We would like to thank Referee 1 for their detailed and constructive comments. In the following to explain how we have modified the manuscript to address their suggestions.

**The original reviewer's suggestions are written in bold font** and our responses with normal font.

**General comments**

**I want to thank the authors for addressing many of my concerns and I commend them for more thoroughly investigating the LSTMs architecture and hyperparameters to try and get the LSTMs to work. I think the new discussion at the end of the paper is helpful for understanding why the LSTM's are fundamentally failing to improve over linear methods, i.e. the implied link is linear. I think that finding, and others from section 4.1, are important enough to warrant inclusion in the abstract since they're broader claims about non-linear methods for CFR that this paper found evidence for. Could the bi-LSTM then be seen as more of a tool for investigating the (non)amenability of CFR to non-linear methods in small data regimes, rather than just an "off the shelf" application that didn't pan out? If that understanding is correct, then I would consider emphasizing that a little more up front.**

A: We think that our original major research scope about this manuscript is for the purpose of testing and verifying our working hypothesis that whether a more complex machine-learning method LSTM would provide better reconstructions for temperature field. At the meantime, we noticed that in this context, we were using a limited amount of dataset to train and validate the neural network method, which could be, on the other side, an investigation of neural network performance as a tool for CFRs especially when only a limited data is available or employed compared to 'big data' and the usually big data-drive deep neural network aspect. We have add sentences into this context to mention this.

**Specific comments**

1. **Line 112-113 — "However, the prior use of information from climate models precludes a posterior critical comparison between simulations and reconstructions, and thereby the resulting reconstructions lose one appeal of climate reconstructions in general." — I do not understand what this is trying to say. Can you please explain this again? Actually I'm struggling to understand the overall reasoning behind DA not being applicable here. DA was compared to PCA in (Steiger et al., 2014) and they claimed their method was computationally efficient. Is there something fundamentally different about the approach presented here that makes DA not applicable**?

A: We wanted to raise two different points regarding DA methods. One is that the DA methods use a lot of information stemming from simulations with climate models. Therefore, the a posteriori comparison of DA-based reconstructions and climate simulations is compromised, and both data sets are not independent. The second point it is difficult to methodologically evaluate DA-methods against other purely statistical methods, since the former use much more information and data (from climate simulations), and thus the comparison cannot be fair.

We have reformulated this paragraph to make these two points more clear.

2. **Table 1 — Why do PCR and CCA improve in the noisy scenarios but LSTM deteriorates**?

A: We believe that maybe because these noise-contaminated data cause obvious overestimations in the amplitude of reconstructed variability for the linear PCA and CCA methods. Some noise signal may deteriorates the reconstructions, while these noise single may also lead to good reconstructions, since the CFR reconstructions are effected by many factors, such as the proxy numbers and its spatial distributions, random noise signal introduced and added to certain important spatial proxy locations which could have significant effect on the overall spatial reconstruction may result in a general better reconstructions. For the nonlinear machine learning methods, it is very sensitive to external noise. Kalapanidas et al., 2003 and Atla

A, et al., 2011, demonstrated that linear regression can perform better results than nonlinear methods considering noise sensitivity studies. And some studies indicated that external interference or noise could damage the ability of neural networks (Heaven 2019).

3. **Line 527: "PCR and CCA exhibit overestimated reconstructions within noisy PPEs, the Bi-LSTM presents relatively robust reconstructions" — What are overestimated reconstructions? Is this indicating some advantage of LSTM**?

A: The overestimated reconstruction here refers to the overestimation in the amplitude of variability. The overestimation is represented by the ratio of standard deviations, as a metric for assessing the reconstruction variance. A  SD ratio close to 1 indicates we achieve perfect reconstructions with the same amplitude as the target. In these noisy PPEs, the linear regression method we employed, especially the PCR method, exhibits obvious overestimations (the value of SD ratio is bigger than 1 over some spatial regions as shown in Fig 6-7). In principle, it is difficult for regression methods to reproduce perfect reconstructions (obtaining SD ratio equals 1).

The LSTM method shows relatively robust reconstructions within these noisy PPEs. However, as we explained in the above comments, neural network is also very sensitive in noisy experiments. In our CFR study, we employed a small sample size and add two type of noise to contaminate these original data. We would conclude that the CFR results based on neural network would be much more dependent on the different scenarios. Since several external factors, such as data set and noise type, and internal factors such as interpretability of neural network, would have significant on drawing a general conclusion about the final reconstructions. But based on our experiments, the LSTM architecture tested in our study seems to show some advantage in achieving reasonably robust reconstructions.

4. **Line 555: "Both ESN and LSTM belong to the family of RNN, yet ESN is much simpler than LSTM (Lukosevicius and Jaeger 2009), and has outperformed the RNN methods in other applications (Chattopadhyay et al., 2019; Nadiga, B. 2020)." — If ESN's are simpler and more promising, then why does this paper stick to the LSTM model, which clearly did not improve over simpler methods, and not just propose and evaluate ESN's, even if alongside the LSTM**?

A: The results regarding the LSTM have been collected along this study, and this experience lead us to think that the ESM could be more promising. This assumption is based on a few preliminary results, but not on a through testing. However, we cannot be sure at this stage that this will turn out to be correct.  Our plan is to test the ESM in a follow-up publication.

The ESN method we mentioned in this manuscript is because we have already implemented further CFR experiments by employing this ESN and also compared it with the LSTM method.

