2nd Review for

METEORv1.0.1: A novel framework for emulating multi-timescale regional climate responses

The new version of this manuscript has improved. The issues that I had raised are mostly solved. The only remaining points are:

- The current justification for timescales with a lower limit of 1 year are not physically/statistically valid. I argue below why the temporal resolution does not preclude lower timescales. I recommend mentioning this as a modelling choice for this statistical model in the Discussion.
- In the Discussion, please point out to the limits in regional performance.

Regarding the local validation / performance of the emulator, the authors combined my suggestions with those of Referee 1 with new figures & sections. While there is no map of the R2, the maps of the difference between CMIP6 mean and emulation mean (Figs 8-9) and the plots that mention R2 (Figs 10-11) still convey the relevant information, albeit harder to extract the limitations. For instance, the R2 ssp534-over is lower over Australia for temperatures (Fig 10h) but it does not really appear on map (Fig 8i), and ssp534-over is not in Figure 14. In my opinion, the last two decades of ssp534-over are not representative of the difficult part of the overshoot, which is right after the peak in warming. I will not ask the authors to add new figures, there are already enough. I simply suggest to discuss a bit more the (local/)regional performance than it is now. Section 3.4 provide many info, but it is rather at a global scale. For now, it is only Line 355-356 of the version with tracked changes where we read about lower performance in some regions. Could these new results make it to the Discussion?

I fully agree that the assumption for the dependency of IRF on preindustrial/current conditions is classic and well understood. I acknowledge that increasing the complexity of the training would represent a massive additional work, and that the work presented in this manuscript is sufficient for publication. Same goes for using PDRMIP experiments.

Regarding the timescales limited to 1 year, I would not interpret that annual resolution for the inputs forces us to timescales higher than 1 year (L408-411). In short, we can infer the pace of processes even if their characteristic timescale is shorter than the pace of observations.

In more details, while it is known that there is a hysteresis, thus timescales higher than 0, we cannot assume that the physical value would be below 1. For instance, a very small hysteresis implies timescales that tend toward 0+, i.e. where the IRF tends towards a Dirac, i.e. where the convolution becomes equivalent to a pattern scaling. Very importantly, imposing a limit to 1 imposes a limit to what the IRF can do. For instance, an IRF $y = e^{-(t-t')/1}$ cannot give less than a value of 0.37 to the forcing at $t' = 1$, while data may say that it could be lower in some

situations. If data shows something closer to 0.14, it may be a timescale around 0.5 year. There is of course a limitation from the data, which is if the timescale is so small that the value at $t' = 1$ becomes non significant.

The fact that so many models hit the set threshold of the timescale at 1 raises a flag, that it is possible that the term with the lowest timescale may be a quick equilibrium, e.g. in the range of 0.1-1. Imposing a spurious limit has the opposite effect, reducing the validity of the approach, while not being less interpretable.

To summarize, the explanations L408-411 are not valid from a physical and statistical perspective. I am not asking the authors to retrain the model, but I encourage them to mention that this modelling choice as a potential limitation.

Regarding all the other points, the manuscript gained in clarity.