Response to Referee #3

We want to thank the three anonymous referees for the very thorough review of our manuscript. In particular, the comments helped us to better articulate the science question of the manuscript, and this hopefully resolves some of the major concerns. We shifted the focus of the paper from general low-frequency variability to multi-annual oscillations, and changed the title to “Multi-annual modes in the 20th century temperature variability in reanalyses and CMIP5 models”.

The comments led to substantial changes in the manuscript. One of the main changes is that we have made is the way the data sets are preprocessed. We have now used a common scaling factor for all the data sets in order to be able to compare the total spectra of the data sets (based on the reasoning of Referee #3). Because of this comment, we have recalculated all results and also made substantial changes to the text, especially in the section describing the Results. Re-calculation did not change the big picture, but the results are now much better justifiable, especially as there is now a new Supplement available.

Because of these substantial changes, we kindly ask the Referees to read the whole manuscript once again.

We hope that these and the changes explained below help to better convey our message. Below are our detailed responses to the reviewer #3 (In the following, our response to each comment is in red font, and the referee’s comment in black).

(1) comments from referees/public
This manuscript focuses on the capability of current climate models to simulate low-frequency climate variability, as determined through a randomised multi-channel singular spectrum analysis (RMSSA) of near-surface air temperature. On the basis of this analysis, the authors conclude that state-of-the-art climate models tend to exhibit variability that is too periodic, under-active at multidecadal timescales, and over-active at decadal timescales. On the positive side, I thought that the manuscript was clearly and smoothly written. However, I was left with many questions about the authors’ choices. The title of the manuscript is very broad and ambitious, but the authors only analyse one variable with one method in only 12 climate models, so any conclusions that are drawn are much narrower in scope than the title would suggest. By focusing on statistically significant periodicities, the authors really do not directly address whether or not models have too much or too little low-frequency variability (particularly since all time series are standardized prior to the analysis). All comparisons between the models and reanalysis are informal and subjective, and all formal significance testing is limited to red noise null hypotheses rather than model/reanalysis differences. Overall, I was hoping that this study would provide a more thorough and objective evaluation of model performance that goes beyond previous studies, or if that was not the intention, that the scope of this study would be more clearly articulated. I describe my concerns more thoroughly below.

(2) author's response
Thank you for this very thoughtful comment. It helped us, in fact, a lot to better formulate our thoughts and scope the revised manuscript better. It is clear that the manuscript title was too general compared to the actual content of our research, and there was a gap or discrepancy. We hope the revision has resolved this issue.

One of the main changes is the way the data sets are preprocessed. We have now used a common scaling factor for all the data sets in order to be able to better compare the total spectra of the data sets. We have re-calculated everything, including all figures, and made substantial changes to the text to accommodate this change.

Major Comments

(1) comments from referees/public
1) Lines 41-48: The attribution of variance at different timescales by the authors is too simple and not entirely accurate. A substantial portion of variance at interannual to interdecadal timescales can be attributed to “climate noise” associated with processes with intrinsic timescales that are much shorter than interannual. That is the nature of red noise. For example, the North Atlantic Oscillation (NAO) is a teleconnection pattern with broad impacts and pronounced interannual and interdecadal variability, and yet much of that can be attributed to internal atmospheric variability (Wunsch 1999; Feldstein 2000). Therefore, it is not accurate to say that interannual variability is primarily attributed to ENSO or that decadal-to-multi-decadal variability is attributed to ocean dynamics. These comments may be true for periodic variability, but then the authors need to explain why they are focusing on oscillatory behavior and neglecting other dominant sources of interannual and multi-decadal variability.

(2) author's response
Thanks for this clarification which we fully agree. We now realize that the original text was not entirely accurate. We have modified the introduction and these statements are not included anymore. Instead, we have added some text on this issue in the discussion on the lines of this comment (c.f. p. 10) and utilized the references.

(3) author's changes in manuscript
Introduction modified, and the statements removed. Text added on this issue in the discussion (c.f. p. 10), references are added.

(1) comments from referees/public
2) Why do the authors choose the 12 models that they choose? Given that there are so many more simulations available, this choice seems arbitrary.

