Interactive comment on "Identifying teleconnections and multidecadal variability of East Asian surface temperature during the last millennium in CMIP5 simulations" by Satyaban B. Ratna et al.

Anonymous Referee #2

General Comments: Ratna et al. examine the influence of transient external forcing (volcanic eruptions) on PDO and AMO variability and teleconnection patterns as they relate to East Asian surface air temperatures (SAT) in three PMIP3/CMIP5 past1000 simulations and paleoclimate reconstructions. This is an interesting study, and the results have interesting implications for how external forcing can impact internal variability and teleconnections. However, more work is needed to compare model output to observations and expand the study to other models.
Reply: We are grateful to the referee for their careful review and that they consider our work to be of interest. We respond below to the suggestions for expanding the scope of the work.

Main Concerns:
1) There are at least ten CMIP5/PMIP3 past1000 (Last Millennium) simulations available on ESGF that span the 850-1849 CE time period (BCC, CCSM4, CSIRO, FGOALS, GISS, Had, IPSL, MIROC, MPI, MRI). The authors exclude several of these simulations (MIROC, FGOALS, GISS) due to spin up/model drift/trend issues and cite Atwood et al for why they exclude these simulations. However, the authors choose not to use the output from CSIRO, HadCM3, IPSL, or MRI (some of which are included in the analysis of Atwood et al). The results therefore seem incomplete and selectively presented- why the exclusion of these other simulations? Please include analyses of these other Last Millennium simulations or at least provide a reason for why these other Last Millennium simulations have been excluded (the data have been available for at least 8-12 months online, so I hope it's not a data availability issue?). As the manuscript is currently written, 1/3 of the models show a completely different result, but this is only one model- is this really 1/3 of all CMIP5 Last Millennium models, or just one outlier in the CMIP5 Last Millennium simulations?
Reply: We originally considered al six models that had CMIP5/PMIP3 Last Millennium and CMIP5 historical simulations and that had data for all the necessary variables in the UK JASMIN facilities. We then discarded three due to drift issues as explained in our manuscript. Following the referee's recommendation, we sought data for the additional models suggested from other parts of the ESGF and we have obtained sufficient data to extend our study to another three CMIP5 models (HadCM3, MRI and IPSL) and we are downloading the necessary data to include CSIRO-Mk3L-1-2 as well. We have already applied many parts of our analysis to these additional models and have included the results in updated figures in this reply (Figs. R1-R4). We agree that the manuscript would be strengthened by revising it to include these additional model results. Some of the results are similar, but there are some differences in the correlations with E Asian temperature that we would discuss in a revised manuscript and we are happy to make these changes.

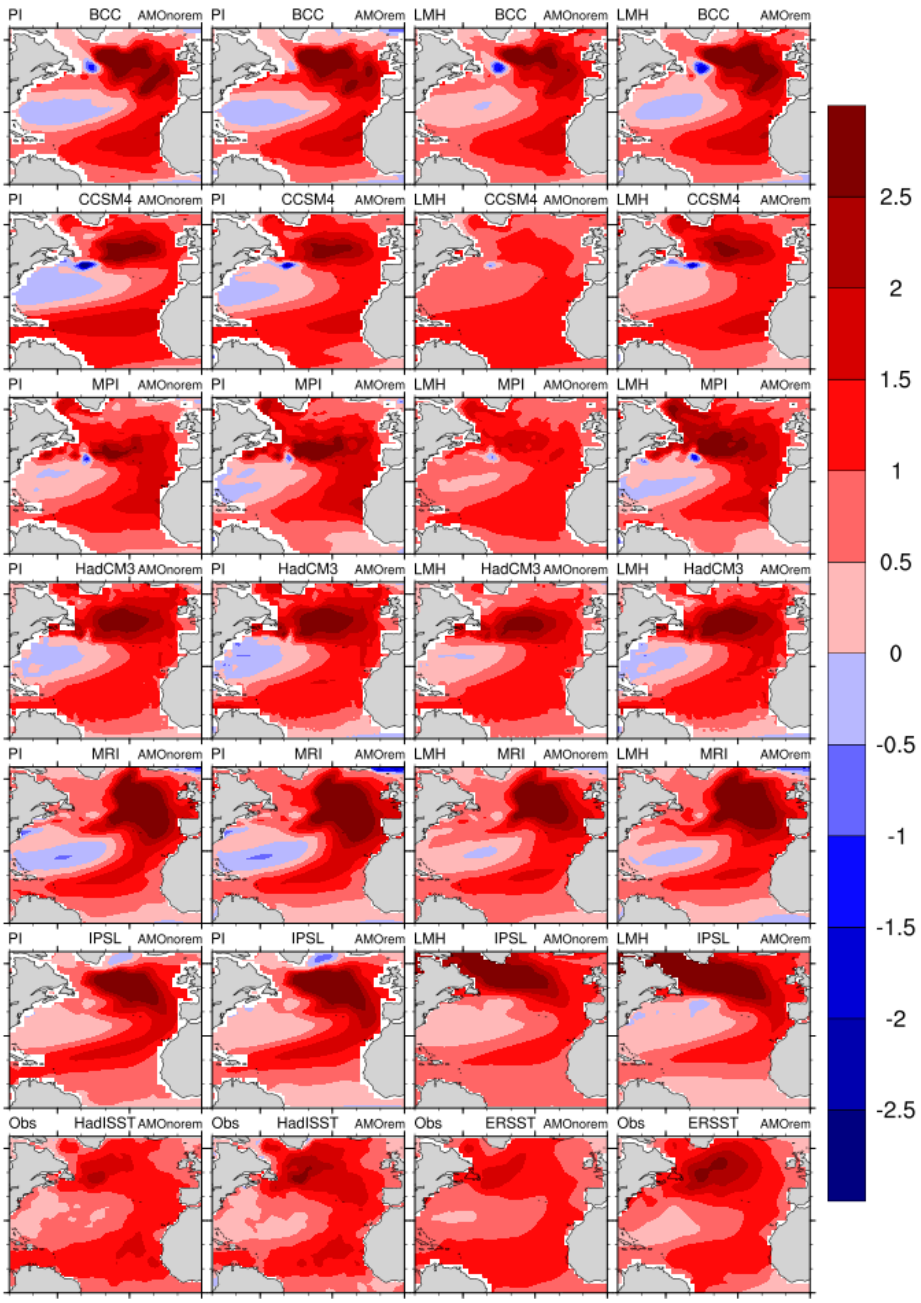

Fig. R1 – AMO SST patterns defined by regression of SST on the AMO index for each GCM including the three additional models now analysed, and for the observations (bottom row). Columns show results for different simulations (with/without forcings) and two definitions of the AMO index.

