Earth System Dynamics Discussions

Earth Syst. Dynam. Discuss., author comment AC2
https://doi.org/10.5194/esd-2021-59-AC2, 2021

Reply on RC2

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-AC2, 2021

General comments

Climate model emulators are becoming more and more useful in assessing climate change through representing complex Earth system model (ESM) behavior and combining many different ESM simulations with other multiple evidence. The MESMER approach, being developed by the authors, is unique as an emulator for generating spatially-resolved forced and unforced climate realizations based on multi-ESM, initial-condition ensemble simulations. This paper describes the newly developed MESMER-M module for monthly downscaling, which is expected to further expand the emulator's applications.

Although the current manuscript adequately describes the structure and performance of MESMER-M, there is room for improvement as follows, which should be appropriately revised for publication.

We thank the Referee for their thorough review of the manuscript and the overall positive feedback. Several good points were made about the need to acknowledge major variability modes such as ENSO and changes in atmospheric circulation patterns. We also found some suggestions in tackling these quite helpful and integral in strengthening our discussion as well as the climatological perspective of the paper. We hope our response addresses these aspects sufficiently. We furthermore value the comments made for better clarification of the methods section and hope that changes made enable better understanding of the methods for the broader non-technical audience.

(1) The calibration and verification results indicate marked dependency on the number of ensemble members, which may raise concerns about the robustness of the methods.

Such dependency appears in most key parameters and performance aspects: the order of the autoregressive process in the temporal variability module (L218 and L239), the stationarity of the shape parameter λ in the spatial variability module (Appendix D), the scale parameter rm for the localization in the spatial variability module (L237, L283, Figure 7), and the bench mark test with GBR (L338). Although each of these results is explained in terms of the amount of training runs as input information, it is not much

convincing. Implications for the robustness of the methods and a possible guideline of an appropriate size of the training runs should be discussed thoroughly.

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". 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. We thus settle for a train-test split of approximately 70-30 as this provides some stability in inference whilst maintaining low training time and leaving aside samples for validation. It should be noted that, even though we explain model calibration results as dependent on training runs, we do not wish to discount such calibration parameters being dependent on the model itself (as referred to in previous MESMER fittings and mentioned in our Conclusion and Outlook, the calibration parameters represent unique model IDs). We will try edit the text so as to make this clearer as well as add a discussion point on the need to choose training set sizes such that stability in inference is stable whilst computational costs are kept low (as according to Castruccio et al. (2019)).

(2) The modules and the calibration and verification results lack interpretation from a climatological point of view. Although the seasonality associated with snow cover is frequently mentioned, this is just one aspect. It is necessary to describe and discuss the validity of the modules from the aspect of major variability modes, such as monsoon, ENSO, and AO.

Deviations from the mean seasonal cycle may not necessarily be biophysical feedbacks, as assumed in L184-185. Internal climate variability leading to some deviations, such as jet meandering and blocking associated with the strength of the polar vortex, is hardly regarded as a biophysical feedback. It is not much convincing that the difference between summer and winter in the mean response verification (5.1) is explained by the snow-albedo feedback only. In the regional-scale verification (5.3), although the increasing deviations in July between the ESMs and the emulator (L294-) is worth being noted, it may also need to be described based on specific natural variability, rather than regarding the tendency as abstract secondary, non-linear responses.

See also the specific comment on the conclusion and outlook below.

Major variability modes such as monsoon, ENSO and AO can indeed drive changes in interannual variability and we acknowledge their relevance in considering monthly temperature responses to GHG induced warming. We propose adding some discussion points within the Conclusion and Outlook on how to tackle their representation as further elaborated in response to the specific comments. We furthermore can modify L184-185 to clarify that we do not expect biophysical variables as the sole explanatory variables and that other atmospheric circulation processes may also play a role. We had only focussed on biophysical variables in investigating residual variability, having considered their advantages and relevance towards the primary purpose for which we employed them (i.e. model benchmarking) as follows:

- These biophysical variables are readily available output within CMIP6 models, providing easier access as explanatory variables within the Gradient Boosting Regressor (GBR) model, with less degrees of uncertainty vs having to deduce jet meanderings, atmospheric blockings, ENSO, AO from for instance SSTs or soil moisture.

