Response to Claudia Tebaldi, Referee 1, 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 Claudia Tebaldi for the open, positive and constructive review. Her and the comments of the other Referee 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

This paper details a modification/extension of the established MESMER-X approach to emulating spatially and temporally resolved impact-relevant variables. The targets of the emulations are historical and scenario outputs from CMIP6-era ESMs, focusing specifically on quantities relevant to wildfire risk, i.e., Fire Weather Index and soil moisture, and in particular measures of their tail behavior.

The paper is clearly written (maybe some occasional oddity due to non-native English usage, I know about that issue myself J), well organized and interesting. It shows the value and promise of the MESMER-X approach, highly relevant for integrating climate information and impact/mitigation analysis, and I think it will be publishable after some minor revision. A lot of what is presented is – as the title states – the extension of a methodological framework that has been established and peer reviewed already, so my comments do not question that part, but are mainly suggestions for further validation/clarification/expansion.
Thank you very much for expressing this opinion! Regarding the oddities, we will read the text another time for some corrections.

I start from perhaps the only significant request: I think it would be good to extend the validation in two directions: the first one is interannual variability, right now obfuscated by considering all realizations together and evaluating only the relative positions of target vs. emulated realizations in those plots showing many time series at once and their envelopes. Since interannual variability may be important in creating and making persistent some of these hazardous conditions, some simple metric that compares it between true and emulated realizations at the grid-point and regional scale would be nice to include. The other analysis that I would like to see is a comparison of the behavior of those time series/envelopes/red dots using the lowest scenario besides the highest (currently the only one presented). I think it will be interesting to evaluate if the differential behavior of true/emulated realizations remains qualitatively the same when comparing different scenarios. Right now, the scenario dimension is a bit downplayed by the choices of the validation metrics and I think it is too important an angle to be shortchanged. Otherwise, I really like the succinct metrics of validation used in the paper, it is never easy to synthesize these emulators' performance and the authors have done a nice job in that regard.

Thank you for these insightful comments. We agree that the two suggested directions would deserve more attention.

- Interannual variability indeed matters for the consequences of these hazards, but also to check how capable is MESMER-X to represent these aspects. The most direct approach we can think of is to represent temporal correlations. We have added an appendix with figures, representing the local correlations for one ESM and its emulated counterpart over three periods: 1851-1900; 2051-2100 of SSP1-2.6 and 2051-2100 of SSP5-8.5. We show that the inter-annual variability is correctly represented over preindustrial, low-warming and high-warming scenarios. This appendix is mentioned line 153.
- We acknowledge that we focused on the higher warming scenarios, to present the evolutions over larger domains of warming. The Anonymous Referee 2 has made similar suggestions. We have added the equivalent of figures 2, 5, 9 and 12 for SSP1-2.6 and SSP2-4.5, though in the appendix.

Page 2, lines 45-47: I know what is meant by this as we are all thinking about the same issues here, connecting our emulation work to the IAM community and their scenarios, but I do not think a random reader would understand/appreciate the problem. Maybe expand a bit, possibly describing a specific example (fires impeding the use of afforestation for carbon capture, for example). Some of this is mentioned on page 3, line 69 and following, maybe connect that discussion to this.

Good point, we have edited the text to include more details, as follows:

"For instance, IAMs mitigate climate change by using bio-energies with carbon capture and storage (BECCS) and afforestation, yet these nature-based solutions would be impacted by droughts and fires (Fuss et al., 2014; Smith et al., 2016; Anderson and Peters, 2016). Thus, accurately replicating regional changes in climate extremes and water conditions of Earth System Models (ESMs) at a lower computational cost would help in exploring mitigation potentials and new emissions scenarios."

Page 2, line 50: TXx is not defined. Later on page 4, lines 93-94, annual maximum temperature is mentioned without a reference to TXx.

The definition of TXx was a bit hidden just before the reference. We have edited the text to give it more visibility:

"The MESMER emulator has been developed with this purpose, first for regional mean variables (Beusch et al., 2020; Beusch et al., 2022), and more recently also extended to the MESMER-X version representing TXx, the annual maximum temperatures (Quilcaille et al., 2022)."