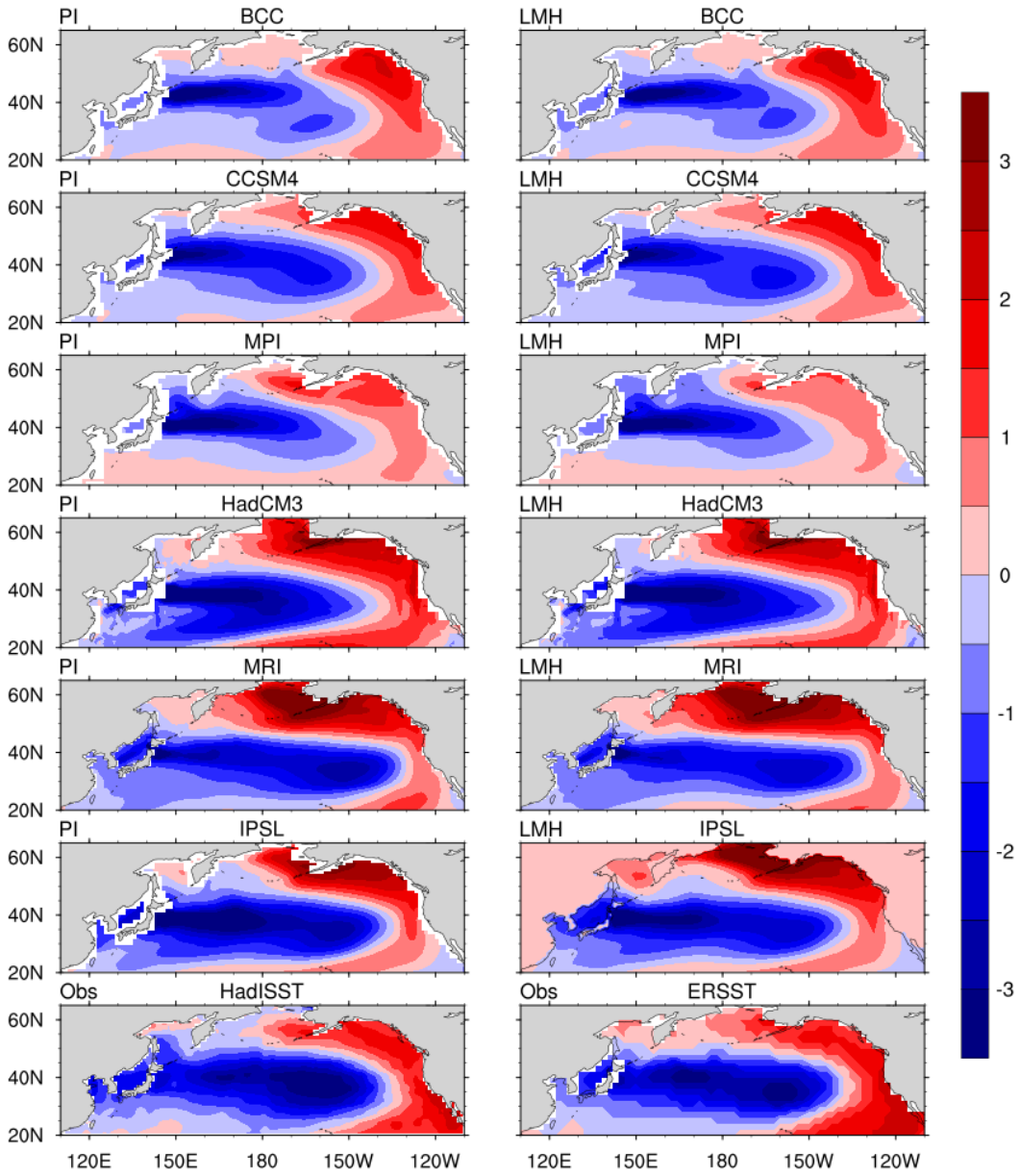

Fig. R2 – PDO SST patterns defined by the first EOF of SST for each GCM including the three additional models now analysed, and for the observations (bottom row). Columns show results for different simulations (with/without forcings).

