Reviewer #1

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

The original comments are written in bold font, our responses with normal font

General comment

C1 The new method and the results are definitely interesting and worthy of publication, albeit I do not exactly see how the manuscript fits into the list of GMD manuscript types. It does not fit into my understanding of the scope for 'model description papers' because while it describes a new method, the implemented model still seems in a rather experimental stage and the README of the code in the Supplement explicitly states "The scripts are taylored to use the provided test data, i.e. they are not written in a general form that would allow to use them with any kind of suitable dataset, yet." This will make it very hard for readers to use the new method outside of the presented PPEs. If it is designed as a model description paper, at least the model must be given an explicit name and version number following the GMD guidelines and some more effort should be put into making the code usable for others. Since the paper develops a new method and is not related to model improvement it is also not directly a 'development and technical paper'. Therefore, I ask the authors to clarify how the manuscript fits into the GMD manuscript types and adapt it accordingly.

R1 This is to some extent a matter of perspective, and we assume that the editor has already perused the manuscript at the submission stage to assess its suitability to the journal. However, we agree to some extent with the reviewer and we will characterize the method more specifically according to its objectives and methodological steps, as also outlined below in response to the specific comments. Indeed the method cannot be totally universal, but it can certainly find applications in other areas of science and technology, where the objective may be to provide a complete field (e.g., image) from sparse information.

Browsing the journal, we find other manuscript that describe a methodological advance but that are not ripe for a general application. Those manuscripts are, as ours, refinements or combinations of statistical methodologies for a particular purpose. We therefore think that the manuscript does fit into the category of "development and technical paper".

2. Major issues

1. **Introduction: The paragraphs l. 27-79 describe the AMV-related research fairly extensively and in my opinion much longer than necessary for a model description/development paper. In contrast, the final part (l. 80-89), where the new method is introduced, is a bit short to help me understand the authors thought process in selecting and developing the described methods.**

We will rebalance the space devoted to the physical motivation and the methodological aspects, placing the model against a broader backdrop of estimation of time series from partial information.

2. C. **Sect. 2.2: The model description varies between long descriptions of general GPR theory / modeling options and fairly short parts on the selected solutions, the motivation behind these choices, and implementation details. I would prefer to focus more on the specific choices for reconstructing the AMVI and why these choices are made. For more specific questions arising from the method description see below.**

R. We will try to strike the right balance between the general GPR description and this particular application. GPR has not been used so far for climate reconstructions and the interested reader might

find it useful to have in this text an introduction about the basic method set-up. Therefore, we think this part, although shortened, can be useful for the reader. We will expand the justification for our particular choices. This justification is mainly based on the need to produce reconstructed time series with the correct amplitude of variability.

C 3. How is σn handled for the AMVI? Is the AMVI just given by f(z) or is ε also added in observed and reconstructed AMVI?

R. The target /reconstructed AMVI is indeed the mean of the GP-posterior, the 'best' estimation of the true value. A realization of sigma_n is not added to the posterior mean. We will include a sentence to clarify this point. Sigma_n is estimated during training. As we estimate only one sigma_n across all dimensions, it is the same for all proxy timeseries and the target AMVI. This is of course a simplification but the estimated hyperparameters (Fig.4) show that the estimated sigma_n corresponds to the mean noise across all records in most cases and is therefore a good first approximation.

C 4. Sect. 3.2 / 3.3: In the evaluation of the PPEs with noisy pseudo-proxies, the authors focus on the ensemble mean time series across all randomized experiments. I do not see why this is a useful quantity for evaluation. It is not a quantity occurring in reality since the authors correctly state that in real-world applications we only have one realization of pseudo-proxies. Therefore, the ensemble members should be evaluated separately (as each of them is a single realization which could occur in reality) and then the mean and spread of the evaluation measures should be reported and analyzed.

R. As per a comment by reviewer #2 We will present summaries of the probability distributions of the evaluation metrics for the different model set ups. This information is already available for the MPI-ESM based Tcnpp ensemble in the appendix (Fig.B1). Note that also the min-max range of the ensemble spread is already given for every ensemble metric throughout the text. We will modify the text to focus on the mean and spread of the metrics. This will however not change our conclusions about the respective performances (see Fig.B1).

