Exploring an Ensemble-Based Approach to Atmospheric Climate Modeling and Testing at Scale

Salil Mahajan[1], Abigail L. Gaddis[1], Katherine J. Evans[1], and Matthew R. Norman[2]

[1] Computer Science and Mathematics Division, Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, USA
mahajans@ornl.gov; gaddisal@ornl.gov; evanskj@ornl.gov;
[2] Oak Ridge Leadership Computing Facility, Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, USA
normanmr@ornl.gov

Abstract

A strict throughput requirement has placed a cap on the degree to which we can depend on the execution of single, long, fine spatial grid simulations to explore global atmospheric climate behavior in more detail. Running an ensemble of short simulations is economical as compared to traditional long simulations for the same number of simulated years, making it useful for tests of climate reproducibility with non-bit for bit changes. We test the null hypothesis that the climate statistics of a full-complexity atmospheric model derived from an ensemble of independent short simulation is equivalent to that from a long simulation. The climate statistics of short simulation ensembles are statistically distinguishable from that of a long simulation in terms of the distribution of global annual means, largely due to the presence of low-frequency atmospheric intrinsic variability in the long simulation. We also find that model climate statistics of the simulation ensemble are sensitive to the choice of compiler optimizations. While some answer-changing optimization choices do not effect the climate state in terms of mean, variability and extremes, aggressive optimizations can result in significantly different climate states.

Keywords: reproducibility, climate simulation, ensemble testing

1 Introduction

Traditional modeling studies addressing climate science questions rely largely on several long, temporally dependent simulations. Relying on single or small long ensembles of simulations is becoming a burden for state-of-the-art high resolution global climate models because the approach does not scale with model complexity. To gain more realism in the model, whether through horizontal grid refinement, additional vertical levels, or new model features, the model

will require more compute cycles to complete because no realistic atmospheric model achieves perfect weak or strong scaling at the levels of parallelism needed for decent throughput.

As a result, the workload per compute node decreases "at scale" and the model spends relatively more time exchanging data between nodes. Also, the MPI messages sent between compute nodes become smaller such that nearly all of the time is spent in latency costs. Futhermore, it appears that latency will not improve significantly with future computing architectures, so this problem will only become worse. The Department of Energy's (DOE) Accelerated Climate Model for Energy (ACME) V1 prototype simulations spend more than 90% of their time waiting for MPI data within the atmospheric dynamical core, and that waiting is dominated by latency costs. Clearly, ACME's traditional science pathway has hit a barrier that cannot be removed without rethinking our simulation strategy. If one can cast the methodology to address key climate science questions in a manner amenable to Short Independent Simulation Ensembles (hereafter, SISE) as opposed to (or in cooperation with) Single, Long Runs (hereafter, SLRs), the negative effects of the throughput constraint can be ameliorated. For example, running 100 independent one-year-long ensembles instead of a single 100-year run, produces a 100x greater workload per node and, therefore, significantly reduced relative MPI and PCI-e overheads (i.e., better parallel scaling).

Given the dissolution of Moore's Law for the throughput of climate simulations with the next generation of computing, we investigate the viability of SISE for testing and scientific analysis in an effort to help climate community achieve its science goals. It is encouraging that there are already scientific studies in climate that have used SISE for scientific benefit. For instance, Verma et al. [9] have conducted an ensemble of fully-coupled sulfate aerosol forced short runs (1 year) starting with initial conditions from various points of a pre-industrial control run trajectory. Also, the Large Ensemble Community Project [3] has provided wealth of data for analysis. Using ensembles for tuning of model parameters is also of interest, for example in [10], where few days short simulation ensembles with perturbed parameters are used.

In harnessing modern hybrid computer architectures, climate simulations generate machine round-off level answer changes from identically configured runs via refactoring source code, compiler changes, compiler optimization choices as well as processor layout changes. The non-linear chaotic nature of the climate system causes these minute perturbations to grow quickly. It is critical to ensure that any of these changes are not systematic, leading to a climate state that is distinguishable from the validated model baseline. Recognizing this, SISE have also been used for verification testing of a new model simulation for these development scenarios [1, 7].