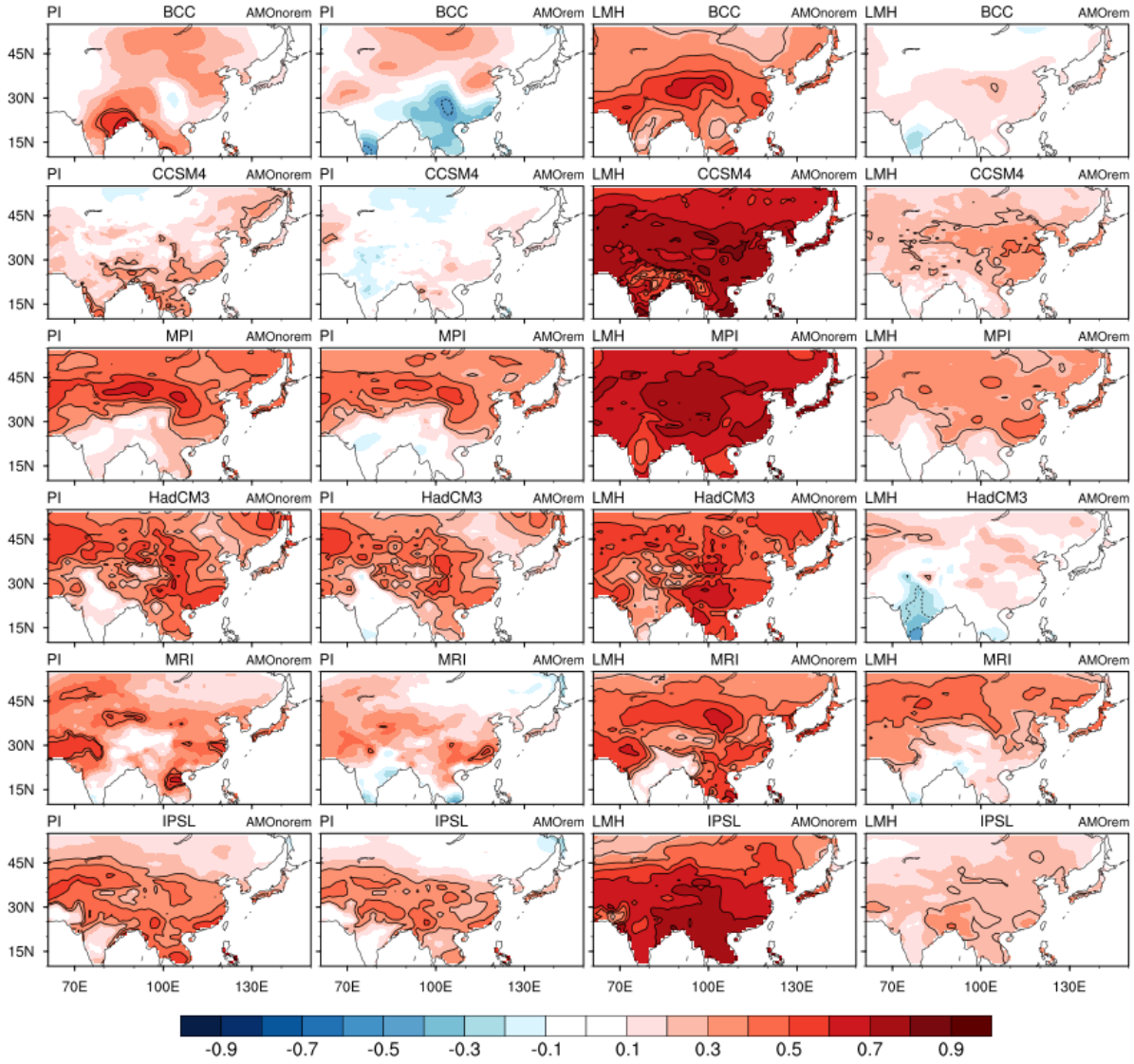

Fig. R3 – Correlations between E Asian temperatures and AMO index for each GCM including the three additional models now analysed. Columns show results for different simulations (with/without forcings) and two definitions of the AMO index.

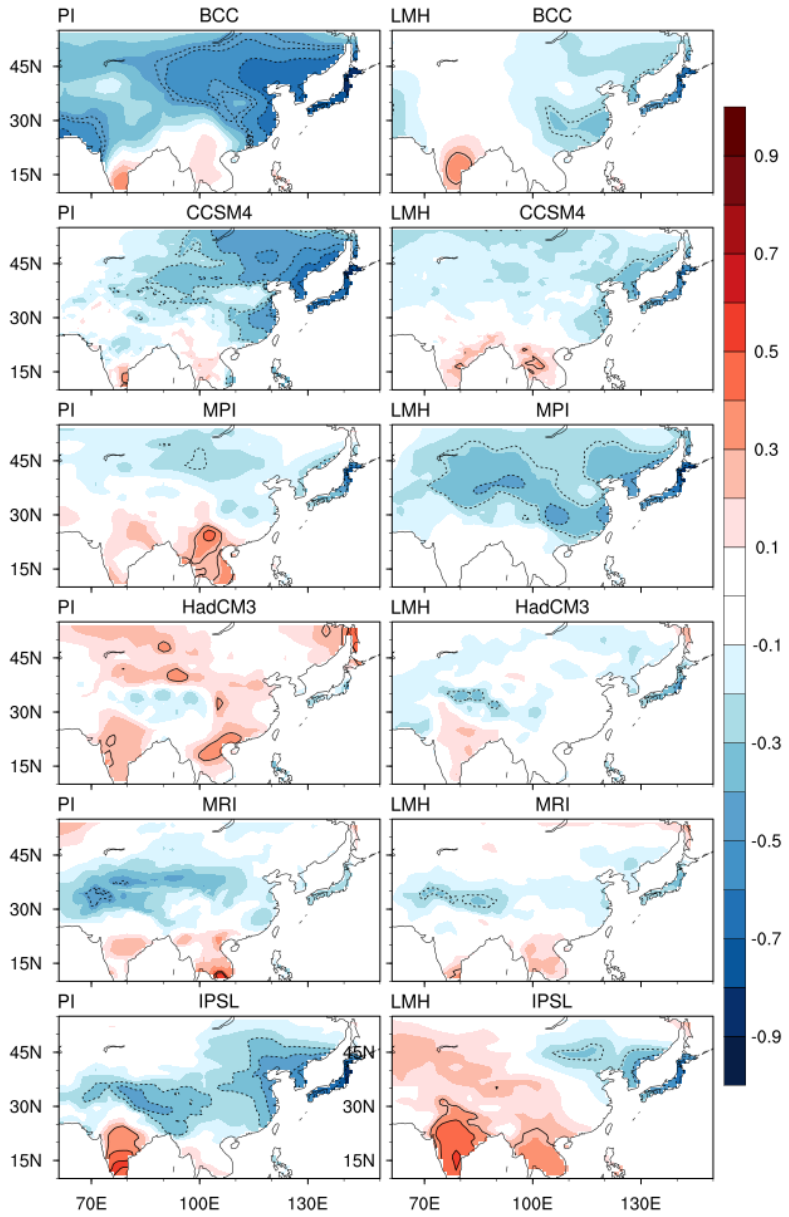

Fig. R4 – Correlations between E Asian temperatures and PDO index for each GCM including the three additional models now analysed. Columns show results for different simulations (with/without forcings).

