Ensuring Statistical Reproducibility of Ocean Model Simulations in the Age of Hybrid Computing

Salil Mahajan
mahajans@ornl.gov
Computational Earth Sciences, Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA

ABSTRACT

Novel high performance computing systems that feature hybrid architectures require large scale code refactoring to unravel underlying exploitable parallelism. Such redesign can often be accompanied with machine-precision changes as the order of computation cannot always be maintained. For chaotic systems like climate models, these round-off level differences can grow rapidly. Systematic errors may also manifest initially as machine-precision differences. Isolating genuine round off level differences from such errors remains a challenge. Here, we apply two-sample equality of distribution tests to evaluate statistical reproducibility of the ocean model component of US Department of Energy's Energy Exascale Earth System Model (E3SM). A 2-year control simulation ensemble is compared to a modified ensemble as a test case – after a known non-bit-for-bit change in a model component is introduced – to evaluate the null hypothesis that the two ensembles are statistically indistinguishable. To quantify the false negative rates of these tests, we conduct a formal power analysis using a targeted suite of short simulation ensembles. The ensemble suite contains several perturbed ensembles, each with a progressively different climate than the baseline ensemble - obtained by perturbing the magnitude of a single model tuning parameter, the Gent and McWilliams κ, in a controlled manner. The null hypothesis is evaluated for each of perturbed ensembles using these tests. The power analysis informs on the detection limits of the tests for given ensemble size allowing model developers to evaluate the impact of an introduced non-bit-for-bit change to the model.

CCS CONCEPTS

• **Software and its engineering → Software verification and validation**; **Empirical software validation**.

KEYWORDS

hybrid computing, solution reproducibility, ocean modeling, climate statistics, machine learning

ACM Reference Format:
Salil Mahajan. 2021. Ensuring Statistical Reproducibility of Ocean Model Simulations in the Age of Hybrid Computing. In ,. ACM, New York, NY, USA, 9 pages.

1 INTRODUCTION

Effective utilization of novel high performance computing (HPC) systems that feature hybrid architectures require large-scale code refactoring to unravel underlying exploitable parallelism. Such redesign is accompanied by the challenge of ensuring solution reproducibility - does the modified refactored code produce the same answers? With hybrid architectures with many-cores, the order of computation cannot always be maintained and bit-for-bit reproducibility is not an easily achievable goal. Thus, machine round-off differences between the original solution and that produced by a refactored module are inevitable. For chaotic systems with inherent non-linearity, like Earth system models simulating the global climate, machine-precision differences in initial conditions are known to grow rapidly by orders of magnitudes compared to a baseline simulation [8, 14]. This is largely associated with the physical parameterizations representing the sub-grid scale processes (like clouds) in the models that result in stronger error growth than predictability error growth associated with non-linear processes in turbulent fluid flow [14]. Even minor perturbations to model tuning parameters are known to yield diverging solutions [5–7]. Errors introduced while refactoring that initially also manifest as near round-off differences, could thus also grow rapidly. The identification of systematic errors from innocuous round-off errors remains challenging, and we attempt to address the problem using statistical techniques on short ensemble simulations for ocean models here.

Typically, answer-changing modifications to Earth system models are addressed by evaluating the model climate by comparing a several years- or decades-long simulation with a control run visually, using a large number of plots and tables of metrics in an ad-hoc manner. Recent work on ensuring statistical reproducibility methodically has largely focused on atmosphere models. These approaches use short (1-year) simulation ensembles to quantify the natural variability of the model system. The simulation ensemble members only differ in their initial conditions that are perturbed by machine-precision-level differences. As these round-off differences grow rapidly, the simulations become independent of each other in a few days. [1, 10] used a principal-components-based (PC-based) approach using globally and annually averaged values of standard model output variables (> 120 variables) from the 1-year long simulation ensembles. Principal components (orthogonal linear combinations) of the output variables were computed from a large control ensemble. Data from a few (about 3) perturbed runs is then projected to these principal components. The individual projected values are evaluated against the density function of the corresponding PCs to evaluate if the projected values belong to the same population using a Z-score (distance of the perturbed data point from the ensemble mean in units of ensemble standard deviation). Lower Z-scores, under a given threshold, indicate that the modified run belongs to the same population represented by the control ensemble and thus the perturbed model reproduces the control model climate. [1] used an ensemble size of 151, and [10] suggested a larger control ensemble of more than 450 members composed of control simulations from different compilers to capture more of the model variability associated with different compiler and optimization changes.

[7] used modern machine-learning techniques for two-sample equality of distribution tests, like the energy test and the kernel test to evaluate the null hypotheses that a 1-year long control ensemble (about 60 members) and a perturbed ensemble (about 60 members) belong to the same population by comparing the multivariate probability distribution functions (pdf). Thus, if the two pdfs are statistically similar for a given significance level, the perturbed run is considered to be reproducing the climate of the control model statistically. These tests were able to identify differences in model simulations due to changes in key tuning parameters, and have been used to evaluate solution reproducibility after code refactoring and code porting. A few studies [6, 7] also conducted formal power analyses to evaluate the detection limits of the tests. Other cost-effective approaches have also been applied for testing reproducibility of the atmosphere model like the perturbation growth test [14] and the time-step convergence test [16], although they are only run for a few time-steps and do not span the full model code.

