Geosci. Model Dev. Discuss., 6, C1559–C1567, 2013
www.geosci-model-dev-discuss.net/6/C1559/2013/

Geoscientific Model Development Discussions

Interactive comment on "PLASIM-ENTSem: a spatio-temporal emulator of future climate change for impacts assessment" *by* P. B. Holden et al.

Anonymous Referee #2

Received and published: 1 October 2013

In this paper the authors present a new climate model and introduce a new emulator on both the parameter and scenario space, which extends previous work and emulates multiple timepoints. Using a principal component decomposition, they develop independent emulators for every PC. The efficiency of the emulator is tested with goodness of fit measures, as well as cross validation, and the climate model is also compared with observations.
This work has elements of interest in developing and extending some statistical methodology to multiple times, but it is not clear how the development of a new climate models has helped in the emulation problem, given that the main topic of the journal is climate models, not statistical methodologies. Also, this technique has some important drawbacks that were not reported and should deserve discussion. To be suitable for publication in this journal, this paper needs major revision.

General comments

- Emulation is a statistical procedure that is independent of the particular climate model that one wants to emulate. Most of this paper is dedicated to the emulator, and in the introduction this is presented as the main contribution. The part that describes the climate model seems in this work only functional to the emulator rather than having an interests on its own. The topic of this paper seems not within the range of this journal, whose focus is in climate model themselves. I would suggest to at least redo the introduction redefining the objectives and trying to stress more the importance of this work in the context of climate model development.

- The description of the statistical model could be improved in terms of mathematical rigor and clarity, there are no clear equations describing what are the assumptions, and a discussion on their validity. For example, at the beginning of Section 4.2, the authors write: "For each PC emulation we build a linear model from all 28 parameters, and then allowed the stepwise addition of ten quadratic and cross terms". This is the main emulator, but is not clear how you added the terms, why you added only ten of them, and why did you pick that particular subset.

- Where was it shown that "parametric error" (which would be perhaps better referred as "parametric uncertainty") is significantly larger than the "code error" (again, "code uncertainty" sounds better) in this study? My guess is that given the ambitious goal of emulating in both parameter and scenario space and given the relatively modest ensemble size, the parameter variability will be very large so this could be justified here. However, this will not be true in a low dimensional calibration problem, where GP emulation is the more suitable approach. So the input dimension of this study (28 in this case), along with a sparse parameter design, justifies a simple statistical model. This should be pointed out as it helps

the reader to understand why the emulator does not rely on a more standard approach.

- The EOF decomposition is based on the assumption that a small number of principal components is able to reproduce the spatial structure of the data, but I don't find this obvious. It is true that the principal components are able to explain the variance of the data, but this does not necessarily imply that that they can also describe the spatial correlation. Closer data in space are expected to be more correlated, and the modeling of their dependence is a wide topic whose review is beyond the scopes of this work. Nevertheless, depending on the nature of the correlation, low rank approximation could not be sufficient to reproduce the statistical features of a spatial field. Some discussion in this regard is needed.

- The use of Chebyshev polynomial (3) and (4) implies some limitations in scenario emulations that were not discussed.

  – This approach allows emulation for only a fixed number of years. More specifically, if one wants to emulate over a specific number of years, he can only use the model output for those years over which the prediction is sought. This was not an issue here, since the ensemble was generated with the same endpoint, but it could severely limit its application to other ensembles where different runs have different endpoints (e.g. some CMIP5 runs end in 2100, some in 2300). The current statistical model will force the user to simply discard all the simulated years beyond the timeframe of interest, possibly losing valuable information. Besides, if one wants to use the emulator to predict at a further time point, he would need to refit the statistical model. This implies SVD of moderately large matrices and to fit a model for every EOF, and that could be computationally demanding. Since the authors have used only decadal averages and a very coarse spatial resolution, I guess they can overcome this limitation with sufficient computational re-

sources, but for finer time scales (years, or months) or with a state-of the-art resolution this will be not trivial. To reduce the size of the spatial field, a different approach could be to consider independent emulators over some predefined regions. This will ignore the dependence between regions but it will be scalable. Also, the fit will be even harder when using statistical models which account for dependence across parameters (e.g. GP emulation). A paragraph discussing the statistical model complexity and its scalability is needed.