Page 3, line 54: maybe specify what periods were used for training? Historical and 21st century/SSPs I assume.
Yes, it makes sense. We have added the period:
"Each one of these emulations account for the spatial and temporal correlations in TXx. MESMER-X was trained on each available ESM of the Climate Model Intercomparison Project Phase 6 (CMIP6) over 1850-2100 (Eyring et al., 2016; O'Neill et al., 2016)."

Page 5, line 118: the reference to Quilcaille et al., 2022 could call out MESMER-X explicitly.
Agreed, we were not sure how to write that. We went for the following sentence:
"Similarly, if $\mathcal{D}$ is a GEV, equation (1) is equivalent to the formalism introduced in the article showcasing MESMER-X (Quilcaille et al., 2022)."

Page 7, lines 172-173: I think the equations referenced should be (8) and (9) not (5) and (6). Also the quantities (Y in particular) in eq. 8 need to be defined.
Thanks for pointing this out, it is corrected.

Page 10, Figure 1 (but this also applies to the other similar figures): It is interesting to see light color cells for some of the ESMs for the reference stationary GEV (and implicitly Gaussian) distribution. **What are we to make of this**? I'm assuming this is a fit over both historical and scenario period, or is it just the historical used for training? Is the light color just a relatively better performance (but still bad) or is it somehow good? **Can the magnitude of the CRPS metric (in this case ~2.5 with little variation, in other figures much different) be interpreted**? In summary: I would like a bit more explanation of how to interpret the magnitude and performance of this baseline metric that is then used to say something about the relative performance of other choices of non-stationary distributions. Also, it is a bit unfortunate that light/dark colors have the opposite meaning in the first row and in the lower rows. Maybe calling this out in the caption could help the reader's not geeting confuse (but that reader is me, feel free to disregard this latter point).
That's right, we didn't explain enough this part. Interpreting the CRPS is complicated by its lack of an upper value, while its minimum is zero. We brought more details when we introduce the metric (lines 179-182):
"A high CRPS for this benchmark means that the differences between the cumulative distribution functions are too big, which implies that a stationary distribution does not correctly reproduce the statistical properties of the training sample, while a distribution reproducing perfectly the training sample would have a CRPS of zero (Hersbach, 2000), as illustrated with ***Error! Reference source not found.*** *in the Appendix."*

We have also edited the Figure 1, 4, 8 and 11 to reverse the colors of the CRPS, with an edit of the caption as follows:
"Figure 1: Selection of the configuration for the seasonal average of the FWI (FWIsa). *For each ESM, the CRPS and CRPSS are averaged over space, time and scenarios. The darker is the colour of a cell, the better is the configuration at reproducing the distribution of the ESM. The upper row (white to black) corresponds to the CRPS of the configuration used as benchmark. A higher CRPS (lighter colour) indicates that the stationary distribution used as benchmark does not reproduce well the distribution of the ESM. The next rows (white to red)*

correspond to the CRPSS of the tested configurations, relatively to the benchmark. A higher CRPSS (darker colour) indicates that the proposed configuration improves the reproduction of the distribution of the ESM."
For information, the Referee 1 also asked for modifications on these figures, more precisely on the font sizes and titles of the color bars.

Page 12, lines 264-265 I did not understand what is meant here: "The median of the ESM remains effectively at the center of the realizations by UKESM1-00-LL."
Thank you for noting that. We corrected this sentence, hopefully now better:
"Over 2014-2100, the realizations by UKESM1-0-LL remain mostly within the range of the emulations, except for the third row that corresponds to a grid point close to Manaus in Amazonia."

Page 15, lines 298-299: here is one spot where the interpretability of the magnitude of the CRPS would be of value.
This is indeed a good spot to illustrate how to interpret this metric. Here is the text we added:
"Because the higher is a CRPS, the worse is the distribution at representing the training sample, two results can be deduced. First, stationary GEV distributions are much better at reproducing FWIsa than stationary Poisson distributions are at reproducing FWIxd. It may be because FWIxd has stronger responses to climate change than FWIsa, meaning that stationary distributions, Poisson or GEV, cannot correctly reproduce these evolutions. It may also be because the shape of a Poisson distribution cannot reproduce as well the shape of the observed FWIxd, compared to a GEV for FWIsa."