2) The authors concatenate the Last Millennium (_850-1849CE) and the Historical simulations (_1850-2005CE) after removing the linear trend from each of these time segments separately. Removing a linear trend from either instrumental or CMIP5 data over the entire 1850-2005CE time period can be problematic if the main component of the 'warming trend' is in the 20th century. Multiple papers choose to remove the linear trend over the 20th century only (e.g., Deser et al., 2010; Messie and Chavez, 2011; Franzke, 2014, Nature Climate Change; Ji et al., Nature Climate Change, 2014). Similarly, many CMIP5 historical simulations appear to show much of the global warming trend starting in the 20th century, so removing a trend over the full historical simulation period (1850-2005) may add in decadal-centennial variability. To avoid this detrending and concatenation problem, could the analysis just be conducted over the 850-1849CE time period (especially because it seems the authors are mostly focused on the impacts of volcanic eruptions on the PDO and AMO in the pre-1850CE time period?). Some recent work even suggests that the dynamics of the system change once GHG forcing becomes dominant (e.g., Song and Yu, 2015, J Clim; Brown et al., 2017, Nature Climate Change), so including this time period could be arguably problematic.

Reply: We recognise the referee's concerns about linearly detrending the Historical simulations, but our findings are not sensitive to this choice.

We note that we are trying to replicate in the models some aspects of what other studies have done using observations (proxy-based reconstructions and/or instrumental) and (a) in some cases linear detrending over the instrumental era is done even though it may not be optimal for the reasons given by the referee; and (b) the timing of the start of the anthropogenic warming can be in conflict with a linear detrending that begins 1850 but this would be ameliorated for the PDO and for the AMOrem indices by the prior removal of global-mean SST from the Atlantic or Pacific SST values.

Nevertheless, we have tested to see whether our findings are sensitive to this issue by restricting the analysis to only the Last Millennium simulation and found that our results are quite similar to those we obtained by the combined detrended LM plus detrended historical simulations (compare columns of Fig. R5 and R6 for correlations with E Asia temperature for AMO and PDO, respectively).

We would report this sensitivity test in the revised manuscript and discuss the few small differences that do occur. We suggest to keep the main results based on the combined LM+Historical analysis because the benefits of having a longer series to analyse outweighs the concerns raised now that we have shown that our findings are not sensitive to this issue.

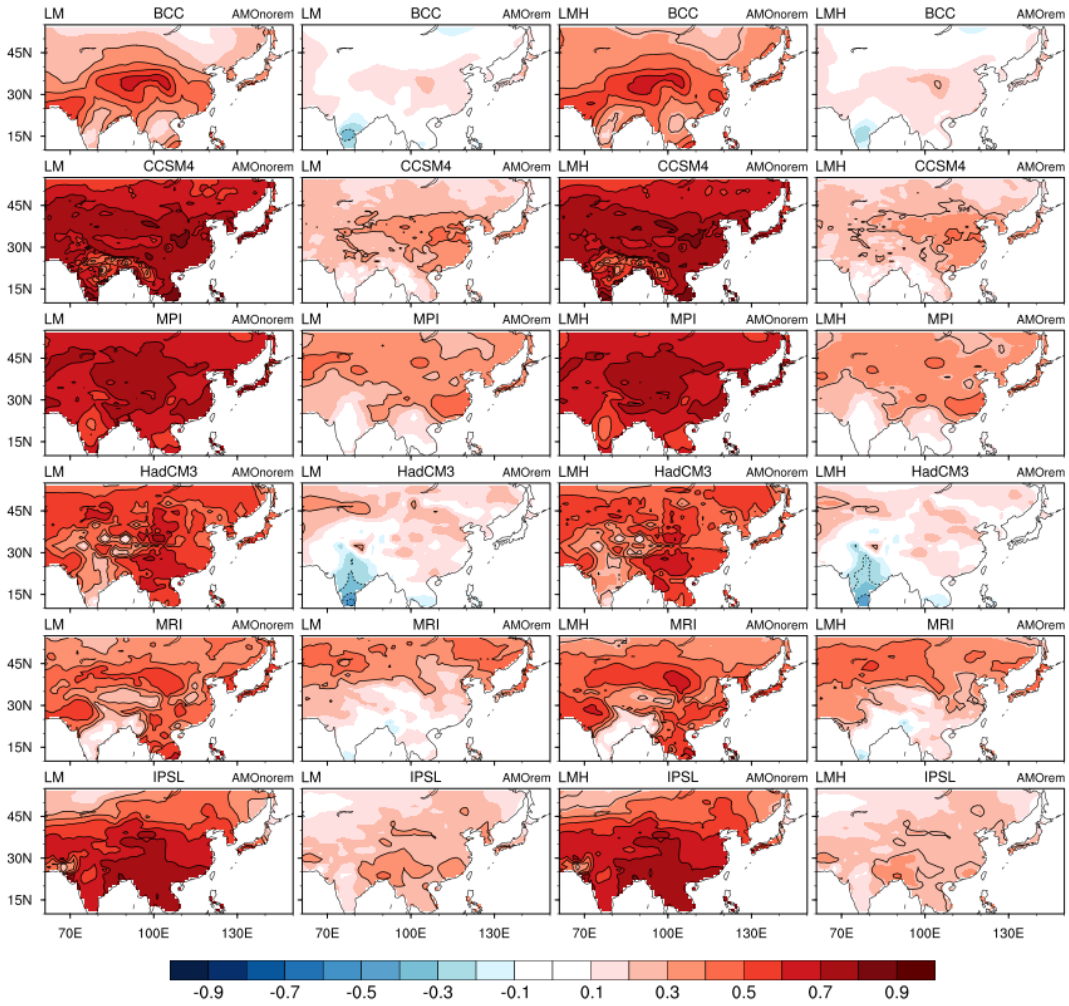

Fig. R5 – As Fig. R3 but for the comparison between Last Millennium (850-1850) simulation (columns 1 and 2) and combined Last Millennium plus historical (850-2000) simulations (columns 3 and 4).

