https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

Enhanced Climate Reproducibility Testing with False Discovery Rate Correction

Michael E. Kelleher[1] and Salil Mahajan[1]

[1]Computational Hydrology and Atmospheric Sciences Group, Oak Ridge National Laboratory, 1 Bethel Valley Rd, Oak Ridge TN

Correspondence: Michael E. Kelleher (kelleherme@ornl.gov)

Abstract. Simulating the Earth's climate is an important and complex problem, thus climate models are similarly complex, comprised of tens to hundreds of thousands of lines of code. In order to appropriately utilize the latest computational and software infrastructure advancements in Earth system models running on modern hybrid computing architectures to improve their performance, precision, accuracy, or all three; it is important to ensure that model simulations are repeatable and robust. This introduces the need for establishing statistical or non-bit-for-bit reproducibility, since bit-for-bit reproducibility may not always be achievable. Here, we propose a short-simulation ensemble-based test for an atmosphere model to evaluate the null hypothesis that modified model results are statistically equivalent to that of the original model. We implement this test in US Department of Energy's Energy Exascale Earth System Model (E3SM). The test evaluates a standard set of output variables across the two simulation ensembles and uses a false discovery rate correction to account for multiple testing. The false positive rates of the test are examined using re-sampling techniques on large simulation ensembles and are found to be lower than the currently implemented bootstrapping-based testing approach in E3SM. We also evaluate the statistical power of the test using perturbed simulation ensemble suites, each with a progressively larger magnitude of change to a tuning parameter. The new test is generally found to exhibit more statistical power than the current approach, being able to detect smaller changes in parameter values with higher confidence.

1 Introduction

Thousands of scientists and engineers work tirelessly in efforts to better understand and model the Earth's changing climate. A large portion of this effort has come from the development of Earth system models at modeling centers around the globe, which seek to simulate the atmosphere, ocean, cryosphere, land surface, and chemistry, among other components of the Earth system. These models are comprised of many millions of lines of code and are enormously complex projects worked on by many individuals, so the need arises to verify that contributions to the model code do not have unintended effects on answers produced. Thorough testing of the output from these models is routinely conducted, assessing if the results are identical (bit-for-bit) or not. If the results are indeed not bit-for-bit identical, statistical checks are also conducted to ensure the simulated climate of the model or component has not significantly changed, unless that is the intended effect. A variety of methods are available (Milroy et al., 2018; Wan et al., 2017; Baker et al., 2015; Mahajan et al., 2017, 2019b), each of which performs a

https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

statistical comparison between a reference ensemble and test simulation or ensembles. Here, an ensemble indicates a set of model runs each initialized with similar but slightly perturbed initial conditions, generally only at machine-precision levels.

The Energy Exascale Earth System Model (E3SM, E3SM Project (2023)) uses a suite of tests which run the model under a variety of configurations and methods using the Common Infrastructure for Modeling the Earth (CIME) software to setup, build, run, and analyze the model. The tests are run at varying frequencies from nightly to weekly, testing the both the latest science and performance updates and maintenance branches, which only receive compatibility updates and should be bit-for-bit identical. A subset of these tests are statistical reproducibility tests (non-bit-for-bit tests), which determine through a comparison of control and perturbed ensembles, whether or not the simulated climate has changed as a result of modifications to the model code or infrastructure. These are the Time Step Convergence test (TSC, Wan et al. (2017)), the Perturbation Growth New test (PGN, related to the work in Rosinski and Williamson (1997)), the multi-testing Kolmogorov-Smirnov test (MVK, Mahajan et al. (2019b)), and the MVK-Ocean test (MVK-O, Mahajan (2021)). The first three test the reproducibility of the E3SM Atmosphere Model, (EAM), and the last tests the Model for Prediction Across Scales-Ocean (MPAS-O), the ocean component in E3SM.

The TSC test evaluates numerical convergence by comparing ensemble differences at two time step sizes (i.e., 1s and 2s) using a Student's t-test on root mean squared difference (RMSD) values, under the assumption that numerical solutions should converge as the time step decreases (Wan et al., 2017). The PGN test, in contrast, assesses stability by comparing the state of the atmosphere after one time step across perturbed ensemble members, identifying whether small initial differences grow inconsistently through individual physics parameterizations. Unlike the short-duration TSC and PGN tests, the MVK and MVK-O tests use year-long and two-year long simulation ensembles respectively, allowing them to assess the cumulative impact of code or configuration changes on the model's climatology after internal variability has saturated, providing a more robust evaluation of long-term climate statistics.

MVK evaluates the null hypothesis that a modified model simulation ensemble is statistically equivalent to a baseline ensemble. It applies a two-sample Kolmogorov-Smirnov test to over 100 output variables and counts how many show statistically significant differences between the two ensembles. If this count exceeds a critical value threshold as expected from internal variability and derived via bootstrapping, the two simulations are considered to have different simulated climates (Mahajan et al., 2017). Further details on the MVK are provided in Section 2.1.

Here, we propose a new testing approach, improving on the MVK, that provides a two-fold benefit over it. Firstly, it is desirable for the usability of a test to reduce the number of false positives (Type I error rates) where two simulations are erroneously labeled as statistically different, without affecting the false negative rates (Type II error rates). Operationally, MVK has been exhibiting a false positive rate of about 7.5%, despite the prescribed significance level of 5%, since its induction into the test suite (as discussed in Section 4.5). Previous works suggest that implementing a false discovery rate (FDR) correction (by adjusting significance thresholds based on the number and rank of p-values, discussed in more detail in Section 2.2), reduces the false positive rate when multi-testing as compared to using bootstrapping-derived critical value threshold for hypothesis testing (Ventura et al., 2004; Wilks, 2006, 2016; Mahajan, 2021), which is the methodology used by MVK. The new testing approach for the atmosphere model thus implements an FDR correction when evaluating multiple variables in the

https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

atmosphere model output. Secondly, it will be useful to reduce the computational cost and time of deriving the critical value thresholds for the MVK. The bootstrapping procedure is time-consuming and computationally expensive because of its need for large control ensembles. And, it would need to be conducted again after a significant enough departure from the original model code to establish a new critical value threshold. Substantial model code changes to numerics and physics can alter internal variability and shift the distributional properties of output fields. FDR correction theoretically asserts the critical value threshold for global null hypothesis evaluation (as discussed in Section 2.2), thus eliminating the need for conducting a large control ensemble simulation to determine it, though one is required to demonstrate this.

This paper thus seeks to answer two questions; can a testing framework using FDR correction reduce the number of false positives in an operational setting as compared to the MVK without impacting its statistical power (false negative rates) and can it eliminate the need for extensive and expensive analysis of large ensembles each time the model is updated significantly? The following section discusses the MVK test and its pitfalls in more detail and describes the FDR approach as implemented here. Section 3 lists the simulation ensembles conducted to evaluate both the MVK and the new testing framework. And, Section 4 discusses our results on the evaluation of the false positive and negative rates of these testing frameworks. Finally, our results are summarized in Section 5 with a brief discussion of caveats of the study and future direction.

