Geosci. Model Dev. Discuss., 6, C458–C461, 2013
www.geosci-model-dev-discuss.net/6/C458/2013/

Geoscientific
Model Development
Discussions

Interactive comment on "Failure analysis of parameter-induced simulation crashes in climate models" *by* D. D. Lucas et al.

D. D. Lucas et al.

ddlucas@alum.mit.edu

Received and published: 1 May 2013

General response to Anonymous Referee #2

Anonymous Referee #2 raises concerns about the applicability of the methodology to other models and the reproducibility of the results. These are important concerns that we take very seriously. To address these concerns, we will revise the manuscript as described in our responses listed below. We believe that these revisions will adequately address the reviewer's concerns, and thus remove all of the potential barriers to publication in GMD cited by this reviewer.

Applicability

Referee: “Its not clear how applicable this technique is for other models and thus if it could be a tool for model development. Most paramater sweep studies don’t contain cases that cause the model to crash and the binary (succeed/fail) nature is crucial to their machine learning technique. If there are such other climate model parameter studies with failures, the authors should include a discussion of them. There is no discussion of any other parameter sweep studies with climate models (such as climateprediction.net)."

Author Response: For one primary and three supporting reasons, we disagree with the referee’s assessment that the method may have potentially limited applicability. First and foremost, the same exact method can be applied to a much broader set of problems than just the “hard" simulation failures described in the manuscript (i.e., binary succeed/fail crashes). The method can be applied to any model output that varies continuously by thresholding or discretizing the output and then classifying the “failures" and “successes" as cases that fall on different sides of the threshold (binary classification) or fall within different bins (multiclass classification). For a single threshold, the method estimates the probability of model output Y exceeding threshold T, denoted by $\mathcal{P}(Y > T)$, and determines the causes for high probabilities of threshold exceedances.

For instance, if a 5 K difference in global average surface temperature between a climate model simulation and a reference case is deemed excessive, then ensemble instances above and below this threshold can be categorized as failures and successes, respectively. The method would provide the probability of a new model case of “failing" by simulating temperatures that exceed the 5 K threshold, information that developers certainly could use to improve their models. We will revise the manuscript to make it more clear that the methodology can be used in this manner.

GMDD
6, C458–C461, 2013

We have three additional supporting arguments related to the utility and applicability of our method.

- Simulation crashes are typically undesirable outcomes that do not directly contribute to the scientific goals and objectives of the simulations. If a failure is encountered, the usual response is to “fix" the model (i.e., prevent the crash) and move on. By virtue of these facts, simulation failures are rarely, if ever, reported. We surmise that a number of failures probably do occur, but go unreported.

- Massive perturbed parameter ensemble studies have been carried out on only a handful of different climate models (e.g., climateprediction.net), which we will cite in our revised manuscript. It is difficult to generalize about the frequency of occurrence of simulation failures in other models from such a limited set of examples. As ensembles of complex geoscientific codes become more commonplace, we fully expect that simulation failures will occur with a greater frequency.

- The few previous examples of parameter sweep studies of climate models that have been conducted were designed for a different purpose. These studies preferentially searched for successful models, where success in this case is defined as a simulation that produces a “good" climate that is reasonably similar to observations or some other desirable target. These studies were not designed to seek out non-physical and undesirable climates. By widening their parameter sampling ranges and including new parameters to sample over, we hypothesize that the likelihood of these models of failing would undoubtedly increase.

Our review of the literature did not turn up other examples of climate model parameter sweep studies that experienced and analyzed systematic simulation failures. The lack of other studies, however, does not mean that “[most] paramater sweep studies don’t contain cases that cause the model to crash." Failures could have been unreported

GMDD

6, C458–C461, 2013

(supporting point 1) or unlikely because the studies were not well designed to sample failures (supporting point 3). We also believe that it would be detrimental to GMD readers to preclude our manuscript from advancing on the basis of the behavior from a few climate model ensemble studies (supporting point 2).

Reproducibility

Referee: "GMD is very concerned with reproducibility. There are no pointers to where others could get either the original data for training and validating the model or the code for calculating the probabilities and other parts of the machine learning model. These must be added for publication to be considered."

Author Response: We strive to provide enough details in our manuscript so that our results can be reproduced by others. To promote reproducibility, our manuscript already lists and cites the SVM software package used to build the probabilistic classifiers (*LIBSVM*, Chang and Lin). The manuscript also provides all of the critical settings and tuning parameters used in *LIBSVM* (i.e., C-support classification, rbf kernels, the kernel width parameter γ, and the cost/penalty parameter C). In order to make our results fully reproducible, we will follow the advice of the referee and make the simulation failure data publicly available. Pending revisions that are accepted for publication in GMD, we will upload our data to a publicly accessible data repository (e.g., the UCI Machine Learning Repository) and provide the details for accessing the data in our final manuscript.

GMDD
6, C458–C461, 2013