Geoscientific
Model Development
Discussions

Open Access

EGU

Interactive comment on "Computationally Efficient Emulators for Earth System Models" *by* Robert Link et al.

Anonymous Referee #1

Received and published: 5 June 2018

General comments I think this idea and motivation of the problem presented in this manuscript is interesting. However, I am very confused about what exactly has been done. The paper is written in a far too technical manner and many parts of the language used leads the reader to be confused. I even found the mathematical and statistical part of the paper confusing. If I as a statistician am finding this confusing there is no way a non-statistician will be able to understand what is going on. At this time, I recommend that the manuscript be rejected in order to give the authors enough time to rewrite it. Specific comments to help with this rewrite are given below.

Major comments

[1] Please don't use 'emulator' at all in your manuscript. What you have implemented is a 'surrogate model' or 'metamodel'. An emulator is a type of surrogate model /

metamodel which is an interpolator and gives a probability distribution for outputs corresponding to inputs it is not trained at. See [1] for details. I understand that others in the earth system modelling community have used 'emulator' in the same way as you, but whoever started using it first and passed down this definition is incorrect for doing so.

[2] Your literature review in the introduction is very limited. This would be fine if the journal was very narrowly focused, but GMD has a very broad appeal. Statisticians from the UQ (Uncertainty Quantification) community have, for example, done a lot research on surrogate modelling methods for very expensive models like an ESM. Names that come to mind are Nathan Urban (Los Alamos), Jonty Rougier (Univ. of Bristol), Michael Goldstein (Durham univ.). A series of workshops were held in Cambridge earlier this year (http://www.newton.ac.uk/event/unq/workshops) which will help you find statisticians working in this field. Non-statisticians like David Sexton at the Met office in the UK are also working on quantifying uncertainty of ESMs. There are probably lots of other research that you can also mention, but above is just a start.

[3] Please use more standard mathematical notation for defining vectors and matrices. In particular, please don't use |x> and <x| for column and row vectors. I have never seen this notation being used before. It is much more standard to use x for a column vector and xT (i.e. the transpose of x) for a row vector.

[4] Please also stick to normal conventions for matrix and vector algebra, e.g. page 2 line 28 you state "Occasionally we will add a matrix and a vector e.g. B = A + |x>. Please do not do this! I know that you explain what you mean when you do this, but mathematically it is not the correct way of doing things because you're effectively defining |x> to be a vector sometimes and a matrix and other times. If you want to add a matrix with a vector in the way you describe then define a new matrix which has as its columns or rows the vector you want. This is a much clearer way of defining things.

[5] Page 3, line 8. You state "We refer to the ESM data as "observations" . . . ". NO, NO,

NO! Please do not do this. I spent about an hour reading your paper thinking you were using real observations and then I realized I had not properly read this very important line in your methods. If you want to use your ESM output as measurements then call them "synthetic measurements". Everyone knows what this is. But if you just state measurements or observations, we think you're using the real thing.

[6] There are lots of subject specific words / jargon used here which you assume that the reader knows the meaning of because you don't define them. Examples include: 'linear pattern scaling', 'discrete Fourier Transform', 'randomizing the phases of G', 'spatial coherence', etc. . .

[7] For things that you define you don't give enough detail. For example, with EOFs (statisticians use principal components but it means the same thing) you need to say that we normally only choose the first n EOFs where n is determined such that most of the variance (or power as you mentioned) is explained. Only when we get to the results section do you talk about the number of EOFs that you're using.

[8] I was also completely lost with section 2.4. You seem This seems to be key part of the methods, so really needs to be explained better. You make statements that make no sense to the non-specialist – e.g. in the lines prior to equation 7 you state that you're making some minor modifications to the standard procedure, but why? And what is a zeroth basis vector and why are you defining it in this way. These things I read and go "okay", but I have no idea why you're doing this. This is just one example of many scattered throughout the manuscript.

[9] There's no motivation or justification for doing what you're doing. At the end of the introduction, you state that there's a need to quantify the uncertainty in ESM output. Fine. But then you jump straight into your approach of quantifying this uncertainty by generating it based on an error covariance matrix that is derived from defining the temporal and spatial correlation of the ESM output. Why is this a good way of defining the ESM uncertainty?

[10] Following on from the previous point, this seems to be a major flaw in this paper. Normally when we do uncertainty analysis, we propagate the uncertainty from the inputs (e.g. uncertainty in model parameters) through to the uncertainty in the ESM outputs. Instead, you seem to be using the spatial and temporal correlations of the ESM output as a means to generate the uncertainty in the ESM output with the metamodel (or "emulator" as you call it) as the vehicle for carrying out the extra ESM runs. Perhaps I'm mistaken and this isn't what you're doing. If I am mistaken then the fact that I have misunderstood this is a major problem. If you want people to be interested in your research, you first need to communicate it clearly and (sometimes) simply to them.

[11] I am confused with how you train your metamodel (or "emulator" as you call it). When I train metamodel, this often involves multiple runs of the computationally expensive model (i.e. the ESM in your case). In your paper, it seems that you just need one run of the ESM to train the metamodel. Is this correct? Again, this isn't how we normally train metamodels so you need to be really clear about this. In fact there should really be a whole section in the methods explaining everything about how you constructed the metamodel. Maybe you feel what you've written is enough, but I'm just generally confused so you need to lay things out much more logically and clearly at the very least.

[12] Table 1 makes little sense to me. If I really concentrate I can probably understand what's going on, but you could help the reader by using less complicated words or phrasing it in a simpler way.

[13] The results section seems to be too short (less than one page). Most papers I read have at least 3 or 4 subsections within the results section. These subsections have their own titles and help navigate the reader through the different aspects of the results. At the moment, the results just seems like a list of things. Your results show flow more like a story. You also don't really give a lot of detail, e.g. just one sentence for figure 5? What's the point of having it in there?

[14] I didn't really read through the discussion in depth, but in section 4.1.1 (and may other subsections) you describe further results that were carried out. Thse should be in the results section. The purpose of the discussion section is: (a) to give an explanation for why your results look the way they do; (b) to put your results in context of other comparable studies. I see a bit of (a) in the discussion but no evidence of (b).

Minor comments

- Panels of figures: label them with letters. E.g. Figure 1a would refer to panel at the top left of figure 1.

- Figures: captions lack enough details.

- When submitting for review, it's more helpful to put all the figures and tables at the end of the manuscript. This makes it easier for the reader to refer to a particular figure when reading a particular part of the results.

- There are probably other minor comments but I think the major changes should be made first.

Reference [1] O'Hagan, A., 2006. Bayesian analysis of computer code outputs: A tutorial. Reliability Engineering & System Safety, 91(10-11), pp.1290-1300.