2 Methods

2.1 Multi-testing Kolmogorov-Smirnov Test

The multi-testing Kolmogorov-Smirnov (MVK) test, as implemented in the CIME for the atmosphere model of E3SM, compares two independent $N = 30$ member ensembles. It is used in nightly testing, comparing a baseline ensemble, generated after each approved "climate changing" code modification, and a test ensemble, newly generated each day. The E3SM model is run at "ultra-low" resolution, $\approx 7.5°$ atmosphere, 240 km ocean, for 14 months. The first two months are discarded as the system reaches quasi-equilibrium. Annual global means of each of the 120 standard output fields of the E3SM Atmosphere Model (EAM) are then computed for each ensemble member. Then, for each field, the null hypothesis (H_1, also referred to as the local null hypothesis here) that the sample distribution function of the annual global mean of that field estimated from the baseline ensemble (estimated from N data points) is statistically similar to that of the new ensemble is evaluated. The two sample Kolmogorov-Smirnov (K-S) test at a significance level of $\alpha = 0.05$ is used for testing H_1. The K-S test is a non-parametric statistical test to compare cumulative or empirical distribution functions. The larger null hypothesis (H_0, also referred to as the global null hypothesis here) that the two ensembles have identical simulated climates is then evaluated for a significance level of α. The test statistic, t, for testing this larger null hypothesis is defined as the number of fields that reject H_1. H_0 is rejected if the number of fields rejecting H_1 is greater than a critical value threshold (found to be 13, Mahajan et al. (2017)), and the test issues a "fail". The null distribution of t, and hence the critical value threshold at a significance level of α, was empirically derived by randomly sampling two N-member ensembles from a 150-member control ensemble of an earlier version of E3SM, and computing t, 500 times (Mahajan et al., 2017).

https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

2.1.1 Potential pitfalls of multi-testing K-S Test

MVK requires a large control ensemble in order to capture the variability of the model and establish proper critical value thresholds for the number of rejected local null hypotheses before rejection of the global null hypothesis is considered (Wilks, 2006). It also, as discussed in Wilks (2006), does not put weight onto fields rejected very strongly, that is those fields with extremely small p-values. If a few fields are rejected with near certainty, this likely indicates global significance, but if the total number of these fields does not exceed the predetermined threshold, the overall result is global null hypothesis acceptance, which results in a lower power to detect differences. Finally, in practice this approach has a larger Type I error (see Sections 4.2, 4.5; Table 2) than desired. While possible to decrease the significance level α to reduce the false positive rates, this in turn decreases the statistical power of the test to detect small changes which is not desirable.

2.2 False Discovery Rate Correction

When conducting simultaneous multiple null hypothesis tests as in the MVK above, Type I error rate inflation occurs, leading to a higher overall probability of false positives. While a significance level (e.g., $\alpha = 0.05$) controls the probability of a false positive for a single test, conducting many tests increases the chance that at least one result will appear significant purely by chance, even if all the null hypotheses are true. This implies that the overall proportion of false discoveries can become unacceptably high, undermining the reliability of the findings. MVK accounts for multi-testing by using a resampling strategy, which was found to give similar results as compared to permutation testing (Mahajan et al., 2019b). In addition to permutation testing and bootstrapping approaches, other approaches for correcting Type I error rate inflation associated with multi-testing include family wise error rate correction (e.g. Bonferroni correction), and the false discovery rate (FDR) (Wilks, 2006; Ventura et al., 2004), which is used here. The Bonferroni correction adjusts the significance level by dividing the desired significance level, α, by the number of tests conducted, but is known to have reduced statistical power (Wilks, 2006; Ventura et al., 2004). Here, we use the Benjamini-Hochberg (BH) FDR correction approach (Benjamini and Hochberg, 1995). The BH FDR approach been shown to effectively control for Type I error rate inflation while also exhibiting more power than other approaches (Ventura et al., 2004; Wilks, 2006). It is widely used in Earth system studies for spatial analysis (Wilks, 2006; Renard et al., 2008; Whan and Zwiers, 2017) and has also been applied to solution reproducibility testing of ocean models (Mahajan, 2021).

Similar to MVK, we use the two sample KS test to evaluate the local null hypothesis (H_1) for each field that their sample distribution functions are statistically identical across the two ensembles. Also, similar to the MVK, the global null hypothesis (H_0) being tested is that the two ensembles are statistically similar. We use the `statsmodels` (Seabold and Perktold, 2010) Python package for applying BH to reproducibility testing. Using the p-values from KS test for all the variables evaluated, and given a global significance level of α, the BH-FDR constrains the critical value threshold for local null hypothesis (H_1) testing as follows:

$$p_{FDR} = \max_{i=1..m} \left[p_{(i)} \leq \frac{i}{m} q^* \right] \tag{1}$$

Here, the p-values from each local null hypothesis test for individual fields are first sorted in ascending order. And, p_{FDR}, is the constrained critical value threshold and as the equation states, it represents the maximum p-value for which the inequality

https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

is satisfied. $p_{(i)}$ is the i^{th} smallest p-value, and m is the total number of p-values (in this case $m = 117$ output fields tested), and q^* parameter is the control on the false detection rate, here chosen to be the same as α, as is typical. A local null hypothesis $H_1^{(i)}$ corresponding to the i^{th} field is rejected only if its $p_i \leq p_{FDR}$. Now, the global null hypothesis, H_0, that can be framed as all H_1 are true, is rejected at the global significance level of α if *any* $H_1^{(i)}$ is rejected using p_{FDR} as the critical value threshold (Renard et al., 2008; Wilks, 2006; Ventura et al., 2004). Thus, the global null hypothesis is rejected if *any* $p_i \leq p_{FDR}$. In other words, the critical value threshold for the test statistic, t, which is the number of fields rejecting H_1, is equal to one for H_0 for BH-FDR. This theoretical critical value threshold of one, when multi-testing with FDR corrections, has been used widely for field significance testing to address the multiple comparisons problem inherent in spatial analyses in the climate and meteorological studies (Wilks, 2006; Renard et al., 2008; Burrell et al., 2020). More details about the BH-FDR as applied to earth system science can also be found in these studies.

This BH-FDR as applied to reproducibility testing will be deemed useful here if it performs at least as well at detecting altered simulated climates (statistical power) as the MVK, while reducing the number of false positives. We investigate these characteristic of the approach using suites of simulation ensembles with controlled modifications in Section 4.

3 Simulation Ensembles

We conduct a suite of multi-member ensembles to evaluate the false positive and false negative rates of BH-FDR approach. Additionally, having undergone significant updates to software (Golaz et al., 2022), we reassess the critical value threshold of MVK at which two ensembles of E3SMv2.1 are statistically distinguishable (the critical value threshold having been previously computed from a control ensemble of an earlier model version) using this ensemble suite. Each ensemble is generated by varying the initial conditions by near-machine precision perturbations ($O(10^{-10})$) to the temperature field at each grid point for each ensemble member. One of these ensembles is the control ensemble which refers to the unmodified default E3SMv2.1 model with default values of all tuning parameters. Other generated ensembles are differentiated from the control ensemble by varying the value of a tuning parameter from its default value by different magnitudes (see Table 1 for details). Here, we vary two tuning parameters separately. Version 1 of E3SM was found to be highly sensitive to a parameter termed `clubb_c1` and weakly sensitive to the `effgw_oro` (Qian et al., 2018). `clubb_c1` is the constant associated with dissipation of variance of $\overline{w'^2}$ (where w is vertical wind speed), and `effgw_oro` is the gravity wave drag intensity (see Table 1 of Qian et al. (2018)). We choose these two parameters to generate ensembles to capture a range of sensitivities in version 2 of E3SM. The simulation ensembles are run at "ultra-low" resolution with a 14-month simulation duration. A 120-member ensemble is generated for each tuning parameter change, along with the control ensemble.

