Earth System Dynamics Discussions

Earth Syst. Dynam. Discuss., author comment AC3
https://doi.org/10.5194/esd-2021-59-AC3, 2021

Reply on RC3

Shruti Nath et al.

Author comment on "MESMER-M: an Earth system model emulator for spatially resolved monthly temperature" by Shruti Nath et al., Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2021-59-AC3, 2021

General summary: This paper extends the existing MESMER approach to include a monthly downscaling module, to enable the generation of large ensembles of spatially explicit monthly temperatures that are representative of ESM behaviors, which could be an useful tool for regional impact assessments. The paper is clearly written for the most part and contributes to the existing literature on ESM emulators. However, I have some concerns that I would like the authors to fully address.

We thank the Referee for their thorough review of the manuscript and the overall positive feedback. The general comments provided a useful compass on what parts of the paper were unclear and we hope that our proposed changes better clarify these parts as well as emphasise aspects of the text that are vital towards the overall analysis. The specific comments were also useful in strengthening the text for the broader audience. We hope that we addressed the comments sufficiently below.

General comments

(1) The training and verification results (such as those described in Lines 218-219, Lines 238-240) show non-trivial dependency on the number of ensemble members available for training, which raises concern not just for the robustness of this method, but also the usefulness of the method presented. If MESMER-M relies on a large number of ensemble members to get robust results, then it defeats its own purpose. A useful component to add into the paper is a sensitivity test to show what is the size of training runs needed to get a robust training result.

General MESMER fitting recommends the training of the emulator on all available ensemble members (see Beusch et al. (2020)). While that means that for some models the emulator will have more training material, it also follows the philosophy of feeding as much information as possible into each model-specific emulator so as to generate the best possible "super-ensemble", without judging the realism of the training sample. There are other approaches in getting the best training set size, such as that employed by Castruccio et al. (2019) so as to balance the stability in the inference (represented for example by variability) of the emulator, and benefits for reduction in computational costs. Such approaches however require the presence of a large ensemble and would mean that

we would be constrained to demonstrating MESMER-M's performance on a smaller subset of the available CMIP6 models.

In this paper we wanted to show the ability of MESMER-M in representing monthly temperature distributions across all CMIP6 models without penalising models with smaller ensemble sizes. While the calibration results show some dependence on ensemble size, we should stress that this is simply a bottleneck and the model itself is the main driver of the calibration results (e.g. even with only one ensemble member MCM-UA-1-0 has a high localisation radius). It should also be stressed that the performance of MESMER-M is not dependent on ensemble size as for instance the pearson correlations within the mean response module as well as the quantile deviation magnitudes within the regional-scale verification are quite similar across all models. Furthermore, even though we attribute the performances of MIROC6 and MPI-ESM1-2-LR within the model benchmarking exercise to the ensemble size, this is not a universal feature as it only appears in January. We will try make clearer that the ensemble size is not the sole determining factor within the text, as well as emphasise that the primary purpose of MESMER-M is to provide the best possible emulations based on training material available. It could be worth doing such sensitivity tests however it goes beyond the scope of this paper and we propose mentioning it in the Conclusion and Outlook section instead.

(2) The technical details need to be better described/clarified, for the potential users of MESMER-M to fully understand the approach taken, the assumptions made, and the procedure to carry out the training, calibration, validation, and generation of 'super-ensemble' using this method. Please see some of my specific comments below.

We see how the technical details may get a bit complicated and hope that the improvements made from the specific comments below as well as those from the other 2 reviewers improves this. We will furthermore experiment with visualisation of the methods by providing X-Y snapshots in Figure 1 as well as a schematic of the power transformer process in Section 3.2.2.

(3) My understanding is that the particular MESMER-M presented here can generate a 'super-ensemble' under the SSP5-8.5 scenario. Although the authors stated that 'MESMER offers the perspective to improve our understanding of the likelihood of future impacts under multiple scenarios', a different version of MESMER and MESMER-M has to be developed per scenario. If so, this needs to be clearly stated, and could the authors comment on how straightforward this process would be to expand this work under multiple scenarios.

