Review of: Extending MESMER-X: A spatially resolved Earth system model emulator for fire weather and soil moisture by Quilcaille et al.

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 ☺), 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.

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.

I have a few other comments which won't be as demanding. I will list those in the order they come up while reading the paper, even if some may be a little more substantial than others.

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.

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.

Page 3, line 54: maybe specify what periods were used for training? Historical and 21st century/SSPs I assume.

Page 5, line 118: the reference to Quilcaille et al., 2022 could call out MESMER-X explicitly.

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.

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).

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."

Page 15, lines 298-299: here is one spot where the interpretability of the magnitude of the CRPS would be of value.

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?

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.

Page 17, lines 336-337: sentence needs correcting: “Using other distributions without distribution…”

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.

Page 26, lines 452-453: I could not understand this sentence: “While the range….assess variations”.