

## Response to Referee #1

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 #1 (In the following, our response to each comment is in red font, and the referee's comment in black).

### **(1) comments from referees/public**

1. The goal of this study is unclear as it falls in between (1) a showcase of an advanced statistical tool (RMSSA) and (2) the evaluation of variability in CMIP5 models. Both goals have already been addressed at length in other publications and it is not clear what is new here.

### **(2) author's response**

Thank you for your remark. We totally agree that the goal was not clearly articulated in the original submission. We hope that the revised manuscript is more of (2) and less of (1). We hope the novel aspects are better conveyed in the revised manuscript so that it no longer unclear what is new here. Although there are a large number of studies on the evaluation of CMIP5 models, we still think that it is worthwhile to have a closer look at the model spectra, especially as the advanced tool (RMSSA) has not been applied never before in this extent in other publications.

### **(3) author's changes in manuscript**

We have modified the title as well as the introduction of the manuscript to clarify the goal (which is to decompose the 20th century climate variability into its multi-annual modes, and to assess how these modes are represented by the contemporary climate models.)

### **(1) comments from referees/public**

2. The title seems to imply the second goal is pursued (model evaluation). Then it is unclear what the precise science question is. Why focus on these specific aspects of variability ? What implications for model use or development?

### **(2) author's response**

The title is changed, and from the revised manuscript it should be now very clear that we produce a reference decomposition from two reanalyses on multi-annual scales and then assess how the model data performs with respect to the reference. The science question is clarified, and we provide hints for model development, but refrain from speculating what exactly may be behind some model deficiencies.

Due to this, and comments by other Referees, we shifted the focus to multi-annual variability because of better statistical confidence of the results. We hope this to provide guidance for model development due

better understanding of the deficiencies in representing reanalysed modes of multi-annual climate variability.

### **(3) author's changes in manuscript**

The comment induced a major revision of the text, especially in the Sections of Introduction and Results.

### **(1) comments from referees/public**

3. The few lines that put in context model errors (p1/119 to p2/17) are quite weak and provide an overly simplistic view of this complex problem. Also, why use only 12 models out of the 40+ CMIP5 model available?

### **(2) author's response**

We agree that the text was too simplistic, even though our goal was not to provide a comprehensive review of the complex question.

A subset of CMIP5 models was chosen to keep the analysis and presentation of results manageable. In selecting the models, a major principle was to use only one model per institution, so to avoid models that are too close relatives. Furthermore, all these models have undergone a long (generally several generations of) history of development, suggesting that the chosen models collectively represent the state-of-the-art.

### **(3) author's changes in manuscript**

The explicitly mentioned lines are removed. A justification to the choice of the models have been added. We want to point out again that it would be advisable to read the manuscript once again, since the revisions have been quite extensive - we cannot simply point to a changed word here and a sentence there.

### **(1) comments from referees/public**

4. For ENSO time scale (and lower frequency), several studies have shown that a minimum of 200-300 years of simulation are necessary to obtain robust statistics (Wittenberg 2009 and Stevenson et al. 2010). This questions the use of historical simulations (140 years).

### **(2) author's response**

We agree that it would be ideal to have time series of 200 - 300 years to obtain robust statistics. In model simulation studies this is of course a possibility. However, the fact of reality is that the longest observation based references only extend over the past century, and this is what there is.

MSSA (and therefore also RMSSA) is especially designed for analysing short time series (see Ghil et al. 2002). By taking lagged copies of the time series, it provides overlapping views of the series and enhances the identification of signals from the noise. We have also estimated the likelihood of the identified patterns being generated only by red noise. This is done by the Monte-Carlo significance test, as described in the paper. The test shows that the multi-annual oscillations, have at most 5% chance of being generated only by red noise in both reanalysis datasets (Figure 4b and c) and most of the climate model simulation datasets (Figure 5). Therefore we can even argue that long time-series are in part needed because weak methods are used to analyse high-dimensional data.

### **(3) author's changes in manuscript**

The focus is shifted to multi-annual scales and abstained from closer scrutiny of the decadal and multi-decadal scales (please see the the text in p.2, l. 4-14.) Proposed references are added.

**(1) comments from referees/public**

5. Spectra are not “objective” measures of model performance (nor any single metric, see IPCC AR5 Chap. 9) as error compensation can lead to the right statistics through the wrong balance of physical processes as shown in many studies.

**(2) author's response**

We agree that ‘objective’ was not the best word to use in this context. We agree that the spatio-temporal modes and their spectra are not objective performance metrics that allow ranking the models based on how different the model spectra are from the reference (reanalyses). However, we see that the total spectra and decompositions of each model provide useful hints of the strengths and weaknesses of the models.

We would like to point out, however, the differences of the method used here and the traditional spectrum analysis. RMSSA separates the variability modes that are independent of each other as orthogonal components, i.e. ST-PCs. We are then using spectrum analysis as a means to show on which frequency each component has most power. These spectra are then summarized to make the comparisons of the variability patterns in different data sets easier. We do not have to calculate any spatial averages to obtain the total spectra and also the regional differences in the variability patterns are included in the spatio-temporal analysis.

**(3) author's changes in manuscript**

The comment has been included and the text is now changed (Section 3.3, 1st para). Total spectra and decompositions (Supplement) of each model are now available and commented in the Results.

**(1) comments from referees/public**

6. The “subjective” discussions are quite vague, unhelpful and don’t provide any perspective either compared to previous studies or for modelling groups.

**(2) author's response**

We agree, and this element has been removed altogether, and instead the revised manuscript now provides some perspective on the strengths and weaknesses of the models in simulating the multi-annual modes of temperature variability.

**(3) author's changes in manuscript**

Text changed as suggested in Results section.