Paper: Correcting a bias in a climate model with an augmented emulator

Authors: McNeall *et al*.

Review:

In this paper the authors present a technique to determine the set of plausible parameter combinations that lead to a credible version of the climate model FAMOUS, in the presence of a significant model bias.

This study follows on from the work described in McNeall *et al*. (2016), using the same perturbed parameter ensemble of model runs from the FAMOUS climate model. McNeall *et al*. (2016) identified the land surface input space of FAMOUS that was consistent with observations of the output 'forest fraction' for different forest regions, and found that there was very little overlap between parameter space that was consistent with observations of the Amazon region and parameter space that was consistent with observations of the other forest regions. This led them to conclude that there is either a local climate bias and/or some missing/incorrect process(es) in the land surface model of FAMOUS.

Here, the authors extend the approach of McNeall *et al*. (2016) to develop a method that accounts for a climate bias in the model-observation comparison, by bias-correcting emulator-predicted model output before it is used in the statistical 'history matching' procedure that determines the plausibility of each tested realisation (combination of land surface model parameters) of FAMOUS . The authors bias correct the climate of the Amazon using an 'augmented' Gaussian process emulator, where temperature and precipitation outputs are treated as model inputs alongside the uncertain land surface input parameters. They find that the forest fraction in a region is sensitive to these climate variables, and by bias correcting the climate in the Amazon region the authors are able to correct the forest fraction in the Amazon to tolerable levels for many of the tested realisations of FAMOUS, including the default parameter set, thus increasing the amount of valid input space shared with the other forest regions (from 1.9% of the parameter space in McNeall *et al*. (2016) to 28.3% here).

This approach is a novel adaptation to current methods for climate model-observation comparison, which shows potential for improving and simplifying the model tuning process for coupled climate models, as well as aiding in the identification of model errors. The manuscript is well-written and definitely falls within the scope of GMD and EGU. I like the concept of the 'augmented emulator' that includes the localised climate outputs as inputs, however, I do have some concerns about the true validity of the augmented emulator, in particular for the bias corrected climates of central Africa and to an extent the Amazon in the application (see Specific comments below). If these concerns, along with the other comments listed, can be addressed then I would recommend the publication of the manuscript in GMD.

Specific Comments:

- **Page 7 Line 11 – Page 8 Line 2:** I'm confused by the role of the 'beta' parameter in the ensemble set-up. I don't think I understand this. The simulations in the ensemble have each been run with one of 10 different configurations of the atmosphere? Was it randomly assigned as to which 'atmosphere configuration' each simulation had? What are the implications of this? Doesn't this introduce biases into the ensemble (into the ensemble outputs) for identifying the parameter combinations that are plausible, if the ensemble members do not have the same starting point in the atmospheric set-up? Please clarify this in the text.
- **Page 8 Line 7-8; Section 3.1:** 'The study only considers regions dominated by tropical broadleaf forest, so as not to confound analysis by including other forests which may have a different set of responses to

perturbations in parameters, rainfall and temperature'. Even though the forest regions are of the same 'type' (tropical broadleaf forest), are there other factors in the model that might affect the forest response between regions that are not accounted for? Such as topography that might affect how the forests respond to the parameter perturbations?

The analysis is based on an assumption that the forests in the different regions will have the same responses to changes in the land surface parameters, rainfall and temperature. How realistic is this assumption? I'm not saying this assumption should not be made (we have to make assumptions for modelling and statistical analysis!), but I think the authors should state more clearly (in section 2 or 3) that they make this assumption (it is implied in the set-up of the augmented emulator, but not openly said), and discuss any possible implications of it on the results. This could be a further reason why the other forests 'do slightly less well' (e.g. Page 16, Line 11; Page 13 Line 12; Table 2).

- **Section 3.2 – validation of the augmented emulator:** The augmented emulator is validated using a 'leave-one-out' approach. However, I am not convinced that this approach fully validates the emulator for its use in the following analysis in Section 4. The emulator looks to be sampled from beyond the range of its training data where it is not validated, which could be affecting the results obtained.

