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

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 $\mathcal{D}(\cdot|\alpha_{s,t,1}, \ldots, \alpha_{s,t,p})$ 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 $\mathcal{D}(\cdot|\alpha_{s,t,1}, \ldots, \alpha_{s,t,p})$, 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 $\mathcal{D}$ and parametric functions $f_{s,p}$ — and demonstrates good performance, with limited quantile deviation.

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.

Questions and comments

- 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"
- L115 : The functions $f_{s,1}, \ldots, f_{s,p}$ 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?
- 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).
- L180 : Probably the case already, but are score spatially averaged by accounting for decreasing cell size toward poles?
- 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 E_0 (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?

- Figure 2, 7 : Could the caption be on the same page as the figure?

- 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.

- L335 : "using other distributions without distribution"?

- L346 : "ESMs mostly provided the total soil moisture" → Suggestion : "a majority of ESMs only provide the total soil moisture"

- L359 : I think it would benefit the reader to provide intuition on why annual averages can be well represented by a normal distribution.

- L509-L514 : slightly redundant with the previous paragraph.

Additional suggestions *I do not expect these suggestions to be integrated in the revised manuscript.*

- 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.

- 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.