Reply to reviewer 2

We thank the reviewer for constructive and careful reading of our manuscript, providing useful suggestions for improvement. Below we answer each point in detail and relay how we have attempted to improve the manuscript according to each of the criticisms.

“My principal point would be on the validation. The current framework is (1) deconvolving the global signals, (2) validation with the Pearson coefficients (Table 1), (3) deducing spatial patterns. The validation is indeed very good globally, but there is for now no validation of the spatial ouputs of the whole modelling chain. To evaluate the spatial decomposition, I would appreciate a map of the R2 for in-sample and out-of-sample scenarios, for comparison of the original data to the emulated data.”

This comment aligns with comments from reviewer 1, and to improve this we have made the suggested maps, and in addition we have made a table of spatial and global errors comparing them to the ClimateBench test set and including the same error metrics for each model and scenario combination.

“L85-90 & L256-257: training responses on single experiments is somewhat of a risk, reducing the domain of validity of the emulator. For instance, although IRFs are reasonable approximations1 , IRFs for the response atmospheric fraction of CO2 to a pulse of CO2 emissions is known to depend on its calibration under preindustrial and current conditions2 . This is mostly because the preindustrial carbon cycle does not behave exactly like a perturbed carbon cycle. Here, some IRFs of METEOR are calibrated with abrupt-4xCO2, thus starting under a preindustrial climate, up to a disturbed climate. Have you tried training on multiple experiments instead, using both experiments under past and current conditions? For instance with the variants of ssp245 as well for GHG, and some variants of historical with only aerosols?”

While we agree that the IRF response assumption is a possible structural weakness of our approach, we also note that the Joos et al 2013 reference (2) also found that the “responses on temperature, sea level and ocean heat content is less sensitive to” CO2 concentration conditions. We note also that the same IRF assumption is made in simple climate models such as FaIR – and so, while it remains an approximation, it is a common and well understood approximation.

In METEOR, the emissions-to-forcing module (provided by CICERO-SCM) also allows for some state-sensitivity and dynamical response to previously emitted carbon, meaning that the forcing strength emulated is also affected by the background conditions to some degree. We have added a small paragraph to clarify this part of the model a bit further. As

for training on more than one scenario, our current setup is done using a single experiment as training input per forcer, and more sophisticated combinations for training would require substantial changes in the training logic of METEOR. We therefore consider it out of scope for this article and leave such experimentation for future work.

“Related question L96-100 and L181-199, in particular equation 9: Why use the difference when there is ssp245-aer? Besides ssp245-GHG, ssp245-CO2, ssp245-stratO3, etc? Quick insight, there may some differences if using different combinations of experiments. For instance, using hist-aer would lead to different results than using historical – hist-PiAer or historical – histpiNTCF. Because the temporal response of a forcing may depend on the other forcings. The response under hist-aer has much less warning, different atmospheric chemistry for aerosols than what we would see under historical – hist-PiAer. Though, I agree that it is a second order effect, tough to include in this framework. Thus I would simply suggest you to mention this limitation.”

We understand the suggestions made here, and in part we agree with them. In fact, in our first versions of the model, we used PDRMIP experiments (1) to train the model to fit multiple forcers one-by-one. One reason why we left this approach, for the use of residual fits, was to be able to use widely available and reasonably up-to-date model data from the CMIP6 ensemble, and to be able to emulate as wide a set of models as possible. For this we wanted to use only experiments which had been widely run for our emulation, at least for this demonstration of our model, choosing versatility and usefulness over the possibility of marginally higher accuracy. However, modelling using more specified experiments such as these to split into more forcer components, possibly even with regional split, is something we hope to do in the future and we touch on this point in the outlook section of the article.

“L101-110: I would be careful about summarizing aerosols with sulfates. Each aerosol would have its own specificities in terms of radiative effects, atmospheric chemistry, lifetimes and transport. Some experiments that would be useful here would be: hist-aer, hist-piNTCF, histstratO3, hist-piAer. I am not asking to recalibrate the model, that would represent a massive additional work to account for different aerosol species. But it could be noted as a potential limitation for future research.”

