Geosci. Model Dev. Discuss., 6, C454–C457, 2013
www.geosci-model-dev-discuss.net/6/C454/2013/

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 #1

We greatly appreciate the highly constructive feedback from Anonymous Referee #1. The most significant issues identified by the referee revolve around 1) training, testing, and validating the classifiers, and 2) comparing to other machine learning methods. The referee also asked for further elaboration on Latin hypercube sampling, which we are happy to provide, and recommended a few minor technical corrections. We address these issues in further detail below and will revise the manuscript for clarity and content as noted below. Again, we thank the referee for the informative comments and for providing interesting leads on alternative sampling methods (i.e., Lovasz and Vempala; Gilad-Bachrach et al).

Training, testing, and validation

There are likely better methods for training, testing, and validating our classification system, though we believe that our approach is satisfactory and follows established practices. We partition the data into a set for supervised learning (studies 1 and 2) and a separate, hold-out set for testing (study 3). The learning set is used to tune the SVM classifiers via bootstrap cross validation, while the hold-out test set is used to quantify the ability to generalize and make predictions from the learned system.

As an interesting anecdote, the primary motivation for publishing our work actually stems from an attempt to predict the outcomes of study 3 before the simulations even began to run! While the 180 simulations were queued up awaiting execution on an LLNL supercomputer, the lead author sent an email to others in the group with predictions of the simulation outcomes. The predictions were based solely on the D_{avg} and D_{sum} criteria from studies 1 and 2 (D_{snr} was added later). The predictions turned out to be ∼97% accurate (as noted in the text), so we knew we had the ingredients for an interesting paper. The results also gave us confidence that the SVM classifiers successfully learned about failures in an 18-dimensional input space from only 32 failure instances.

Apart from the implementation of the machine learning procedures, we recognize that there is room for clarifying and improving the training and testing discussion. As one potential point of confusion, during the learning phase we also hold out 20% of the training data (studies 1 and 2) per pass through the bootstrap loop. The parameters of the SVM classifiers are tuned using only the bootstrap hold-out data (i.e., not the hold-out data from study 3). This process is similar to N-fold cross validation techniques that are commonly used to tune regression and classification models (we will include references). As another potential point of confusion, we use all of the data together (studies 1-3) for the sensitivity analysis, though we feel justified because it is used for diagnostic purposes, not making predictions. We will revise the text and add columns in Table 2 to clarify our use of the data throughout the various stages of analysis. We will also include portions of the anecdote above, to make it clear that learning and prediction occur in two distinct phases.

Comparing to other machine learning methods

The referee noted that it would be informative to compare our approach to other machine learning methods or SVM kernels. Although not described in the manuscript, we tested other SVM kernels during training (i.e., specifically the hyperbolic tangent and polynomial kernels). For this and other climate machine-learning problems we have worked on, we find that SVMs using Gaussian kernels tend to be easier to train and generalize better. Though we would be happy to repeat the full analysis using all of the alternate SVM kernels, limited resources prevent us from pursuing this goal. As a compromise, we will summarize our training results using the hyperbolic tangent and polynomial kernels in the revised manuscript (in Sect. 4.1). In a similar vein, we are interested in cross comparing different classification methods (e.g., logistic regression, random forests, etc), but time restrictions prevent us from performing a rigorous comparison. In our revisions, we will mention and cite other methods that can be instead of SVMs.

Specific comments and other corrections

- Latin hypercube sampling – We will gladly add a brief paragraph with references describing the space-filling LHS method.
- Fig. 8 – This figure is indeed busy. Predictions are based on the magnitude along

the vertical axis, while observations are color coded. We will add a legend and make other alterations to make it easier to interpret the information. Points 2 and 6 (and others) change positions between the plots because there are different ways to combine the predictions from the committee. For each simulation, the committee can be averaged (upper figure) or the average and standard deviation of the committee can be combined (middle and lower figures).
- We thank the reviewer for suggesting the missing and important word “parallel" in front of the description of the SVM hyperplanes. We will add it.
- We thank the reviewer for identifying the misleading description of the locations of the support vectors. We will revise the description to state that the support vectors lie on the margin for cases that are linearly separable and lie on or within the margin for cases that are not.
- We thank the reviewer for finding the mis-spelled word in the title for Table 1.