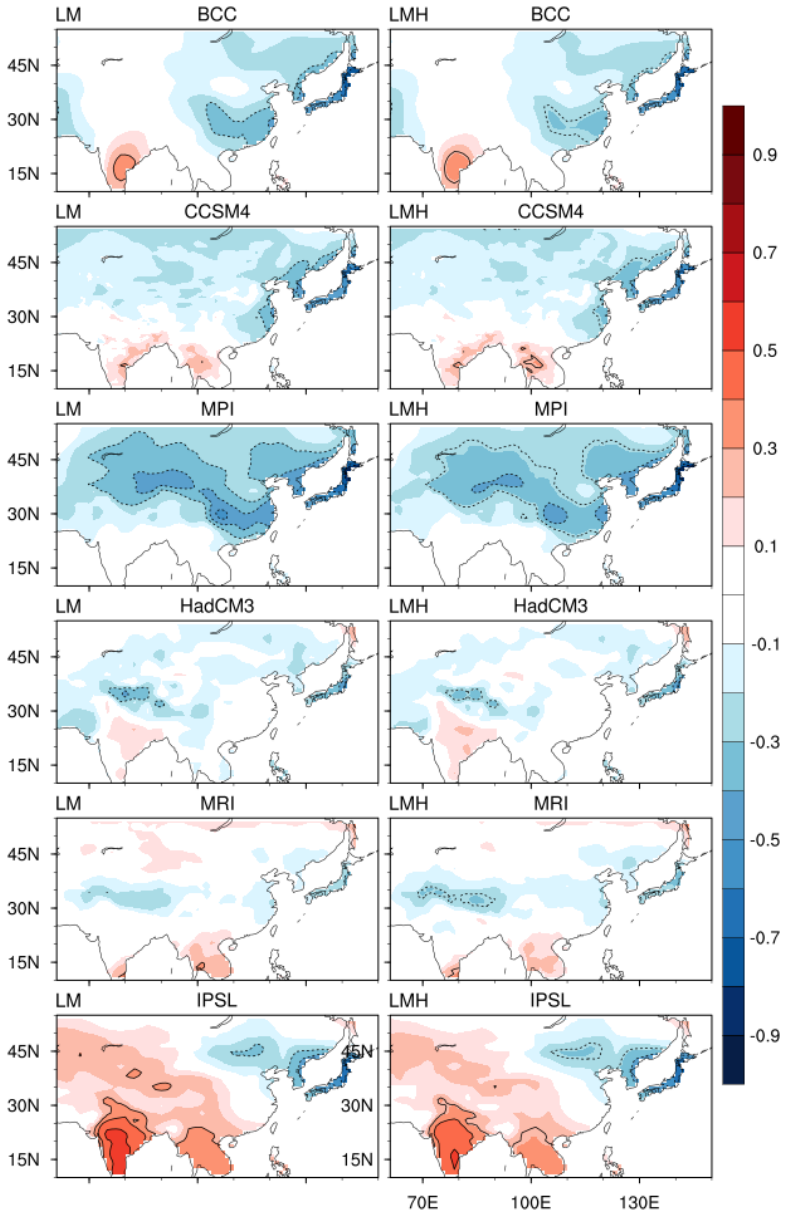

Fig. R6 – As Fig. R4 but for the comparison between Last Millennium (850-1850) simulation (column 1) and combined Last Millennium plus historical (850-2000) simulations (column 2).

3) There is no comparison between the spatial patterns of the AMO and PDO in instrumental-based reconstructions and the three models used here- perhaps some of these simulated spatial patterns are more realistic than others? The authors state that the model results are realistic, but never show this in the manuscript. The Climate Variability Diagnostics Package (http://webext.cgd.ucar.edu/Multi-Case/CVDP_ex/CMIP5-Historical/) shows that the spatial expressions of the AMO (and PDO) can be quite different in the various CMIP5 Historical simulations. Interpretation of the model results may be viewed through a more informed perspective if the models are compared to instrumental-based observations.
Reply: We will include a comparison of the AMO and PDO patterns with the instrumental data (similar comment from referee 1), see Fig. R1 and R2 above.

4) Varying significance levels are used in the paper (90% vs 95%). Please use a consistent 95% or 99% significance level- as the paper stands, it appears that the significance level has been lowered to show 'significance spectral peaks' (e.g., Fig 10), but the spectra barely surpass this 90% level- why not use 95% or 99% everywhere? At least please include some discussion of the sensitivity of the results to significance level if the results don't pass this higher threshold (significance levels are admittedly is arbitrary, but the current, inconsistent use of 90% runs the risk of appearing selectively low to attempt to present a 'significant' result).
Reply: Significance tests are reported for three types of analysis in the manuscript. (1) For correlations between area-averaged temperature and driving factors (bar charts) we used the 'standard' 95% level. (2) For correlations between temperature fields and driving factors (contoured maps) we lowered this to the 90% level because the additional noise at the grid cell level increases the risk of a type II error (wrongly failing to reject the null hypothesis that there is no correlation). (3) For power spectra of AMO and PDO, we used a 90% level, but actually our interest is not really in the significance of the individual spectral peaks (and whether they pass an arbitrary level or not) but in the overall shape of the spectra, their redness and broad multi-decadal power, and whether these are similar between models, with/without forcing, and between AMO index definitions. We will explain this better in a revised manuscript, we will also either remove the significance lines from the spectra or change them to 95%, and justify the use of 90% for the fields as explained above.