First, we verify the utility of SISE for detecting model changes that lead to statistically disparate climates as demonstrated by [1], with a different strategy. This entails testing the null hypothesis that two SISE belong to the same population. The computational benefits of using SISE in lieu of SLRs in practice are explained. We investigate whether SISE can in fact replace SLRs to represent the natural variability within the atmospheric model, given that the non-linear growth of small state differences limit deterministic predictability to time scales of 10-20 days. James and James [5] show that non-linear atmospheric dynamics can also generate variability of planetary-scale features on decadal time scales, which will not be captured by SISE. Atmospheric internal variability is also understood to generate low-frequency modes of variability like the North Atlantic Oscillation (NAO). NAO variability ranges from sub-seasonal to multidecadal timescales and atmospheric models forced with prescribed SST climatology alone can generate NAO variability across these timescales with little role for the ocean [8, 4].

2 Experiments and Computational Benefits of SISE

We use a version of ACME that predates the first release of the model, used as a baseline for verification. It matches the configuration used in [6] in terms of the model tag and specifies the same active atmosphere and land surface components with a data-only ocean model. However, the ocean data is cycled annually, valid at year 1850 rather than the present day. This component set with this resolution has been tuned for global energy balance and satisfies our requirements for climate behavior during pre-industrial conditions. For our experiments, we use a model configuration with a cubed-sphere spectral element atmospheric grid (with 16x16 elements on each of the six panels) that results in an average equatorial grid spacing of about 208km, which is equivalent to about 2^o in more traditional latitude-longitude coordinates. The model configuration also uses 30 vertical levels. We briefly describe the SLR and SISE using this setup (named *2deg*), and Table 1 provides a brief summary of our experiments:

1. SLR is the baseline control simulation of a single 100 year run, representing a traditional approach to establishing a climate baseline. The land is initialized with a cold start. We discard the first 20 years of the simulation as spin up. These simulations used the PGI 15.7 compiler with "-O2 -Kieee -Mvect=nosse -tp=istanbul-64" flags passed in and used 96 nodes on the Oak Ridge Leadership Computing Facility (OLCF)'s Titan machine with an all-MPI decomposition (16 MPI tasks per node, no GPU). This gives an average of 16 elements per node.

2. SISE experiments comprise about 60 simultaneous one-year-long simulations. Each ensemble member perturbs the initial 3-D temperature field using a Psuedo Random Number Generator (PRNG) with a uniform distribution to a relative magnitude of 10^{-14}, which is near machine precision. The computational setup per job is identical to SLR, and the ensembles are run in parallel within the same job submission script. We could easily run each of these simulations at much lower node counts than the SLR and still utilize Titan's favorable queueing system policy for large jobs, because there are so many executables running in parallel. For this conceptual study, we assigned each simulation 48 nodes (32 elements per node). These experiments illustrate the benefit of additional parallelism for ensembles; when extending this strategy for high-resolution configurations, the workload is so much greater that there is more flexibility in the layouts to achieve favorable queue status on Titan. Consider what we would be able to do with 1^o grid ensembles: using 128 elements per node, a 100-member ensemble would be able to use over 4,000 nodes in a single job submission, which is considered "capability scale."

   The SISE-DEFAULT experiment uses exactly the same compiler options as SLR. SISE-O1 uses a reduced optimization of "-O1 -Kieee -Mvect=nosse -tp=istanbul-64", whereas SISE-FAST applies a very aggressive optimization, "-fast -Mvect -tp=bulldozer-64".

3 Testing Methodology

3.1 Equality of Distribution

To test the null hypothesis that two simulation ensembles represent the same climate state ($\mathbf{H_0}$), we implement a testing framework based on testing the equality of distribution of each variable in the standard model output (158 variables) between the two simulations. The test statistic (t) for $\mathbf{H_0}$ is the number of variables that reject the null hypothesis of equality of

Table 1: List of experiments. Failures of several members, as expected using a multi-petascale computer, create slightly different ensemble sizes for SISEs job submissions.