To further evaluate the BH-FDR approach, we conduct two additional ensemble simulations to test model sensitivity to compiler optimization choices. All ensembles are run on Argonne National Laboratory's Chrysalis machine and built using the Intel compiler v20.0.4. The default optimization flag for E3SM is "-O3", and is used to generate the control and other ensembles with tuning parameter changes. The two different optimization test ensembles are titled `opt-O1` and `fastmath`,

https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

Parameter	% 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

Table 1. List of simulations for each tuning parameter

and are compiled with optimization flags "-O1" and "-O3 -fp-model=fast" respectively. Previous work suggests that using optimization "-O1" is expected to not produce a significantly different simulated climate than the default (Baker et al., 2015; Mahajan et al., 2017). Mahajan et al. (2017) found that using optimization flag "fast" with "Mvect" resulted in a statistically different climate compared to the default, which used an optimization of "-O2" using the PGI compiler.

The "-O1" optimization turns off most of the aggressive optimizations used for the "-O3" level, including loop vectorization, loop unrolling, and global register allocation, while enabling the "-fp-model=fast" fast floating point model allows the compiler to be less strict in its handling of floating point arithmetic. (Intel Corporation, 2023) This means using "-O1" in place of "-O3" ought to result in slower operation but similar results, while using "-fp-model=fast" could result in different answers under specific conditions, including if there are "NaN" or not-a-number values present. Though in the case of E3SM, it is already used in the compilation of several source files, thus adding it as a global option only changes those where it is not in use already.

4 Results

4.1 Estimating Critical Value Threshold for MVK

E3SMv2.1 has undergone several scientific feature changes as well as software infrastructure changes since the release of E3SMv0 which was used to estimate the critical value threshold for null hypothesis testing using MVK. Here, we estimate the null distribution of the test statistic of MVK which is more representative of E3SMv2.1. We use a bootstrapping (resampling) strategy to derive the null distribution and the empirical critical value threshold of the test statistic, t, the number of variables falsely rejecting the true null hypothesis H_1 at $\alpha = 0.05$, for the global null hypothesis using for the MVK test for E3SMv2.1.

https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

For this analysis, two 30 member ensembles were drawn, without replacement, from the 120 member control ensemble (with $> 10^{18}$ possible ways of drawing a 30 member ensemble). t is computed for each such draws of 30-member ensemble pairs and then this procedure is repeated 1000 times. This procedure was then applied to each parameter adjustment ensemble separately, comparing two random draws from each large ensemble, in an effort to expand the sample size to estimate the null distribution. It is also applied to the ensembles with changes to compiler optimizations. The null distribution of t is representative of the internal variability of the model as the drawn ensemble pairs are part of the same population with ensemble members differing only in the initial conditions at machine precision level perturbations. Figure 1 illustrates the null distribution of t with a box and whiskers plot derived from each 120-member ensemble. The critical value threshold is also estimated as the 95th percentile of t, which ranges from 10 to 13 for different ensembles. We thus set the critical value threshold for MVK at α = 0.05 as the median of those values, found to be 11.

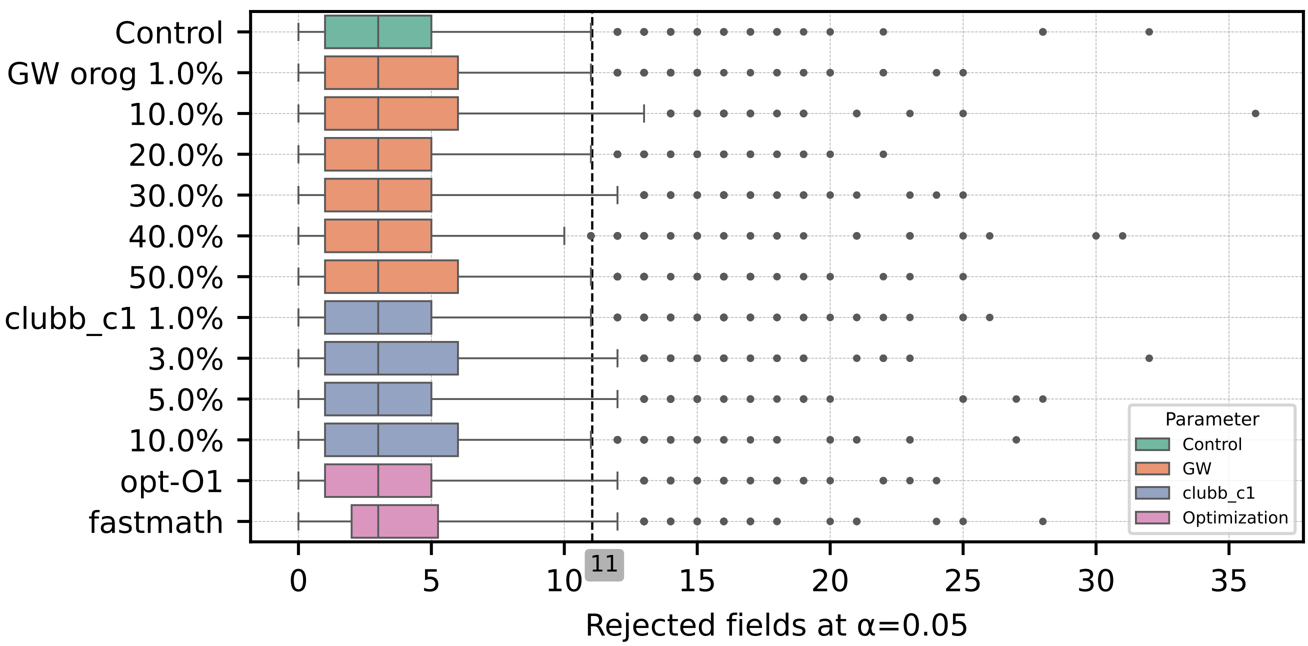

Figure 1. Box plot of number of rejected output fields for each ensemble self-comparison. The solid center line of each box represents the median, the box represents the inter-quartile range (25^{th}-75^{th} percentiles), the whiskers are plotted at the 5^{th}-95^{th} percentiles, and outliers marked beyond that range. The dashed vertical line is the median of all 95th percentiles which is 11 fields.

4.2 False positive rates: MVK and BH-FDR approach

The bootstrapping method described above in Sec. 4.1 to determine the critical value threshold for global null hypothesis testing for MVK can also be used to estimate the false positive rates. Since ensemble pairs are drawn from the same population, each drawn pair that rejects the null hypothesis is a false positive. For the MVK test, the critical value threshold of 11, estimated

https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

Simulation Ensemble	MVK	BH-FDR
Control	0.049	0.028
GW orog 1.0%	0.054	0.032
GW orog 10.0%	0.044	0.022
GW orog 20.0%	0.061	0.039
GW orog 30.0%	0.048	0.032
GW orog 40.0%	0.035	0.031
GW orog 50.0%	0.053	0.028
clubb c1 1.0%	0.041	0.049
clubb c1 3.0%	0.053	0.034
clubb c1 5.0%	0.049	0.031
clubb c1 10.0%	0.044	0.029
opt-O1	0.060	0.032
fastmath	0.059	0.034
Mean	0.050	0.032
95th %tile	0.060	0.039

Table 2. False positive rates for self-comparison bootstraps (per 1000 iterations for each ensemble) at α=0.05, and the mean and 95th percentile false-positive rate over all self-comparison bootstraps.

above, is used for null hypothesis testing at $\alpha = 0.05$. For the BH-FDR approach, any field rejecting the null hypothesis after false discovery rate correction, implies a rejection of the larger null hypothesis.