Page 17, lines 321-324. I would argue that also the first-row panel (regional average) shows the same tendency. I think you mention it in the discussion, but would the model allow different choices (linear or quadratic) in different locations? If that is a possibility, I understand not giving it a try not to make things more complicated, but I think it could be mentioned here explicitly as a capability/powerful feature of the model…but maybe I'm wrong and it would be difficult to apply a mix of linear/non linear links?
That's right, the top row shows such an effect, although to a lesser extent, because that's averaged over a region where not all grid points have this effect. We are adding that:
"The same effect appears on the first row, although to a lesser extent."

At the moment, MESMER-X assumes that all grid points share the same configuration (distribution & functions on parameters). Theoretically, MESMER-X would support having the configuration tailored to the grid point, and we already did some technical steps in this direction (looping over configurations & choice on CRPS or BIC), but not all of them (interpreting parametrizations that depend on the grid point). However, the full implementation of this feature isn't planned for this paper. Though, we are still unsure about the marginal gain in performances. We are adding that to the conclusion:
"Making parametrizations dependent on the grid point would be a solution, but wasn't implemented for this article."

Page 17, lines 335-336: I did not see this detailed point made at the beginning of the session. In this regard, could you also discuss if the use of a discrete distribution poses any challenge to the probability transform, or if it all works seamlessly? Evidently it worked but it is not obvious to me how.

We should have announced this point from the beginning of section 3.3, here is what we added lines 299-304:
"Using this distribution implicitly assumes that the events are independent of each other, which is not exactly the case here. Assuming that a day matches the criteria for extreme fire weather (Quilcaille et al., 2023) for instance during the fire season, there are higher chances to have the next days also matching this criteria, compared to a period out of the fire season. Nevertheless, we choose this distribution because of its relative simplicity."

You are right, a discrete distribution through a probability integral transform (PIT) isn't straight-forward. We added an appendix with a figure and some explanations, introduced in the manuscript, lines 135-136:
"Equation (2) applies as well if $\mathcal{D}$ is a discrete distribution, as illustrated in Appendix 6.2."

We do not copy here the full text of the appendix, that is a bit too long. In a nutshell, it works seamlessly because of two reasons. The first one is that a discrete distribution at a value X is representative of the interval [X-0.5; X+0.5[, there is an underestimation over [X-0.5; X[, but an overestimation over [X; X+0.5[, thus leading to partly compensating errors. The second one is that this process occurs in one way during training, then the other way round during emulation. We acknowledge that this is not a rigorous demonstration, but we are also planning to write down all the statistics behind to ensure that.

Page 17, lines 336-337: sentence needs correcting: "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."

Page 20-21: This behavior if soil moisture is really interesting! I'm impressed that just the lagged temperature is helping to account for this behavior. In that regard, is this lagged temperature produced by the emulator? Is it global temperature produced according to the scenario by a simple model? Need a bit more elaboration of how this is actually implemented. I'm also wondering if a more robust derivative measure (implicit in the use of the lagged T of course) could be a good auxiliary variable.
We tried this effect to give a sense of the local trend in temperature. With ΔT being the change in global mean temperature, one may rewrite the terms as follows:

$$\lambda_{s,1}\Delta T_t + \lambda_{s,2}\Delta T_{t-1} = \left(\lambda_{s,1} + \lambda_{s,2}\right)\Delta T_t + \left(-\lambda_{s,2}\right)\left(\Delta T_t - \Delta T_{t-1}\right)$$

Thus, the second term can be associated with the first derivative in time. We decided to write it down this way, because ΔT_t is our main driver, and introducing its derivative may confuse the readers. The lagged temperature is not produced by the emulator, it is simply the one at the former year that the ESM provides. For a scenario in 2015, we are using the corresponding historical in 2014. For the value in 1850 for its former year, we tried either using the average over 1850-1899 or the value of 1850 itself, but it does not make much difference, for it is just 1 point, with a preindustrial period long enough to account for this period. In this regard, we could start training in 1851 instead. To elaborate while not being too technical, we edited lines 395-397 as follows:
"Here, as a first attempt to reproduce this effect, we will test in the configuration a lagged variable using the ΔT at the former year. This lagged variable is obtained by shifting the ΔT of the ESM by one year. From a modeling perspective, having both ΔT_t and ΔT_{t-1} is equivalent to having the value at year t and its first derivative."