C 5. As the authors state, GPR is a Bayesian method. Thus, it naturally produces uncertainty estimates through sampling of the posterior distribution. Currently, only the posterior mean is evaluated throughout the manuscript if I see it correctly. Uncertainty quantification is an important part of climate index reconstructions and has been the subject of intense debates over the last decades. Therefore, I would like to see some evaluation on how useful the uncertainty estimates provided by the posterior distribution are.

R The manuscript in its original form does address, albeit partially, the uncertainty in the estimation of important hyperparameters, such as the local noise represented by sigma_n, and the scale parameters in the kernel. This is represented in Figure 4. We acknowledge that the figure caption is not clear enough in this respect and we will state this point more clearly. The confidence intervals given in Fig. 2 and 3 correspond to the posterior uncertainty (2sigma of the posterior distribution). The displayed uncertainty ranges in Fig.3 are an average of the 2sigma ranges of the individual ensemble members.

We will also evaluate whether the true AMV index does lie within the 95%- uncertainty ranges in 95% of the time steps, or whether the GPR-derived uncertainty ranges too liberal or too conservative are. We thank the reviewer for this good point.

C 6. Sect. 3/4: While the authors report several situations where one or several of the reconstruction methods give unreasonable results or fail to reconstruct the underlying truth, explanations for why the models are better in some aspects and worse in others are mostly missing. The authors speculate

on some potential reasons but I would like to see some more sensitivity tests to give the reader a better feel for the strengths and weaknesses of the different methods.
R. We have tried in the original manuscript to interpret the results obtained with the different set-ups of the GPR model, but at some point its difficult to point out to particular reasons for their behaviour, as it happens in many other applications of machine learning. Perhaps the most relevant interpretation is why the different behaviour in the skill of the Sparse GPR and standard GPR with respect to the presence of noise in the proxies. One assumption, as suggested by the two reviewers, is that the number of proxy records and the presence of noise in the proxy records may affect the overfitting tendency of the model. we will test this hypothesis by conducting a targeted experiment with larger and smaller proxy networks.

Specific comments

l. 12: Please reformulate one of the 'relevants'
We believe the use of "relevant" to be fine in both cases.

● l. 17: The last 1000-2000 years are normally named the 'Common Era' and not the 'preindustrial period'. 'Preindustrial' could lead to confusion with simulations using fixed preindustrial boundary conditions. Focusing here on the Common Era as a 'must' seems a bit arbitrary since also other periods would be of interest.
R. We will change pre-industrial period to Common Era

● l. 58: Which non-linear methods have shown promise? Is non-linearity really the main advantage here or could other factors also be important?
R. Data assimilation methods based on k-nearest neighbor (Pfister et al., 2020 doi:10.5194/cp-16-663-2020). However, these require the use of climate simulations. Also random forest (Michel et al 2020, https://doi.org/10.5194/gmd-13-841-2020). We will include a brief disscussion of these methods in the introduction.

● l. 103-105: Which climate variable do you use to construct the pseudo-proxies (e.g. SST, surface temperature, near-surface air temperature)?
R. Surface-air temperature. This is the standard variable that for instance temperature sensitive tree-rings represent. As noted in line 100.

● l. 109-111: How do the SNR and the construction of pseudo-proxies compare to other pseudo-proxy studies?
The amount of local interannual noise in real proxies usually assessed by the correlation between the proxy time series and the instrumental time series. These correlations may be in the range 03 to 0.7. The amount of local noise used in other pseudo-proxy studies is within this range, as ours. We will include a sentence in the revised version.

● Figure 1a: From the color scale in (a), it is very difficult to distinguish the correlation of the different records. Maybe you can improve the color scale.
We chose this color scale to match the one used for the cross-correlation matrix in Fig.1c. But we agree, the respective values are not easily distinguished in this case. We will use a non-continuous color scale, so that the correlations can be more easily inferred by eye.

● Figure 1c-e: Over which period are the cross-correlation, STD radio, and embedding distance computed?
For the pseudoproxy records the entire 2000 years of the simulation, for all pairs including the AMVI only the most recent 150 years. We will add this information in Section 2.2.3.

● Figure 1d: The 100 could be removed.
Noted and it will be amended

● l. 132: Are Matern functions really 'very complex' kernels?
We will change to 'more complex kernels'

● l. 139-140: The inference strategy is described very briefly here. Some more explanation could be useful for readers not that familiar with Bayesian inference
We will add more explanation about inference

● l. 160: Is there a reason why the abbreviation SVGP is not adopted in the manuscript and 'sparse GP' is used instead?
This was done because we later only use "full and sparse *em*GP" when we refer to the GP in embedding space. In section 2.2.1 we will stick instead to SVGP. In the remainder of the manuscript we will stick to the full and sparse emGP nomenclature.