There are not many formal reproducibility tests for the ocean models. A simple test has been used by the CESM community to validate ports of its ocean model to new machines that evaluates the rate of growth of the root mean squared error (RMSE) during the first five days between a ported run and a control run. This RMSE growth is compared to the RMSE of a simulation where the convergence criteria for the solver is changed by an order of magnitude [2]. This approach has been found to be insufficient in identifying climate-changing modifications associated with the model's linear solver [4].

Recent work on formal comprehensive statistical reproducibility of ocean models [2] is based on comparing statistics of a single run of the perturbed model against a large control ensemble (about 40 members), at a 12 month time-slice (snap-shot) point of the model trajectories after initialization. To establish reproducibility, [2] evaluate the Z-score for each of the five independent prognostic ocean model variables: (temperature (T), zonal velocity (U), meridional velocity (V), sea surface height (SSH) and salinity (S)). Unlike the atmosphere model tests that use global averages, [2] evaluate the Z-score at each grid point for the ocean climate reproducibility test. Ocean variability is much more spatially heterogeneous than in the atmosphere because water is a denser fluid than air and has a much shorter Rossby radius of deformation. Thus, global averaging may dampen differences that arise over certain areas and thus may not be an ideal approach for evaluating the reproducibility of the ocean model climate. Also, [2] use a model configuration that is forced with prescribed annual cyclical (1-year repeating forcing) atmospheric fluxes of heat, moisture and momentum, so longer-term variability associated with atmosphere-ocean interactions is suppressed in the simulations.

Here, we take a different approach than [2] and compare a control ensemble against a perturbed model ensemble (instead of a single run) to test the null hypothesis that the two ocean model simulation ensembles belong to the same population. This is similar to two-sample equality of distribution approaches applied to the atmosphere in some previous studies [6–8]. But, following [2], we conduct these two-sample equality of distribution tests at each grid point of the ocean model instead of global averages.

The next section describes our testing approach in more detail. Section 3 describes the ocean model set-up we use. Section 4 and 5 discuss the false positive rates and the false negative rates of our testing approach. Finally, we summarize our results in Section 6.

2 TESTING APPROACH

Following previous studies [1, 2, 6–8], we consider the variability of the simulated ocean climate across short ensembles as a representative distribution of natural variability (noise statistics) of the model system. We generate short simulation ensembles for both the control and the modified model and evaluate the distribution of the annual average (of one year of simulation) at each grid point after the models reach steady state. Thus, there are no significant trends in the 1-year period used for evaluation and the annual averaging removes the effects of the seasonal cycle induced by the atmosphere flux forcings. To evaluate if the control and modified model are statistically equivalent, we pose the problem as a multivariate two-sample equality of distribution test on the two simulation ensembles following previous studies [6, 7] across all model grid points. We apply two testing frameworks described below for this purpose to evaluate the null hypothesis that the ensembles of control and modified model simulations belong to the same population (produce the same climate statistically). The testing frameworks are applied to the five prognostic ocean variables separately, following [2].

2.1 Kolmogorov-Smirnov Testing Framework

For our first testing framework, we use the widely known Kolmogorov-Smirnov (KS) two-sample equality of distribution test. It is a non-parametric test for univariate cases and is independent of the underlying distribution of the samples. Its test statistic is the maximum distance between the empirical cumulative distribution functions of the two samples, D_{nm}, and is defined as:

$$D_{nm} = \max_x |F_{1,n}(x) - F_{2,m}(x)| \tag{1}$$

where, n and m are the sample sizes of the two samples and $F_{1,n}$ and $F_{2,m}$ are their empirical cumulative distribution functions, respectively. The null distribution of D_{nm} is known and its critical values for hypothesis testing are widely available in statistical software libraries.

We conduct the KS test at each of the M grid points of the model for the null hypothesis, H_0, that the data at the corresponding grid points from the two ensembles are statistically indistinguishable. For the larger null hypothesis ($\mathbf{H_0}$) that the two ensembles belong to the same population, our test statistic (t) is the number of grid points that reject the null hypothesis H_0. The null distribution of t

and its critical value for conducting the larger hypothesis test, $\mathbf{H_0}$, is computed using resampling techniques as well as using correction approaches to critical values when conducting simultaneous multiple tests at all grid points (Section 2.3).

2.2 Cucconi Testing Framework

The Cucconi test is also a non-parametric two-sample equality of distribution test for univariate cases and is rank-based. It has been shown to exhibit more statistical power than other similar tests [9]. For the test, the two samples, say **A** and **B**, are pooled together and ranked. The Cucconi test statistic, C, is then computed as:

$$C = \frac{U^2 + V^2 - 2\rho UV}{2(1-\rho^2)} \tag{2}$$

where, U is the standardized sum of squared ranks of the data of sample, **A**, and V is the standardized sum of contrary-rank of the data of sample **A**. ρ is the correlation coefficient of U and V. Its null distribution is derived using permutations [9]. Under the permutations approach, the two samples are pooled together. They are then randomly relabeled into the two samples, and the test statistic is computed for this permutation. Repeating the relabeling a finite number of time yields an approximate null distribution of the test statistic and is used to compute the critical value for null hypothesis testing for the original samples. If all possible relabelings are conducted then the derived null distribution is exact. Large values of C indicate that the two samples are statistically distinguishable [9].

Similar to the KS testing framework, we conduct the Cucconi test at each grid point for H_0, and the test statistic (t) for the larger null hypothesis, $\mathbf{H_0}$, is the number of grid points that reject H_0, whose null distribution is computed by resampling techniques and correction approaches to critical values for multi-testing.