We preferred here a backward difference operator for the first derivative, to give more weight to the past. we agree that other measures may be more appropriate. Without going into all the details, one could model the interannual change in the variable instead of the variable itself. Also, ΔT_{t-1}, ΔT_{t-2},… and/or ΔT_{t-n} with a BIC criteria may help for different timescales. Extending this principle could be done using impulse response functions.

To investigate the proper modeling approach, we think that we have to identify adequate variables, and to try them on adequate scenarios, e.g. with overshoots. This is in our ToDo list, and we are convinced that such a work would benefit to all spatial emulators :)

Page 26, lines 452-453: I could not understand this sentence: "While the range….assess variations".
Thank you for noting that. We cleaned this sentence and the one after, leading to:
"The spatial patterns of the ESM shown here on the top row, CNRM-CM6-1, are correctly reproduced by the emulations on the three following rows. The right column shows that the regional responses are correctly reproduced, with a majority of the ESM points being within the range of the emulations."

Anderson, K. and Peters, G.: The trouble with negative emissions, Science, 354, 182-183, 10.1126/science.aah4567, 2016.
Beusch, L., Gudmundsson, L., and Seneviratne, S. I.: Emulating Earth system model temperatures with MESMER: from global mean temperature trajectories to grid-point-level realizations on land, Earth Syst. Dynam., 11, 139-159, 10.5194/esd-11-139-2020, 2020.
Beusch, L., Nicholls, Z., Gudmundsson, L., Hauser, M., Meinshausen, M., and Seneviratne, S. I.: From emission scenarios to spatially resolved projections with a chain of computationally efficient emulators: coupling of MAGICC (v7.5.1) and MESMER (v0.8.3), Geosci. Model Dev., 15, 2085-2103, 10.5194/gmd-15-2085-2022, 2022.
Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937-1958, 10.5194/gmd-9-1937-2016, 2016.
Fuss, S., Canadell, J. G., Peters, G. P., Tavoni, M., Andrew, R. M., Ciais, P., Jackson, R. B., Jones, C. D., Kraxner, F., Nakicenovic, N., Le Qu\'e, r\'e,,, Corinne, Raupach, M. R., Sharifi, A., Smith, P., and Yamagata, Y.: COMMENTARY: Betting on negative emissions, Nature Climate Change, 4, 850-853, 2014.
Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather and Forecasting, 15, 559-570, 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.
O'Neill, B. C., Tebaldi, C., Van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J. F., Lowe, J., Meehl, G. A., Moss, R., Riahi, K., and Sanderson, B. M.: The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6, Geoscientific Model Development, 9, 3461-3482, 10.5194/gmd-9-3461-2016, 2016.
Quilcaille, Y., Batibeniz, F., Ribeiro, A. F. S., Padrón, R. S., and Seneviratne, S. I.: Fire weather index data under historical and shared socioeconomic pathway projections in the 6th phase of the Coupled Model Intercomparison Project from 1850 to 2100, Earth Syst. Sci. Data, 15, 2153-2177, 10.5194/essd-15-2153-2023, 2023.
Quilcaille, Y., Gudmundsson, L., Beusch, L., Hauser, M., and Seneviratne, S. I.: Showcasing MESMER-X: Spatially Resolved Emulation of Annual Maximum Temperatures of Earth System Models, Geophysical Research Letters, 49, e2022GL099012, https://doi.org/10.1029/2022GL099012, 2022.
Smith, P., Davis, S. J., Creutzig, F., Fuss, S., Minx, J., Gabrielle, B., Kato, E., Jackson, R. B., Cowie, A., Kriegler, E., Van Vuuren, D. P., Rogelj, J., Ciais, P., Milne, J., Canadell, J. G., McCollum, D., Peters, G., Andrew, R., Krey, V., Shrestha, G., Friedlingstein, P., Gasser, T.,

Gr\"u, bler, Arnulf, Heidug, W. K., Jonas, M., Jones, C. D., Kraxner, F., Littleton, E., Lowe, J., Moreira, J. e., , Roberto, Nakicenovic, N., Obersteiner, M., Patwardhan, A., Rogner, M., Rubin, E., Sharifi, A., Torvanger, A. o., rn, Yamagata, Y., Edmonds, J., and Yongsung, C.: Biophysical and economic limits to negative CO2emissions, Nature Climate Change, 6, 42-50, 2016.