● l. 163: What is the 'Adam Optimiser'?
R. The Adam optimiser is an algorithm to search the best model parameters according to a prescribed cost function. It belongs to the family of stochastic gradient descent algorithms and it is widely used in machine learning applications due to is favorable properties. Essentially, it takes into account the mean and standard deviations of the gradient of the cost function in previous optimization iterations to propose a new value of the parameters. We will include this sentence to complement the reference.

● l. 171: How is the AMVI formulated in proxy space since it is defined over a different spatial scale than the individual proxies and how does GPR-based climate index reconstruction work when the GP is formulated as function of the proxy values (=temperature?)?
R. We interpret 'spatial scale' in this comment as 'amplitude of variability' In this particular application, the pseudo-proxies and the AMV index are all defined as near-surface air temperature, so that the range of variability is roughly similar for all of them. The reviewer is right that this condition cannot be generalized and for other reconstructions the proxies and target time series would need to be standarized to unit variance. We will include a brief explanation in this regard.
.

● l. 177-192: I struggled to understand why the embedding space needs to have the given dimension and the idea of how the embedding space is constructed did not become clear to me until I finished reading Sect. 2.2.3.
R. Indeed the new set-up of the GPR is not easy to visualize. We will reformulate this important part of the manuscript more clearly. The number of necessary dimensions is most easily conceived from the easiest setup with equal distances between all possible pairs of records. Imagine a hypothetical case with four records (e.g. three proxy records and one AMVI). To be able to arrange all four records such that there is an equal distance between all respective pairs of records, one needs a three dimensional

embedding space. Thus for q timeseries, the embedding space must have a dimension of q-1. Time then is added as an additional dimension.

● l. 203-205: What are the properties that need to be fulfilled by the distance matrix / distance measure? Why is a positive correlation between records needed between the records?
R. In our set-up, the distance matrix does not require any specific properties, other than that the distance should be a monotonous function of the 'similarity' between time series and symmetric. In our case we have chosen the correlation between the time series as a measure of distance, augmented by the amplitude of variations, both combined so that the resulting metric is symmetric. We think this is a reasonable choice.

● l. 209: I guess that equidistant coordinates perform worse because than all records influence the AMVI roughly equally whereas for CC-based coordinates records with a high correlation with the AMVI become more important. But why does including the SR improve the result compared to just using CC?
R. The reviewer is correct about the case with the equidistant coordinates, When performing the reconstructions only with CC-based coordinates, we found that the reconstruction is dominated by northern hemisphere records which have a much larger range of variability. This became especially important for networks with realistic proxy availability. Further back in time, only NH records are available and the reconstructed variability for earlier periods was strongly overestimated in the CC-based case. The consideration of the SR in the metric ensures that records that may have larger variability are considered farther away from the target, this improved the magnitude if the reconstructed variability. An alternative approach would be the normalization of all records. In that case, CC-based coordinates would be sufficient.

● Equation (3): Is RA = SR?
R Yes, it will be corrected

● l. 213-218: The description was a bit short here for me to really understand what is happening and why.
R. The normalisation of the time axis is done to ensure that the variations along the time axis are comparable to the variations along the other embedding dimensions. This is a necessary step due to how the kernel k2 is formulated, distance in time and embeddin pace must comparable, to allow for interactions across tim and records. Without the normalisation, either the distance between records or the timescale would dominate the lengthscale of k2. If the kernel would be formulated such that time and embedding dimensions were treated separately, this normalisation would likely not be necessary. But as stated in the manuscript, the current kernel formulation outperforms the separated kernel. We will explain this point more clearly.

● Equations (4/5): The Gaussian kernel functions lead to very smooth (infinitely differentiable) functions, likely much smoother than most processes actually observed in climatology. Does this lead to overly smooth predictions on certain timescales and did you test the procedure with kernels that lead to less smooth posteriors? It also might be useful to write down the final covariance model as an equation.
R. The temporal component of Gaussian kernel leads to predictions that are indeed smooth in time, as it acts as a low-pass filter on the predictor time series. However, this does not mean that the model is

not able to represent rapid changes of the target variable if the proxy records also change rapidly. The case that the reviewer raising - a non differentiable temporal behavior- is in our opinion extremely rare. We will also complement equations 4 and 5 with a full expression of the covariance model.