For each of the thirteen 120-member ensembles conducted (control, tuning parameter changes and optimization change ensembles), a 1000 iteration bootstrapping analysis is performed separately, and the false positive rates are computed for each under both MVK and the BH-FDR testing approaches. Table 2 details the false positive rates derived from each analysis. The mean of these 13 values for the MVK is 0.05, which can be expected to be near the prescribed α since the critical value threshold was estimated from the same population (set of ensembles), although with different bootstrap samples. The results may be different if an ensemble size of 120 does not capture the true internal variability. The mean false positive rate under the BH-FDR approach is lower at 0.032. Also, the 95th percentile of false positive rates of the 13 values is 0.06 for the MVK, and is reduced to 0.039 for the BH-FDR approach. The above indicate that the application of the FDR correction works as intended and based on the Lemma of Theorem 1 in Benjamini and Hochberg (1995), the FDR puts an upper limit to the level of false discovery at q^* (which here is chosen as $q^* = \alpha = 0.05$), thus false positive rates lower than α are expected.

https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

4.3 False Negative Rates and Statistical Power: MVK and BH-FDR

To evaluate the magnitude of change that the tests can detect confidently, we again rely on bootstrapping following previous work (Mahajan et al., 2019b, a; Mahajan, 2021). For each tuning parameter change, 30 ensemble members each are drawn from the control and that tuning parameter change ensemble. The test statistic, t, is then computed for both MVK and BH-FDR tests and the two tests are conducted on the ensemble pair. This procedure is then repeated 1000 times. To illustrate the impact of progressively increasing the tuning parameter on the model climate, Fig. 2 shows the 95th percentile of t from the 1000 hypothesis tests for each tuning parameter change for both effgw_oro and clubb_c1. As the magnitude of the tuning parameter change to the model increases, t increases for both MVK and BH-FDR approach. To reiterate, an increase in t indicates an increase in the number of fields rejecting the local null hypothesis, H_1. Fig. 2 also re-iterates the lower sensitivity of E3SM to the orographic gravity wave drag parameter than the C1 parameter from the CLUBB cloud parameterization scheme. Smaller percentage changes in clubb_c1 result in large changes to t as compared to effgw_oro, where larger percentage changes are needed for similar changes in t.

Fig. 2 also points towards the detectability of modifications to the model by the tests. For both parameters, a 1% change in their value, does not result in a change in the simulated climate that is detectable by the two tests, since >95% of the bootstrapped ensemble pairs have t less than the critical value threshold of the tests (11 for MVK and one for BD-FDR). Increasing effgw_oro to 10% and clubb_c1 to 2.5% results in some of the bootstrap iterations having exceeded or met the critical value threshold for both MVK and BH-FDR. As the magnitude of tuning parameter change increases, the number of bootstrap iterations crossing the critical value threshold also increases.

A formal estimate of the statistical power (P, rate of correctly rejecting a false null hypothesis), also representative of the false negative rates $(1-P)$, incorrectly accepting a false null hypothesis, of these tests is illustrated in Fig. 3. It shows the number of bootstrap iterations where the tests correctly reject H_0 at a significance level of $\alpha = 0.05$ and is indicative of the likelihood of the tests detecting a modification to the model. P can be computed by dividing the ordinate (y-axis) values by the total number of bootstrap iterations (1000). Similar to Fig. 2, Fig. 3 shows that as the magnitude of a tuning parameter is increased, the number of bootstrap iterations rejecting H_0 also increases. For a 1% change to effgw_oro or clubb_c1, only about 30 bootstrap iterations reject h_0, which implies that there is only a 30 out of a 1000 (3%) chance that a change of this magnitude could be detected by the tests. Increasing effgw_oro to 10% results in more than 50 bootstrap iterations (5%) rejecting H_0, indicating that it unlikely to be caused by random chance at the 0.05 significance level, it still exhibits a low likelihood of being detected by the tests (about 60 and 80 out of a 1000 chance for MVK and BH-FDR tests respectively). As the magnitude of change to effgw_oro is increased, the likelihood of detecting a change by the tests increases. For a change of 40% to effgw_oro there is a >90% chance of being detected by both the tests and it reaches nearly a 100% for a larger change. Similarly, for clubb_c1, a change of 5% results in greater than 80% (90%) chance of being detected by the MVK (BH-FDR) test, and nearly a 100% chance of detection for a change of about 10% change to its magnitude.

Overall, BH-FDR approach exhibits greater statistical power than MVK for almost all tuning parameter changes, allowing increased confidence in detecting changes, adding to its advantages. This is consistent with previous works (Benjamini and

https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

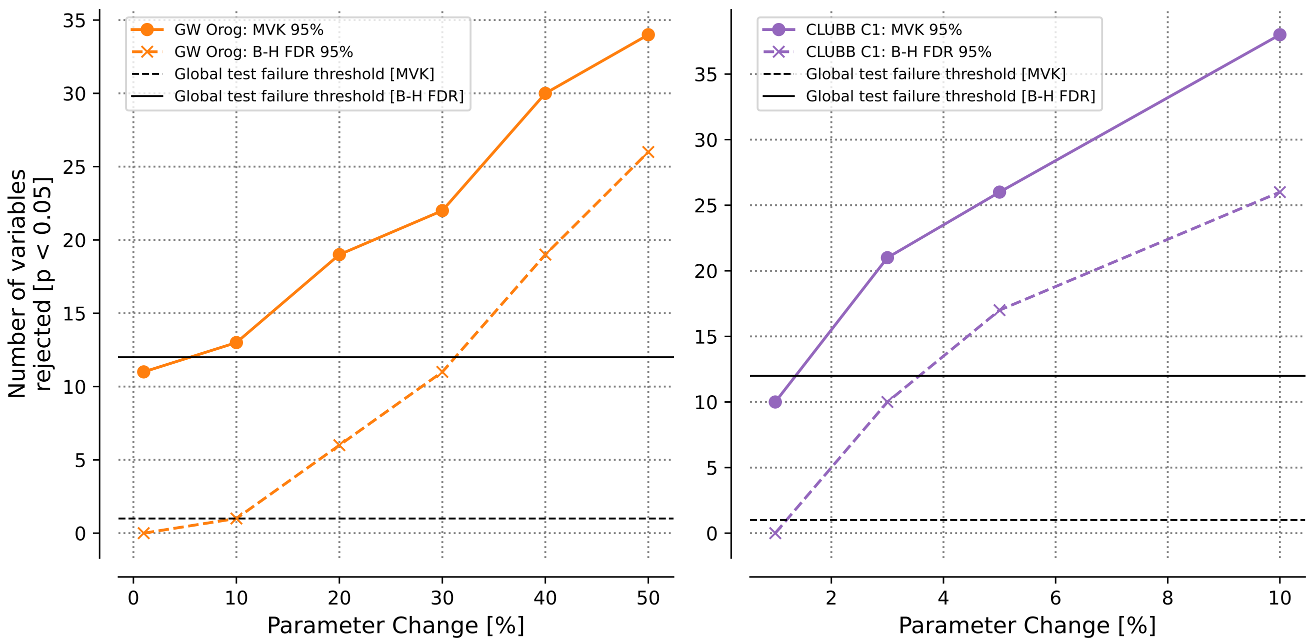

Figure 2. 95th percentile of the number of fields with statistically significant differences from the control ensemble (reject the local null hypothesis, H_1 at the $\alpha = 0.05$ significance level) on the y-axis by percentage change in tuning parameter along the x-axis. Solid line represents MVK test, dashed represents BH-FDR. The horizontal lines represent global null hypothesis critical value thresholds for MVK (dashed) and BH-FDR (solid).