General minor issues:
Many authors abbreviate pre-industrial Control as PI (e.g., Atwood et al., 2016, J Clim, among others)- in an effort to maintain some sort of standard abbreviation that may be quickly recognized, I would encourage the authors to employ more commonly used acronyms (e.g., PI or piControl).
Reply: We will modify the manuscript to use the abbreviation PI for pre-industrial Control.

Also, when reading through the figures, it is difficult to interpret the acronyms used in each figure without searching through the other figure captions or the text for the definitions of the acronyms- please define the acronyms used in each figure in each figure caption (or at least reference where they are defined) so readers can quickly understand the figure without searching for what they mean.
Reply: Now we will define the acronym for each Figure captions.

Specific comments:
Page 1, Line 12-13: The simulated PDO and AMO spectra and spatial patterns are never compared to instrumental-based patterns or spectra (or even to proxy-based spatial patterns). Please include figures/analysis that support this statement in the main text or remove it.
Reply: We will include a comparison of the AMO and PDO patterns with the instrumental data , see Fig. R1 and R2 above.

Page 2, _line 10: The previous paragraph critiques the instrumental and proxy-based records, but little attention is paid to potential model deficiencies- can you at least briefly discuss or cite a few papers that may critique or even acknowledge that CMIP5/PMIP3 models have their own biases and problems as they relate to low-frequency SAT variability (e.g., Laepple and Huybers, 2014; Parsons et al., 2017 J Clim;) or 'modes' of internal variability (or their responses to stratospheric aerosol loading from volcanic eruptions)? Alternatively, directing the reader to where these model deficiencies, and their implications for your results, are going to be discussed later in the paper could be helpful.

Reply: we will cite these references and add a few sentences to describe their implications for potential model deficiencies. We will discuss that Laepple and Huybers (2014) found potential deficiencies in CMIP5 SST variability, with model simulations diverging from a multiproxy estimate of SST variability (that is consistent between proxy types and with instrumental estimates) toward longer timescales. Parsons et al. (2017) found very different pictures of natural variability between CMIP5 models, including the North Atlantic, and between models and paleoclimate data in the tropics, in terms of the magnitude and spatial consistency of climate variance across interannual to centennial timescales.

Laepple, T., and P. H. Huybers, 2014: Ocean surface temperature variability: Large model–data differences at decadal and longer periods. Proc. Natl. Acad. Sci. USA, 111, 16 682–16 687, https://doi.org/10.1073/pnas.1412077111.

Page 3, lines 5-6: please see general comments in previous section- why were the bulk of the CMIP5/PMIP3 Last Millennium simulations excluded? Analysis of results would appear much more robust if an attempt is made to present more than 1/3 of the Last Millennium simulations, or if reasoning can be given why the other simulations were excluded. Also, what is the cutoff used for a drift that is 'too strong'? Is this a global or local drift? All the CMIP5/PMIP3 past1000 simulations appear to show some sort of trend/drift at many grid points- the question is what is too much for the purposes of this AMO/PDO teleconnection study. Please clarify.

Reply: This comment has been addressed under the first "Main Concerns" earlier: we will extend our analysis to include four more models CSIRO, HadCM3, MRI and IPSL and we will revise the manuscript to include these models and compare the additional results.

Our decision to exclude three CMIP5 models (MIROC-ESM, FGOALS-s2 and GISS) was based on the results discussed in Atwood et al. 2016, Fleming and Anchukaitis, 2016; Bothe et al., 2013. This is mentioned in our original manuscript (Page 3, Line 7). Atwood et al (2016) and Bothe et al (2013) discussed long term drift in global mean surface air temperature in their PI simulations. Similarly, Fleming and Anchukaitis (2016) found drift in the Last Millennium simulations, which are apparent in the initial several centuries and excluded from their PDO analysis.

Page 3, Line 20: please see general comments in previous section- removing one linear trend over the full 1850-2005CE time period seems like it may add in low-frequency variability, and I am still not even sure why the historical simulations have been included if the focus is on the impact of volcanic eruptions in the pre-historical simulation time period.

Reply: This comment has been addressed under the second "Main Concerns" earlier. Our findings are not sensitive to this choice and we will include this in a revised manuscript. Our focus is broader than just the impact of volcanic eruptions, we are interested in the influence of external forcings in general on the diagnosis of the role of internal variability from observational evidence. During these simulations, volcanic forcing plays a major role so we did some additional analysis of that but it is not our only result.

Page 4, Line 8: ‘we don’t see much differences’- this is a subjective statement. What criteria are used? Perhaps something like a pattern metric or Euclidean distances metric could be used to say something more quantitative?
Reply: we will revise this sentence to be less subjective, noting which key features (position and strength of the loading maxima and loading gradients) of the PDO patterns are present in the simulated and observed fields.

Page 4, Line 10: Please explain how the TAS time series is made- I assume annual mean (Jan-Dec?) temperature at each grid box, latitude-weighted, and masked ocean grid boxes? Over what latitude and longitude range is this area average made (is it the whole region used in the maps in the figures showing East Asia?)? Please provide more details in the text.
Reply: For the TAS time series, the annual mean (Jan-Dec) TAS is calculated over the land grid points only and area averaged over the region 60E -150E and 10N-55N. We will add this additional information to a revised manuscript.

Page 5, line 5-7: There are other PDO reconstructions- fine to not include them, but can you state why this one is selected over others?
Reply: We have used the selected PDO reconstructions based on the availability of the data for a longer period that covers at least 1000 years of our main analysis period 850-2000.