• l. 230: Is σn 2 the same as a nugget effect in statistical modeling? If so, it might make sense to mention it here for the statistically-inclined readers.
R. Yes, we thank the review for this suggestion.

• l. 235/236: Do you have an explanation for why this slightly unusual formulation (two kernels acting in time but only one kernel acts in the "embedding space/distance") performs better than models without k1 or with separated kernels acting in time and embedding distance? Does this indicate that the system (AMVI) is better described by two characteristic timescales instead of one similar to two-box energy balance models outperforming one-box energy balance models in predicting sea surface temperatures?
R The rationale is to allow for different time scales in the autocorrelation (k1) and in the cross-correlation between proxy records (k2). Also, the lengthscale of k2 cannot be interpreted as a pure timescale as it discribes the typical lengthscale for the influence of the respective records across time and embedding space. We will explain this point more clearly. We think that this is not related with the issue raised by the reviewer, as our model does not include any type of forcing or reservoirs for the Atlantic Ocean. If we had included in the predictors a forcing time series, then the interpretation of the time scale would indeed be closely related to a thermal inertial timescale of the ocean.

• Equations (6) - (9) did not help my understanding. Either they should be embedded better in the text / explained better or they could be removed.
R These equations were meant to better explain the embedding process, but perhaps they should be better introduced. We will also formulate these equations in symbolic terms as product of matrix and vectors.

• l. 259: There is a typo in 'obtain'
R Noted

• l. 299/300 (and similar parts in subsequent sections): The difference between the MPI- ESM- and CCSM4-based results could be explored a bit further. What are the main differences between the simulations that might explain the differing behavior of the reconstruction methods?
R. We suspect that the main reason must be the different spatial correlations within the temperature field. We will explore this in terms of the spatial variability modes (e.g. EOFs)

• l. 313/314: Is there an explanation for why the sparse GP performs better for noisy than for perfect pseudo-proxies on multi-decadal timescales?
R. This is an interesting question (also posed by reviewer #2) , but it is not easy to address. We assume that the sparse GP is less impacted by local noise when using only a portion of the data available for the training, and thus it can behave differently considering overfitting. As mentioned before, we will include a targeted pair of experiments with a large and asmall proxy network.

• Fig. 4: This is an interesting figure to explain some of the differences between methods displayed in Fig. 2/3 but unless I missed something, it is barely discussed in the text.

R. Actually large parts of the discussion refers to Fig 4 but we will make sure to discuss Fig.4 even more deeply.

• l. 359: Could the overestimation of variability in periods with few available proxy records be explained by relying too strongly on a small number of proxy records which tend to be more variable than the AMVI due to integrating over a smaller spatial scale? This could maybe be tested by comparing hyperparameters fitted separately for periods with high and low record availability.

R. This is certainly a reasonable explanation. In fact, dendroclimatologists apply a statistical tool specially designed for this purpose called 'variance stabilization'. Our objective is, however, to compare the methods in different situations , i.e. is method A or B closer to the truth, and not so much the correction of the effect of sparser proxy networks. That would be a different methodological paper.

In our specific case, however, we think that the suggestion of the reviewer can only be part of the explanation. On the one hand we try to amend for the effect of few records wtwithh large variability through the additional SR scaling in the distance matrix. On the other hand, if this were the only explanation, we would expect the effect to get stronger the further we go back in time as the number of records decreases further. An additional explanation could also be non-stationary cross-correlations between records, this is however difficult to test for. Nonetheless, as we are planning to conduct experiments with also a smaller number of records in response to other comments regarding overfitting, we will also compare hyperparameters for the different data availability, as suggested by the reviewer.

• l. 393-398: Is the introduction of a 'noise variance' parameter similar to error-in-variables approaches for frequentist regression models?

R. Yes, it is conceptually similar in the sense that uncertainty in the 'predictors' is also taken into consideration. The EIV model, however, requires the knowledge of the ratio of noise in the independent and dependent variables, whereas here the noise variance is estimated along with with hyperparameters, We will refer to the EIV model to make this link explicit to the reader.

• l. 408: Is the model really using one tenth of the available training data if you use every tenth time step but also an (optimized) subset of the original locations?

During the entire optimisation/fitting procedure, the model always uses a slightly different subset in every optimisation step, so in a sense it uses more than a tenth of the data~~probably uses more than that~~. But the resulting co-variance matrix is based on an optimised subset which always corresponds to a tenth of the data.