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

distribution (H_0) at a given confidence level (say 95%). $\mathbf{H_0}$ is rejected if $t > \alpha$, where α is some critical number (threshold). We use the non-parametric Kolmogorov-Smirnov (K-S) test as the univariate test of equality of distribution of global means. The critical number (α) is obtained from an empirically derived approximate null distribution of t using resampling techniques. This KS-test based testing structure provides an alternative to [1], which involves rotation into orthogonal principal component space and requires a large control simulation ensemble. This could be useful in situations where the generation of a large ensemble is not feasible (particularly for high resolution model configurations) or readily available.

Significant correlations exist between global means of different variables. Thus, α cannot be determined simply from the significance level of the equality of distribution hypothesis (H_0) test. So, one cannot reject $\mathbf{H_0}$ if more than 5% of the variables reject H_0 at the 95% confidence level, because correlations between variables will increase the likelihood of correlated variables rejecting the null hypothesis simultaneously. Instead, we empirically derive the approximate null distribution of the test statistic from random resampling. In a Monte-Carlo (random) permutation-test approach, simulations from the two ensembles of size n and m are pooled together and the simulations from the pool are the randomly assigned to one of two groups of sizes n and m. The t-statistic is then computed for the random drawing. If all possible such random drawings are made, the null distribution of t is exact. Here, we only conduct 500 re-samplings for all cases, which yields an approximate null distribution. Separately, we also use a bootstrapping approach, where sub-samples are drawn from the same simulation ensemble.

3.2 Climate Extremes

Classical non-parametric distance-based tests of equality of distributions like the K-S test are not robust for distributions with different tails because the cumulative distribution functions converge at the tails. So we evaluate extremes separately using the generalized extreme value (GEV) theory. The GEV theory postulates that the maxima or minima of a process can be represented by the three parameter GEV family of distributions asymptotically, irrespective of the underlying distribution of the process. The GEV family of distributions can be represented as:

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

where μ, σ and ξ represent the location, scale and shape parameter respectively of the distribution, and z is a normalized maxima or minima of a process. For situations where $\xi = 0$, the function is interpreted as the limit of the equation as $\xi \to 0$ [2]. Return periods of extremes, more familiar to the risk assessment community, can be computed by inverting the distribution

[2]. Presently, we evaluate the location parameter of the GEV model for annual maxima of daily average surface temperature and precipitation rate. Annual maxima of temperature or precipitation for each year (each ensemble member in case of short simulation ensembles) for each grid point are fit to a GEV distribution using a maximum likelihood approach. Maximum likelihood maximizes the probability of the occurrence of the values provided for fitting. The estimated GEV parameters belong to a multivariate normal distribution approximately [2]. We thus use the Student's t-test to test the null hypothesis (G_0) that the location parameter estimated from two samples belong to the same population, at each grid point, based on the standard error estimates of the parameter from the maximum likelihood method.

Our larger null hypothesis ($\mathbf{G_0}$) is that the simulation of extremes of a variable between two simulation ensembles is statistically indistinguishable. We choose the test statistic (g) for $\mathbf{G_0}$ to be the number of grid points that reject G_0. As with the mean values, even extremes at grid points cannot be considered to be independent of each other, as will be shown in Section 4.1.2, because significant spatial correlations exist in the climate system locally and remotely via teleconnection mechanisms. We again use a resampling approach as in Section 3.1 to determine the null distribution of g.

4 Results

4.1 Differences arising from choice of compiler optimization

4.1.1 Equality of Distributions

We apply the KS-test based testing framework to different pairings of SISE experiments, each with a different level of optimization. Table 2 lists the values of the test statistic (t) for the KS test at the 95% confidence level for the three pairings. Also listed are critical values (α) based on the null distribution derived from the 500 permutations (Monte Carlo permutation test) for each pairing at the 95% confidence level. Table 2 also lists the results of null hypothesis $\mathbf{H_0}$ test, that the two SISE belong to the same population, based on α obtained from the Monte Carlo permutation test approach (accept if $t < \alpha$). Bootstrapping, where 65 ensemble members of the SISE-DEFAULT experiment are randomly pooled without replacement into two groups of 30 members each 500 times, reveal that 95% of the random drawings have 13 variables (critical value) that reject the null hypothesis. The results, thus, do not change if the bootstrap critical value is used for hypothesis testing.