It has already been demonstrated that one MESMER can be used to represent different climate scenarios, if it is trained across a representative sample of scenarios (See Beusch et al. (2021): https://gmd.copernicus.org/preprints/gmd-2021-252/). From preliminary testing, we expect the monthly to yearly temperature relationship to be fairly scenario independent as well. To clarify the process needed to expand MESMER-M to other scenarios we will elaborate on the existing expansion of MESMER as well as the need of a training set representative of all scenarios within the Conclusion and Outlook section.

(4) The fact that the emulator solely relies on annual temperature as input, and the assumption that other forcings have very little impact on the local monthly temperature response, makes the applicability of the monthly temperature probability distributions derived from the emulator limited. There should be more discussions around in what applications would the emulator results be particularly useful in Section 6.

We took this emulation exercise as a challenge in seeing how much of the monthly temperature response could be reproduced using yearly temperatures as the sole input and in the simplest manner possible. We do acknowledge that the assumption of little

impact from other forcers limits the applicability of MESMER-M, particularly when other forcings have a strong impact on the monthly cycle (e.g. changes in tree cover due to deforestation in WAF for MPI-ESM1-2-LR see Figure 4 and refer to comments 14 and 15 for referee 1) that goes beyond the harmonic form represented from yearly temperatures. Within the Conclusion and Outlook section we propose future MESMER-M developments (e.g. of a module to represent land-cover effects) and emphasise the need that representation of other forcings should be sufficiently decoupled from the GHG induced temperature response. We propose prefacing this part with the limitations of MESMER-M's applicability i.e. it will fall short when the overall mean response in monthly temperatures is dominated by other forcings that are decoupled from the GHG induced temperature response.

Specific comments

(1)In the Abstract, it should be 'model projection uncertainty' instead of 'model uncertainty. Model uncertainty comes from incomplete representation of physical processes, uncertain/unknown physical parameterisations, structural uncertainty.

We will change "model uncertainty" to "model projection uncertainty"

(2) Line 5, it should be 'selected climate variables' instead of 'select climate variables'.

This change will be implemented accordingly

(3) Line 10, what does 'mean response' refer to here? Also this is an odd sentence structure, consider reframe to 'represent the monthly temperature cycle in response to the yearly temperatures'.

Mean response refers to the direct response to evolving yearly temperatures, we will rephrase to:

'represent the mean monthly temperature response to yearly temperatures'

(4) Line 50 & in the Abstract, it's important to be clear which one the authors are trying to refer to, internal variability or natural variability, to me these are different things.

We will edit such that we consistently use natural variability

(5) Line 64, is there any particular reason why a spatial resolution of 2.5 by 2.5 is used? Could the authors comment on how they expect the results to change if the analysis is done at a lower or higher resolution, and does the training of the emulator have any requirement or restriction when it comes to spatial resolution?

A 2.5° by 2.5° grid is used simply as it is the best compromise to getting all climate models to the same resolution, thus allowing comparison of emulator applicability between models. Unfortunately fully investigating the effect of spatial resolution is beyond the scope of this study, we wouldn't expect it to have much of an effect however apart from slowing down or speeding up the emulator training time.

(6) Line 66, please specify why this reference period is chosen.

This was for consistency with the paper describing the MESMER emulator (Beusch et al. 2020). We will specify this within L66.

(7) Line 101, the authors need to explain Bayesian Information Criterion to the readers (the significance of BIC), and why 8? Figure 2 suggests you are using 6 instead of 8.

We will modify L101 to read as:

"... we quantify the balance between the model complexity and accuracy using the Bayesian Information Criterion…"

We performed the BIC till n=8, however n greater than 6 was never chosen. We thus show an upper limit of 6 in the top panel's colorbar for Figure 2. For sake of simplicity we can however change n = 6 in L101 also.