Hochberg, 1995; Wilks, 2006), that suggest that BH-FDR approach generally exhibits greater power than bootstrapping methods. For the smallest parameter change to effgw_oro, where MVK test exhibits greater power than the BH-FDR approach, it is possible that at the 1% change to effgw_oro, the simulated climate is not different in a meaningful way. Sampling errors may also be playing a role, given the small magnitude of change. Increasing the ensemble sizes can significantly increase the power of the tests as shown for MVK (Mahajan et al., 2019b), but operationally add to the computational cost. This trade-off between false negative rates and ensemble sizes is a decision left to model developers and code integrators. The power analysis here, provides them with some reference to interpret the test results. For instance, if a non-bit-for-bit change passes the test (with the ensemble size of say, 30), developers can infer that its impact is likely smaller than a 5% change in C1 parameter of CLUBB, which can be detected at a high confidence by the tests. This contextual comparison helps determine whether to accept or investigate a change further and guides the selection of ensemble size needed to detect changes of interest. In the future, we will expand our power analysis to include other tuning parameter changes to better inform developers, integrators and domain scientists using the tests.

https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

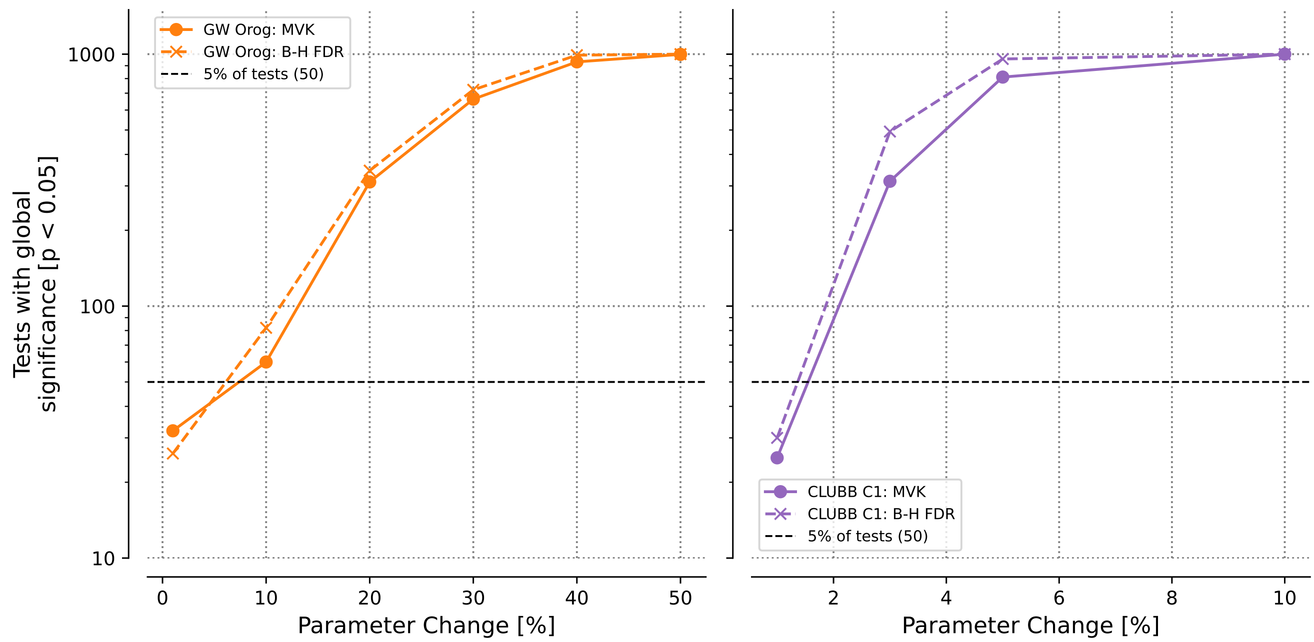

Figure 3. Power Analysis. Number of bootstrap iterations that reject the global null hypothesis (H_0) out of 1000 bootstrap iterations for changes in tuning parameter effgw_oro (left) and clubb_c1 (right).

4.4 Optimization changes

Two simulation ensembles ("opt-O1" and "fastmath") were performed to apply the MVK and BH-FDR approach to evaluate sensitivity of model results to compiler optimization flags. Bootstrapping procedures, similar to those described in the last section indicate that these optimizations do not have a significant impact on solution reproducibility (Figure 4). Only 28 and 22 out the 1000 bootstrap iterations reject H_0 using the MVK for "opt-O1" and "fastmath" respectively at the significance level of 0.05. Even fewer bootstrap iterations reject H_0 using the BH-FDR test. Reducing optimizations has been shown to yield climate reproducibility in previous studies as well (Baker et al., 2015; Mahajan et al., 2017), similar to our result that the simulated climate of "opt-O1" is statistically indistinguishable from that of the control ensemble that uses "-O3" optimizations.

The aggressive optimizations enabled by "-fp-model=fast" are expected to decrease solution accuracy (Mielikainen et al., 2016; Büttner et al., 2024), however, in this set of ensembles they do not significantly alter the simulated climate. As previously mentioned, the "-fp-model=fast" option is already a default for several source files, thus this simulation ensemble tests its use only for those additional source files in EAM. This is different from testing the impact of using it across model as a whole as in Mahajan et al. (2017). Acceptance of H_0 by the tests indicate that including the option for all of EAM, instead of selectively as is the default, does not result in statistically distinguishable solutions. This indicates that the "-fp-model=fast" optimization could be applied more generally throughout EAM code base. However, for higher resolutions or fully-coupled ensembles, this

https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

result may not apply as these configurations may respond to perturbations differently than ultra-low resolution model. Thus, further examination of its applicability may be required.