- Given that within our monthly emulations we are also interested in intra-annual variability, we expect biophysical variables to most effectively "kill two birds with one stone" providing information on the drivers behind both changes in intra and inter annual variabilities that the GBR can effectively sift through.
- Atmospheric circulation patterns can be expected to leave their fingerprints on for instance latent & sensible heat fluxes, such that even though we cannot directly attribute temperature variability to atmospheric circulation patterns we do still have a good representation of temperature variability against which to benchmark MESMER-M.

Unfortunately, given that we mainly focus on the technical aspects of MESMER-M we intentionally chose not to delve into a full climatological investigation as this would add an extra layer of complexity that is out of scope of our aim: providing simple statistical emulations of monthly temperatures from yearly temperatures. Furthermore, following studies such as Schwingschackl et al. (2018) we saw the added benefit of pursuing a biophysical based GBR representation of temperature variability as this allows extraction of biophysical IDs, which are of relevance to the Earth System Modelling community (e.g. for model benchmarking as proposed within the Conclusion and Outlook). Our analysis within L294- mainly focuses on pinpointing shortcomings within the emulator performance so as to highlight caveats of its usage, and in such we did not fully elaborate on the underlying processes but refer to papers where they have been thoroughly investigated. We do agree that it would be a better idea however, to add to the aforementioned lines that it is not just biophysical feedbacks but also atmospheric drivers at play.

(3) From the standpoint of potential users of the series of MESMER modules, who are not necessarily familiar with technical details, it is recommended to devise some descriptions for better understanding.

For example, in Figure 1, adding X-Y plots illustrating a typical seasonal cycle and its variability and skewness would help understand the local variability module. Visual materials would be useful for making sense of technical concepts like the multivariate trans-Gaussian process and the Gaspari-Cohn function.

In terms of consistency between MESMER and MESMER-M, it may also be useful to verify whether the annual average of each element of MESMER-M is consistent with corresponding elements of MESMER.

MESMER-M is planned to be made open-source with its own Github page, we agree that there it will be of relevance to provide visual representations when describing each module. We will furthermore experiment with adding some in text grid point X-Y snapshots to Figure 1, for ease of readers understanding as well as visualisation of the power transformation process in section 3.2.2.

Specific comments

L59. A brief explanation about limiting the scenarios to high emission SSP5-8.5 and applicability to low emission scenarios would be helpful.

Our rationale between training on SSP5-8.5 was to get the extreme end of monthly temperature response to yearly temperatures. In general, we would expect the yearly to monthly temperature downscaling relationship to be relatively scenario independent such that training on the extreme SSP5-8.5 scenario allows a rough capturing and validation over the whole spectrum of monthly temperature response types. To explore inter-scenario applicability however, the emulator should of course be trained across all scenarios. To elaborate on this we could add:

"... so as to first explore the emulator's applicability to the extreme end of GHG induced

warming”

L61-62. The 70-30 train-test split is not consistent with actual split shown in Table A1. It appears that the 70-30 ratio is rather exceptional, and that some models with a large number of members have 50-50.

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.”

L65-66. It should be clarified how the anomalies are calculated, i.e., whether they are deviations from the annual climatological mean or from the monthly climatological mean.

To make this clearer we can modify L65-66 to read as

“... annual climatological mean over the reference period of 1870-1899.”

L82, 85. The use of the term "forcing" can be a bit confusing. As "other external forcings" imply an underlying primary forcing, "a certain forcing" may be better rephrased in a specific way. Changes in land cover can be anthropogenically forced or induced by climate change and variability. A more specific wording may be necessary to avoid misunderstanding.

We accept this and will change the use of "other external forcings" accordingly.

L90. The term "monthly cycle's mean response" is a bit, confusing considering the subsequent "seasonal cycle". "Monthly mean response" may communicate its intention without ambiguity.

We accept this and will replace "monthly cycle's mean response" accordingly.

L96-97. It appears that the need for high-order harmonic terms is not convincing. My understanding is such that up to the second order term representing a bi-modal cycle is enough for the mean monthly response. Are the month-to-month correlations, which is the case for some natural variability modes, out of scope for the mean response?

More flexibility in the order of harmonics chosen was allowed, as even though only 2 orders would be needed for bimodal representation, this could be too simple in terms of representing complex changes in the amplitude of the seasonal cycle with evolving yearly temperatures. In such, we deliberately allowed more degrees of freedom for order of harmonics, whilst using the BIC to ensure model complexity still made sense with respect to accuracy in mean response representation.

L111. Check "time-dependent and space-dependent components" is correct wording. They are functions in terms of month, space, and year. Maybe, tempral-variability and spatial-variability components.