  The augmented emulator is only trained to predict the forest fraction for climates (P and T variable combinations) that occur in the original simulations of the ensemble for the 3 regions. Figure 3 shows that coverage of the climate variables [P,T] 2-dimensional state-space is not uniform, and has sparse (if any) coverage in many areas of that 2-d space. In particular, there is rather limited coverage around the 'observed' climate (P,T combination) for the Amazon, and for Central Africa there looks to have no training points particularly close by. How the information for P and T augment on to the 7-dimensional space-filling design of the land-surface parameters to produce the final 9-dimensional input design with which the emulator is constructed is not shown (I imagine the actual coverage is some weird and complicated shape with some potentially large gaps) and so I wonder what the training data coverage, and hence emulator skill, is like for the areas of that 9-dimensional parameter space that are used in the bias correction analysis for these observed climates? The leave-one-out validation approach is only testing the emulator in the areas of space that have training runs, and so although the validation plots in Section 3.2 seem reasonable, it looks to me that the emulator is not tested (and so not validated) for the 'bias- corrected' climate of central Africa, and the Amazon, where the emulator is densely sampled for the analysis in Section 4. Outside the trained area of parameter space (e.g. where the observed climate for Central Africa is) prediction from the emulator becomes extrapolation from the emulator, the emulator prediction uncertainty can quickly increase and prediction values from the emulator will return back towards the form of the prior specification of the mean functional form from the GP emulator construction (the emulator mean response surface will bend/shape back towards that form), here a linear function of the inputs [stated in the supplementary information]. Hence, how do we know that the bias corrected predictions used in the analysis for these regions are sensible?

  Can the authors provide some further validation for the emulator predictions for the observed climates of central Africa and the Amazon? If not, then the authors should explicitly state this limitation of the emulator in the paper and discuss the possible consequences of this on the figures and results presented in Section 4, and in the discussion and conclusion Sections 5 and 6.

  Also, could the emulator predicted responses to climates not covered by the training data (including the observed climate for Central Africa and the top left area in Fig 3) be dependent on the emulator's prior (linear) form, and change if this specification was changed? If yes, how confident are the authors in the prior emulator form (linear) being representative of the climate model's actual behaviour in parameter space beyond the training data? If they are not confident in it, then it either shouldn't be used or it needs to be more carefully specified so that it can be used.

In particular:

- The results shown in Fig 8 are obtained by sampling with the climate variables set at the observed values (shown in Fig 3). Hence, this means that for central Africa (and the Amazon to some extent), the sampled predictions come from extrapolating from the emulator beyond the extent of the training data. The responses to the climate variables are reasonably linear and I wonder if this response is at least partially driven by the form of the prior specification of the GP emulator mean function?
- Are the results in Fig 9 from the FAST99 algorithm generated by sampling the across the full 2-dimensional climate [P,T] space? How are these results affected by sampling from the emulator where there is no training data, particularly for a cool/wet climate at the top left of Fig 3? Could the sensitivity to the climate variables be over-estimated here? (I cannot easily tell from the plot, but do the individual main effect sensitivities sum to <1? (Main effect + interaction should sum to 1.) I've seen instances where the algorithm produces main effect values that sum to >1 in the presence of noise in the emulator fit.)
- In Figure 10 the emulator is used to simulate across the entire range of simulated temperature and precipitation with all other inputs fixed at the default setting. How might this result be affected?
- How might this issue affect the results of the retained parameter space from the history matching (Figures 13-16) for each forest region? In Figure 16, are the regions not covered by the training data more likely to be retained as emulator error is larger, reducing the value of the implausibility metric so that it cannot be ruled out?

- **Page 15 Line 3-4, Figure 15 (and discussion):** guiding further runs, choosing high density regions to run new ensemble members. This is a useful outcome. My question here really relates to how the results trace back to improve model performance... How does the bias correction information feed back for the modeller to know what 'atmospheric configuration' should be used (the 10 atmospheric parameters, or beta?) with the inputs selected as good for any new runs? Can a good representation of forest fraction for all forests simultaneously be obtained in new runs at these parameter combinations without bias correction? Or, would any new runs always need to be bias corrected too, until further work to understand the true cause of the climate bias is completed and the climate model updated? As obviously, just running the model at more combinations in this identified joint space will induce climates as shown in Figure 3, away from the observed climates for each forest. Could the authors comment on this in the discussion?

Further Comments:

- **Page 3 Line 18:** 'Without strong prior information...' What is meant by 'prior information' here? What kind of information? On observations? On model skill? On both? This is a bit vague and needs more clarity.
- **Page 5 Line 9:** What reasons? Please give more details here: '...whereas there were a number of reasons one might reject the proposed parameter space, **including...**'
- **Section 1.3 (Page 5 Line 11):** It might give more context to the first listed aim on line 11, and for the detail coming in the second paragraph of the section (discussing the results of McNeall et al (2016), which are not 'aims of this paper') to connect that this study is extending the analysis of McNeall et al. (2016) at the start of the section in the first line (first aim?)?