Page 5, lines 11-15: As far as I can tell, the model-based PDO indices are made from monthly data, and the paleo-based PDO indices are made ‘annual’ data (or seasonally sensitive proxy records)- would a better comparison be to make annual means of SAT for the model data, then construct the PDO index, so the index is more comparable to the annual proxy-based index? (or can you show that the annual and monthly modelbased PDO patterns and time series are similar?)
Reply: The model simulated monthly mean SST data were, in fact, already converted to annual mean data before applying the EOF analysis to get the PDO pattern and its time series (i.e. as suggested by the reviewer – we will make this clear by adding ‘annual-mean SST anomalies’ to the Fig. 2 caption). We have done this because all our analysis is based on the annual mean data, which also compares with the annual mean reconstructed data. We have also confirmed that model based PDO patterns for annual and monthly data are similar (figure R7).

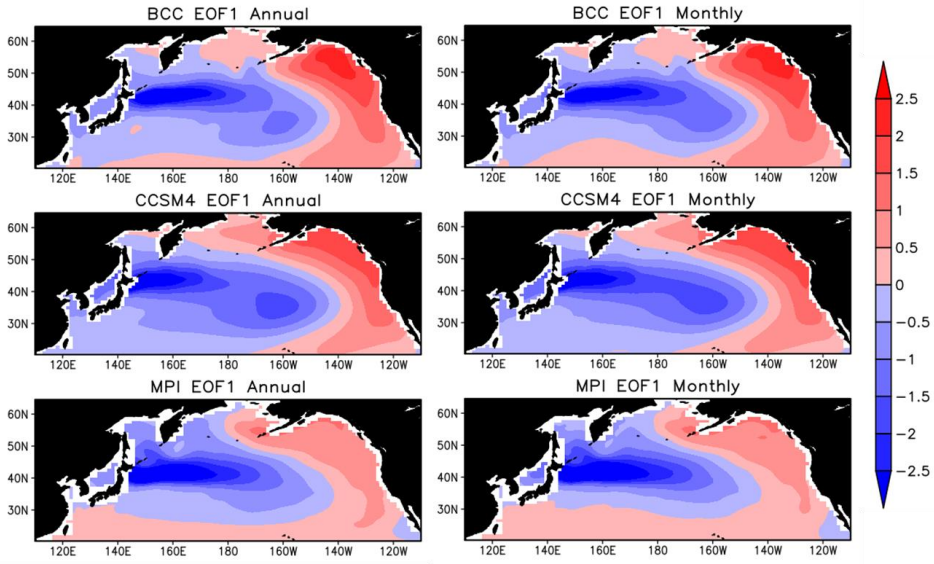

Figure R7: Comparison of PDO pattern calculated from annual-mean SST (left) and monthly-mean SST (right).

Page 5, Line 21, line 25: The 90% significance level seems oddly low, and arbitrarily used in only certain cases- do your results consistently pass a 95% significance test (both the regions in the maps and the spectra)? For example, the 'significant' spectral peaks in Figure 10 appear quite close to the 90% significance level- if you made this a 95 or 99%, are these 'significant spectral peaks' at all significant?
Reply: Please see our response to "main concern (4)" above.

Page 8, line 26: the authors discuss a weak response in BCC to volcanic eruptions- is this a finding that has been noted previously (e.g., Driscoll et al., JGR, 2012, or some sort of similar CMIP5 comparison to observations?)? How realistic is this model's response relative to the other models' responses to volcanic eruptions (especially compared to observations of more recent eruptions and their impacts)? I ask because this difference seems to be important to the results- for example, should the BCC changes (or lack thereof relative to the other models) in PDO, AMO, and associated teleconnections with E Asia be viewed as just as realistic as the other models' responses? Or is it an outlier because it doesn't respond at all to volcanic eruptions when it should?
Reply: By analysing the CMIP5 *historical* simulations, Driscoll et al. (2012) found largest anomaly in the reflected SW radiation in the BCC model. Here, we show that the weak response in BCC to volcanic response only exists in the *last millennium* simulations, where we have analysed three major volcanic eruptions that happened in the last millennium. We also have analysed the same for the major volcanic events during the historical period but didn't find such weak response in BCC model compared to the other two models. So, we feel that the weak volcanic forcing and response only exist in BCC in their last millennium simulations.

Page 9, Line 25-26: It would be helpful to show results from the other four CMIP5 Last Millennium simulations here to put these results in context- right now, 1/3 of the models show a completely different result, but this '1/3 of models' is just the BCC model.
Reply: As noted earlier, we have now analysed this for three other models (HadCM3, MRI and IPSL), with a fourth underway (CSIRO) and would include these results in a revised manuscript.
Page 9, Line 26-29: Would this result imply that the models show an unrealistically large response to eruptions? Or that there is too little internal, low-frequency variability (e.g., Laepple and Huybers)? Or does this suggest both, or something else?
Reply: The potential reasons for the stronger volcanic signal in some models compared with some reconstructions are varied and could include those stated by the referee alongside other reasons (notably errors and biases in the reconstructed temperatures, AMO, PDO and/or forcing histories). We would prefer not to over-speculate at this point and instead present the findings.

Page 10, line 2-3: the authors state that 'all models display red spectra'- in the methods (and in the time series in the figures), it seems that the data have been low-pass filtered, so by definition, the high-frequency variability has been reduced relative to the low-frequency variability (thus reddened)- I'm not sure that 'redness' really means anything in this case. If 'redness' does mean something after the data have been filtered, or if the data have not been low-pass filtered before spectral estimation, please clarify/explain- for example, if the authors mean to say that one model has more lowfrequency variability than another, that may be more accurate.
Reply: Here, the spectral estimation is calculated without applying the low-pass filter to the data. (i.e. the data have not been low pass filtered before spectral estimation). We will ensure that this is clear in a revised manuscript.

Furthermore, the 'pronounced multidecadal variability' barely surpasses the 90% significance threshold, as do most of the 'significant' peaks referenced in this paragraph.
Reply: As noted above, the presence of individual periodicities is of less interest than the overall shape of the spectra (to repeat here for convenience: "overall shape of the spectra, their redness and broad multi-decadal power, and whether these are similar between models, with/without forcing, and

between AMO index definitions"). This paragraph discusses some of these features and not individual significant periodicities, so it is not affect by the choice of the significance threshold. We may remove the significance lines to avoid this confusion.