This point has now been clarified in our manuscript, we are in fact not conflating all aerosols with sulfate in the modelling. Instead, what we call aerosol patterns are only driven by and mapped using sulfate forcing (which also includes all aerosol-cloud interactions as they are sulfate driven in ciceroscm). All other forcing terms are mapped using the GHG-forcing responses. Also, the ciceroscm emissions-to-forcing modelling used as part of driving METEOR does, though in extremely simplified ways, account for

lifetimes, radiative effects and some atmospheric chemistry on a per forcer basis for both aerosols and other individual forcer components. We have added a sentence to the text here to further clarify how we map the different forcing patterns.

“L144-150 + 199-201: I appreciate the technique of separating the timescales into these bins, and great work to evaluate the adequate numbers of timescales in Appendix A. Though I have a question on the choice of bounds. For now, the minimum τ is 1 year. It assumes that there is a non-immediate stabilization of the response to the forcing, which is physically quite robust in this context. Though, it is not unlikely that there may be one mode for the response below 1 year, for a very rapid stabilization. Of course, the model runs at an annual resolution, but it would give some flexibility to the response at t=1 and the asymptote (equation 2). Maybe even more relevant for aerosols? So my question would be: would there be a significant gain in performance by including another mode below 1 year? By the way, looking at Table B1, there are lots of τ1 at 1.0 year, which could be a sign that the minimization algorithm forced the τ1 at the very limit of what it could do given the user-defined bounds, in other words that the error function could be minimized by relaxing these bounds. Overall, about half of the tau of the table seems to hit their bounds. I think that it should be investigated. ”

This point is valid, and we have done testing of the model using both overlapping timescales and shorter timescales than this. The solution presented in this paper was the best trade-off in terms of interpretability and performance. Given that the data and modelling are limited to annual resolution, the overall validity and interpretability of sub-yearly timescales would be somewhat questionable. We have now expanded the discussion a bit on this point to clarify why we have chosen this setup.

“Suggestions: L24: can add as well RCMIP phase 2, probably more relevant than RCMIP phase 1.”

Thank you for the suggestion, we have added this reference.

“L26-27: the uses for fast spatial climate modelling frameworks is broader than that, see for instance3-6”

Thank you for the suggested widening of scope and proposed references which we have now included.

“L35-38: Important point on probabilistic spatial climate information & pattern scaling. Pattern scaling provides only a deterministic response, the mean of the climate field. Having a probabilistic information may come either from the uncertainty in modelling or through natural variability. Pure pattern scaling like PRIME does not include natural variability. MESMER does represent the natural variability obtained through temporal auto-

regressions with spatially correlated innovations. Then, regarding precipitations, this is obtained through a more elaborated approach7 . Finally, pattern scaling is generalized with non-stationary distributions with MESMER-X 8,9, while also removing the assumption of linearity, eg for soil moisture. For the sake of transparency, I'm the author of the two latter papers, and I'm not asking the authors to add them as references. My point is that there is a need for clarification, that pure pattern scaling is limited to uncertainties in modelling for probabilistic assessment, and going further with natural variability requires additional statistical tools. "

We have taken out the word probabilistic here, and have a longer discussion of probabilistic extensions later in the text where we mention the works suggested here (although we were in fact already citing them in our first version of the article).

"L38-41: In STITCHES, the portions of existing simulations are found not only through the median in global mean temperature, but also by its derivative. "

Thanks for pointing this out, we have now fixed this in the text.

"L56-58: The feedbacks of JULES don't feedback in FaIR or PRIME (yet)."

We have now made it clearer that such feedback is a potential extension for prime and not an existing feature.

"L17-137: At resolution of the ESM? Is there any rescaling?"

METEOR works on the resolution of the ESM model. For the comparison and plots in the manuscript we have done rescaling for multi-model comparisons, but the model natively emulates at the model resolution, and updated error score tables have also been calculated at model resolution. We added a note to clarify this to the text.

"L140-143: decomposing $xglobal$ with IRFs is a good idea, it simplifies the modelling10. Another approach is to decompose the local effects with IRFs, e.g. $\boldsymbol{TGHG}$, as conducted in Womack et al 2025 11 . In my opinion, both approaches have pros and cons, I'm not sure which one performs best, but Womack et al, 2025 should be mentioned. "

Thank you for pointing out this very relevant reference, we have included references to this now.

"L211: the link between the Moore-Penrose pseudoinverse and Barata and Hussen, 201212 should be clearer. In the text, the paper was mentioned without the method, and here the method is mentioned without the paper. "

We have tried to improve this link now by citing in both places and mentioning the technique where the citation was already in place.