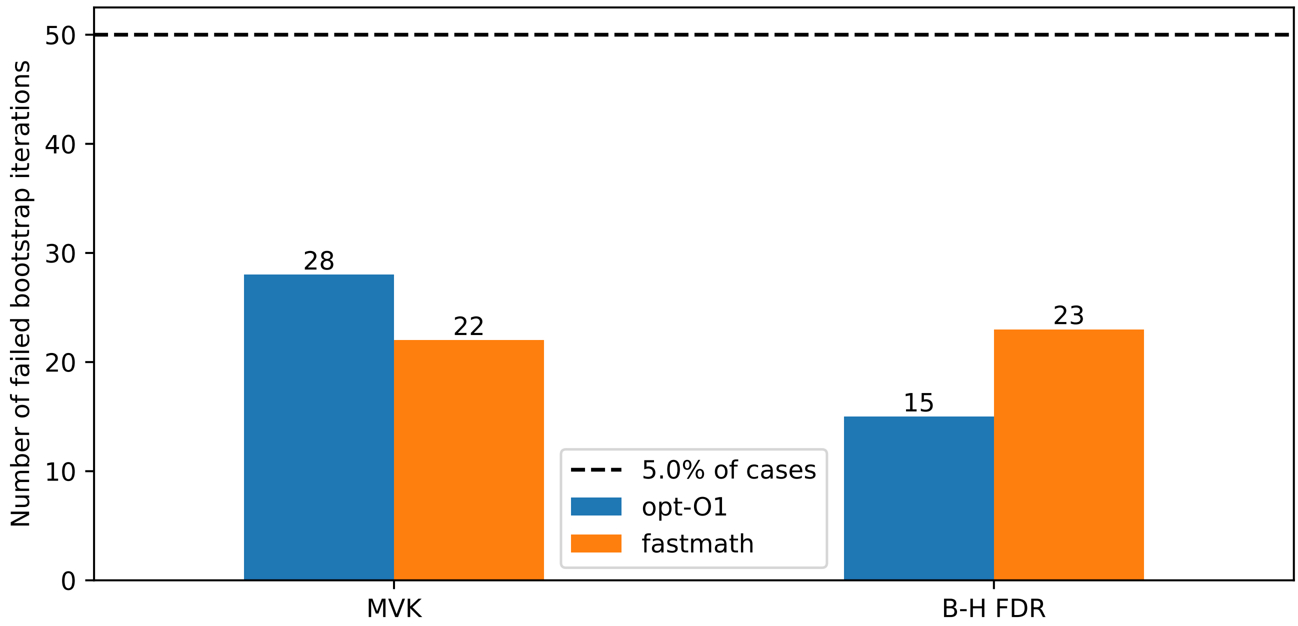

Figure 4. Number of bootstrap iterations that reject the global null hypothesis (H_0) for opt-O1, and fastmath ensembles, when compared to the control ensemble, out of 1000 bootstrap iterations for the MVK test (left) and B-H FDR (right). The dashed horizontal line denotes 5% of all iterations.

4.5 Operational results

Fig. 5 shows the time series of t, or the number of variables rejecting H_1, for MVK and BH-FDR, for a period of a few weeks after BH-FDR was implemented and included in nightly testing in late September last year. The model maintained bit-for-bit reproducibility during these weeks. Bit-for-bit reproducibility is ascertained by a suite of bit-for-bit tests that are run each night, testing the model under a variety of conditions. When these pass, the model is bit-for-bit with previous results, and a test fail by MVK and BH-FDR on these days is thus known to be a false positive. Fig. 5 shows that operationally BH-FDR has a reduced false positive rate of 1.9% (1 of 53 tests) as compared to 7.5% (4 of 53 tests) for MVK. t varies between 0 to a maximum of 15 (on two individual days) for MVK, while t is either 0 or 1 (on one occasion) for BH-FDR. This solitary global null hypothesis rejection using BH-FDR does not occur at the same time as a global rejection using uncorrected p-values, indicating that a systematic change in its statistics may not have occurred, in which case both tests may be expected to fail. The rate of global rejections using MVK (7.5%) is higher than the targeted 5%, which is reduced to under 2% using corrected p-values, as is expected as the correction controls for the false discovery rate when using BH-FDR. The higher false positive rate of MVK can

https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

be associated with the small sample size (53 days of testing). However, it may also indicate that the 120-member ensembles used to derive the null distribution of t may not be sampling the internal variability of the system sufficiently well. In the future, we plan to expand the ensemble size of the control ensembles to derive the null distribution and evaluate its impact on operational false positive rates of MVK.

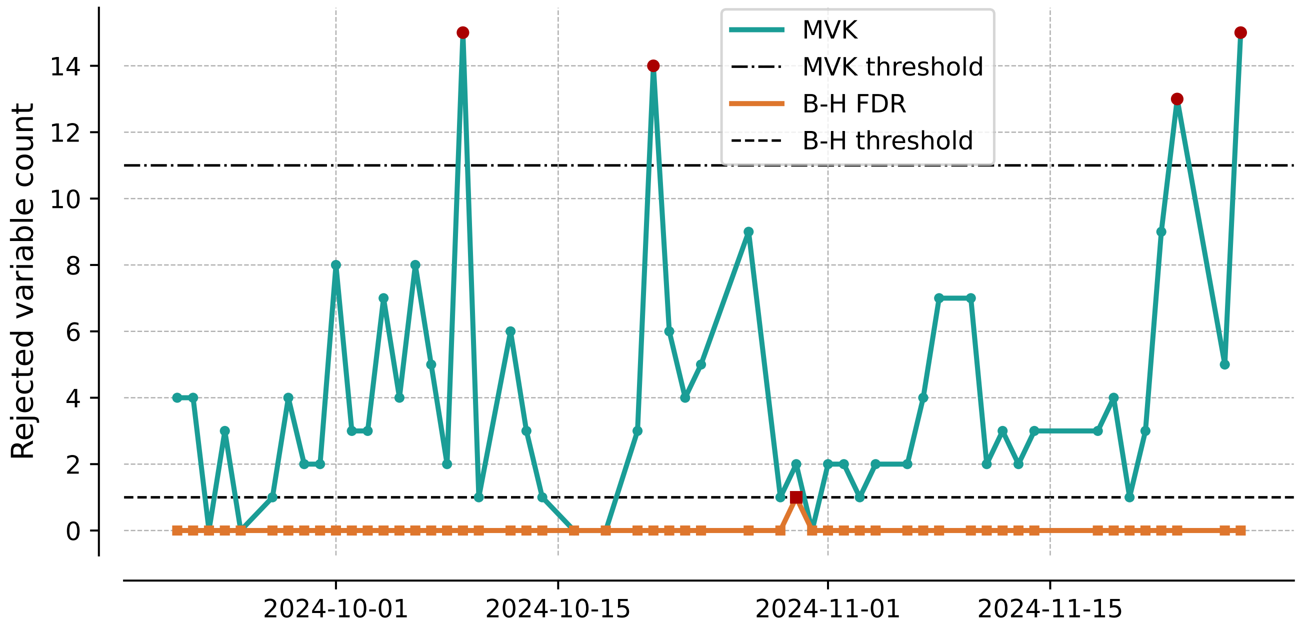

Figure 5. Number of rejected variables (of 120) for nightly tests of E3SM. The teal line shows rejection based on MVK p-values, orange shows number of rejections based on B-H FDR p-values.