(2) author's response
A subset of CMIP5 models was needed to keep the analysis and presentation manageable. In selecting the models, we used only one model per major institution to avoid models with close common ancestors. Furthermore, all these models have undergone a long history of development covering several model generations, suggesting that the chosen models collectively represent the state-of-the-art. We admit that the subset could be selected in many different ways.

(3) author's changes in manuscript
The choice of models is justified in the revised manuscript, and some text is added (Section 2.4, 1st para).

(1) comments from referees/public
3) Line 166: Again, perhaps this relates to my misconception about what the authors are trying to address, but the decision to standardize the data sets has made it challenging for me to interpret the authors' results. The climate models may have very different temperature standard deviations, which would impact the temperature variability from interannual to multidecadal timescales (e.g., Thompson et al. 2015). However, by standardizing the data, the authors essentially are artificially adjusting the climate models and reanalyses to have common variance at every grid point. Therefore, the authors are erasing potentially important differences in variance between the models and reanalyses that would impact reanalysis/model differences at all timescales. The motivation for this decision and the consequences for interpretation should be discussed.

(2) author's response
Thank you for this thoughtful remark. After quite some internal discussions we concluded that the way we normalized the data is not the best choice on the viewpoint of comparing the total spectra. We therefore

decided to recompute everything according to this comment about the standardisation, and use a common normalisation factor (the average standard deviation of all the data sets). This better retains comparability of total spectra. The revision of the manuscript is thus extensive (also including the new focus on multi-annual modes exclusively).

The data processing steps after the revision are:

- linear trend fitted and removed,
- annual cycle estimated using Seasonal-Trend Decomposition (STL; Cleveland et al., 1990) and removed,
- resulting values mean-centered and divided by the average standard deviation of all the data sets (see Figure 1). Average standard deviation is obtained after removal of the trend and the annual cycle.

These changes in the preprocessing has led to changes in the results, analysis and conclusions (not so much in the leading modes of variability but more of the modes of smaller eigenvalues). We note that the common normalisation factor may not be the optimal for each data set, but it supports better the aim of this study, which is to compare the multi-annual modes in reanalysis and climate model data sets.

(3) author's changes in manuscript
Text revised extensively.

(1) comments from referees/public
4) Lines 230-234: The decision to evaluate model performance subjectively is unsatisfying. It is difficult to compare power spectra with short records, and visual inspection can be deceiving. Combined with my previous comment, I have difficulty interpreting the authors' results. There may be truth in the authors' conclusions in lines 252-254, but I would like more support.

(2) author's response
We agree with this difficulty, and are not completely satisfied with the subjectivity either. The revised manuscript is our attempt for more objective conclusions. We have also included Supplementary material to better support the analysis. We have removed Table 3 and changed/removed the associated discussions. The conclusions are modified for objectivity, and solely focusing on the multi-annual variability.

(3) author's changes in manuscript
Table 3 is removed, and the associated discussions changed/removed. The conclusions are modified, and focus changed to multi-annual variability. A Supplement added.

(1) comments from referees/public
5) Line 272: How are "false alarms" defined? Again, the authors did not determine if there are significant differences between the reanalyses and models, and so I do not see how the determination of false alarms was made.

(2) author's response
Thanks for this comment - it helped us to realize that the original choice of making a subjective evaluation of model performance inevitably leads to this cascade of problems. We agree that 'false alarms' were not defined at all.

We do not use the term 'false alarm' in the revised manuscript. In addition, we have decided to remove Table 2 and show the significant multi-annual modes in Figure 5 (thin vertical lines) and also in the Supplementary material (S2). We think that these figures are more reader-friendly than Table 2, and the discussion more objective.

(3) author's changes in manuscript
The term ‘false alarm’ is removed, Table 2 is removed, and a Supplement is added.

(1) comments from referees/public
6) Line 289: Why did the authors subjectively choose the Nino3.4 region to base the composites? Although it seems reasonable that the 3.5-yr mode would be related to ENSO, by basing the composites on a subjectively chosen region, the authors seem to be predisposing the analysis to highlight ENSO-like variability. More generally, I am not sure why Section 3.4 is entirely focused on ENSO and its teleconnections, given that these topics have been covered extensively in other studies and that the authors argue that five periodicities exist. It would seem less arbitrary to let the analysis direct the content and to focus on all identified periodicities.