The results of the KS-test (Table 2) indicate that SISE-DEFAULT (with -O2) and SISE-O1 simulations represent the same climate state. Thus, the simulated climate with an optimization choice of -O2, although answer changing, is virtually indistinguishable from -O1 optimized simulations with machine precision round-off level differences in initial conditions. Simulations using the -fast optimization however, produce a climate state that is statistically distinct from both the SISE-DEFAULT and the SISE-O1 experiments. Therefore, aggressive compiler choices with the PGI compiler on Titan can result in climate-changing simulations. This result is similar to the conclusions of previous studies using different testing strategies, compiler choices, and machines [1, 7, 11], indicating that the result is robust.

4.1.2 Climate Extremes

Fig. 1 a,b show the location parameter (μ) of the GEV fit to the maximum annual daily surface temperature and precipitation rate at each grid point for SISE-DEFAULT ensemble. Their difference in μ between SISE-DEFAULT and SISE-O1 ensembles are presented in color

Table 2: KS-test Results. H_0 represents the univariate null hypothesis of equality of distribution for each of the 158 variables, whereas $\mathbf{H_0}$ represents the larger null hypothesis that two simulation ensembles simulate identical climates. The test statistic (t) is the number (%) of variables that reject H_0 based on the KS-test. The critical value (α) is derived from Monte Carlo Permutations. $\mathbf{H_0}$ is accepted if $t < \alpha$.

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

in Figures 1 c,d if they are statistically distinct. The spatial coherence of grid points that are statistically different between the two simulation ensembles in Fig. 1 c,d suggests a spatial correlation of extremes. Table 3 summarizes the results of climate extremes test. It lists the test statistic value (g) - the percentage of grid points that have statistically distinct location parameter at a 95% confidence level - for surface temperature and precipitation for each comparison. It also lists the critical value (α) derived from the approximate null distribution from Monte Carlo permutations. The result of the test of the null hypothesis ($\mathbf{G_0}$) that two simulation ensembles simulate identical climate extremes at the 95% confidence level based on g are also presented. Only land grid points are considered for temperature extremes, since the sea surface temperatures are prescribed in these simulations.

Bootstrapping, where the 65 SISE-DEFAULT are randomly pooled to one of two groups of 30 each and evaluated for their μ, reveals that 95% of these random samples have 6.8% of grid points rejecting the null hypothesis for annual maxima of daily precipitation. For annual maxima of daily surface temperature, the critical value is 10.8%. These values of β are close to values obtained from the Monte Carlo Permutations and thus the results listed in Table 3 do not change if the critical value computed from bootstrapping is used instead of that from the Monte Carlo permutations.

The results thus indicate the all SISE simulations are identical to each other in terms of their simulation of climate extremes, although surface temperature extremes between SISE-O1 and SISE-FAST are marginally statistically distinct. This is in contrast to the result of the KS-testing framework which indicates that climate as defined by the distribution of the variables in the standard model output in SISE-FAST is distinct from SISE-DEFAULT and SISE-O1. This either suggests that optimization choices do not effect climate extremes or that climate extremes are not a good metric to evaluate answer changes that might effect the simulation of the climate, with 60 ensemble members. We will explore climate extremes in this context more in future work.

4.2 Long simulation vs. short simulation ensemble

The long control simulation (SLR) of 65 years is broken down into an ensemble of one year simulation segments for comparing against SISE. Table 4 lists the results of the comparison of 65 one-year segments of SLR and the 65 member SISE-DEFAULT ensemble using the KS-test based testing framework. 80 (50.6%) variables fail the KS-test. Applying the Monte Carlo permutation approach on SLR and SSE-DEFAULT yields a critical value of 15 at the 95% confidence level. The critical value obtained from bootstrapping the 65 member SISE-DEFAULT ensemble into two groups of 30 is 13 as discussed in Section 5.1.1. The SLR simulation is clearly distinct from the SISE-DEFAULT simulation based on these critical values. We thus conclude

Table 3: Climate extremes test results. G_0 represents the null hypothesis that GEV location parameter (μ) is statistically identical at the 95% confidence level for each grid point, whereas $\mathbf{G_0}$ represents the larger null hypothesis that two simulation ensembles simulate statistically identical extremes at the 95% confidence level. The test statistic (g) is the percentage of variables that reject G_0 based on the Student's t-test. The critical value (β) is derived from Monte Carlo permutations. $\mathbf{G_0}$ is accepted if $t < \alpha$.