This is a good suggestion, and we will implement the changes accordingly.

L112-114. Is this an appropriate explanation for adopting a autoregressive order-one

process model? AR(1) may be suitable when the autocorrelation function of the stochastic process has significant components up to lag three or so.

In this case, as the variability module is month specific, the autocorrelation function for a month is built independent of that of other months. Hence, we would expect that accounting for lag-1 autocorrelations would only represent the month-to-month correlations for subsequent months and not more i.e. up to lag one only. It should be noted that here we are also mainly referring to the deterministic part of the AR(1) process.

L130-136. The purpose of localization should be explicitly stated, which would be helpful for the relevant issue described in the paragraph starting L302.

The need for localisation if elaborated in the Beusch et al. 2020 MESMER paper and hence we chose not to repeat the reasoning. To clarify this, we can modify L133 to read as:

"...and covariance matrix, Sigma_{nu_m}. Similar to previous MESMER fittings Sigma_{nu_m} is rank deficient (Beusch et al. 2020), and is thus 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."

L140. In equation (7), is there a case where the magnitude of γ1 is greater than 1? Figure 2 shows that some models have means close to ±1.

The absolute value of γ1 is constrained to less than or equal to 1 during fitting of equation (4). We can specify this in L115 as:

"... where gamma_{1,m,s} is between -1 and 1."

L145-146. Specify whether area weighting is processed or not.

To clarify the above we can add:

"... across the whole globe for each month with each grid point weighted equally."

L168-170. This quantile comparison procedure is unclear.

We propose providing a numbered step-by-step procedure of this within this section.

L194-197. This sentence is complicated and should be clarified more.

We propose dividing the sentence as follows:

"Pearson Correlations over all months, between ESM test runs and harmonic model test results augmented by biophysical variable, T_{yr} and month based physical model predictions are calculated. As a measure of performance the aforementioned correlation values are given relative to those obtained when augmenting using only T_{yr} and month based physical model predictions."

L207. The symbols in Equation (8) are not fully described. Instead of this equation, the integral of difference between two CDFs would be more understandable as a definition of the energy distance.

The energy distance is not exactly the same as the difference between 2 CDFs but was

proven to be equivalent to the square root of twice the distance between 2 CDFs (see Székely: The energy of statistical samples (2002)). This relationship however is dimension dependent and to be consistent with the actual method used to calculate the energy distance as well as emphasise the non-parametricity of it, we chose to use the formula as given in equation (8).

L224-225. Is there anything to be added? Readers would be curious about what kind of characteristics of the two models result in such outlier results.

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 documentation itself, which is beyond the scope of this study.

L233. The description about the equatorial region appears to be limited to January, if so, it should be stated as such.

We can rephrase L233 to read as:

"Around the equator (-5° to 5°) we see lambda^{tilde}_{m,s} values consistently higher than 1 especially in the month of July, with INM-CM5-8 and INM-CM5-0 displaying significantly high values.

L248-254. Readers would be curious about the MIROC6 results, in which relatively many ESM points appear outside of the emulator range in WAF. Is there anything to be mentioned in this regard?

MIROC6 in general shows higher interannual variability in WAF as compared to other ESMs for which the addition of variability terms tends to be under dispersive. We are not sure about the exact causes of this but suspect an influence of cloud feedbacks that do not directly relate to changes in yearly temperatures. Since we focus on more technical aspects of the emulator in this paper however, we refrain from going into climatological conjectures of individual regions and ESMs.

L258-260. The authors mention different timing of changes in snow cover among ensemble members for lower correlations in April and October, but this is not the case for several ESMs that have only one training run and zero test run.

We agree with this and realise that the low correlations also arise from inter-annual differences within each run in the timing of snow cover increase and decrease, hence we will modify these lines to read as:

"Such low correlations could result from the inter-annual spread in the timing of snow cover decrease and increase, such that the mean response extracted does not always match individual years."

L275. Meaning of "such higher power spectra" appears unclear.

For more clarity we will replace "such higher power spectra" with "higher order temporal patterns".

L291. It looks mixed positive and negative rather than "generally low values."

To clarify we will replace "generally low values" with "generally low magnitudes."

L288-289. Is it correct for "deviations of the ESM training (and where available test) runs from the full emulations"? Figure 8 caption writes quantile deviations of the monthly emulated quantile from that of its ESM training and test runs although the embedded figure title is different. Which is the base for the deviation?