2.3 False Discovery Rate Approach for Simultaneous Multi-testing

Within our testing frameworks, we conduct hypothesis tests at each grid point of the ocean model simultaneously. For a single hypothesis test (one grid point), the null hypothesis is rejected if the p-value of the test statistic is less than a critical value, which is the given significance level (α, say 0.05). While conducting multiple tests (M grid points) simultaneously, one can expect $M\alpha$ grid points will reject the null hypothesis by chance even if they all are true, by design [e.g. 17]. There are several approaches to correct for such false rejection of true null hypotheses while multi-testing. We use one such approach called the false discovery rate (FDR) method [12, 15, 17], where we constrain the critical value for a given global significance level, α, as follows:

$$\alpha_{FDR} = \max_{j=1,2,\ldots,M} \{p_j : p_j \leq \alpha(j/M)\} \tag{3}$$

where, α_{FDR} is the new constrained critical value used for hypothesis testing at the global α significance level and p_j are the p-values of the M test statistics from each of the grid points sorted in ascending order. The above implies that α_{FDR} by design is less than α and thus is a constraint on it. We refer the reader to [15] (section 2a and Fig. 2) for a more comprehensive discussion on α_{FDR}. The null hypothesis at each grid point now is rejected only if the p-value of the test statistic is less than α_{FDR}. By design, the FDR approach also works as a field significance test [17], which refers to the global null hypothesis that the null hypothesis is true at all grid points. This is a manifestation of our larger null hypothesis, $\mathbf{H_0}$. Thus, with the FDR approach, if the null hypothesis, H_0, is rejected for even one grid point when using α_{FDR} as the critical value, the global null hypothesis $\mathbf{H_0}$ is rejected at the α significance level [e.g. 15, 17]. Thus, the critical value of the test statistic for $\mathbf{H_0}$ is zero at α significance level. Furthermore, the FDR approach has been shown to be effective even in the presence of significant spatial correlations as related to climatological data [15, 17] as applied here.

3 MODEL SET-UP

We use the Model for Prediction Across Scales - Ocean (MPAS-O) model [13] within Department of Energy's Energy Exascale Earth System Model (E3SMv1) [3] framework for developing our testing methodology. To adapt statistical tests for the problem of ocean model reproducibility, here, we use a low computational cost low-resolution version of the model with an effective horizontal resolution of about 240km. It contains 7153 ocean cells in the horizontal and 60 vertical layers. This version of the model is currently only used for regression testing and has not been validated for scientific use. This low-resolution model version, however, includes all of the model structure and source code including the dynamical core and physical parameterizations used in the scientifically validated higher-resolution versions of the model that are used for large scale climate simulations including those conducted for climate change studies. Our purpose here is not to understand ocean climate variability or its response to external forcing using this model version. We are solely comparing the statistics of the model against a perturbed version of the same model here, for which this low-cost model provides a cost effective option as we develop a testing infrastructure.

Nonetheless, we integrate the out-of-the-box low-resolution model forced with climatologically averaged 1-year cyclical atmospheric forcings (called normal year) for several decades to achieve quasi-equilibrium before branching out for ensemble simulations. This ensures that ensemble simulations remain largely isolated from artifacts of large model drifts as the model spins-up and produces a stationary climate. The presence of trends associated with model drifts could generate spurious similarities between the control and perturbed ensemble simulations, limiting the statistical power of the testing frameworks. Normal year cyclical atmospheric forcing also limits atmosphere-ocean coupled dynamics and thus long term variability associated with them like El Nino Southern Ocean (ENSO) are absent in these runs. Fig. 1 shows the daily time series of the global average ocean temperature for the run after default initialization. Even after a simulation length of about 90 years, a trend in global ocean temperatures can still be discerned. At about 50 years, the model seems to attain a quasi-equilibrium with stationary statistics for about a decade. We chose to branch out from this point of the model's transient trajectory for generating model ensembles for testing. Since we only run for a couple of years for statistical testing, our choice helps to minimize the effects of model drift on our results. We plan to run this simulation longer in the future to obtain an equilibrium model state before branching out.

4 EVALUATION OF FALSE POSITIVE RATE

With our choice of initial conditions from the trajectory of the long transient run, where the model is at a quasi-equilibirum state, we conduct a 150-member simulation ensemble. For each ensemble member, j, the three-dimensional initial condition temperature field is perturbed by machine-precision perturbations as follows:

$$T_i^j = T_i(1 + x_i^j) \tag{4}$$

where for ensemble member, j, a random perturbation, x_i^j is added to the initial condition temperature field T_i only at each of its grid points, i, to generate a perturbed initial condition, T_i^j. x_i^j is randomly picked from an open uniform distribution, $U(-10^{-14}, 10^{-14})$.

Fig. 2 shows the daily time series of the L1-Norm of sea surface temperature for each of the 150 ensemble members from the initialization time to day 2000 on the log-scale. L1-Norm is computed as the sum of absolute difference at each grid point between a perturbed initial conditions run and the control run trajectory. Machine precision differences grow quickly by orders of magnitude in the first few days. The error growth trajectory after the first few days is thus similar to that of an ensemble with initial conditions perturbed by a stronger perturbation of the order of 10^{-7} (machine precision of 32-bit architectures; not shown). After the end of the first year, the L1-Norms tend to become asymptotic, appearing to attain a quasi-steady state and become completely independent and uncorrelated of each other, and are essentially random solutions of the system to the imposed boundary condition of atmospheric fluxes. The L1-Norm grows very gradually thereafter through day 2000. Thus, we choose to use the annual average of the prognostic variables over the second year of simulation at each grid point for applying our statistical tests, when the perturbation growth has settled and perturbation growth trends do not lend spurious relationships between and within the control and test ensembles. We consider the ensemble spread of the second year annual mean to represent the distribution of the sample for the purpose of two-sample equality of distribution testing.

