Answer to Anonymous Referee #1

We thank the anonymous referee for the constructive feedback which will help to improve the quality of our manuscript. In the following, we provide a point-by-point answer to the reviewer whereby we show the reviewer's comment in black and our response in blue.

General comments
The authors propose a statistical model for emulating output from Earth System Models (ESMs). The model is composed of deterministic and stochastic components that are intended to capture the forced trend and variability, respectively. There is clearly a lot of work to be done in developing cheap tools like emulators to get more information from our climate model archive, and I am glad to see another contribution to this field. However, I have a number of concerns about the model formulation and, echoing Comment 1 from Robert Link, the validation of the emulator output.

We are happy to hear that the reviewer agrees that developing computationally cheap tools to get more information from the climate model archive is important. In the following, we will address the concerns the reviewer expresses.

Specific comments
1. One of the challenges of fitting emulators to data or climate model output is separation of the forced and internal components (under the common assumption that they are linearly separable). The authors propose the use of a common approach of regressing onto a smoothed version of the global mean temperature (plus volcanic bursts), but do not provide evidence that this approach is successful. The method can and should be tested within one or multiple initial condition ensembles.

We thank the reviewer for this comment. There seems to have been a misunderstanding caused by a naming convention we chose, which resulted in comments from both this reviewer and R. Link. In fact, we do test the emulator using multiple initial-condition ensembles. While we calibrate the emulator with a single run per climate model, for all models where several initial-condition members are available, we evaluate the performance of the emulator using initial-condition members not employed during training in Sect. 6.3.2 of the original manuscript. For illustrative examples, please additionally consult Fig. 5 of the original manuscript. Throughout the manuscript, we referred to this type of evaluation as "out-of-sample" testing.
To improve the readability of the paper, we will exchange the "out-of-sample" terminology with explicitly referring to "independent initial-condition ensemble members not employed during training".

2. The spatial model for the innovations is presented with minimal justification. How was the exponential covariance model chosen versus one that is smoother in space? More importantly, given that the spatial structure of temperature variability depends on the prevailing wind direction, the geometry of the coasts, land surface type, etc., is an isotropic covariance model even appropriate?

A misunderstanding occurred here. We do not employ an exponential covariance model as a spatial model for the innovations, instead we sample from a regularized empirical covariance matrix which is detailed in Eq. 9 of the original manuscript. For the regularization, we employ an approach referred to as localization which is well established in the field of data assimilation (Carrassi et al., 2018).
To convey this point more clearly, we will dedicate more text to the justification of our spatial model in the revised manuscript, highlighting that we employ an approach which is common in data assimilation and which is able to retain anisotropy in the underlying data on regional scales.

3. Identifying parsimonious but sufficient metrics for validation of model ensembles is a challenging and unsolved problem. However, the authors are too qualitative in their evaluation of their emulator skill, which is composed primarily of visual inspection of emulated fields and plots like Figs. 9 and 10. Given the choice of spatial model discussed in (2), it would be helpful to see validation metrics on both the spatial and temporal correlation structure. In addition, the assumption of Gaussianity is built-in but never checked. Finally, validation metrics should be provided with respect to a reasonable null hypothesis, otherwise it is difficult to assess whether a certain error value is meaningful. For example, how large would a given error metric be if different realizations of an actual ESM were resampled, and then the metric of interest was calculated?

To address the reviewers call for a more quantitative validation of the emulator, we plan to extend the space-time verification of our emulator in the revised manuscript. Specifically, to address the concerns raised by this reviewer: we will (1) expand the verification section in the paper to include both verification of the deterministic trend and the variability around it, (2) include results from a Shapiro-Wilks test to demonstrate the validity of the Gaussianity assumption of the innovations of the local residual variability in the supplementary material, (3) extend Figs. 9 and 10 of the original paper to contain information on the degree to which true ESM runs are indistinguishable from single emulated runs.

4. The writing could be improved to make the manuscript flow more smoothly. In particular, Section 2 could be reworked to more clearly identify what is missing in the current literature that the authors aim to ameliorate with this manuscript.

The text (in particular Sect. 2) will be carefully revised and re-structured as needed to improve the readability of the manuscript. In particular, we will focus on highlighting the added value of our study compared to existing literature more explicitly.

5. Lines 437-439 make strong statements about replacing single model ensembles with emulators such as the one proposed. Without further validation, I don't think the authors can say "the latter can be readily mimicked by our emulator based on a single ESM run."

As highlighted in our answer to the specific comment 1, there seems to have been a misunderstanding regarding our validation of the ability of the emulator in reproducing initial-condition ensembles. The diagnostics we have provided in Sect. 6.3.2 of the original manuscript address these concerns. But we will improve the clarity of the manuscript with respect to this point and also provide additional validation metrics. This specific sentence will be replaced with a more in-depth assessment of the potential of an emulator such as the one we have developed to achieve this goal.

Technical corrections/minor points

1. There are minor grammatical and spelling errors throughout.

The manuscript will be carefully revised with focus on grammatical and spelling errors.

2. In the discussion of the forced component, the authors should additionally reference the various LIM-based methods (e.g. Frankignoul et al, 2017, Estimation of the SST Response to Anthropogenic and External Forcing and Its Impact on the Atlantic Multidecadal Oscillation and the Pacific Decadal Oscillation), signal to noise maximizing EOFs (e.g. Ting et al., 2009, Forced and Internal Twentieth-

Century SST Trends in the North Atlantic), and low frequency component analysis (e.g. Wills et al., 2018, Disentangling Global Warming, Multidecadal Variability, and El Niño in Pacific Temperatures).

We thank the reviewer for directing us towards LIM-based methods, and we will consider including these discussion points in the revised manuscript.

3. The citation of McKinnon and Deser (2018) is slightly misleading. The longer timescales related to coupled modes are explicitly modeled, such that the remaining variability has near-zero memory, and so can be block bootstrapped.

We thank the reviewer for noting this, and we will revise the text accordingly.

4. I was somewhat mystified by the comment on Line 378 that CMIP5 models do not reproduce the large-scale temperature response to atmospheric waves, which is incorrect. Any reasonable atmospheric model produces Rossby waves and is reasonably accurate at simulating the temperature response.

We thank the reviewer for pointing out that the corresponding line was not formulated clearly enough. In the revised manuscript, we will clarify that: “We hypothesize that localization radii below 1500 km are not selected in any of the 40 CMIP5 model emulators, because such localization radii create too small-scale stochastic temperature variability which cannot mimic typical temperature responses induced by planetary-scale atmospheric waves in climate models.”