8th International Workshop on Climate Informatics
September 19-21, 2018
CI 2018
Hosted by the National Center for Atmospheric Research in Boulder, CO

ENSEMBLE CONSISTENCY TESTING USING CAUSAL CONNECTIVITY

Dorit Hammerling[1], Imme Ebert-Uphoff[2], Allison H. Baker[1]

***Abstract*—Understanding differences in climate model output can be challenging in light of the chaotic nature of the climate system and its inherent variability. While progress has been made in automatically detecting small changes, making quantitative assessments of when changes truly affect the climate state in the model - and thus indicate a potential problem - has remained elusive. We present a first step in this direction based on evaluating changes in the connectivity structure among key model variables. We use a collection (ensemble) of climate model simulations to obtain a graphical model of the relationships of 15 key climate variables and then build a statistical model to probabilistically describe the occurrence of these relationships. This statistical model forms the basis of a test to evaluate new runs. We illustrate our methodology using data from a large publicly available ensemble of climate model simulations.**

I. MOTIVATION

Climate simulation models are key to furthering our understanding of the complicated interactions between Earth systems (e.g., oceans, atmosphere, ice, land) and to make sense of past and present climate variability, and future climate scenarios. The complexity of the Earth system itself typically results in climate model codes that are similarly complex, very large, and the result of years (or decades) of development from a large community. Our work focuses on the Community Earth System Model (CESM™) [1], which is a popular and fully coupled simulation code. Many climate model codes, such as CESM, are in a continuous state of development in order to incorporate new processes, permit finer resolutions, or optimize high-performance computing (HPC) performance, for example. The continual development of the CESM, combined with the potential importance of societal implications of conclusions drawn from its output, highlight the importance of quality assurance testing during CESM code development. In particular, maintaining user confidence in the CESM is critical.

The CESM Ensemble Consistency Test (CESM-ECT) [2] was recently developed to address the difficulty of verifying the "correctness" of a modification or update to the CESM hardware/software stack when the simulation results after the update are *not* bit-for-bit (BFB) identical with the original result. Non-BFB results are common to the CESM development cycle because the chaotic nature of climate model simulations means that a tiny change (e.g., at machine-rounding level) can propagate rapidly. The original CESM-ECT tool developed in [2] evaluates the statistical indistinguishability (i.e., consistency) of new climate simulations against an accepted ensemble, characterizing the distribution by performing principal component (PC) analysis on the global area-weighted means of one-year annual averages. Notably, a more recent variant "ultra-fast" CAM-ECT) [3] allows evaluation against an ensemble of simulations of only nine time steps in length, greatly reducing computation costs. The suite of tests in CESM-ECT, which also includes an ocean-specific variant [4], has proved useful in practice for detecting errors in the CESM hardware/software stack with minimal cost and expertise required. However, an open question is whether a modification that is flagged as "statistically distinguishable" really affects the longer term state of the climate in a meaningful way, as we have yet to evaluate time scales longer than one year with this methodology. Further, there are applications where changes purposefully go beyond minor modifications (e.g., a new release) and where larger changes are expected. Even in these applications, one would want to be aware of changes in the nature of the connectivity structure between key climate variables, which is the question we address in this work.

II. EXPERIMENTAL DATA

We use publicly available 1-degree CESM simulation data from the CESM Large Ensemble (LENS)

Corresponding author: D Hammerling, dorith@ucar.edu
[1]National Center for Atmospheric Research, Boulder, CO.
[2]Electrical and Computer Engineering, Colorado State University, Fort Collins, CO.

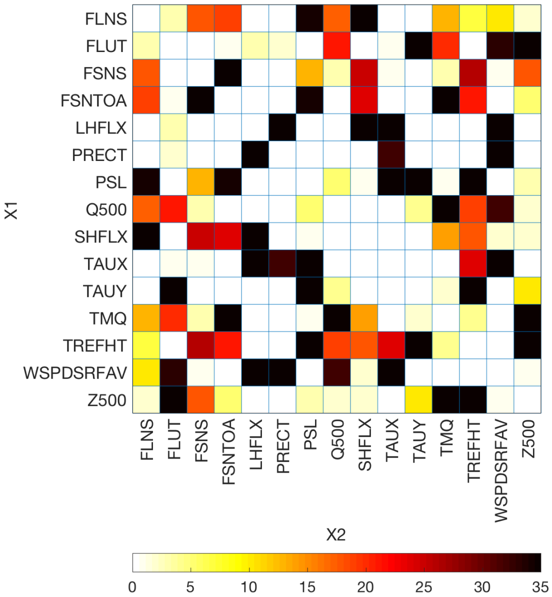

Fig. 1. Number of edge occurrences between 15 key climate variables among the 35 ensemble members evaluated. White indicates none of the ensemble members featured a given connection, while black indicates that every ensemble member had this connection.

