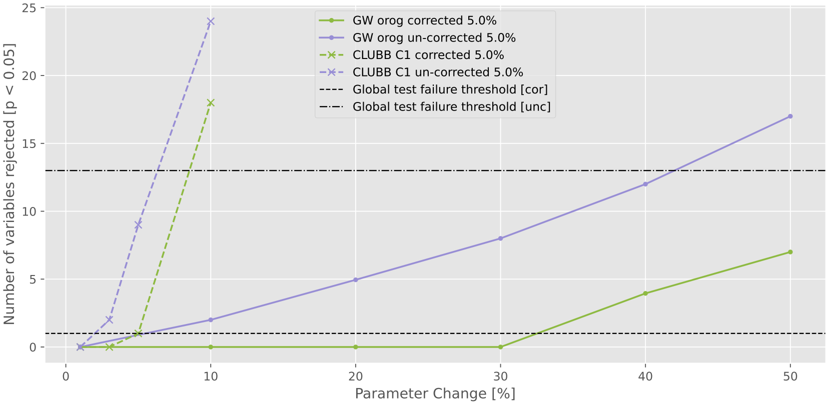

Figure 5: Confidence interval for number of rejected variables by change in tuning parameter

Bootstrap comparison method

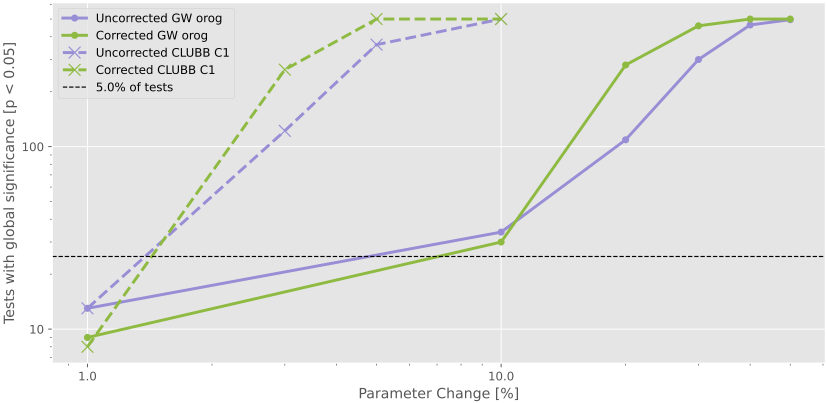

Figure 6: Number of tests with a global rejection by change in tuning parameter

Conclusions

- What can FDR do for our nightly testing?
  - Increases confidence: test false detection (erroneous failures) at a rate of α (here set at 5%)
  - Remove need for bootstrapping to find global failure threshold
- What can it not do?
  - So far, does not make testing able to detect smaller changes in parameters

References

Mahajan, S., K. J. Evans, J. H. Kennedy, M. Xu, M. R. Norman, and M. L. Branstetter, 2019: Ongoing solution reproducibility of earth system models as they progress toward exascale computing. The International Journal of High Performance Computing Applications, 33 (5), 784–790, https://doi.org/10.1177/1094342019837341.

Qian, Y., and Coauthors, 2018: Parametric Sensitivity and Uncertainty Quantification in the Version 1 of E3SM Atmosphere Model Based on Short Perturbed Parameter Ensemble Simulations. Journal of Geophysical Research: Atmospheres, 123 (23), 13,046–13,073, https://doi.org/10.1029/2018JD028927.

Wilks, D. S., 2016: "the stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. Bulletin of the American Meteorological Society, 97 (12), 2263 – 2273, https://doi.org/10.1175/BAMS-D-15-00267.1.

Oak Ridge National Laboratory

Additional information

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

U.S. Department of Energy

Simulation table

Parameter	Pct Change	Parameter value
effgw_oro	0.0	0.375
	1.0	0.3788
	10.0	0.4125
	20.0	0.4500
	30.0	0.4875
	40.0	0.5250
	50.0	0.5625
clubb_c1	0.0	2.400
	1.0	2.424
	3.0	2.472
	5.0	2.520
	10.0	2.640

1 Month Simulations

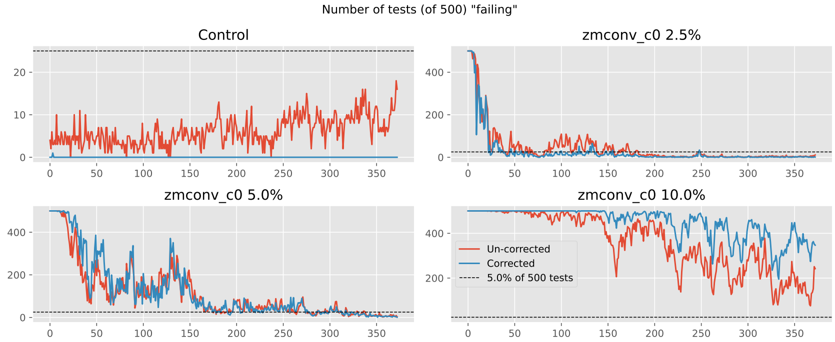

Figure 7: Number of tests with a global rejection for 1 month simulations, changing *zmconv_c0_lnd* and *zmconv_c0_ocn*