• l. 416-418: Can you expand on these length scales and magnitudes? What are expected values and where do the parameters rank in the range of reasonable values?

R. We will expand the discussion on the value of the hyperparameters, as per comment on Figure 4

• l. 444: The GPR-model seems to handle white noise proxies very well, in parts due to the inclusion of the parameter σn . What would happen if the proxy noise would be auto-correlated? Is there a way to adapt the model accordingly?

R. We are not aware of an application of of GPR for the case the reviewer is suggesting, but the GPR model could be modified to account for temporal autocorrelation in sigma_n. Actually, it seems that the

GPR model would need to be cast in an embedding scape similar to the one implemented here (explicitly including the temporal dimension). This augmented model could be even expanded to account for spatial and temporal correlation of the sigma_n - in that case sigma_n would not be just a a random variable but a gaussian random field itself, described by three additional hyperparameters (the local noise, one temporal decorrelation and one spatial decorrelation). The calibration would become certainly more complex.
The case of spatial correlation of the proxy noise would be somewhat special, and not that common for paleoclimate reconstructions, but reasonable for certain type of proxies, so we will briefly also discuss this possibility in the revised version.

● Conclusions: Since this paper develops and tests a new methods for climate index reconstructions, it would be very useful for the reader to get some more guidelines for future applications of the method and how it might be applied to other indices.
R. The current GPR model is applicable with slight modifications in a broad set of situations in paleoclimate, in which just one index is reconstructed. The reviewer has hinted in other comments at other possible situations in which the GPR method does require a more complex model. For instance, in the case of reconstructions of a spatially resolved field or with autocorrrelated noise terms. There are still others that we have not mentioned, like the use of proxy records of different nature - e.g. tree-rings and lake sediments, which display different statistical properties.
We will expand the discussion and the conclusions to present those cases.

Reviewer #2

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

The original comments are written in bold font, our responses with normal font

C The quality of the paper is generally high, the analysis appears accurate from my knowledge, and there are no major points which I think should prevent its publication. There are however several minor points and suggestions which I identified and reported below, which I think could improve the paper further.

R. We thank the reviewer for the general positive assessment

Lines 33-34: "however" is awkward in the middle of the negation.

Noted

C Line 134: Might be useful to describe the term 'hyperparameter' for those outside the machine learning field, how is it different from a normal 'parameter'?

R. The reviewer is right. This notation is usual in the Gaussian Process literature, but confusing outside the machine learning community

C.Lines 148: So the batch size is the total number of observations across records (i.e. 5 records with 100 observations plus 2 records with 200 observations would mean a batch size of 900)?

R. The batch size corresponds to number of training observations given to the algorithm in. In our case it is defined as suggested by the reviewer.

C. Lines 180-183: How are the irregular resolution of the proxies handled (for the realistic P2k case)? Or are the pseudo-proxies created at annual resolution? In which case it would be helpful to briefly hint in the discussion how realistic irregular proxies could be used in the future and how it might affect the results.

R. The case considered here is when all proxies are annually resolved, and all have similar statistical properties (differing only in their amplitude of variability). the case of proxies with different statistical properties is much harder to handle with a GPR model. Assuming that the resolution is known, it would need to be incorporated in the Gaussian process prior, and the inference of the posterior will not be given by the standard equations any more. In this case, it seems to us that the more immediate approach would be to subsample all records according to the one with least resolution, although much information would be lost. This is the standard approach for long time scale paleoreconstructions. Other possibility is to set up two GPR models, one for the records with high-frequency resolution, applying a high-pass filter to those) and one GPR model for the records with low-frequency resolution (low-pass filtering the records with high resolution). This approach is sometimes also used in paleoclimatology, but one cveat is the difficulty to calibrate a model for low-resolution records, since the observational period is too short. We believe that this difficulty is common to all methods, and not a unique faeture of GPR.

We will include this point in the discussion.

C. Equation 3: Should RA be SR?

Yes, it will be corrected.

C. Lines 248-249: So the embedding coordinates are calculated from the distance matrix via multidimensional scaling? Might not be obvious for the layman (including me) how the matrix is obtained, could it be made more explicit in an Appendix?
Yes, the reviewer is correct. We will explain this point more clearly.