We assess the false positive rate - falsely rejecting a true null hypothesis - of the tests empirically, using the bootstrap resampling technique with this larger ensemble where we know that all members belong to the same population. We draw two samples of size 30 (arbitrary sample size choice, see below for the impact of sample size) each without replacement from this larger control ensemble of 150 members and conduct the KS and Cucconi testing frameworks on them for the five diagnostic variables (T, U, V, SSH, S) separately. We compute the test statistic, t - the number of grid points that reject the null hypothesis, H_0 - that the annual average of the second year of the simulation is statistically equivalent between the two ensembles at the 5% significance level. Repeating this resampling 500 times provides 500 values of t, yielding an approximate empirical null distribution of t. The critical value of the t can be obtained by computing percentiles. The results presented here are for SSH. Other diagnostic variables also yield qualitatively similar results and we do not discuss them further here. Ranking t (the number of grid points that reject the null hypothesis) in ascending order for the 500 resamplings, we find that for the KS test with the FDR correction, the 95th percentile value (critical value at the 5% significance level) of t is zero, as expected after FDR corrections. This implies that for 95% of the drawings the KS test rejects zero grid points for the null hypothesis, H_0, thus accepting $\mathbf{H_0}$ that the ensembles belong to the same population. For higher percentiles, the values of t are greater than zero. The above imply that the KS test rejects the null hypothesis that the two ensembles belong to the same population ($\mathbf{H_0}$) if even a single grid point between the two ensembles is statistically distinguishable, consistent with FDR theory. For Cucconi test, the 95th percentile value of t with FDR correction is also close to zero with 0.2% of the grid points, again largely consistent with FDR theory. Without FDR corrections the critical value for the KS test and Cucconi test are 5.9% and 8.9% of the grid points, respectively. We obtain qualitatively similar results with ensemble sizes of 40, 50 and 60. [2] also found that an ensemble size of 40 was sufficient to provide a low false positive rate, and a further increase in ensemble size provided little benefit.

5 TEST CASE

To assess if the KS and Cucconi testing frameworks can detect climate changing differences, we conduct a simulation ensemble with known changes to the climate statistics. This is achieved by perturbing the tuning parameter κ of the Gent and McWilliams (GM) eddy transport parameterization, that is commonly used to tune the simulations of the Atlantic meridional overturning circulation and Southern Ocean transport. The default value of the parameter is 1800 m^2/s for the 240km resolution model version. We change this value of κ to 600 m^2/s, which is known to change the model simulations dramatically and is the default value at other MPAS-O resolutions [11]. We conducted a 2-year 100-member ensemble run with $\kappa = 600 m^2/s$ using initial conditions at year 50 of the transient run, and perturbing the initial conditions with machine precision differences similar to the control ensemble. Fig. 2 shows the 2-year time series of the L1-Norm with respect to the transient run for SST for each of the ensemble members of this test ensemble, similar to the control ensemble. Visually, it is clear that the climate statistics of this ensemble significantly differs from that of the control ensemble, with no overlap across the spread of the time series of the two ensembles after the first year of simulations. We apply our testing frameworks to the annual average of the second year of the simulations when they appear to exhibit stationary climate statistics. Both the KS and Cucconi testing frameworks correctly reject the null hypothesis that the two ensembles belong to the same population at the 5% significance level using 100 members of each ensemble. More than 80% of the grid points rejected H_0 for both the testing frameworks for SSH and other variables. The result remains the same for ensemble sizes of 30, 40, 50 and 60. The null hypothesis is also rejected without using FDR corrections and using critical values obtained in Section 4. The above indicates that our testing frameworks can successfully detect known changes to climate statistics.

6 POWER ANALYSIS

We conduct a formal power analysis to assess the detection limits of the testing frameworks and the required ensemble size to detect a given magnitude of change. For this purpose, we integrate a suite of 2-year 100-member ensembles, where the value of κ is increasingly different from its default value of 1800 m^2/s for each ensemble in the suite. Values of κ range from 1800 to 600 m^2/s and each ensemble uses one fixed discrete value of κ within this range for all of its 100 members. The x-axis of Fig. 3 lists the fixed discrete values of κ used. For each of these values of κ, a 100-member ensemble simulation was conducted. The suite thus yields an increasingly different climate for each ensemble, compared to the control ensemble, as the difference between its κ and the default value increases. This controlled change in climate statistics allows the analysis of the false negative rate (β) - probability of accepting a false null hypothesis (Type II error rate) - and the detection limits of the testing frameworks.

