Response to the Anonymous Referee 2 for the manuscript

Extending MESMER-X: A spatially resolved Earth system model emulator for fire weather and soil moisture

Yann Quilcaille[1], Lukas Gudmundsson[1], Sonia I. Seneviratne[1]
[1]Institute for Atmospheric and Climate Science, Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland.
Correspondence to: Yann Quilcaille (yann.quilcaille@env.ethz.ch)

We would like to sincerely thank the Anonymous Referee 2 for the positive and constructive review. Its and Claudia Tebaldi's comments were completely integrated to the manuscript, which we believe has improve its quality.
We detail in this document the modifications brought to the manuscript. Referees' comments are shown in black. The authors' response is shown in green text. The text quoted from the manuscript is shown between quotation marks in italics. Numbers of lines correspond to the version including tracked changes.
Summary of modifications:

- Corrections in the main text following Referees' recommendations, as detailed in the responses to the Referees.
- Updated figures 1, 4, 8 & 11 (selection of configuration): reversed colormap on CRPS, labels on color bars and bigger font sizes
- New appendices:
  - 6.1: Application of the Probability Integral Transform to discrete distributions
  - 6.2: Representation of the interannual variability
  - 6.3: Interpretability of the CRPS
  - 6.4: Maps of negative log likelihood for the retained configurations
  - 6.5 to 6.8: Equivalent of figures 2, 5, 9, 12 but for SSP1-2.6 and SSP2-4.5

Summary This work introduces an extension of MESMER-X to produce regional emulation of fire weather and soil moisture indices. This is an important direction for emulation because these variables play a key role in the assessment of nature-based solutions to mitigate climate change. The proposed emulators follow a shared construction. First, a local observation process $D(\cdot|\alpha_{s,t,1}, \ldots, \alpha_{s,t,p})$ is prescribed to define the distribution of a variable of interest at time t and spatial location s. Second, the parameters $\alpha_{s,t,1}, \ldots, \alpha_{s,t,p}$ of this observation process are estimated using (possibly non-linear) functions $f_{s,1}, \ldots, f_{s,p}$, which take as input global climate variables at time t. Internal variability is introduced by gaussianising the distribution $D(\cdot|\alpha_{s,t,1}, \ldots, \alpha_{s,t,p})$, and introducing spatially-correlated innovations in this gaussianised space. The model is evaluated for multiple indices of fire weather and soil moisture — which induce different choices of conditional distribution D and parametric functions $f_{s,p}$ — and demonstrates good performance, with limited quantile deviation.
Thank you very much for this review. We cannot help but notice the term that you introduce, "gaussianising". We appreciate it so much that we may borrow it from you, if you may! The first referee asked for details about this transformation, we added an appendix about it where we use this term, we hope that you don't mind.

Strengths and Weaknesses Well written and presented paper, easy to follow. The model is formulated with great generality, which is a strength. Then, every single emulator is a particular case of the general formulation proposed. Extensive experiments are conducted on diverse variables which display different statistical properties and allow to fully appreciate the capacity of the model. Great effort is put into visualising the model performance.
In general, the emulators outperform the baseline and are capable to faithfully reproduce spatial patterns, even though the models local parametrisation are fairly simple, which is in my opinion a strength. The main weakness of the paper lies in the difficulty to assess the quality of models calibration on different future forcing scenarios. I would have appreciated to visualise emulation outputs on low and medium forcing scenarios.
I think the paper is very interesting and will be publishable after minor revision.
We are grateful for your very positive comment. Regarding the outputs on other scenarios, we agree with you and Referee 1 that the scenario dimension may not have been clear enough in this manuscript. We used high warming scenarios to show the evolution over larger domains of warming, but showing their performances on low to mid warming scenarios do matter. We are adding the equivalent of the figures 2, 5, 9 and 12 for SSP1-2.6 and SSP2-4.5 to the Appendix.

- L41 : ”For their part” → I’m assuming this is a literal translation from french ”Pour leur part”, might sound better to reformulate with for example ”On the other hand”
Thank you for noting that, this is likely what happened… We corrected into:
“Besides, changes in the water cycle are more challenging to represent than changes in temperature (Allan et al., 2020).”