C Line 259: "obatain"
R Noted

C Equation 6,7,8,9: Shouldn't the matrix be equal to something for an equation? Make explicit which one is the input and which the output.
R . Yes, we will include a version of this equation in a more symbolic form

C Lines 299-300: Could it be because CCSM4 is more homogeneous (less spatial degrees of freedom) since the mean embedding distance between the records is smaller? Just a thought.
R. This is certainly a possibility. As per a comment by the reviewer one, we will explore the degrees of freedom in both models , for instance by an EOF analysis., and look at the structure and length scales of the resulting spatial patterns.

C Figure 2: Unclear to me what the 95% confidence intervals represent in the temporal domain? Are they the spread of the unsmoothed data? Also for the spectra it is not explained what the confidence intervals are, simply chi-square CI with 2 degrees of freedom? It is a question of style and not necessary, but I personally like smoothing the spectra to make for a clearer comparison (e.g. Using a Gaussian smoothing kernel as in JW Kirchner, Aliasing in 1/f(α) noise spectra: Origins, consequences, and remedies. Phys Rev E Stat Nonlin Soft Matter Phys 71, 066110; 2005).
R We will explain this point more clearly. The uncertainty ranges correspond to the 2 sigma range of the posterior GP distribution.. The ranges displayed in the Figure are an average over all the ensemble members. The Spectral uncertainty given for the target spectrum and the spectrum of the mean does indeed correspond to the chi-square CI.

C. Caption Figure 2 and 3: "powerspectra" -> "power spectra"
R Noted

C Figure 3: I don't understand the indicated 95% CI number. Do those correspond to the same CI shown on Figure 2?
R Yes, we will make it explicit

C Section 3.2: I would generally favour calculating the statistics for individual ensemble members and reporting the mean+/- standard deviation rather than calculating them with respect to the ensemble mean. Similarly, I would show the mean of the spectra rather than the spectrum of the mean for Figure 3 b,d,f as it is more representative of the real result one would obtain.
R. This information is partly already available for the MPI-ESM based Tcnpp ensemble in the appendix (Fig.B1). Note that also the min-max range of the ensemble spread is already given for every ensemble metric throughout the text. We will modify the text to focus on the mean and spread of the metrics. This will however not change our conclusions about the respective performances (see Fig.B1). We will also modify the spectrum figures to show the mean of the spectra rather than the spectrum of the mean.

C Figure 3 b,d,f: I wonder whether the PCR has the right high-frequency amplitude for the

right reason? Are the high-frequencies just noise and thus PCR doesn't perform better than the other methods or are they actually correlated with the real series?
R This a behaviour that has been found in other previous analysis. Yes, the PCR reconstructions are indeed correlated with the target at high frequencies.

C Lines 329-331: Do the authors have an idea why the sparse can outperform the full GP? Could it be a case of overfitting when noise is present? Such that the sparse one is less sensitive to overfitting?
R This is an interesting question (also posed by reviewer #1). The suggestion by the reviewer is a reasonable explanation, As mentioned in previous comments, we will perform a pair of targeted experiments with a large and a small proxy network to test this hypothesis.

C Lines 335-339: How are data resolution handled? If there is 5 years resolution, then are there gaps between the years or are the values interpolated? Or are the annual data used and only clipped at the end of the record?
R The latter case. In the manuscript only proxies with annual resolution are created and clipped at the end of the record.. This is the standard case for past past millennium reconstructions. The case with proxies with different temporal resolution is much more difficult to handle (see previous comment). We will include a discussion on this case, but we will not be able to find an explicit solution.

C Line 387: I would remind the reader for the discussion what AMV-relevant timescales are. Maybe write in parenthesis something like (decadal to multi-decadal).
R. Noted

C One issue I would like better discussed is the loss of variance on longer than centennial timescales in the sparse emGPR for the full 2k run. To me this is quite an important limitation since I don't think it makes sense to restrict AMV-relevant timescales to decadal to multi-decadal; there is a continuum of processes and I don't think there are reasons to believe that it would flatten out on longer than centennial timescales or be related to a separate non-AMV relevant process right?
R. The length of the available simulations is 1000-2000 years, so that it is difficult to assess the behavior on multicentennial timescales - the degrees of freedom is considerably reduced. For instance, the correlation between reconstruction and target woud be estimated using, say, 10 degrees of freedom, assuming that centuries are independent samples.

C Line 465-466: Looking forward to seeing how it compares to the traditional PCR method!
R. We are also curious to see how it will compare to existing reconstructions. This is to be the focus of a follow up study.