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/l19 to p2/l7) 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.

Response to Referee #2

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 #2 (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. Fig. 1 shows that the greatest variance is explained by decadal-multidecadal variabilities (after detrending). However, the decadal-multidecadal variabilities are not examined in this paper, including their spatial patterns and potential mechanisms as well as model biases.

(2) author's response
The comment is exactly right: we did not provide many details about these slower modes although it would be very interesting to see some more details. The revised manuscript is even scarcer in this respect since due to the review comments, the scope is now firmly on multi-annual modes. We think the Referees’ comments were justified (that there is no statistical significance in the results related to the slow modes) and followed the advice in scoping the manuscript anew. We only note briefly in the revised manuscript that the models behave quite differently regarding the variability in decadal and multi-decadal scales

(3) author's changes in manuscript
These are changes throughout the revised manuscript due to the refined scope, especially in Section 3.3, 2nd para.

(1) comments from referees/public
2. Table 2: Some of the periods identified by the RMSSA are very close to each other (for example, 2.2, 2.3, and 2.5; 3.5 and 3.6). It is unclear whether those identified periods truly represent significantly different physical modes or they could merely represent the artifacts of the RMSSA method.

(2) author's response
Thank you for this remark which is now addressed in the revised text. The identified modes in the reanalysis data and CMIP5 models are quasi-periodic, meaning that the oscillation is wobbly and within some neighborhood of a given frequency. Thus, more than one frequency in this neighborhood will be identified as significant. This seems to explain Table 2 of the original submission. In addition, the method itself has a certain spectral resolution depending on the analysis window and temporal resolution of the original data set (monthly data in this case).

Based on the review comments, we realised that Table 2 is not very reader-friendly, and is now removed. The information is now incorporated in Figure 5 instead, which is more compact regarding the significant multi-annual periods. Figure S2 in the Supplementary material provides the test results exhaustively.

(3) author's changes in manuscript
Text has been changed (Section 3.4), Table 2 is removed and the information is incorporated in Figure 5. Supplementary material added.

(1) comments from referees/public
3. Figs. 1 and 3: In addition to ENSO, it will be useful to display the spatial patterns of other significant periods and examine the models' performance in simulating them.

(2) author's response
We totally agree with this comment. The snag with this option is that there would soon be an excessive number of figures. In the revised manuscript, we selected to visualize a mode that is simulated reasonably well by most models, and the 3-4 yr variability pattern was the best option for this purpose.

(3) author's changes in manuscript
New figures available in the Supplement.

4. Specific comments:

(1) comments from referees/public
a) Last paragraph of Page 1: Atmosphere's memory is too short to explain the signal with a period of 1.7 years.

(2) author's response
This is true. In the revised manuscript, the explanatory power of the 1.7 yr mode has become weaker, presumably because of the new normalization with the common variance, and the mode no longer pop up so dramatically. There is thus no longer discussion about this mode in the revised manuscript.

(3) author's changes in manuscript
Text removed about the 1.7 yr mode.

(1) comments from referees/public
b) First paragraph of Page 2: Ocean dynamics responsible for the decadal-multidecadal variabilities needs to be discussed.

(2) author's response
This is completely true. However, with the focus on multi-annual modes, the revised manuscript no longer have this issue.

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

(1) comments from referees/public
c) Page 5, Line 19-20: Components 15-17 of ERA-20C appears to capture the decadal variability of ENSO.

(2) author's response
The visualisation of all modes would be very nice. The new normalization with the common variance

changed the results somewhat especially regarding the modes with low explanatory power. Therefore the components 15-17 of ERA-20C were affected.

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

(1) comments from referees/public
d) page 7, Line 16-25: Replace "a warm pool" and "a cold pool" by "a warm anomaly" and "a cold anomaly"

(2) author's response
Thanks, this is now corrected.

(3) author's changes in manuscript
Text is corrected.

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_i) for which the variance is the largest, and calculate its time derivative (RC'_i). The phase of the mode at each time step is determined by calculating the angle between the vector (RCI, RC'_i) 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

List of changes in the manuscript:

We want to point out 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. Here are the major revisions that we have made:

- We have changed the title to: “Multi-annual modes in the 20th century temperature variability in reanalyses and CMIP5 models”

- We have modified the Introduction to clarify the goal of the work (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.)

- 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.). There are changes throughout the revised manuscript due to the refined scope.

- 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. This is described in 2.5, p.4-5. We have re-calculated everything, including all figures, and made substantial changes to the text to accommodate this change.