(2) author's response
It is true that the choice of the Nino3.4 region to base the composites directs the analysis (which is not nice), although the choice was made “post mortem”, i.e., after inspecting all the individual figures, which seem to illustrate ENSO-type variability.

Inspired by your comment, we used a completely different and fully objective approach in the revised manuscript, which results in “phase composites”. The compositing procedure now follows the one described in Plaut and Vautard (1994). The idea is to choose the grid point time series (RC_l) for which the variance is the largest, and calculate its time derivative (RC'_l). The phase of the mode at each time step is determined by calculating the angle between the vector (RCl, RC'_l) and the vector (0, 1). These phases, in the interval (0, 2π), are then classified into eight categories, each occupied by equal number of “maps”. Composite maps are then constructed from the maps in each category. This description is included in a new subsection (2.6 Data visualisation).

In the revised manuscript we have identified significant multi-annual periods in the reanalysis data sets at 3.5/3.6 and 5.2-5.7 yr. A variability mode with a period between 3 and 4 years was identified significant (at 5% level) in majority of the climate models (excluding models c, g and k) and therefore we decided to illustrate this particular mode.

There are indeed interesting but different patterns in several CMIP5 models that would be worth studying, but inclusion of all these would make this paper excessively long, and are therefore not included. We tend to think that the model development groups should do this for sake of their own model development.

(3) author's changes in manuscript
Results are recomputed and text changed extensively.

Minor Comments

(1) comments from referees/public
1) Lines 137-138: This relates to my first comment, but the authors are not really addressing whether the models “capture the observed temperature distribution.”

(2) author's response
As far as we understand correctly this comment, the total spectra are now better inter-comparable, and therefore one can better assess the “capture” of variability.

(3) author's changes in manuscript
No action taken.

(1) comments from referees/public
2) Line 179: Is there any sensitivity to this choice of lag window?

(2) author's response
The sensitivity was studied in Seitola et al. (2015), c.f. Fig 7. In that paper it was concluded that the choice of the lag window does not have major effects on the significant periods (on multi-annual scales). We did not redo this sensitivity test here.

(3) author's changes in manuscript
No action taken.

(1) comments from referees/public

3) Line 202: I do not understand why those components are called "trend components" if the data were detrended.

(2) author's response
Sorry, calling the slow component as 'trend components' is a convention that has been used in some of our more statistics oriented references. We agree that this is misleading and the term has been replaced in the revised manuscript.

(3) author's changes in manuscript
Terminology is changed.

(1) comments from referees/public
4) Lines 208-209: How is it determined that ENSO variability has a decadal component in 20CR?

(2) author's response
In the original manuscript, Figure 1, the components 5 and 6 of 20CR have also spectral power between 10 and 20 yr periods, in addition to power on multi-annual time-scales. In the revised manuscript, Figure 2, a similar pattern is seen in components 7 and 8.

(3) author's changes in manuscript
No action taken.

(1) comments from referees/public
5) Line 210: I would not consider the similarity of the 20CR and ERA-20C spectra to be striking, given that the reanalyses assimilate similar data.

(2) author's response
We agree to some extent, but also think that this shows that the data assimilation systems of 20CR and ERA-20C extract observed information in a very similar manner (which is of course good news).

(3) author's changes in manuscript
The word "striking" does not appear in the revised manuscript.

(1) comments from referees/public
6) Line 245: I am not convinced of five key periodicities. Physically, it seems that all identified periods may relate to one phenomenon (ENSO), and these five frequencies just happened to pass the significance threshold.

(2) author's response
Thanks for the remark. We agree that the identified periods may relate to ENSO. Since it is a quasi-periodic oscillation, the ENSO-variability is captured by several near-by frequencies in the significance test.

(3) author's changes in manuscript
Text has been amended (Section 3.4).

(1) comments from referees/public
7) Discussion: Isn't it possible that the existence of too many significant periodicities in the climate models could be due to ENSO being too periodic in some models, which has been discussed previously?

(2) author's response
We agree that too strong / periodic ENSO may result in a large number of significant periodicities in climate models, and the significance test then to pick these up.

(3) author's changes in manuscript
The text has been amended in Section 3.4