280 Operationally, the MVK test has recently been successful in identifying two bugs committed to the model's code, which were thought by developers to be non bit-for-bit, but not climate changing. The test returned a "fail" result following both changes, and the changes were reverted and eventually re-worked into non climate changing code alterations. This unintentional climate changing update was also captured by the BH-FDR approach and indicates its ability to capture erroneous alterations in addition to its ability to capture tuning parameter changes. In the first case, developers introduced changes related to aqueous-phase chemistry, specifically involving reactions tied to cloud water and trace gases. These updates were considered minor adjustments to internal model behavior and were not expected to alter the simulated climate. However, the tests flagged the update as a baseline failure, revealing statistically significant differences across many atmospheric variables. Further investigation showed that the change affected cloud-aerosol interactions and subsequently altered radiative balance, demonstrating how a seemingly minor, but non-bit-for-bit change in chemical processes led to broader simulated climate impacts. In the second case, ozone chemistry configuration changes were merged, aiming to improve numerical consistency in offline and online chemistry calculations and were also expected to be non-climate changing. However, MVK and BH-FDR detected a statistically signif-

https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

icant difference, which was traced to elevated tropospheric ozone concentrations resulting from the changes. The alteration impacted radiative forcing enough to shift the ensemble mean state, and was also subsequently corrected.

5 Summary and Discussion

This study presents a new approach, BH-FDR, to evaluate statistical solution reproducibility of EAM after unintended non-bit-for-bit changes are introduced. BH-FDR improvises on the existing MVK test and applies a false discovery rate correction to control Type I error inflation in multi-testing scenarios. While the original MVK approach relies on computationally expensive bootstrapping to determine critical value thresholds, BH-FDR offers a theoretically grounded and operationally simpler alternative. Our evaluation using a comprehensive suite of ensembles, including both parameter perturbations and compiler optimization changes, demonstrates that BH-FDR approach maintains or improves the statistical power of the MVK test, while significantly reducing false positive rates. Notably, the BH-FDR approach eliminates the need for re-calibrating critical value thresholds after major model revisions. Operational implementation of this method in nightly E3SM testing has further validated its utility, showing a reduction in false positives from 7.5% to 1.9%. Overall, the BH-FDR approach enhances the robustness, accuracy, and efficiency of statistical testing for climate model reproducibility.

A caveat of our study here is that BH-FDR (and MVK as well) has been applied to and evaluated for the ultra-low resolution version of the model, which is not used in practical applications. Further exploration is required to examine the consistency of test results across model resolutions and evaluate if test results with the ultra-low resolution model hold for standard and high resolution model configurations used for production runs. Higher resolution models resolve finer scale processes, which can effect numerical sensitivity, internal variability and process feedbacks of the model. In the future, we plan to identify the underlying reasons and scenarios in which test results may or may not remain consistent across resolutions. Nonetheless, an earlier unpublished work found that the results of MVK applied to ultra-low resolution ensembles and MVK applied to standard resolution ensembles were consistent when evaluating a port of an earlier version of E3SM to a new machine at the National Energy Research Scientific Computing Center (NERSC). Further, we will explore enhancing computational feasibility of the tests by using shorter run times of simulation ensembles (Milroy et al., 2018), allowing for routine testing with higher resolution models.

In addition to its application in traditional Earth system models, the BH-FDR-based statistical testing framework has potential for assessing the reproducibility of AI-based climate models, which are rapidly gaining prominence in climate science. Recent developments such as ClimateBench (Watson-Parris et al., 2022), FourCastNet (Pathak et al., 2022), and Pangu-Weather (Bi et al., 2022) demonstrate the capabilities of deep learning models to emulate or replace components of physics-based models with significant computational advantages allowing the generation of very large ensembles at very low computational costs. However, verifying the reliability and reproducibility of these models presents unique challenges. AI models often involve stochastic elements in training, sensitivity to floating-point precision, and reliance on hardware-specific optimizations, all of which can lead to variability in output across runs. Standard bit-for-bit reproducibility tests are inadequate in this context, and statistical frameworks like MVK or BH-FDR could serve as robust alternatives to assess whether differences in AI model outputs are statistically meaningful or within expected variability. Prior work has highlighted the need for principled evaluation methods tailored to the probabilistic nature of machine learning in scientific applications (Rasp et al., 2020; Dueben et al., 2021), and integrating ensemble-based hypothesis testing into AI model workflows could be a step toward more rigorous, interpretable, and trustworthy deployment of AI systems in operational climate modeling. Ultra-low resolution models, like the one used here, can also be used to create very large ensembles at low computational cost allowing comparisons with their AI surrogate large ensembles. We plan to conduct such evaluations in the near future.

Code availability. The code for this work can be found at https://github.com/mkstratos/detectable_climate (Kelleher and Mahajan, 2025a)

Data availability. The bootstrap data from each comparison is available in (Kelleher and Mahajan, 2025b)

Author contributions. MK and SM developed the methodology. MK wrote the code, conducted the simulations, and wrote the first draft of the manuscript. SM supervised the project and contributed writing to the final manuscript.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Acknowledgements. This research was supported as part of the Energy Exascale Earth System Model (E3SM) project, funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research (BER).

The authors also gratefully acknowledge the computing resources provided on Blues, a high-performance computing cluster operated by the Laboratory Computing Resource Center at Argonne National Laboratory.

The authors also acknowledge the numerous open-source libraries on which this work depends, Harris et al. (2020); Virtanen et al. (2020); Hoyer and Hamman (2017); Dask Development Team (2016); Wes McKinney (2010); Hunter (2007); Waskom (2021); Seabold and Perktold (2010).

References

Baker, A., Hammerling, D., Levy, M., Xu, H., Dennis, J., Eaton, B., Edwards, J., Hannay, C., Mickelson, S., Neale, R., et al.: A new ensemble-based consistency test for the Community Earth System Model (pyCECT v1. 0), Geoscientific Model Development, 8, 2829–2840, 2015.

Benjamini, Y. and Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society: Series B (Methodological), 57, 289–300, https://doi.org/https://doi.org/10.1111/j.2517-6161.1995.tb02031.x, 1995.

Bi, K., Zhang, Z., Lin, T., and et al.: Accurate medium-range global weather forecasting with 3D neural networks, Nature, 610, 87–93, 2022.

Burrell, A. L., Evans, J. P., and De Kauwe, M. G.: Anthropogenic climate change has driven over 5 million km^2 of drylands towards desertification, Nature Communications, 11, 3853, https://doi.org/10.1038/s41467-020-17710-7, 2020.

Büttner, M., Alt, C., Kenter, T., Köstler, H., Plessl, C., and Aizinger, V.: Enabling Performance Portability for Shallow Water Equations on CPUs, GPUs, and FPGAs with SYCL, in: Proceedings of the Platform for Advanced Scientific Computing Conference, PASC '24, Association for Computing Machinery, New York, NY, USA, ISBN 9798400706394, https://doi.org/10.1145/3659914.3659925, 2024.

Dask Development Team: Dask: Library for dynamic task scheduling, http://dask.pydata.org, 2016.

Dueben, P. D., Chantry, M., Scher, S., Weyn, J. A., Camps-Valls, G., and Bauer, P.: Challenges and design choices for global weather and climate models based on machine learning, Geoscientific Model Development, 14, 4829–4845, 2021.

E3SM Project, D.: Energy Exascale Earth System Model v2.1.0, [Computer Software] https://doi.org/10.11578/E3SM/dc.20230110.5, https://doi.org/10.11578/E3SM/dc.20230110.5, 2023.