  – This method violates causality. The decadal temperature for, say, 2055-2065 is emulated by also using the values of the forcing scenario after that date. This can be fixed by doing independent emulation of every decade, but I don't see any easy modification of this method to generate a time series which avoids this.

  – With this parametrization it is not possible to emulate scenarios with abrupt jumps or drops. This is of course not an issue for impact assessment, but it could be if one were to use emulation with a different goal in mind, such as understanding the physics of the system by increasing the signal to noise ratio.

- I found Section 4.4 quite unclear. My understanding is that the EOF were obtained via SVD of $\mathbf{Y}$, so they represent the decomposition of the whole space/time field for all the runs (or a subset of them for crossvalidation). What does it mean that here the EOFs are averaged over space at each time slice? I would like to have more details in this section before I can further comment.

Specific comments

- p.3350 l.15. The results in this paper are of two types: simulator results and emulator results. The former are compared with empirical data while the latter with other higher-complexity models. It is worth pointing out this difference when mentioning validation.
- p.3351 l.10. What needs to be emulated here is not only a "high dimensional output", but an output whose dimensional indexes are space and time and are therefore physically meaningful. This was not acknowledged throughout this work and the resulting statistical methods violate causality in time and might not be the best approach for spatial dependent output as well (see general comments).
- p.3351 l.25-28. The emulator is a simple statistical model, especially in this work where only linear models were used, so it is possible to have gradients in a simple form. However, it is in principle possible to compute gradients directly from the primitive equations underlying the climate model. Besides, this feature of the emulator was not used here, so there should be some references on works where this was.
- p.3351 l.28. The word "calibrated" is correct here, but should be augmented with references to the MPEF ensemble on p.3357-3358, since the reader has no information on the ensemble at this point.
- p.3352 l.11. What is the meaning of "self-consistent" here?
- p.3353 l.15-17. Section 5 compares the HDDs and the CDDs from the climate model with observational data, it is not using any emulator. Also, Section 6 is validating the emulator with the CMIP3 and CMIP5 ensembles, not the actual climate model.

- p.3356 l.26. It is perhaps more appropriate not to refer to climate runs with different physical parameters as "realizations", but simply as "runs". Realizations imply independence across the different values in the parameter space. Although this might be true for an initial condition ensemble, this is not the case in a perturbed physics experiment.
- p.3357 l.22. Specifying that the key model outputs are five helps the reading in line 27.
- p.3357 l.23-26. This part could benefit from more details, i.e. how many parameters were eventually used, is there any evidence that some cross-terms are more significant than others, etc..
- p.3358 l.4. Why does the word end looks slanted?
- p.3358 l.1. A reasonable model state is meant with respect to preindustrial conditions here, since the plausibility test on line 19 has a higher temperature range, is that correct?
- p.3358 l.23-25. On lines 6-7 you mentioned that only 10 of the 500 parameter set where classified as plausible, yet here there are 188 simulations which pass the modern day plausibility test. I don't understand the transition from the MPEF to the MPSF set.
- p.3358 l.22. Why are the energy balance bounds larger than on line 1?
- p.3359 l.17. Have you tested how much the results of Section 6 would change with more Chebyshev polynomial? Or with another functional basis?
- p.3360 l.6. The Maxmin Latin Hypercube is a particular design for the $(A_{1e}, A_{2e}, A_{3e}, A_1, A_2, A_3)$ which maximizes the distance across the parameters in the design. In this case however, we are interested in the functions resulting

from these parameters, not in the parameter themselves. Does this design result in some particular property for the functions as well? Besides, what would be the meaning of maximizing the distance across parameters that generates two different inputs in the model (CO_2 and CO_{2e})?