(8) Line 121, please explain the Yeo-Johnson transformation, for the benefit of the readers who are not familiar with this, which I expect would be the case for many readers.

We accept that further explanation is desirable and will add the equation for the Yeo-Johnson transformation before equation (5).

(9) Line 127, I don't understand how and what fitting is being done using maximum likelihood here?

The epsilon coefficients which the Yeo-Johnson's lambda parameter is defined from are being fitted for. Equation 5 defines lambda and we state that epsilon_{0,m,s} and epsilon_{0,m,s} are the coefficients fitted for using maximum likelihood.

(10) Line 134, again, for the benefit of the readers, please specify what a Gaspari-Cohn function is here and why you choose this function to apply here.

The Gaspari-Cohn function was used in previous MESMER developments (see Beusch et al. 2020, equation 8) and allows for exponentially vanishing correlations with distance such that anisotropy of spatial cross-correlations on regional scales is still retained. We chose it for consistency within MESMER as well as the aforementioned property. Since it was already elaborated on in the previous MESMER paper we do not go into its specifics. To make this clearer however we can add to L134

"... localized by point-wise multiplication with the smooth Gaspari Cohn correlation function (Gaspari and Cohn, 1999) which has exponentially vanishing correlations with distance r_m."

(11) Line 156, I don't understand the authors' decision to only look at the top 50 highest power spectra. Please elaborate your thinking and reasoning behind this.

We initially wanted to check how well the emulated power spectra corresponded to that of the ESM runs by looking at how well we represent the frequencies where power within the signal is most concentrated (thus focussing on the 50 highest power spectra). From insight of reviewer 1 however, we will change this to look at how well we represent the spectra of the highest frequencies that occur within the ESM runs (see comment 10 of reviewer 1 for more details).

(12) Line 166-168, please consider rewriting this bit to clarify what's exactly being done to create these emulated quantiles. This part reads very confusing as it is now. In the following sentences, the quantile comparison description also lacks clarity.

This was also identified by referee 2 and we propose providing a numbered step-by-step procedure of this within the section.

(13) Section 3.4. Please explain why these particular biophysical variables (as listed in Table 1) are considered (chosen over other variables) and used in this study.

We mainly wanted a representation on changes in radiative and thermal fluxes for which these variables de

(14) Line 195, I don't quite understand how this procedure is done, how did the authors use the physical model to augment the harmonic model results. Please elaborate.

By "augment" we simply mean that we add the phyiscal model predicted variability to the harmonic model results.

(15) Line 224-225, the authors should consider adding some discussions here on why these two models show such outlier behaviour.

Unfortunately we could not think of any obvious reason for such model behaviour. We additionally refrained from going into too detailed analysis of each model as that would require a more in-depth analysis of the model formulation itself, which is beyond the scope of this study.

(16) Section 4.2, please explain why these 4 ESMs are presented here, how they are representative (e.g. span across some projection range), and why WNA and WAF regions are chosen here.

We will add that these models represent divers genealogies according to \cite{Knutti2013} and \cite{Brunner2020}.

(17) Figure 3 & Figure 4. The labelling on the upper left corner should be 4 ESMs.

In this case we mean the number of ESM runs we are showing and not ESMs. We indeed show 4 different ESMs but for each ESM 3 runs are plotted.

(18) I would suggest changing the whisker colors in Figure 6 & 7 so that it's easier to see them.

This is a good suggestion and we will try darken the whiskers

(19) Table A1 clearly shows the split between training/test runs is not always 70/30, as opposed to what's stated in the text. Please check and confirm what's being done

We roughly followed a 70-30 split, however for MIROC6 and CanESM5 a 50-50 split was done, as training on more than 10 ensemble members led to significant training time with no real gain in model performance. We can modify L61-62 to read as:

"...is done in a roughly 70-30 manner, and for models with more than 20 ensemble members a 50-50 manner so as to maintain a good balance between training time and model performance."