These power spectra (AMO and PDO power spectra figures) are shown without any error bars- when the spectra are compared and declared similar/different, some sort of spectral estimation confidence bound/error bar on the figure could show if these differences fall within the confidence bounds of the spectral estimates.
Reply: We will compute and add confidence intervals to the spectra.

Page 10, Lines _5-15: Perhaps this is the first time that this analysis has been done, but I would be surprised- has anyone else compared the power spectra across these simulations before? For example, Cheung et al., (2017) compares instrumental-based AMO and Pacific variability to CMIP5 historical simulations (and also how the spatial patterns associated with these modes can change through time). Parsons et al., 2017 (J Clim) compares instrumental, AR1, and CMIP5 Last Millennium, and CMIP5 Control spectra over the North Pacific and North Atlantic, and Fredriksen and Rypdal (2016, J Clim) compare spectra over ocean basins in CMIP5 models.
Reply: As per our understanding, we didn't find any study that compared the power spectra across the simulations in detail. As the reviewer mentioned, Parsons et al. (2017) discussed the power spectra in terms of ensemble mean of CMIP5 models but not the details of the power spectrum of individual CMIP5 models. Similarly, Cheung et al. (2017) did mention about the power spectrum of ensemble mean for the historical period and not the details of the individual members nor the last millennium runs. Fredriksen and Rypdal (2016) compared the power spectrum of CMIP5 control runs with instrumental records and not compared with last millennium simulations. So, we focused on the power spectra of individual CMIP5 models used in our study and compare the results between control and last millennium simulations.

Page 10, Line 23: the authors claim that the spatial patterns of AMO and PDO are similar to the patterns from observations. I see no comparisons among modelled and observed spatial patterns of variability. In fact, it would be helpful if the authors would show the spatial patterns from observations (of course acknowledging that the instrumental-based data have their own limitations) in Figures 1 and 2- this would help put the model results in context.
Reply: Now we have added the spatial patterns of AMO and PDO using observation data. See Figure R1 and R2.

Page 10, line 25: again, it's unclear if the data have been low-pass filtered before spectral analysis. Also, see my above comments- saying the spectra are 'red' seems meaningless if the data have been low-pass filtered. Again, the significant peaks barely surpass a 90% threshold- please discuss or mention if this significance is sensitive to threshold level.
Reply: The data have not been low pass filtered before the spectral analysis. We will make this clear in the revised manuscript.

Also, as stated above it would be good to include error bars/lines on the spectra to know if the 'significant' differences from the background spectrum significant given uncertainties in the power spectral estimation?
Reply: We will compute and add confidence intervals to the spectra.

Page 11, _Line 25: good point.
Reply: Thank you.

Page 12, _line 4: OK, so other recent methods have been used to reconstruct SAT fields (e.g., Last Millennium Reanalysis from Hakim et al., 2016, JGRA and Tardiff et al., in review at CP)

Reply: We will cite the most recent Last Millennium Reanalysis paper and note that this does provide a surface temperature field that could be used to define an index based on the difference between the regional and global SST (though it is not independent of the climate model used to produce the reanalysis, so there may be some circularity in using the resultant AMO reconstruction to evaluate climate model behaviour).

Figures:
Figures 1, 2, 3, 5: please include panels showing similar analyses from instrumental-based data products.
Reply: Now we have included panels based on the data from the instrumental period for Fig1 and 2. We have not included the instrumental data analysis for Figure 3 and 5, because the data length is not enough to calculate the correlation which is passed through 30-year low pass filtered.

Figure 4, Figure 6: it is interesting to see the PDO-E Asia and AMO-E Asia differences, but it would be nice to see some confidence bars on the control run values. For example, Coats et al. 2013 show that teleconnections can change from century to century. Could you do some sort of running correlation or subsample the control run to see how variable this E Asian relationship is (or is there enough data?)
Reply: Confidence intervals could be included instead of the indicator of statistical significance (which occurs when the confidence interval does not include zero) and we will consider how best to graphically present this when we revise the figure (because it also needs revision to include additional GCM results). A running correlation or equivalent is not appropriate here because we are working with 30-year smoothed data (so that we can assess multi-decadal variability rather than the interannual variability that Coats et al., 2013, considered) and dividing it century by century would leave insufficient independent 30-year samples in each century.

Figure 9: Is there a way to put these results in context? For example, if you include the post-Pinatubo response in these models, could you show how the models compare to obs? Which models are more realistic? (CCSM4/MPI or BCC?)
Reply: The issue with BCC appears to be confined to the Last Millennium simulation and not to the Historical simulation, so a comparison with observations post-Pinatubo would not help.

Figures 10 and 11: inclusion of instrumental-based spectra could be helpful here too how realistic are these reconstructions?
Reply: We can include the instrumental based spectra although the data length is short.

Compact listing of purely technical corrections (typing errors, etc.).
Page 1, Line 13: ‘and their spectral characteristics’- remove ‘their’
Page 3, line17: change sentence to: ‘Each model version was the same across all the simulations.’
Page 4, line 8: ‘much differences’- please re-word (e.g., ‘A pattern correlation statistic shows minimal differences among : : :’
Page 5, Line 7: ‘largely suffer from the influence of external forcing’
Page 6, Line 4: ‘no time-varying (transient?) external radiative forcing’
Page 6, line 31: ‘This situation is equivalent to (that?) of Fig.’ – there appears to be a missing word here
Page 7, Line 13-15: ‘in the southern parts’: : :’in all three models’: : :’with the strongest correlation in the northeast region’
Page 7, line17: ‘though it varies’
Page 9, Line 10: the sentence starting with ‘Despite’ appears a bit awkward- suggest rewording.
Reply: Thank you for the careful checking – we will address these minor technical/wording errors in our revised manuscript.