Community Project [5], specifically, daily output from CESM-LENS ensemble members 1-35. (Note that we use the original output from simulations 31 and 33, not the compressed variants, e.g. [6].) As with CESM-ECT, our initial focus is on output from the atmospheric component of CESM, where daily time series are available for the global fields of more than 50 atmospheric variables. As a first step, following the approach in [7], we only use global spatial averages of those variables from 1920-2005, and also selected (with the help of domain scientists) a subset of 15 variables that are most representative of the climate state and have little redundancy. Thus, we are investigating the relationship between the time series of 15 scalar variables, which appear in the columns and rows of Fig. 1. Additional information, including variable descriptions, on the CESM-LENS project can be found on the CESM-LENS website (http://www.cesm.ucar.edu/projects/community-projects/LENS/.)

III. CAUSAL CONNECTIONS IN CESM

We use the framework of probabilistic graphical models established by Pearl [8], [9] to identify pairs of variables that are likely to have a direct causal interaction. The key idea of the approach is as follows. It is well known that it is impossible to *prove* the existence of causal connections based on observed data due to the potential existence of hidden common causes. However, it is possible to *disprove* causal connections based on conditional independence tests. This fact is used by algorithms for *constraint-based structure learning*, which start out assuming that each variable is a cause of all other variables, then apply statistical tests to eliminate as many such connections as possible. The result is the set of *potential* cause-effect relationships, often visually represented as a graph, or simply as a list of variable pairs that are connected.

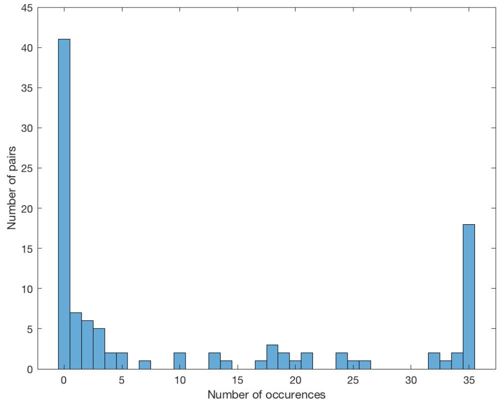

Fig. 2. Histogram of the number of occurrences of the connections among the 35 ensemble members. For the 15 key climate variables considered, there are a total of 105 possible connections for each pair of variables. Of those, 41 connections never occurred within the ensemble, 18 always occurred, and 46 occurred some of the time.

In contrast to our previous study on this data set [7], we develop a static model here, i.e. we only consider instantaneous connections between the 15 variables (no lags). As implementation we use the *PC stable* algorithm [10] with a significance value of $\alpha = 0.05$. Note that the acronym“PC” in PC stable is *not* related to Principal Component analysis, but to the first name initials of its inventors. For this particular study we only calculate which pairs of variables are connected, and do not attempt to assign directions to the edges.

Applying this method to all 35 ensemble members results in 35 different sets of connections. Fig. 1 shows the frequency of occurrence for each edge, i.e. the count of how often that edge appeared across all 35 ensemble members. The count value for each edge can thus range from 0 to 35. We note that the values on the main diagonal, where the “pairing” would be with each variable itself, are by specification set to 0. Fig. 2 shows the distribution of the count values. As can be seen toward the left side, there are many edges that occur in zero (or very few) ensemble members. Likewise there are some edges that occur in all (or almost all) 35 ensemble members.

IV. BERNOULLI MODEL

We model each edge connection as an independent Bernoulli variable with parameters estimated from the ensemble. A Bernoulli distribution is a single parameter discrete distribution to model occurrence of an event. The maximum likelihood estimate for the parameter is simply the number of occurrences divided by the number of trials. We limit our model to the n edges that have some variability, i.e. that are neither always zero nor always exist within the ensemble, which leads to $n = 46$ (out of 105) in our example application (Fig 1). Those edges are then modeled using a joint distribution of independent Bernoulli variables

$$f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} \theta_i^{x_i}(1-\theta_i)^{1-x_i} \quad (1)$$

with $x_i \in \{0, 1\}$, and where θ_i corresponds to the parameter estimated from the ensemble for edge connection x_i.

V. TESTING FRAMEWORK

The first step is to verify that the edges which are always zero in the ensemble (41 in our example) and those that always exist throughout the ensemble (18 in our example) behave the same way in a new run to be tested. If that is not the case, the run is immediately flagged for further investigation. Assuming that condition is satisfied, we then assess how the remaining edge connections of a new run compare to those estimated from the ensemble.

For illustrative purposes, we assess how new runs with all edges existing or no edges (among the 46) fare compared to the distribution of a set of 10,000 random samples drawn from the estimated distribution. Given the high dimension of the potential outcomes of the model (2^{46}), the probability of any specific given outcome is very small. A natural way to assess the probability of a given run is to compare its probability to the distribution of a sample drawn from the estimated distribution. Here, this could potentially be done exhaustively, while in cases with more edges an exhaustive comparison might be truly impossible.