We randomly draw N (arbitrarily, say 30) members without replacement from the 100-member ensemble with a perturbed κ and from the control ensemble and apply our testing frameworks on the second year of the simulations to evaluate if the null hypothesis is rejected. We conduct 500 such draws and compute the number of times the testing frameworks correctly reject the null hypothesis for SSH and the other variables. Fig. 3 shows the power (1 - β) of the KS and Cucconi testing frameworks for each of the perturbed κ ensembles when evaluated against the control ensemble for $N = 30, 40, 50, 60$. For $N = 30$, the KS testing framework can detect changes to the climate statistics induced by a small change to κ from 1800 to 1799 m^2/s with 100% probability. As the climate statistics become more dissimilar with increasing differences in κ, the KS testing framework is able to reject the null hypothesis consistently at near 100% probability. For smaller changes to κ the KS testing gradually loses power, correctly rejecting the null hypothesis only 40% of the times when $\kappa = 1799.95$. The statistical power of the framework increases with sample size. For $N = 60$, the framework can detect changes to climate statistics caused by a small change to κ from 1800 to 1799.5 m^2/s at near 100% probability.

The Cucconi testing framework can also detect changes to the climate statistics as κ is changed from 1800 to 1799 m^2/s with 100% probability for all ensemble sizes. For smaller changes to κ, it exhibits weaker power than the KS testing framework. For example, for $N = 30$, Cucconi testing framework only detects changes caused by changing κ from 1800 to 1799.5 about 55% of the times, whereas the KS testing framework detects the change with a higher probability of about 90%. Also, for testing without FDR corrections, the power curves are qualitatively similar to those with FDR corrections.

The above power analysis provides model developers using the tests with information on the magnitude of changes that the tests can detect for a given sample size. For example, in a non-bit-for-bit scenario, the acceptance of the null hypothesis with an ensemble size of 30, would imply that the change introduced is smaller than a change when κ is changed from 1800 to 1799 m^2/s. Such insights with the tests will allow developers and integrators to make an informed decision when accepting or rejecting an unintended non-bit-for-bit change to the model. It also provides them with information on the ensemble size needed to detect a given magnitude of change.

7 SUMMARY AND DISCUSSION

Novel HPC systems that feature hybrid architectures require large-scale code refactoring to unravel underlying exploitable parallelism, which may not always result in a bit-for-bit reproducible solution. Strong non-linearities in ocean models cause these round-off level differences to grow rapidly as noted in Fig. 2. Isolating these harmless round-off level differences from systematic errors, however, remains a challenge. Here, we demonstrate the application of two-sample equality of distribution tests, namely the KS and Cucconi testing frameworks, to evaluate statistical reproducibility of MPAS-O. The tests are conducted on the second-year average of two-year simulation ensembles on independent diagnostic variables of the ocean model like SSH. Our tests correctly reject the null hypothesis when a strong climate-changing modification is made by changing GM κ from 1800 to 600 m^2/s. To quantify the false negative rates of these tests, we conduct a formal power analysis using a targeted suite of short simulation ensembles. The ensemble suite contains several perturbed ensembles, each with a progressively different climate than the baseline ensemble - obtained by perturbing the magnitude of the GM κ, in a controlled manner from 1800 to 600 m^2/s. The null hypothesis is evaluated for each of perturbed ensembles using these tests. The tests appear to have strong statistical power and can identify differences induced by changing GM κ by a small amount from 1800 to 1799 with a high confidence with 30 or more ensemble members. The power analysis informs on the detection limits of the tests allowing model developers to evaluate the impact of an introduced non-bit-for-bit change to the model. It also informs on the ensemble size needed for the tests to detect a change of a given magnitude.

The difference in the ocean climate between a simulation with $\kappa = 1800$ and $\kappa = 1799$ may not be scientifically significant. But such a difference would almost always be rejected by our testing frameworks when using ensemble sizes of 30 or larger, as indicated by our power analysis. It is possible that our tests are too restrictive with a low false negative rate. We plan to engage with ocean modelers and scientists to investigate the required power of these tests and adjust the false positive rates of the test to allow accepting the null hypothesis when the magnitude of changes are physically justifiable.

We are expanding our test cases to various scenarios including other model tuning parameters (e.g. within K-profile parameterization (KPP) for vertical mixing, Redi diffusion along isopycnal surfaces), and known changes to the code base like innocuous changes to the order of operations to evaluate the robustness of our testing frameworks. Furthermore, we would expand on the power analyses by conducting additional ensemble suites with controlled gradual changes to model, by perturbing other model tuning parameters (KPP, Redi diffusion, etc.). Also, we have not formally compared our results with the approach of [1]. But, we intend to conduct the known climate changing experiments in their study to evaluate if our tests can detect those climate changes confidently.

Here, we conduct our tests on a low resolution model. The resolution sensitivity of our testing frameworks remains an open research question. However, the FDR approach is designed to account for multiple testing and performs robustly in the presence of spatial correlations. Thus, as the tests are applied to higher resolution data, the FDR corrections would adjust accordingly to reduce false positives. However, we have not tested this explicitly. We plan to investigate the resolution sensitivity of the robustness of our testing frameworks by evaluating their false positive and false negative rates in the near future, as we apply our tests to model configurations with increasing resolution up to eddy resolving resolutions where eddy parameterizations are not needed.