- p.3360 l.21-22. A reference on Section 5 might help here.
- p.3361 l.28. What to you mean by "DJF SAT variability"? You mean the variance across the three months for daily temperature? Or the variance for monthly temperature? If this is the case, why not calling it simply "variance"? It would make the reading easier.
- p. 3361 l.23. The citation to the R development team should be for the year 2013.
- p. 3363 l.10-12. PC components are linear combination of the model output, and it is generally hard to give them an interpretation in terms of the original data. What kind of physical processes are reflected by linear combinations of high order PCs?
- p.3363 and Table 2. How was the fitted R^2 computed? Was it computed independently for all the simulations (i.e. across all rows of the $\mathbf{Y}$ matrix in (5)), and then averaged out across the 564 elements of the ensemble? Also, my understanding is that for each PC a different emulator was build and an R^2 was computed, but shouldn't it be more of interest to see the incremental contribution of the different PCs? In other terms, for the kth emulated PC pne could consider the field reconstructed with the $1, \ldots, k$ PCs. This should presumably have a increasing R^2.
- p.3363 l.20. I don't fully understand this discussion about adding orthogonal terms. It is true that every PC adds predictive power, but this is true when you consider the joint contribution of all PCs, which you haven't done here (and even

in this case, the contribution can be marginal anyway). Do you use the term "emulator" to indicate the reproduction of the space/time fields from the single PC or all of them?

- p.3364 l.2-5. Reproducing EOF might be a necessary condition to approximate the space/time output, but it is not obvious why this is also a sufficient condition (see general comments). A more interpretable summary could be the spatial or temporal variance of the emulator vs the simulator, or a comparison of the temporal autocorrelation (assuming maybe a simple AR(1) model after some detrending). It is hard to understand what does of this underestimation of the PC variability implies in terms of the lack of characterization of the space/time properties of the field.
- p.3366 l.9. Remove "as".
- p.3368 l.3 "CO_2 is only an input". Without the comma.
- p.3368 l.8. The future transient period should be 2000 to 2100 AD.
- p.3369 l.5. What is the "CW05" fit?
- p.3369 l.18-19. This paper has shown that a model with very coarse spatial and temporal resolution can be emulated and has shown better results than pattern scaling, but pattern scaling does not need large matrix computations. If we consider for example the RCP experiments for the CMIP5 CCSM4 ensemble, $\mathbf{L}$ and $\mathbf{R}$ will have $192 \times 288 \times 251$ rows and 24 columns, which means that in double precision each matrix will be approximately 2.5 Gb. It will not be trivial to store these matrices in the RAM, and make scalable computations, and it will be even harder in the future with higher resolution models or for statistical models which account for parameter dependence. I am not suggesting that this method is not a good first step, but scalability in the context of climate models should always be taken into account.

- Fig.1. For subplot a, why there are no labels on the x and y axis? For subplots b-f, the resolution of the model is 64×32, as stated in p.3352 l.24 and in p.3360 l.25, so why is there a 36×18 grid here? The data look constant over some grid, but that is the one that is represented. I would suggest plotting the original T21 grid. Also the reported data min and max are too small to be readable, larger fonts are needed. Why they are reported only for subplots e-f? What does the label "data 0" on subplots c-f means? Subplot b has fonts which look smaller than the others.

- Fig.2. A reference on the names on this table could help (I know they come from SIAM-WORLD, but refer to a certain page or table on a literature reference). Also the underscores in the x labels should be deleted, and there must be a labeling on the y axis. The figure is too small compared to the size of the caption, can it be made larger?

- Fig.3. Every row has the same $x - y$ scale, and this might hide some of the polynomial misfit for RCP 4.5, 6.0 and 8.5. I would suggest using the log scale. Since the temporal scale is the same for all columns it is not necessary to repeat the labels for every row, this redundancy looks unnecessary. Larger fonts for the labels and legend might help, too.

- Fig.4. See comments for Fig.1.