Comparison	**Variable**	**Test statistic (g)**	**Critical value (β)**	**$\mathbf{G_0}$ Test**
SISE-DEFAULT vs. SISE-O1	Precipitation Rate	5.1%	6.5%	Accept $\mathbf{G_0}$
	Surface Temperature	5.0%	9.6%	Accept $\mathbf{G_0}$
SISE-DEFAULT vs. SISE-FAST	Precipitation Rate	4.7%	6.3%	Accept $\mathbf{G_0}$
	Surface Temperature	3.6%	9.6 %	Accept $\mathbf{G_0}$
SISE-O1 vs. SISE-FAST	Precipitation Rate	5.2%	6.5%	Accept $\mathbf{G_0}$
	Surface Temperature	10.3%	9.8%	Reject $\mathbf{G_0}$

that the null hypothesis - that the climate statistics between SISE and a long simulation are statistically equivalent - does not hold.

All the simulations also include an active land, which could influence the internal variability of the system. The segment of the SLR experiment considered had equilibrated land conditions, which were then allowed to evolve. Whereas, the SISE simulations were each initialized with the same land conditions obtained from the last year of the equilibrated SLR simulation. The SISE-LND-INIT experiment, identical to SISE-DEFAULT except the land initial conditions for each of the 65 years of simulation is borrowed from a year of the control run sequentially, captures some of the impact of land conditions simulated in SLR. However, the SISE-LND-INIT ensemble is also found to be statistically distinct from the long control simulation (Table 4), based on the critical value from bootstrapping, indicating that the atmosphere is largely responsible for creating the differences in climate statistics between SLR and SISE.

Table 4: KS test results: SLR vs. SISE.

Comparison	**Test Statistic (t)**	**Critical Value (α)**	**$\mathbf{H_0}$ Test Result**
SLR vs. SISE-DEFAULT	80 (50.6 %)	15	Reject $\mathbf{H_0}$
SLR vs. SISE-LND-INIT	74 (48 %)	13	Reject $\mathbf{H_0}$

Individual simulations in the SISE become independent of each other in 10-20 days as illustrated in Fig. 2, where the error growth of temperature appears to be correlated only for the first few days of the simulations. Our tests thus indicate that the spread of this ensemble does not accurately represent the inter-annual variability of the SLR simulation as the atmosphere is allowed to evolve. It is known that free atmospheric internal variability also includes variability on longer time-scales. We illustrate this fact in Fig 3a, which shows the spectrum of the time-series of 80 years of global average monthly surface temperature after removing the seasonal cycle. The SLR experiment shows large variability on time-scales longer than a year. A spectrum of a time-series generated by sequentially joining all 65 ensemble members of the SISE-DEFAULT experiment conspicuously shows the absence of low frequency variability