The base of the deviation is the emulated quantile. We will correct Figure 8's caption to read as:

"... quantile deviations (colour) of the ESM training run values from their monthly emulated quantiles…"

L311. Around here, it is hard to trace relations between descriptions in the text and corresponding parts in the figures. Even if the complexity of the figures is unavoidable to some extent, main points and their implications should be clearly indicated.

We see how this can be confusing and will replace L309-314- such that the main points are emphasised -as follows.

"For January over the time period of 1870-2000, lowest magnitudes in 5th and 95th quantile deviations is observed for Southern Hemispheric regions (e.g. AMZ, NEB, WSA, SSA) , along with slight overdispersivity (see the blue, respectively red values in the left, respectively right panels of Figure 8). Over the period of 2000-2100, this behaviour for January switches to Northern Hemispheric regions (e.g. CEU, ALA, ENA, WNA, TIB) and is mostly apparent for the 95th quantile, possibly due to a decrease (increase) in January variability in the Northern (Southern) Hemisphere with increasing yearly temperatures \cite{Holmes2016}. In contrast, over both the periods of 1870-2000 and 2000-2100, July consistently displays lowest magnitudes of 5th and 95th quantile deviations (with even slight overdispersivity) in Northern Hemispheric regions (e.g. WNA, ENA, NAS, WAS). Regional dependent overdispersivity indicates that the vanishing of spatial cross correlations with distance is contingent on the region itself, such that some regions have less spatial cross-correlations with distance than others. Since the spatial covariance matrix is localised with a single global number, the variability terms for such regions therefore account for more spatial cross-correlations than necessary leading to the observed overdispersivity."

We will furthermore add some discussion points on tackling improving the localisation of the spatial covariance matrix, for example by expressing r_m as a function of latitude.

L341-344. Figure E2 shows an exceptionally bad performance of NorESM2-MM. Is there anything to be mentioned?

We are not sure why NorESM2-MM has such bad performance, it is possible that there is some key biophysical information lacking and the biophysical variables investigated within this study don't play much of a role in explaining the variability for this ESM. We do not go into detail about this however as it would be beyond the scope of this study.

L335-338. Regarding greater CDF distance in January, Figure E2 may not clearly indicate such distinctive difference between January and July. Grid lines in the plot space would be helpful for identifying the difference.

For bringing out the difference we will add grid lines to Figure E2

Figure E2 also shows that CanESM5 has greater distance in January as well as MIROC6 and MPI-ESM1-2-LR.

We are aware of this but do not mention it as we focus primarily on MIROC6 and MPI-ESM1-2-LR within section 5.4.

Conclusion and outlook. While the emphasis is on future developments that take into account biophysical variables, there remains the question of how similar the variability components by each ESM are to observations, although the latter is out of scope in the current manuscript. What needs to be focused on in this regard would include properly emulating how major variability modes, such as ENSO, modulate with warming, considering any dependence on ESMs. If a certain variability mode includes some memory effects associated with , for example, land soil moisture they may need to be modeled by higher-order autoregressive processes, and if the variability affects remote climate on a global scale through teleconnections, the localization of the spatial covariance structure may need to be improved. In any case, future developments will need to be described in a broader scope.

We acknowledge that the representation of higher order AR processes possibly arising from major variability modes are not modelled for. We had looked at AR processes up to order 3 but the coefficients' magnitudes decreased significantly after order 1 hence we opted for an AR(1) process (we could include a demonstration of this within the supplementary materials). We did also include soil moisture within our biophyiscal variable exploration, however it proved unimportant when latent heat flux was included and hence we chose not to go further with it. It could be of interest to investigate the lag correlation between soil moisture and the temperature variability, however this seemed out of scope for us given that we wanted to emulate monthly temperatures from yearly temperatures only. We will however add a discussion about this, as well as link it as a possible method of pursuing representation of variability modes (e.g. ENSO) as possible future avenues in MESMER development.

We are aware of a monthly emulator developed by Mckinnon and Deser 2018 which deals with modes of ENSO, PDO and AMO, requiring SSTs as input. Methods of including bits of their approach within MESMER-M were also considered, however we were limited by yearly temperature being the sole input. We can however add a point on this within the Conclusion and Outlook, highlighting the shortcoming of the representation of global-scale circulation patterns within MESMER-M, but also emphasising the need for combined emulator approaches as each emulator has its own strengths in representation.