- Change in compositing procedure: 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). Please see the Figure 6 in the revised manuscript and also the supplement (S3).

- We have added a supplement including decompositions (S1), significance tests (S2) and spatial patterns of 3-4 yr mode for each data set (S3).

- A justification to the choice of the models have been added (section 2.4 data sources)

- Table 2 is removed and the information (significance test) is incorporated in Figure 5 and supplement (S2).

- Table 3 is removed, and the associated discussions changed/removed. The discussion (Section 4, p. 10) and conclusions (Section 5, p. 10) are modified, and focus changed to multi-annual variability.

- Figures 1, 3, and 6 are new

- Figures 2 (fig.1 in original version), 4 (fig. 2 in original version) and 5 (fig. 3 in original version) are modified

~~An evaluation of current capabilities of modelling low-frequency climate~~ Multi-annual modes in the 20th century temperature variability in reanalyses and CMIP5 models

Heikki Järvinen[1], Teija Seitola[1,2], Johan Silén[2], and Jouni Räisänen[1]

[1]Department of Physics, University of Helsinki, Finland
[2]Finnish Meteorological Institute, Helsinki, Finland

Correspondence to: Heikki Järvinen (heikki.j.jarvinen@helsinki.fi)

Abstract. A ~~crucial performance test of~~ performance expectation is that Earth system models ~~is their ability to simulate~~ simulate well the climate mean state and ~~variability. Here we concentrate on representation of inter-annual to multi-decadal variability in 12 CMIP5 climate model simulations. Reference climate is provided by~~ the climate variability. To test this expectation, we decompose two 20th century reanalysis data sets ~~of~~ and 12 CMIP5 model simulations for years 1901 – 2005 of the monthly mean near-surface air temperature ~~. The spectral decomposition is based on~~ using Randomised Multi-Channel Singular Spectrum Analysis (RMSSA). Due to the relatively short time span, we concentrate on the representation of multi-annual variability which the RMSSA method effectively captures as separate and mutually orthogonal spatio-temporal components. This decomposition is a unique way to separate statistically significant quasi-periodic oscillations from one another in high-dimensional data sets.

The main results are as follows. First, the total spectra for the two reanalysis data sets are remarkably similar in all time scales, except that ~~spectral power of decadal variability (10–30 yr) differ in these data by about 30~~ the spectral power in ERA-20C is systematically slightly higher than in 20CR. Apart from the slow components related to multi-decadal periodicities, ENSO oscillations with approximately 3.5 yr and 5 yr periods are the most prominent forms of variability in both reanalyses. In 20CR, these are relatively slightly more pronounced than in ERA-20C. Since about the 1970's, the amplitudes of the 3.5 yr and 5 yr oscillations have increased, presumably due to some combination of forced climate change, intrinsic low-frequency climate variability, or change in global observing network. Second, none of the 12 coupled climate models closely reproduce all aspects of the ~~reference~~ reanalysis spectra, although some models represent many aspects well. For instance, the ~~IPSL-CM5B-LR model is close to reanalyses but has too little multi-decadal variability, and the HadGEM2-ES model is close to reanalyses except the notable over-activity at periods at and around 10 yr~~GFDL-ESM2M model has two nicely separated ENSO periods although they are relatively too prominent as compared with the reanalyses. There is an extensive Supplement and Youtube videos to illustrate the multi-annual variability of the data sets.

Keywords: ~~climate model assessment, dimensionality reduction~~spatio-temporal modes, climate variability, climate model simulation, random projection, ~~20th century reanalysis, significance testing, RMSSA algorithm~~RMSSA algorithm, ENSO oscillation, Youtube video

1 Introduction

The ultimate goal in developing Earth system models (ESM) is to ~~exploit the inherent predictability of the Earth system to~~ enable exploitation of the inherent Earth system predictability, and hence reduce weather and climate related uncertainties in our daily life, and guide societies in making sustainable choices (e.g., Slingo and Palmer 2011; Meehl et al. 2014). ~~Prediction tools are very complex and their testing goes hand-in-hand with their development. A crucial performance test of ESMs is related to their ability to simulate well the observed~~ For the predictions to be useful and usable, the expectation is that the climate mean state and ~~the variability around the mean.~~