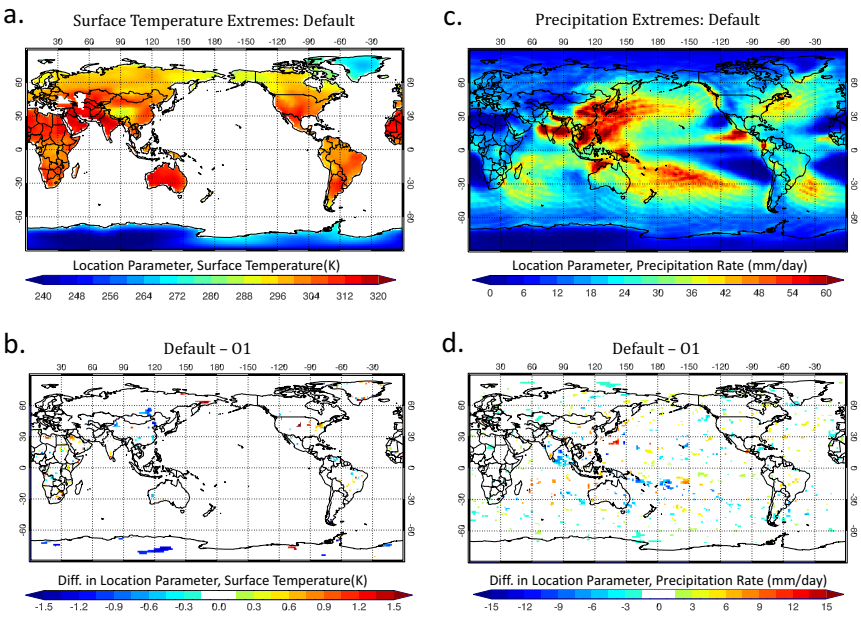

Figure 1: Climate Extremes. Location parameter of (a) surface temperature (K) over land areas and (b) precipitation rate (mm/day) for the default short simulation ensemble. Difference in location parameter between SISE-DEFAULT and SISE-O1 experiments for (c) surface temperature and (d) precipitation rate. Colored areas represent grid points where the extremes are statistically distinguishable at the 95% confidence level.

- with some artifacts on one to three-year time-scales (Fig. 3b). This low frequency variability in the SLR experiment will significantly influence the variability of the ensemble derived from breaking the SLR simulation into one year segments. This is consistent with the rejected null hypothesis for the derived SLR and SISE-default ensemble set. We plan to investigate the impact of low frequency variability on climate statistics in this context more in future work.

5 Summary and Discussion

Motivated by computational efficiency in high performance computing environments, we investigate SISE as a potential framework to provide model verification and testing and for scientific analysis.

To evaluate if two global climate model simulation ensembles simulate the climate statistically, we implement a KS equality of distribution test based framework. We also apply the GEV theory to test if extremes simulated by two simulation ensembles are statistically equivalent. On Titan, the use of aggresive optimizations with PGI compilers can result in statistically distinct climate simulations. We also find that climate extreme statistics are not sensitive to such optimizations, based on about 60 one-year simulation ensembles. Surface temperature extremes are found to differ between SISE-O1 and SISE-fast but only marginally. Climate statistics derived from short 1-yr simulation ensembles are also found not be equivalent to that

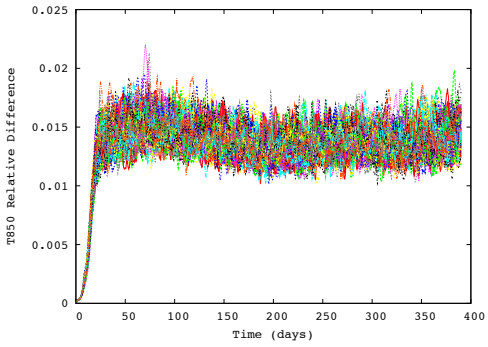

Figure 2: For each ensemble pair of the SISE simulations, the L1-norm of the absolute differences for hourly 850mb temperature in Kelvin is plotted. The initial 10^{-14} K differences grow rapidly over the first 23 days at which point they plateau at roughly 10^{-2}.

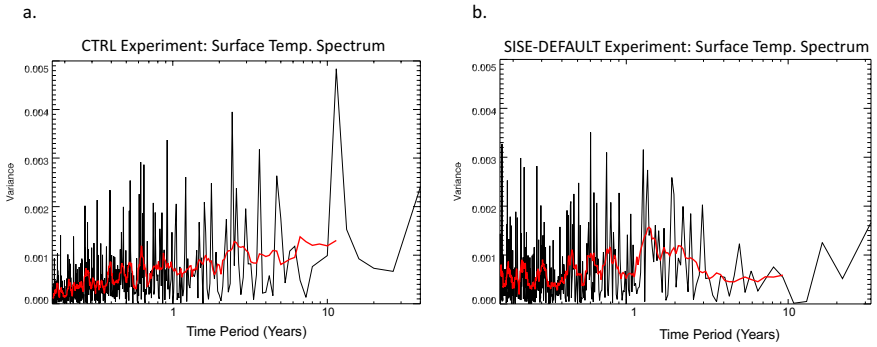

Figure 3: Climate variability. Temporal variance spectrum of monthly global-average surface temperature after the seasonal cycle is removed, for (a) the SLR long simulation of 80 years and (b) SISE-DEFAULT one year simulation ensemble (65 years) with the global averaged surface temperature time series joined together into one long time series. The spectrum is normalized by the total variance and the area under the curve is proportional to the variance. The red lines represent the spectrum smoothed using a moving window of size 11 in the frequency domain, and the black line represents the un-smoothed spectrum.

of a long control simulation for the same number of simulated years, perhaps due to the absence of low frequency atmospheric internal variability in the short simulation ensembles. This is an important limitation when considering SISE for scientific analysis of atmospheric features.