- **Page 6 Line 9-11:** Sentence starting 'Parameter perturbations...'. Are there any references for examples of such findings?
- **Page 6 Line 19:** '...was sensitive to perturbations in parameters,'. This is vague... What kind of parameters? Edit to say '...was sensitive to perturbations in parameters **such as**...'
- **Page 6 Line 31 – Page 7 Line 1:** I realise that the details of the ensemble are in McNeall et al (2016), but I think it would be useful to give minimal details of the parameters perturbed in this paper also. Their effects are being compared to the climate variables in Section 4.1, with parameter names (acronyms) given in the text, and yet I have to go to a completely different paper to find a description of them /what they correspond to in the model. Please add a small summary table (to the supplementary file, if not to the main paper) that lists the parameters with short descriptions.
- **Page 7 Line 4:** What is meant by 'global values'? Global values of what? Please clarify.
- **Page 8 Line 30:** Please provide a reference for GP emulation.
- **Page 9 Line 13:** '...the 10 atmospheric parameters perturbed in a previous ensemble, summarised by the β parameter'. I'm struggling to picture how the effects of 10 parameters can be summarised by 1 parameter. A lot of information is being condensed here? This needs more explanation. (Also see first comment above under 'Specific comments'.)
- **Page 9 Line 14:** This sentence: 'We cannot control them directly and thus ensure that they lie in a latin hypercube configuration' is confusing and could be interpreted in different ways. I first read 'and thus ensure' as that you **do** ensure that they lie in a LH configuration. But on second read I see the meaning you want is that you can't ensure this. Please re-phrase.

  Also, the 'latin' in 'Latin hypercube' should have a capitol L. Please update here and elsewhere.
- **Page 10 Line 12:** '3% of the maximum possible value of the ensemble.' What does this really statement tell us? The maximum forest fraction is 1 so the error of 0.03 is 3% of this forest fraction value, but this is the minimum percentage of an output value that it could be. The majority of predictions will be less than 1, and so the error of an 'average prediction' is in general a larger percentage than this. It seems a bit of a misleading statement, and I suggest removing it here and on line 14.
- **Page 10 Line 25:** I don't understand how the 'rank histogram' indicates that we have 'reliable' uncertainty estimates? It shows the predictions are a mixture of over and under-estimations of the actual model, but it gives no indication of the size of errors, which could be large and therefore not reliable? Maybe I misunderstand this.
- **Section 4.2 and Figure 10:** The interpretation of this plot needs clarification. Would the contours be similar at different parts of the land-surface parameter space? (Fixing at a different simulation to the default?) On Page 11, Line 26, it suggests that for central Africa, moving any ensemble member (small triangle point) to the observed (big triangle) would not cross many contours, but the central Africa points with wetter climates (normalised T at approx. 0.4, normalised P at approx. 0.5 to 0.6) would cross between 3 and 4 contours to be in the same one as the observed (big triangle), so I'm not sure this is true? Please clarify this.
- **Page 12 Line 8-11:** Are the values given in this paragraph mean absolute error between model and observations? The first line says 'difference', but this must be an average? Please clarify. Also, could the lack of training data near the observed climate for central Africa be a contributing factor to why central Africa is worse (Line 9) in this metric?
- **Section 4.4:** It might be better for the flow of the results section if the first part of Section 4.4 (up to Page 13 Line 2) describing the history matching methodology was moved into Section 3 on methods?

- **Page 13 Line 12:** Could more detail be given as to why the implausibility value at the default settings rises for central Africa and SE Asia on bias correction?
- **Page 17 Line 14-16:** Summarising the 10 parameters using 2 outputs is useful, but this is likely to have little traceability back to the original 10 input values? Many different combinations of the 10 inputs could lead to a similar combination in the 2 outputs, so how would one know what combination of the 10 inputs is best when setting up any further runs of the model? Also, the '*O(10xp)*' rule is when training points are space-filling across the parameter space being emulated. There is no guarantee (as seen in this example) that this property will hold when outputs are used as dimension reduction, so more points, or even less points, could easily be required. This should be acknowledged.