We believe our approach would have wide-ranging use aiding developers in model porting and refactoring for different architectures to improve model efficiency, for example for solver algorithms. Adhoc approaches where domain experts evaluate if a non-bit-for-bit change is climate changing using long simulations is too demanding for every minor change associated with refactoring. Simple approaches tracking error growth using global metrics like RMSE may also be inadequate in identifying climate changing behavior of ocean models, where the perturbations do not have a significant immediate effect [4]. Comprehensive statistical testing approaches, like [2] and our proposed alternative approach here that evaluate the model climate more fully as the perturbations propagate through all components of the ocean model, would be needed to maintain the trust-worthiness of climate models. Our approach would also be useful for evaluating changes to the source code that are not part of the time-stepping loop like infrastructure code for domain decomposition and parallel processing commands or writing output files [6]. We hope that these testing frameworks would soon become part of routine testing of models - complementing other regular tests, like regression tests and unit tests - for unavoidable non-bit-for-bit modifications allowing us to maintain confidence in climate model simulations following development work.

Also, machine-learning approaches are quickly being adopted by the climate community. Bit-for-bit reproducibility to the original solution will not be possible where machine learning tool-kits and modules are included in the model code base to replace current modules. Our tests provide a framework for a systematic evaluation of the simulated ocean climate as these new paradigm shifts occur in global Earth system modeling.

ACKNOWLEDGMENTS

The E3SM code that produced the simulations is open-source and can be accessed from https://e3sm.org/model. All the data used are listed in the references. 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. This manuscript has been authored by UT-Battelle, LLC which is supported by the Office of Science of the U.S. Department of Energy under Contract No.DE-AC05-00OR22725. This research used the resources of the Oak Ridge and Argonne Leadership Computing Facilities at the Oak Ridge and Argonne National Laboratories, respectively, and the National Energy Research Scientific Computing Center, which are supported by the Office of Science of the U.S. Department of Energy under Contracts DE-AC05-00OR22725, DE-AC02-06CH11357, and DE-AC02-05CH11231.

REFERENCES

[1] A. H. Baker, D. M. Hammerling, M. N. Levy, H. Xu, J. M. Dennis, B. E. Eaton, J. Edwards, C. Hannay, S. A. Mickelson, R. B. Neale, D. Nychka, J. Shollenberger, J. Tribbia, M. Vertenstein, and D. Williamson. 2015. A new ensemble-based consistency test for the Community Earth System Model (pyCECT v1.0). *Geoscientific Model Development* 8, 9 (2015), 2829–2840. https://doi.org/10.5194/gmd-8-2829-2015

[2] A. H. Baker, Y. Hu, D. M. Hammerling, Y.-H. Tseng, H. Xu, X. Huang, F. O. Bryan, and G. Yang. 2016. Evaluating statistical consistency in the ocean model component of the Community Earth System Model (pyCECT v2.0). *Geoscientific Model Development* 9, 7 (2016), 2391–2406. https://doi.org/10.5194/gmd-9-2391-2016

[3] Jean-Christophe Golaz, Peter M. Caldwell, Luke P. Van Roekel, Mark R. Petersen, Qi Tang, Jonathan D. Wolfe, Guta Abeshu, Valentine Anantharaj, Xylar S. Asay-Davis, David C. Bader, Sterling A. Baldwin, Gautam Bisht, Peter A. Bogenschutz, Marcia Branstetter, Michael A. Brunke, Steven R. Brus, Susannah M. Burrows, Philip J. Cameron-Smith, Aaron S. Donahue, Michael Deakin, Richard C. Easter, Katherine J. Evans, Yan Feng, Mark Flanner, James G. Foucar, Jeremy G. Fyke, Brian M. Griffin, Cécile Hannay, Bryce E. Harrop, Mattthew J. Hoffman, Elizabeth C. Hunke, Robert L. Jacob, Douglas W. Jacobsen, Nicole Jeffery, Philip W. Jones, Noel D. Keen, Stephen A. Klein, Vincent E. Larson, L. Ruby Leung, Hong-Yi Li, Wuyin Lin, William H. Lipscomb, Po-Lun Ma, Salil Mahajan, Mathew E. Maltrud, Azamat Mametjanov, Julie L. McClean, Renata B. McCoy, Richard B. Neale, Stephen F. Price, Yun Qian, Philip J. Rasch, J. E. Jack Reeves Eyre, William J. Riley, Todd D. Ringler, Andrew F. Roberts, Erika L. Roesler, Andrew G. Salinger, Zeshawn Shaheen, Xiaoying Shi, Balwinder Singh, Jinyun Tang, Mark A. Taylor, Peter E. Thornton, Adrian K. Turner, Milena Veneziani, Hui Wan, Hailong Wang, Shanlin Wang, Dean N. Williams, Phillip J. Wolfram, Patrick H. Worley, Shaocheng Xie, Yang Yang, Jin-Ho Yoon, Mark D. Zelinka, Charles S. Zender, Xubin Zeng, Chengzhu Zhang, Kai Zhang, Yuying Zhang, Xue Zheng, Tian Zhou, and Qing Zhu. 2019. The DOE E3SM Coupled Model Version 1: Overview and Evaluation at Standard Resolution. *Journal of Advances in Modeling Earth Systems* 11, 7 (2019), 2089–2129. https://doi.org/10.1029/2018MS001603 arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018MS001603

[4] Y. Hu, X. Huang, A. H. Baker, Y. Tseng, F. O. Bryan, J. M. Dennis, and G. Yang. 2015. Improving the scalability of the ocean barotropic solver in the community earth system model. In *SC '15: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–12. https://doi.org/10.1145/2807591.2807596

[5] H-L. Liu, F. Sassi, and R. R. Garcia. 01 Jan. 2009. Error Growth in a Whole Atmosphere Climate Model. *Journal of the Atmospheric Sciences* 66, 1 (01 Jan. 2009), 173 – 186. https://doi.org/10.1175/2008JAS2825.1

[6] Salil Mahajan, Katherine J Evans, Joseph H Kennedy, Min Xu, and Matthew R Norman. 2019. A multivariate approach to ensure statistical reproducibility of climate model simulations. In *Proceedings of the Platform for Advanced Scientific Computing Conference*. 1–10.