A potential scientific application of SISE is to quantify the impact of various forcings on short time-scales - problems for which traditionally long control and single forcing simulations are the norm. For example, they can be used to study the impact of dust aerosols on Atlantic hurricanes. One could potentially integrate short (few months of the Atlantic Hurricane season) forced runs starting from various points on a control run trajectory following [9]. Initializing from a control trajectory ensures that the climate state would have evolved identically without the forcing, allowing direct attribution. These short simulations also do not allow the low

frequency signals to damp climate response signals and thus enhance the signal-to-noise ratio.

The number of ensemble members available governs the natural variability (stationary noise) of the system that can be captured by the ensemble. Larger ensembles would yield a better quantification of the noise structure and increase the robustness of these tests, and we plan to investigate the role of ensemble sizes on our results in the near future. We also plan to implement several advanced tests of equality of multivariate distributions, developed by the statistics and machine learning community, particularly for problems with high dimensions and small sample sizes as an alternative testing suite to assess climate changing effects.

5.1 Acknowledgments

The authors are grateful for support from the U.S. Department of Energy ACME and CMDV-Software projects. The extremes methodology was initiated through an ORNL Laboratory SEED level Laboratory Research and Development Grant. This research used resources of the OLCF, which is supported by the Office of Science of the DOE under Contract No. DE-AC05-00OR22725.

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. A new ensemble-based consistency test for the community earth system model (pycect v1.0). *Geoscientific Model Development*, 8(9):2829–2840, 2015.

[2] S. G. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

[3] J. E. Kay et al. The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability. *Bull. Amer, Meteor. Soc.*, 96:1333–1349, 2015.

[4] Edwin P. Gerber and Geoffrey K. Vallis. On the zonal structure of the north atlantic oscillation and annular modes. *Journal of the Atmospheric Sciences*, 66(2):332–352, 2016/11/17 2009.

[5] I. N. James and P. M. James. Ultra-low-frequency variability in a simple atmospheric circulation model. *Nature*, 342(6245):53–55, 11 1989.

[6] T. Jiang, K. Evans, M. Branstetter, R. Neale, P. Rasch, Q. Tang, and P. Worley S. Xie. Northern hemisphere blocking in a high-resolution ACME atmosphere prototype. *JGR Atmos.*, Submitted., 2017.

[7] Daniel J. Milroy, Allison H. Baker, Dorit M. Hammerling, John M. Dennis, Sheri A. Mickelson, and Elizabeth R. Jessup. Towards characterizing the variability of statistically consistent community earth system model simulations. *Procedia Computer Science*, 80:1589–1600, 2016.

[8] R. Saravanan. Atmospheric low-frequency variability and its relationship to midlatitude SST variability: Studies using the NCAR Climate System Model. *Journal of Climate*, 11(6):1386–1404, 1998.

[9] Tarun Verma, Salil Mahajan, W. C. Hsieh, Ping Chang, and R. Saravanan. Oceanic feedback in regional climate response to sulfate aerosol forcing. In *AMS Annual Meeting*, page 10A.2, 2016.

[10] H. Wan, P. J. Rasch, K. Zhang, Y. Qian, H. Yan, and C. Zhao. Short ensembles: an efficient method for discerning climate-relevant sensitivities in atmospheric general circulation models. *Geoscientific Model Development*, 7(5):1961–1977, 2014.

[11] H. Wan, K. Zhang, P. J. Rasch, B. Singh, X. Chen, and J. Edwards. A new and inexpensive non-bit-for-bit solution reproducibility test based on time step convergence (tsc1.0). *Geoscientific Model Development*, 10(2):537–552, 2017.

———-