Technical corrections:

- **Page 1 Line 13-14:** 'This might be due to...'. I think this sentence might be easier to read if the number/list format is replaced with 'This might be due to either ..., or...., or a combination of both.'
- **Page 1 Line 16-17:** '...alongside **regular** land surface input parameters.'. Is the term 'regular' needed here? Maybe remove this word. [The 'regular' suggests to me that there may be other types of land surface input parameters that are 'not regular' which are not included, which I don't think is the case.]
- **Page 1 line 15:** Should '...**a** climate model...' be '...**the** climate model...'? This is now the specific climate model used by McNeall et al (2016).
- **Page 1 Line 17-18:** For readability, please change 'is nearly as sensitive to climate variables as changes in **its** land surface parameter values.' to 'is nearly as sensitive to climate variables as **it is to** changes in land surface parameter values.'
- **Page 2 Line 9:** Should '...processes sufficiently to trust...' be '...processes sufficiently **and** to trust...'?
- **Page 2 Line 28:** The sentence here is hard to read. Change: '...practices, there appear no standard procedures for climate model tuning however - as the authors...' to '...practices, there appear **to be** no standard procedures for climate model tuning**. H**owever**,** as the authors...'.
- **Page 2 Line 32:** Missing word? Edit: 'It might start with single column version...' to be 'It might start with **a** single column version...'
- **Page 3 Line 2:** Change word order? Edit: '...might be then tuned...' to '...might **then be** tuned...'
- **Page 3 Line 8:** Change 'Golaz et al. (2013) Show...' to 'Golaz et al. (2013) **s**how...'.
- **Page 3 Line 20-21:** This sentence needs plural 'candidates' at the start, and the second part should be given as more of a negative to make the point? Revise to: 'This means that **good candidates** for input parameters might be found in a large volume of input space, **but** projections of the model made with candidates from across that space might **diverge to** display a very wide range of outcomes.'
- **Page 3 Line 23:** 'individual parts' Of the tuning process? Or the model? Please clarify.
- **Page 4 line 14:** Missing full stop after Vernon et al. (2010).
- **Page 4 Line 20, Line 23:** References in bracketed format when should be in in-line format.
- **Page 4 line 33:** Remove the second 'used'.
- **Page 5 line 1:** Missing words. Change to: '...structural bias in **the** ocean component of **the** climate model HadCM3 could'
- **Page 5 Line 33:** Should this be '...use the **augmented** emulator to estimate the sensitivity...'?

- **Page 7 Line 2:** Move the reference to Fig 1 to the next sentence, which is the sentence that is describing what is shown in Fig 1.
- **Page 7 Line 7-8:** '...parameter settings which the **emulator** suggested should lead to **an** adequate simulations of...'. The emulator provides the predictions of model output but does not indicate adequacy – this comes from the history matching process (as the authors have described). Change to: '...parameter settings which the **history matching process** suggested should lead to adequate simulations of...'.
- **Page 8 Line 9, 12:** References to Jones et al, and Adler et al are in in-line format when should be in bracketed format?
- **Page 9 Line 10:** Change to: '...each of the forests**:** the Amazon, central Africa and Southeast Asia...' or put the forest region names in brackets?
- **Page 9 Line 11-12:** For readability, move the sentence 'Regional extent of ... supplementary material.' so that it is the second sentence in this paragraph. (The next sentence follows better from the one before it!)
- **Page 9 Line 24:** Here the work by M16 is referred to as being by the authors of this work ('we built'), but in all previous references to this point it has been referred to as a separate study (e.g. M16 argue..., or M16 speculated...). Update as needed to be consistent.
- **Page 11 Line 3:** Should this paragraph start with: 'The **augmented** emulator...'
- **Page 11 Line 5:** '...predict changes in forest fraction as each variable is changed from the lowest to highest setting in turn...'. Should 'variable' in this sentence be replaced with 'input'? – as this is done for the land surface input parameters as well as the climate variable inputs?
- **Page 12 Line 22:** Remove the second 'the'.
- **Page 13 Line 21:** The term 'the climate-bias forest' sounds weird? Should this be 'the climate-bias-corrected forest'?
- **Page 16 Line 34:** Remove the second 'could'.
- **Page 17 Line 14:** 'O(170)' should be in italics?
- **Page 18 Line 6:** Change: 'If trained an ensemble...' to 'If trained **on** an ensemble...'. Also, remove the second 'which'.
- **Page 18 Line 10:** Remove; '(e.g.)'.
- **Page 18 Line 19:** Should 'learned' be 'learn'?
- **Page 18 Line 32:** Missing word. Change: '...in leave-one-out...' to '...in **a** leave-one-out...'.
- **Page 19 Line 13:** Change 'finding' to 'findings**'**.
- **Page 27, Fig 5 caption:** The brackets for g_1 are in the wrong place? For g_n and y_n, should 'n' be replaced with '3', as is shown in the diagram, and written for 'C' in the next sentence?
- **Supplementary information Line 17:** Reference in bracketed format when should be in in-line format.
- **Supplementary information Line 18:** Missing word. Change: '...in section 3 the...' to '...in section 3 **of** the...'