[7] Salil Mahajan, Katherine J Evans, Joseph H Kennedy, Min Xu, Mathew R Norman, and Marcia 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 (2019), 784–790.

[8] Salil Mahajan, Abigail L Gaddis, Katherine J Evans, and Matthew R Norman. 2017. Exploring an ensemble-based approach to atmospheric climate modeling and testing at scale. *Procedia Computer Science* 108 (2017), 735–744.

[9] Marco Marozzi. 2013. Nonparametric Simultaneous Tests for Location and Scale Testing: A Comparison of Several Methods. *Communications in Statistics - Simulation and Computation* 42, 6 (2013), 1298–1317. https://doi.org/10.1080/03610918.2012.665546 arXiv:https://doi.org/10.1080/03610918.2012.665546

[10] Daniel J. Milroy, Allison H. Baker, Dorit M. Hammerling, John M. Dennis, Sheri A. Mickelson, and Elizabeth R. Jessup. 2016. Towards Characterizing the Variability of Statistically Consistent Community Earth System Model Simulations. *Procedia Computer Science* 80 (2016), 1589–1600. https://doi.org/10.1016/j.procs.2016.05.489

[11] Mark R. Petersen, Xylar S. Asay-Davis, Anne S. Berres, Qingshan Chen, Nils Feige, Matthew J. Hoffman, Douglas W. Jacobsen, Philip W. Jones, Mathew E. Maltrud, Stephen F. Price, Todd D. Ringler, Gregory J. Streletz, Adrian K. Turner, Luke P. Van Roekel, Milena Veneziani, Jonathan D. Wolfe, Phillip J. Wolfram, and Jonathan L. Woodring. 2019. An Evaluation of the Ocean and Sea Ice Climate of E3SM Using MPAS and Interannual CORE-II Forcing. *Journal of Advances in Modeling Earth Systems* 11, 5 (2019), 1438–1458. https://doi.org/10.1029/2018MS001373 arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018MS001373

[12] B. Renard, M. Lang, P. Bois, A. Dupeyrat, O. Mestre, H. Niel, E. Sauquet, C. Prudhomme, S. Parey, E. Paquet, L. Neppel, and J. Gailhard. [n.d.]. Regional

methods for trend detection: Assessing field significance and regional consistency. *Water Resources Research* 44, 8 ([n. d.]). https://doi.org/10.1029/2007WR006268 arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2007WR006268

[13] Todd Ringler, Mark Petersen, Robert L. Higdon, Doug Jacobsen, Philip W. Jones, and Mathew Maltrud. 2013. A multi-resolution approach to global ocean modeling. *Ocean Modelling* 69 (2013), 211 – 232. https://doi.org/10.1016/j.ocemod.2013.04.010

[14] J M Rosinski and D L Williamson. 1997. The accumulation of rounding errors and port validation for global atmospheric models. *SIAM Journal on Scientific Computing* 18, 2 (3 1997). https://doi.org/10.1137/S1064827594275534

[15] Valérie Ventura, Christopher J. Paciorek, and James S. Risbey. 2004. Controlling the Proportion of Falsely Rejected Hypotheses when Conducting Multiple Tests with Climatological Data. *Journal of Climate* 17, 22 (2004), 4343–4356. https://doi.org/10.1175/3199.1 arXiv:https://doi.org/10.1175/3199.1

[16] H. Wan, K. Zhang, P. J. Rasch, B. Singh, X. Chen, and J. Edwards. 2017. A new and inexpensive non-bit-for-bit solution reproducibility test based on time step convergence (TSC1.0). *Geoscientific Model Development* 10, 2 (2017), 537–552. https://doi.org/10.5194/gmd-10-537-2017

[17] D. S. Wilks. 2006. On "Field Significance" and the False Discovery Rate. *Journal of Applied Meteorology and Climatology* 45, 9 (2006), 1181–1189. https://doi.org/10.1175/JAM2404.1 arXiv:https://doi.org/10.1175/JAM2404.1

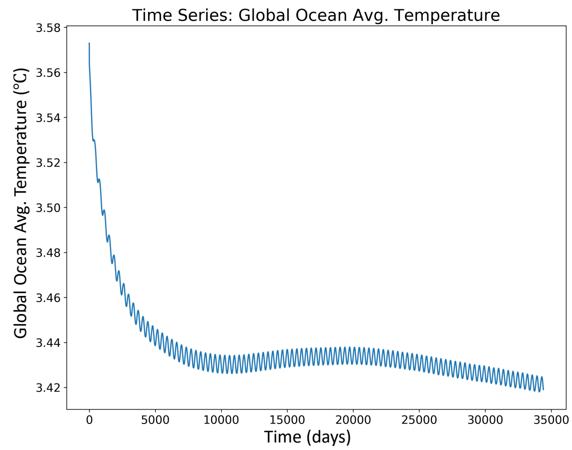


Figure 1: Time series of the global average ocean temperature of a low resolution version of MPAS-O, the ocean model component of E3SMv1, forced with a cyclical normal year atmospheric forcing initialized with default initial conditions.