In case of investigating global scale teleconnections, we had attempted a Principal Component Analysis based approach to including teleconnections. However, the added degree of freedom did not add much information as the main local monthly temperature response followed directly from local yearly temperatures and otherwise corresponded to local changes (for instance in land cover). Hence, disentangling the variability modes arising from teleconnections and their manifestations on a local level proved difficult and we decided to follow a simpler approach of generating spatio-temporally correlated residuals instead. While exhibiting the added value of being consistent with previous MESMER developments (Beusch et al., 2020), we hope to demonstrate in this manuscript that it achieves a good performance.

Acknowledgement. Refer to
https://pcmdi.llnl.gov/CMIP6/TermsOfUse/TermsOfUse6-1.html to confirm whether acknowledging CMIP6 is appropriate, and requirement for citing CMIP6 model output.

The link indeed suggests an acknowledgement following language such as:

"We acknowledge the World Climate Research Programme, which, through its Working

Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF."

We will also double check the citation requirements.

Figure 3 and 4. Units are missing. The figure legend implies that plotted data are three members for the ESMs, the three patterns for the sum of the yearly and seasonal cycle, and 50 realizations for each of the three patterns for the emulation, but should be fully described in the caption text. The second sentence of the caption text should be clarified about whether the reference temperature is monthly or yearly.

This is a good point and we will add this to the caption test. We will clarify that a yearly reference temperature is used as well.

Figure 6. Although the text in 5.2 (L271) implies that the box plots in Figure 6 show distributions across different realizations by the emulator, the figure caption should be fully described about how different bands (50 elements), different emulator realizations (50 members), and different training and test runs (ESM-dependent ensemble number) are processed to represent the data distribution. Also, if the whiskers in this figure (Figure 7 as well) are drawn as in Figure 2, indicating a min-max range, explicitly state so, otherwise describe it accordingly.

This is a good point and we will include the different bands and emulator realisations used as well as the range represented by the whiskers within the caption text.

Figure 8. It would be good to have a guide so that readers easily identify representative geographical zones corresponding to the individual regions alighted on the horizontal axis.

We were afraid that providing a visual guide would take up more space and add an extra layer of complexity to the figure. However, we will explicitly refer to to Figure B1 to provide visual guidance.

Editorial comments

L70. Inline Mathematical symbols should be italic.

We will change this accordingly

L73, L95. Indent is unnecessary.

Indent will be removed

L105. Section 3.2.1, not section 4.1.1, but this indication within the parenthesis is redundant.

This reference will be removed

L110, L115, etc. Add comma to the end of the preceding expression and lowercase "Where".

This change will be implemented accordingly

L136-137. "based on", not "based off."

This change will be implemented accordingly

L169, L175, L202, etc. Long dashes (em dashes) are not typesetted correctly.

We will correct this

L183. "properties of the monthly temperature response" (maybe "of" is missing)

We will add ‘’of”

L203, 204, 220. "added ontop" and "ontop of" may not be common wording.

We will change “added ontop” to “added” and “ontop of” to “additional to”

L224. HadGEM3-GC31-LL, not HadGeM3-GC31-LL. Put it after ACCESS-CM2 if alphabetically ordered.

This change will be made

L234. "this" in "The source of this" is unclear.

We will change “this” to “the aforementioned”

L237-238. In this context, "boreal winter", not just "winter". Pay attention to whether inappropriately referring to specific seasons in terms of the Northern Hemisphere throughout the manuscript.

We will add “boreal” to clarify this

L241. "four selected ESMs", instead of "a select 4 ESMs." See the journal's English guidelines for numbers, and, if needed, spell out numerals less than 10 throughout the manuscript.

Change will be implemented accordingly

L244. It is unclear "all ESMs" are the four selected ESMs or all the ESMs used in this study.

We will change “all ESMs” to “the four selected ESMs”

L328-329. "latent and sensible heat fluxes", not "latent heat fluxes."

Change will be implemented accordingly

L163. Spell out SREX here. Also in Figure B1 caption.

SREX will be spelt out in the 2 places

Figure C1 and C2 captions (also, Figure D1 and D2 captions). For Figure C2, "Same as Figure C1, except for January" would be fine. Referring to the Benjamini/Hochberg correction may be required in the Figure C1 legend.

We will implement this suggestion

Figure E1. It might be better to replace X and Y axes for comparison with Figure 12. A more concise caption would be "Same as Figure 12, except that all CMIP6 models are

shown for the global land."

We chose the X and Y axes for aesthetic purpose as otherwise the score numbers would have to be either squashed or rotated. We will implement the caption suggestion.