Answer to Robert Link

We thank Robert Link for taking the time to write such a detailed short comment on our paper and for raising several points which will help to improve the quality of the manuscript. In the following, we provide a point-by-point answer whereby we show R. Link's comment in black and our response in blue.

This paper does some interesting work toward systematizing the way we construct climate model emulators, which could be very useful for comparing different kinds of emulators and for designing interoperable components for emulating climate models.

We are happy to hear that our effort to systematize the design of emulators and provide a modular framework is appreciated.

I would have liked to see a little more depth in section 6.3, "Quantitative verification". The authors show plots comparing the quantiles of the emulator-generated ensemble to the corresponding quantiles of the CMIP ensemble, for three regions, and they remark that "the median [of the CMIP ensemble] is successfully emulated, but the emulations are a bit underdispersive", but this assessment seems to be based entirely on visual inspection of Figure 8. This analysis would be a lot more compelling if it included quantitative statistical tests, such as a t-test for equality of the means and the Kolmogorov-Smirnov test for equivalence of the overall distributions. If underdispersion is a particular concern, tests for equality of variances could also be applied. Better still would be to develop measures of differences in key properties of the distribution and to derive confidence intervals for those difference measures. Such measures would give prospective users the tools they need to evaluate whether an emulator is fit for whatever use they intend to put it to.

A misunderstanding seems to have occurred here. Figure 8 is still part of Sect. 6.2 "Example realizations". Figures 9 and 10, however, provide a quantitative verification for all climate models in all SREX regions for the 5^{th}, the 50^{th}, and the 95^{th} quantile with Fig. 9 depicting the performance on the training runs and Fig. 10 the performance on independent initial-condition ensemble members not seen during training. These analyses address the main concerns of the reviewer and show that the developed emulator has a satisfactory performance in the representation of an initial-condition ensemble despite being trained on a single run.
Furthermore, we agree on the usefulness of additional quantitative verification and will adapt the revised manuscript accordingly. More specifically, Figs. 9 and 10 will be extended to contain information about the degree to which true ESM runs are indistinguishable from single emulated runs.

In addition to concerns about how these marginal distributions are evaluated, the marginal distributions appear to be the only dimension along which the authors evaluate the emulator performance. There is no mention at all of testing the spatial correlation or time correlation properties of the emulator. This is a significant omission because the marginal distributions are surely the easiest part to get right when designing an emulator. Capturing the space and time correlations is the true test of the algorithm. In particular, we know that both ESMs and the real climate system display long-range teleconnections and quasi-periodic oscillatory behavior with periods ranging from years to decades. In order to truly evaluate the emulator algorithm, the authors need to explore its ability to produce these phenomena.

As noted by R. Link, there was no explicit evaluation of the space-time characteristics of the emulator in the original manuscript. In the revised manuscript, we will include additional analyses showing to what extent serial and spatial correlations of ESM runs can be retained by our simple statistical emulator.
However, it should be pointed out that Figs. 9 and 10 already constitute a very aggregated form of a space-time verification. These figures show the skill of the emulator in capturing the underlying trend (as indicated by 50th percentile) as well as the variability around it (as indicated by the 5th and the 95th percentile) on regional scales. Large parts of the regional variability are reproduced but the emulations are slightly underdispersive compared to true ESM runs. Without accounting for spatial correlation in the innovations of the local residual variability module, the results would be far more underdispersive. Furthermore, we would like to note that the scope of the present study was to focus on the generating stochastic fields that resemble temperature at regional scales based on a single training run, not on emulating far-reaching teleconnections and multi-decadal variability. The considered local residual variability module was chosen accordingly and is by design not able to reproduce teleconnections at the planetary scale or multi-decadal dependencies. In the revised manuscript, we will clarify these design choices alongside limitations with respect to spatial teleconnections and multi-decadal dependencies.

The authors' choice to do out of sample validation was interesting, but I am unsure as to whether I agree that it's a useful step in this sort of work. Out of sample validation is normally done when developing models that provide point estimates of the system they are modeling. The theory is that the fitting data is a combination of features that area deterministic function of the covariates and random features that are idiosyncratic to the sample data. Out of sample validation provides a way to ensure that the model is capturing the former and ignoring the latter.
The goal of this kind of emulator, however, is something different. Instead of trying to provide a point estimate that reflects the influence of certain covariates, we are trying to simulate random draws from the probability distribution implicitly defined by the ESMs, including all components, both random and deterministic. Therefore, it is not clear what it is that we are trying to exclude by doing out of sample validation. In other words, normally overfitting is caused by the presence of noise (i.e., random response) in the fitting data, but if the noise itself is what we are trying to fit (i.e., we are trying to produce a stochastic variable with similar properties to the noise), then what is it that we are potentially overfitting?

Based on the comments of R. Link and reviewer 1, we realized that the chosen "out-of-sample" terminology may have been misleading for some readers. To increase clarity, we will no longer employ the "out-of-sample" term in the revised manuscript, instead we will explicitly refer to "independent initial-condition ensemble members not employed during training".
Note also that we regard testing the emulator's performance on independent initial-condition ensemble members, where possible, essential because training on a single run might results in overfitting to this specific realization. Our analyses reveal a satisfactory performance of the emulator on regional scales when evaluated against both the training run and independent initial-condition ensemble members not seen during training. Please refer to Sect. 6.3.2. of the original manuscript for a detailed discussion on this topic.

In equation (3) the authors split the global mean temperature time series into a deterministic component and a stochastic variable component. Their purpose in doing this is to allow the local temperature to respond differently to the two components, an innovative approach that makes some sense theoretically. However, they do not take the next step of evaluating the local mean temperature model to see whether the additional coefficient is supported by the data. Either the deviance information criterion (DIC) or Watanabe-Akaike information criterion (WAIC) would be a good choice for such an analysis.

In the paper, we have decided to follow the approach of proposing a specific implementation for our emulator without testing individual components against alternatives as there would be countless alternative implementations which could be considered. In the revised manuscript, we will emphasize that the presented modular framework allows to exchange individual components to accommodate specific user needs.
Note also that we have tested both regression against the full global mean temperature time series and regression against the split-up global mean temperature time series in the exploratory phase of this study. The chosen configuration led to less underdispersive results on regional scales.

The more I read of the literature in the this area of including variability in climate model emulators, the more I am convinced that designing a plausible emulation algorithm is the easy part of this kind of research. What is hard is proving that the statistical properties of the distribution of the emulator outputs are consistent with those of the emulated system. The big frontier in this research area lies in finding ways to characterize similarities and differences between the joint probability distribution of the variables produced by the emulator and that of the system being emulated. Such methods should be fully quantitatvie (i.e., they should produce a measurement of how much the emulator distribution might deviate from the distribution in the real system). Determining what properties of the joint distribution should be reproduced will be an important step in this sort of evaluation. These properties should include, at a minimum, not only marginal distributions, but also space and time correlation properties.

We appreciate this philosophical input on emulator research. We agree that it is challenging to find the most suitable validation metrics for emulators which contain variability. Especially, because it heavily depends on the application in mind which validation metrics are even suitable to look at. For example, for researchers interested in impacts on regional scales, such as the ones caused by heat waves, it may be more important to reproduce local correlations than far-reaching teleconnections. Additionally, even though fully quantitative validation metrics can be very informative, one should not underestimate the importance of intuitive validation strategies, which can be easily understood by potential users that may have a less technical background. For example, demonstrating an emulator’s capability to produce visually consistent output with ESMs can be very helpful in communicating an emulator’s worth to a broader audience of potential users.