Refereces:

Kalapanidas E, Avouris N, Craciun M, Neagu D. (2003). Machine learning algorithms: a study on noise sensitivity. In Proc. 1st Balcan Conference in Informatics. pp. 356–365.

Atla A, et al. Sensitivity of different machine learning algorithms to noise. J Comput Sci Coll. 2011;26(5):96–103.

Heaven, D. Why deep-learning AIs are so easy to fool. Nature 574, 163–166 (2019).

We would like to thank Referee 2 for their detailed and constructive comments. In the following to explain how we have modified the manuscript to address their suggestions.

**The original reviewer's suggestions are written in bold font** and our responses with normal font.

**General comments**
**The authors have done a good job of revising the manuscript in response to my original review. The article is now more comprehensive, contextualized, and described. I support publication after the items I list below are addressed. As a side note, I do not point out many of the typos and grammatical errors, but the paper would benefit from detailed language editing**.
A: We have iterated the manuscript on correcting typos and grammatical errors.

**Specific Comments:**
**Ln. 14: field,.**
A: we have corrected this typo.

**Ln 117: My point about the AMV is that the observational AMV is not defined exclusively as the "decadal filtered surface temperature anomaly." It is not even necessarily the decadally filtered anomaly, but the index of average north Atlantic SSTs after removing the forced signal in that average (whether by removal of a linear trend or otherwise). I am not suggesting that the authors use a different definition to isolate the AMV in the longer last-millennium runs, but to better define the AMV in this location**.
A: As the reviewer points out, this is a terminology issue, the use of which is not quite clear through the literature. For instance, the GFSL site on the AMV defines the purely internal variability of the North Atlantic SST as Atlantic Multidecadal Oscillation (AMO) whereas the AMV will include natural and externally forced variability. Other authors indeed refer with AMV to the internal variability only. The new version is now more specific on this terminology.

**Ln 127-150: This is useful information and should be included in the Data and Methods sections. This is nevertheless a strange collection of information to include in the Intro. The authors may want to summarize some of this information as part of a roadmap in the Intro, but much of it should be incorporated into the Data and Methods sections (which does not include some of this important info, i.e. it is not just repeated here)**.
A: We agree with this comment, and have made changes correspondingly, moved the proxy and climate model information from Introduction to Data and Method section.

**Ln 172: The use of CESM1 and CAM5 is strangely garbled here and elsewhere in the manuscript. In all cases that the authors are talking about the CESM-LME results they should refer to CESM. CAM5 is the atmospheric model used in the CESM1 coupled model. Use of CESM1-CAM5 is strange (CESM1 by definition uses CAM5 in its architecture), while all of the figures identify the CESM-LME results as CAM5. This should all be remedied**.
A: We have corrected all the CESM1, CAM5 and CESM1-CAM5 into CESM in the text, and corrected all CAM/CAM5 caption into CESM in all figures.

**Ln 180: "We use ensemble member 13 from the CESM-LME as the basis for our CESM pseudoproxy experiments." Note also that LME is defined in this sentence but it is used in line 174 without definition**.

A: We have corrected this.

**Ln 378-9: This does indeed support the stationarity of the teleconnection patterns, but also says something about the physical nature of the patterns, i.e. they are to some degree localized and do not share significant amounts of covariance outside of the regions where they are sampled**.
A:

**Ln 409: This is only true if the EOFs in the training interval are stationary, i.e. well represent the EOF patterns in the reconstruction interval as well**.
A: In our manuscript, we have indicated that we assume the EOF patterns derived from training interval remain constant in time, which is stationary with time.

**Ln 424: It is not clear whether these assessments were done for EOF patterns over the reconstruction interval, the training interval, or both. Interpretation of the results depends on this choice, namely whether stationarity is part of what is being assessed. Reduced skill in the recon-interval EOFs could be both associated with deficiencies in the methods or non-stationarity in the EOF structure. This should be made more clear**.
A: We corrected this. Here in Line 424 the EOF patterns were derived from the reconstruction interval.

**Ln 434: The stated explanation of the Yun et al. methodological choice and its potential statistical artifacts is not clear here. The few sentences that start here should be more clearly articulated**.
A: We integrated an additional paragraph to articulate this more clearly.

**Ln 454: I believe Figure 8 should be Figure 11 here.**
A: We have corrected this typo.

**Ln 466: I believe Figure 11 should be Figure 12 here.**
A: We have corrected this typo

**Ln 482: There is a general and non-quantitative discussion that starts here about how one distribution describes the target distribution "better." This is generally vague language that begs to be quantified. The KS tests are the quantitative part of this discussion and are more sufficient for describing things as better or worse. I would combine the language, or just move directly to the KS tests as a means of characterizing how well the distributions compare**.
A: We think that a detailed text description of histogram may be necessary for people to better distinguish the capability of each different method on capturing the extremes – lower or upper tails intuitively. The KS statistic would be then cable of better describing the detailed differences of each reconstruction methods in a quantitatively way. Considering the advice of the reviewer, we have changed the ordering of these two paragraphs. Now the quantitative results derived from the KS tests are presented first, followed by our more qualitative description of the behavior of the distributions at their tails The reader encounters first the quantitative tests and then better appreciate the more subtle differences a the fringes of the distribution.

**Ln 527: There are multiple places in the discussion where the authors use overestimated reconstructions or similar constructs. I think what they mean is overestimated variance, which should be used to be precise**.
A: We have corrected the overestimated related reconstructions to overestimated variance.