Golaz, J.-C., Van Roekel, L. P., Zheng, X., Roberts, A. F., Wolfe, J. D., Lin, W., Bradley, A. M., Tang, Q., Maltrud, M. E., Forsyth, R. M., Zhang, C., Zhou, T., Zhang, K., Zender, C. S., Wu, M., Wang, H., Turner, A. K., Singh, B., Richter, J. H., Qin, Y., Petersen, M. R., Mametjanov, A., Ma, P.-L., Larson, V. E., Krishna, J., Keen, N. D., Jeffery, N., Hunke, E. C., Hannah, W. M., Guba, O., Griffin, B. M., Feng, Y., Engwirda, D., Di Vittorio, A. V., Dang, C., Conlon, L. M., Chen, C.-C.-J., Brunke, M. A., Bisht, G., Benedict, J. J., Asay-Davis, X. S., Zhang, Y., Zhang, M., Zeng, X., Xie, S., Wolfram, P. J., Vo, T., Veneziani, M., Tesfa, T. K., Sreepathi, S., Salinger, A. G., Reeves Eyre, J. E. J., Prather, M. J., Mahajan, S., Li, Q., Jones, P. W., Jacob, R. L., Huebler, G. W., Huang, X., Hillman, B. R., Harrop, B. E., Foucar, J. G., Fang, Y., Comeau, D. S., Caldwell, P. M., Bartoletti, T., Balaguru, K., Taylor, M. A., McCoy, R. B., Leung, L. R., and Bader, D. C.: The DOE E3SM Model Version 2: Overview of the Physical Model and Initial Model Evaluation, Journal of Advances in Modeling Earth Systems, 14, e2022MS003 156, https://doi.org/10.1029/2022MS003156, 2022.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E.: Array programming with NumPy, Nature, 585, 357–362, https://doi.org/10.1038/s41586-020-2649-2, 2020.

Hoyer, S. and Hamman, J.: xarray: N-D labeled arrays and datasets in Python, Journal of Open Research Software, 5, https://doi.org/10.5334/jors.148, 2017.

Hunter, J. D.: Matplotlib: A 2D graphics environment, Computing in Science & Engineering, 9, 90–95, https://doi.org/10.1109/MCSE.2007.55, 2007.

Intel Corporation: Intel Fortran Compiler Developer Guide and Reference, https://www.intel.com/content/www/us/en/docs/fortran-compiler/developer-guide-reference/2023-0/compiler-options-001.html, 2023.

Kelleher, M. and Mahajan, S.: Detectable Climate, https://doi.org/10.5281/zenodo.15343054, 2025a.

https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

Kelleher, M. and Mahajan, S.: Detectable Climate Bootstrap Data, https://doi.org/10.5281/zenodo.15344575, 2025b.

Mahajan, S.: Ensuring statistical reproducibility of ocean model simulations in the age of hybrid computing, in: Proceedings of the Platform for Advanced Scientific Computing Conference, PASC '21, Association for Computing Machinery, New York, NY, USA, ISBN 9781450385633, https://doi.org/10.1145/3468267.3470572, 2021.

Mahajan, S., Gaddis, A. L., Evans, K. J., and Norman, M. R.: Exploring an Ensemble-Based Approach to Atmospheric Climate Modeling and Testing at Scale, Procedia Computer Science, 108, 735–744, https://doi.org/https://doi.org/10.1016/j.procs.2017.05.259, international Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland, 2017.

Mahajan, S., Evans, K. J., Kennedy, J. H., Xu, M., and Norman, M. R.: A multivariate approach to ensure statistical reproducibility of climate model simulations, in: Proceedings of the Platform for Advanced Scientific Computing Conference, pp. 1–10, 2019a.

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

Mielikainen, J., Price, E., Huang, B., Huang, H.-L. A., and Lee, T.: GPU Compute Unified Device Architecture (CUDA)-based Parallelization of the RRTMG Shortwave Rapid Radiative Transfer Model, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9, 921–931, https://doi.org/10.1109/JSTARS.2015.2427652, 2016.

Milroy, D. J., Baker, A. H., Hammerling, D. M., and Jessup, E. R.: Nine time steps: ultra-fast statistical consistency testing of the Community Earth System Model (pyCECT v3.0), Geoscientific Model Development, 11, 697–711, https://doi.org/10.5194/gmd-11-697-2018, 2018.

Pathak, J., Subramanian, S., Harrington, P., and et al.: FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators, Nature, 611, 459–464, 2022.

Qian, Y., Wan, H., Yang, B., Golaz, J.-C., Harrop, B., Hou, Z., Larson, V. E., Leung, L. R., Lin, G., Lin, W., Ma, P.-L., Ma, H.-Y., Rasch, P., Singh, B., Wang, H., Xie, S., and Zhang, K.: 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, 13,046–13,073, https://doi.org/10.1029/2018JD028927, 2018.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N.: WeatherBench: A benchmark dataset for data-driven weather forecasting, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 203, 2020.

Renard, B., Lang, M., Bois, P., Dupeyrat, A., Mestre, O., Niel, H., Sauquet, E., Prudhomme, C., Parey, S., Paquet, E., et al.: Regional methods for trend detection: Assessing field significance and regional consistency, Water Resources Research, 44, 2008.

Rosinski, J. M. and Williamson, D. L.: The Accumulation of Rounding Errors and Port Validation for Global Atmospheric Models, SIAM Journal on Scientific Computing, 18, 552–564, https://doi.org/10.1137/S1064827594275534, 1997.

Seabold, S. and Perktold, J.: statsmodels: Econometric and statistical modeling with Python, in: 9th Python in Science Conference, 2010.

Ventura, V., Paciorek, C. J., and Risbey, J. S.: Controlling the Proportion of Falsely Rejected Hypotheses when Conducting Multiple Tests with Climatological Data, Journal of Climate, 17, 4343 – 4356, https://doi.org/10.1175/3199.1, 2004.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods, 17, 261–272, https://doi.org/10.1038/s41592-019-0686-2, 2020.

https://doi.org/10.5194/egusphere-2025-2311
Preprint. Discussion started: 28 May 2025

Wan, H., Zhang, K., Rasch, P. J., Singh, B., Chen, X., and Edwards, J.: A new and inexpensive non-bit-for-bit solution reproducibility test based on time step convergence (TSC1.0), Geoscientific Model Development, 10, 537–552, https://doi.org/10.5194/gmd-10-537-2017, 2017.

Waskom, M. L.: seaborn: statistical data visualization, Journal of Open Source Software, 6, 3021, https://doi.org/10.21105/joss.03021, 2021.

Watson-Parris, D., Szekeres, C., Boukouvalas, A., et al.: ClimateBench: A Benchmark Dataset for Data-Driven Climate Projections, arXiv preprint arXiv:2206.07038, 2022.

Wes McKinney: Data Structures for Statistical Computing in Python, in: Proceedings of the 9th Python in Science Conference, edited by Stéfan van der Walt and Jarrod Millman, pp. 56 – 61, https://doi.org/10.25080/Majora-92bf1922-00a, 2010.

Whan, K. and Zwiers, F.: The impact of ENSO and the NAO on extreme winter precipitation in North America in observations and regional climate models, Climate Dynamics, 48, 1401–1411, 2017.

Wilks, D. S.: On “Field Significance” and the False Discovery Rate, Journal of Applied Meteorology and Climatology, 45, 1181 – 1189, https://doi.org/10.1175/JAM2404.1, 2006.

Wilks, D. S.: “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, 2263 – 2273, https://doi.org/10.1175/BAMS-D-15-00267.1, 2016.