- L115 : The functions fs,1, . . . , fs,p are time-independent, which introduces a time stationarity assumption for these mappings. From the results, this seems to be a robust assumption, but could you briefly comment on the grounds for this assumption?
Thanks for this insightful comment. The initial reason why we use this assumption is because of pattern scaling. At a regional scale, we observe this linear scaling at a regional level, for instance between global mean temperature and regional temperatures. It was extended to the grid cell level for different applications, e.g. MESMER. What we did with MESMER-X is to extend this principle to conditional distributions. Though, even if this assumption is based on what we observe with the scaling in CMIP outputs, it does not explain the actual ground.
One way to look at the question would be to reason on its contraposition. If we assume that the mappings were not stationary in time, then it means that we could not isolate all dependencies of the parameters of the distribution being driven, here, by global mean temperature.
A potential reason may be that global mean temperature isn’t enough. For instance, there may be some inertias in the water cycle, as seen with the soil moisture. Adding a lagged temperature helped there, but one may think of other forms of lagged variables or other variables like radiative forcing or ocean heat contents or fluxes. If we manage to rewrite the $f_{s,p}$ with these new global drivers, then we verify the initial assumption.
If this is still not enough, then it is not because of additional global drivers or dependencies on the global pathway. It means that at a prescribed global climate, the parameters of the distributions are not fixed. Because the global climate pathway would also be prescribed, it is not due to changes in regimes / bifurcation / hysteresis effects. As far as we can think of it, the only remaining effect would be local drivers, such as changes in land use or different “mix in

radiative forcings", for instance more greenhouse gases but also more from cooling aerosols. Such effects would lead to similar global climate pathways, but without the same local effects. As a summary, this assumption allows to account for a large range of global effects, but its major issue would be changes in local drivers that would compensate at the global scale. Though, fixing that isn't straight-forward, as the modified model may need local inputs as well, thus hindering its use for coupling with simple climate models. A step in this direction has been made in (Nath et al., 2022).
We think that this discussion may be of interest to some readers, so we added the following paragraph lines 121-128. We are afraid that it may not have been as "briefly" as you asked for.
"Equation (1) offers a large flexibility in terms of modeling. Using variables such as global mean surface temperature, radiative forcing or ocean heat content facilitates the modeling of interplays in the Earth system. Using lagged variables such as the global mean temperature at ΔT_{t-n} *or accumulated warming over the past* n *years would also help in representing more advanced dynamics such as inertias in the water cycle. Such a capacity is of particular interest for overshoot scenarios. Yet, equation (1) has also its limits: any changes in local climate drivers (e.g. land-use, combination of individual radiative forcings) that would compensate at a global scale would not be accounted for. Such effects may still be modeled (Nath et al., 2022), but are not integrated in this framework."*

- L122 : Could you comment on why would it be an appropriate model of internal variability to introduce innovations over the gaussianised variable? In particular, would be interested to hear about potential issues that could arise from the fact that the Normal distribution is not heavy tailed while the GEV distribution is (my intuition is that some tail events in a GEV density might not be properly represented by a Gaussian density, but that might be wrong).
Thanks for this excellent comment. The first reason is that representing internal variability on the non-gaussianised variable would be much harder. Detrending is feasible, but then we would also need to deal with changes in spread, eventually in the shape of the distribution, that is not trivial. That is why we made this modeling choice.
Then, regarding the choice of this specific model of internal variability over the gaussianised variable, we think that it would work because this probability integral transform is a projection of the cumulative distributive function. It implies that the tails would be correctly reproduced. By construction, the distribution of gaussianised events follow a Normal distribution, which ensures that generating events using this model of internal variability, based on Normal distributions, would normally give a proper Normal distribution, then a proper reconstruction of the tails. More precisely, the tail events at 99% or higher in the GEV would occur in the 99% or higher region of the Normal distribution as well. When generating realizations, about 1% would be in this region of the Normal, thus only 1% in the GEV.
The Referee 1 had also a question regarding this transformation, though on discrete distributions. This is why I added an appendix, with a figure for some visual material.

- L180 : Probably the case already, but are score spatially averaged by accounting for decreasing cell size toward poles?
Thanks again for this comment. Yes, we are accounting for the cell size. The CRPS is calculated in each grid point and at each time step of each scenario, for each its member. Then

it is averaged over ensemble members, so that scenarios run more than others would not have a higher weight in the final performances. Then, we average over time, accounting for the length of the historical or scenario. Finally, we average over space, accounting for the size of the grid cells, as you said, so that the cells at the poles don't artificially bias the performances. The CRPSS is calculated in each grid point, time step, each member, then averaged the same way.