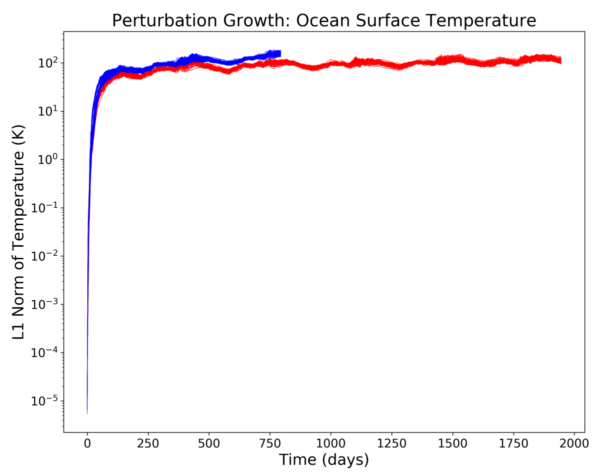
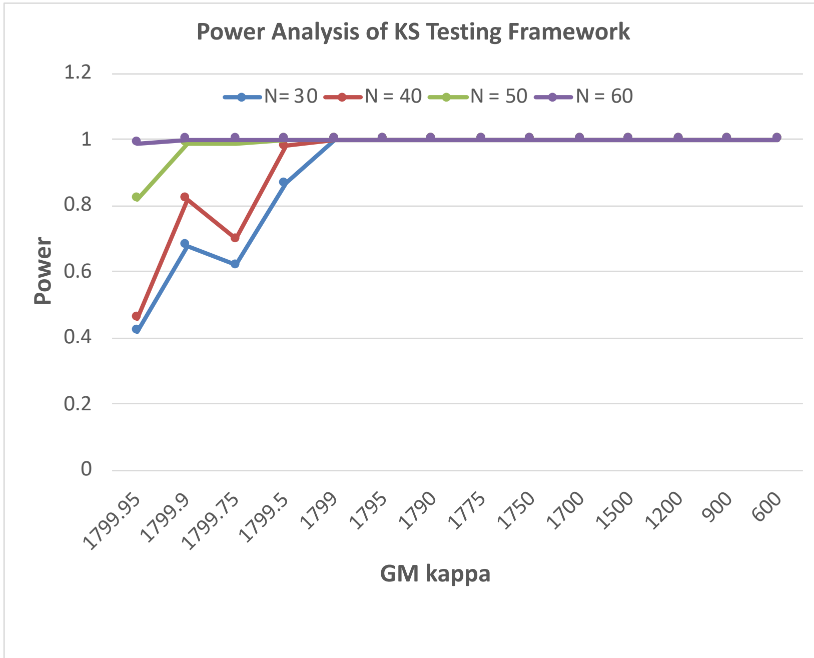


Figure 2: Growth of machine-precision differences in MPAS-O and ensemble spread: L1-Norm (sum of absolute difference at each grid point, log-scale) of SST of each of the 150 ensemble members with round-off differences in initial conditions compared to a reference run for the control (κ = 1800 m^2/s, red lines) ensemble. The same is also shown for the 100 members of a perturbed (κ = 600 m^2/s, blue lines) ensemble. The control run was integrated for several years to assess its quasi-equilibrium. The perturbed ensemble run is integrated for 2 years.

a.



b.

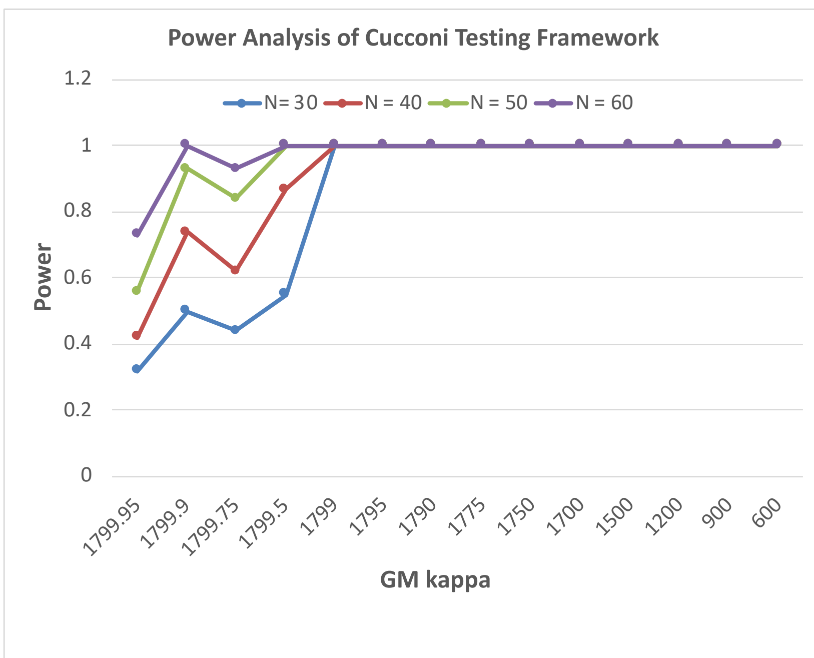


Figure 3: Power Analysis. Probability of correctly rejecting a false null hypothesis (power) of the test in detecting changes to an MPAS-O tuning parameter from a control case (GM κ = 1800 m^2/s) for different ensemble sizes (N).