Figure 3 shows the comparison results for the hypothetical runs with all edges and no edges present. The same $10,000$ random draws from the estimated edge model are used in both comparisons. Both of these scenarios are highly unlikely given the edge structure estimated from the ensemble. However, while the scenario with all edges present has a lower probability than any sample from the reference distribution, the scenario with no edges is slightly more likely, with 28 samples from the reference distribution having a lower probability of occurrence. This corresponds to the 0.28 percentile. Depending on the application, one can define a cutoff such as 2.5^{th} percentile, where a new run falling below this value will trigger further investigation.

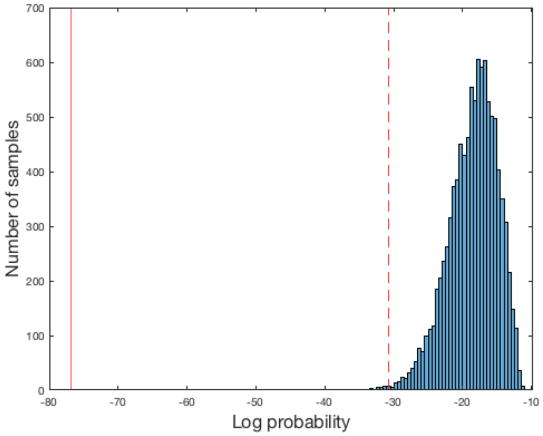

Fig. 3. Comparison of the log probability of hypothetical runs with all edges (solid red line) and no edges (dashed red line) to 10,000 samples (blue histogram) drawn from the estimated joint distribution for the 46 edges which featured some variability within the ensemble.

VI. CONCLUDING REMARKS AND FUTURE WORK

By modeling the causal connectivity of CESM variables, our goal is to provide a complementary tool for evaluating ensemble consistency. An important next step is to generate examples that CESM-ECT has found to be statistically distinct and examine their causal connectivity with the methodology presented in this work. The new method may be particularly well-suited when assessing modifications that purposefully go beyond optimization (e.g. new releases).

We are currently modeling the Bernoulli variables describing the occurrences of connections as independent, but ideally we would also like to incorporate their dependence structure in the model. This would, however, require a much larger ensemble size and is subject of future work.

ACKNOWLEDGMENTS

We thank Haiying Xu for providing subsets of CESM-LENS data for our use. Support for this work was provided by NSF grant AGS-1445978 (Ebert-Uphoff) of the CLD program.

REFERENCES

[1] J. Hurrell, M. Holland, P. Gent, S. Ghan, J. Kay, P. Kushner, J.-F. Lamarque, W. Large, D. Lawrence, K. Lindsay, W. Lipscomb, M. Long, N. Mahowald, D. Marsh, R. Neale, P. Rasch, S. Vavrus, M. Vertenstein, D. Bader, W. Collins, J. Hack, J. Kiehl, and S. Marshall, "The Community Earth System Model: a framework for collaborative research," *Bulletin of the American Meteorological Society*, vol. 94, pp. 1339–1360, 2013.

[2] 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," *Geoscientific Model Development*, vol. 8, pp. 2829–2840, 2015.

[3] D. J. Milroy, A. H. Baker, D. M. Hammerling, and E. R. Jessup, "Nine time steps: ultra-fast statistical consistency testing of the community earth system model (pycect v3.0)," *Geoscientific Model Development*, vol. 11, pp. 697–711, 2018.

[4] A. H. Baker, Y. Hu, D. M. Hammerling, Y.-H. Tseng, H. Xu, X. Huang, F. O. Bryan, and G. Yang, "Evaluating statistical consistency in the ocean model component of the community earth system model (pycect v2.0)," *Geoscientific Model Development*, vol. 9, no. 7, pp. 2391–2406, 2016.

[5] J. Kay, C. Deser, A. Phillips, A. Mai, C. Hannay, G. Strand, J. Arblaster, S. Bates, G. Danabasoglu, J. Edwards, *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," *Bulletin of the American Meteorological Society*, vol. 96, no. 8, pp. 1333–1349, 2015.

[6] A. H. Baker, D. M. Hammerling, S. A. Mickelson, H. Xu, M. B. Stolpe, P. Naveau, B. Sanderson, I. Ebert-Uphoff, S. Samarasinghe, F. De Simone, F. Carbone, C. N. Gencarelli, J. M. Dennis, J. E. Kay, and P. Lindstrom, "Evaluating lossy data compression on climate simulation data within a large ensemble," *Geoscientific Model Development*, vol. 9, no. 12, pp. 4381–4403, 2016.

[7] D. Hammerling, A. Baker, and I. Ebert-Uphoff, "What can we learn about climate model runs from the causal signatures?," in *Proc. of The Fifth International Workshop on Climate Informatics (CI2015)*, 2015.

[8] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman, 2nd ed., 1988.

[9] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.

[10] D. Colombo and M. H. Maathuis, "Order-independent constraint-based causal structure learning," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3741–3782, 2014.