- Figure 1 (and other configuration selection plots): Could you add a label on the colorbars and make the fontsize larger for labels (the model description labels in particular are a bit difficult to read). It seems at first glance that CRPS values are quite high for E0 (even lowest is around 2.2). That means that even for best CRPSS values, we still get a relatively high CRPS score for the emulator (this is also true for the other experiments). Could you comment on this?
We acknowledge that these labels were too small, and that titles for the color bars would help. We have edited the figures 1, 4, 8 and 11 to integrate these requests. For information, the Referee 1 also asked to revert the colors on the CRPS, for a better interpretability.
Regarding the interpretability of the CRPS, our understanding is that it is due to the large natural variability in the obtained values. For instance, you point to $FWIsa$, with a lowest average CRPS of 2.2. The $FWIsa$ has an approximate range of 10 for an approximate value of 20. We highlight that these approximations are very rough, only for the sake of explanation. Given this spread, a realization by the ESM would likely have a high CRPS value. The average over time steps would still lead to a high CRPS. To give some statistical ground to this explanation, the equation 8.55 p.353, of (Wilks, 2011) for the CRPS a Normal distribution of location μ and scale σ and an observation X can be written:

$$CRPS = \sigma \left(X(2\,\mathcal{F}_{\mathcal{N}}(X,\mu,\sigma) - 1) + 2f_{\mathcal{N}}(X,\mu,\sigma) - {}^{1}/_{\sqrt{\pi}} \right)$$

With f and $\mathcal{F}$ the respective probability and cumulative distribution functions of the Normal distribution. Thus, assuming variables with different scales but with points at the same quantiles of the distribution, the variables with the higher scales would have the higher CRPS. The Referee also asked for information on the interpretability of the CRPS, so we are adding the appendix 6.1.

- Figure 2, 7 : Could the caption be on the same page as the figure?
This is a good point, we will make sure that it happens during the editing of this manuscript before publication.

- Figure 3 (and other quantile deviation tables) : Is this computed jointly over multiple SSPs? I would assume so, but could benefit the manuscript to write this explicitly somewhere.
Yes, they are calculated together over the available scenarios, and then averaged. We are adding this information in the caption of the Figure 3:
"The deviation is calculated on all available scenarios."

- L335 : "using other distributions without distribution"?
Thanks for noting this error, here is the correction:
"Using other distributions that would not assume independent events may improve these results but would require a higher degree of complexity."

- L346 : ”ESMs mostly provided the total soil moisture” → Suggestion : ”a majority of ESMs only provide the total soil moisture”
Changed accordingly to your suggestion.

- L359 : I think it would benefit the reader to provide intuition on why annual averages can be well represented by a normal distribution.
We have simply added a mention to the theorem that was implicit:
“As an annual average, *SM* may be represented by a normal distribution according to the central limit theorem.”

- L509-L514 : slightly redundant with the previous paragraph.
We decided to keep this paragraph, because even if it is slightly redundant, it summarizes the former paragraph, which is what some readers may want to read.

It would be interesting to evaluate the fitness of the proposed conditional distributions with a statistical test (e.g. Kolmogorov-Smirnov test) or by evaluating the loglikelihood of ESM outputs under the proposed conditional distribution. I would expect the statistical test to fail because the fit would really need to be perfect, but the obtained p-value would nonetheless be a useful indicator.
That is an excellent point. We have appended a section with figures for the negative log-likelihood, averaged over the size of the training sample. We chose the NLL instead of the Kolmogorov-Smirnov, because distributions are trained by minimizing the NLL. We are adding a reference in the appendix the first time that we identify a configuration, likes 259-260:
“We point out that the local performances for this configuration are shown in the Appendix 6.7, along with those of the other variables emulated.”

It would be useful to evaluate the soundness of the emulator with a calibration score (e.g. take the 95% confidence interval of your emulated distribution, and see what fraction of the observations fall within — it should be 95% of them). That should provide a concise and intuitive assessment of the emulators calibration, whilst the CRPS is a distance between cdfs which may be robust, but hard to develop practical intuition upon.
We agree that showing a concise, intuitive but accurate as well assessment of the emulator is the golden target. This is complicated by the number of dimensions (space, time & scenarios, quantiles, configurations, ESMs). This is why we had created the figures 3, 6, 10 and 13, for the deviations of the quantiles. We think that the readers may use these figures to evaluate the performances of the emulator.

Allan, R. P., Barlow, M., Byrne, M. P., Cherchi, A., Douville, H., Fowler, H. J., Gan, T. Y., Pendergrass, A. G., Rosenfeld, D., Swann, A. L. S., Wilcox, L. J., and Zolina, O.: Advances in understanding large-scale responses of the water cycle to climate change, Annals of the New York Academy of Sciences, 1472, 49-75, https://doi.org/10.1111/nyas.14337, 2020.
Nath, S., Gudmundsson, L., Schwaab, J., Duveiller, G., De Hertog, S. J., Guo, S., Havermann, F., Luo, F., Manola, I., Pongratz, J., Seneviratne, S. I., Schleussner, C. F., Thiery, W., and Lejeune, Q.: TIMBER v0.1: a conceptual framework for emulating temperature responses to tree cover change, EGUsphere, 2022, 1-36, 10.5194/egusphere-2022-1024, 2022.
Wilks, D. S.: Statistical methods in the atmospheric sciences, Academic press2011.