~~Here we focus on ESMs of today and how they represent inter-annual to multi-decadal climate variability . This is a very broad range of temporal scales and it is associated with a multitude of spatial scales. Generally speaking, spectral misrepresentations appear either due to lack of variability in a model or over-activity of a model in some temporal scales. Conclusions about model deficiencies based on spectral differences are very scale dependent, and some general guidance can be obtained by thinking about the mechanisms of natural climate variability (e.g., Ghil 2002). Essentially, short time scale variability (below 2 yr) in the model spectrum of near-surface air temperature is most likely related to the representation of internal variability of the atmosphere. Associated model deficiencies, such as low resolution, can explain most of these weaknesses. Inter-annual variability (2–7 yr) is prominently related to the ENSO phenomenon, and simulation weaknesses point more towards deficiencies in atmosphere-ocean feedback processes and ocean model dynamics. Decadal-to-multi-decadal variability can be thought of being driven by ocean dynamics. There is however clear indication of multi-decadal variability, such as Atlantic multi-decadal oscillation (AMO), that may be driven by stochastic forcing of mid-latitude atmospheric circulation on ocean, and changes in ocean circulation may rather be a response rather than driver of the variability (Clement et al., 2015). This interpretation widens the scope of possible root causes from model errors in ocean dynamics to coupling issues~~climate variability are well simulated by these tools. Due to the complexity of the models and the data they produce, testing the expectation poses a challenge: many aspects of the model performance are gathered under the variability concept and no single diagnostic alone is sufficient to exhaust its all facets. Yet, understanding the discrepancies between the observed and simulated variability is crucial feedback for model development.

Representation of ~~inter-annual to multi-decadal~~ climate variability among models participating in climate model inter-comparisons~~(~~, such as CMIP5~~)~~ , has been studied by e.g. Bellenger et al. (2014), Knutson et al. (2013), Ba et al. (2014), and Fredriksen and Rypdal (2016). We will add to this literature by ~~applying a recently developed powerful spectral analysis tool in this field. We identify the spectral signatures by~~ interfacing a representative set of contemporary coupled climate models with reanalysis data focusing on spatio-temporal modes of climate variability. One century covered with global reanalysis data is naturally very short for this purpose and severely constrains inter-comparison studies (e. g. Wittenberg 2009 and Stevenson et al. 2010). First, time series should cover a sufficient number of recurring "events" for obtaining significance for the findings. Therefore, decadal-to-multi-decadal variability is of interest but not as informative as focusing on shorter cycles of variability. Second, the applied methods have to be very effective in extracting information from the short but high-dimensional data sets. For these reasons, we concentrate on the representation of multi-annual variability in reanalyses and coupled climate models

applying Randomised Multi-Channel Singular Spectrum Analysis (RMSSA; Seitola et al. 2014, 2015) which ~~is an advanced~~ effectively separates mutually orthogonal spatio-temporal components from our high-dimensional data sets.

The aim of this study 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. We hope this to provide guidance for model development due better understanding of the deficiencies in representing reanalysed modes of multi-annual climate variability. Ultimately, interpreting the hints about model deficiencies as development topics are due for the development teams themselves. Our role is to point towards the potential error sources. For reassuring the teams that high-dimensional time series analysis ~~method for~~ is possible today, we emphasise the methodological aspect of this study. RMSSA can, under very weak assumptions on the data, decompose high-dimensional ~~problems. The strength of RMSSA lies in the fact it is able to~~ data sets in a unique way and separate statistically significant quasi-periodic spatio-temporal oscillations from one another. This is in contrast to many other approaches which either make assumptions about the oscillation structures, such as Fourier or spherical decomposition, or resolve only either spatial or temporal aspects of variability. RMSSA can detect spatially evolving "chains of events~~in high-dimensional systems by~~" through resolving eigenmodes of spatio-temporal covariance data. This is a significant advantage, say, over PCA which only resolves eigenmodes of spatial covariances ~~. This can lead to undesirable projection of~~ and often projects temporal evolution of an ~~event~~ "event" onto a number of different eigenmodes. ~~Additional benefits of RMSSA are: (i) dimension reduction via~~ In addition, the novel data compression based on random projections enable ~~applications in extremely high dimensional problems, (ii) convergence properties of the eigenmode decomposition are very good allowing better physical interpretation of fewer components, and (iii) the resulting spectrum is straightforward to test for significance. The paper is organised as follows: data and methods are explained in Section 2, results in Section 3, followed by discussion and conclusions~~here a vast increase in tractable problem size (i.e., data dimension) - even multi-variate decomposition is now possible, although not included here.

2 Methods and Data

2.1 Randomised multi-channel singular spectrum analysis

Multi-channel singular spectrum analysis (MSSA; Broomhead and King, 1986a,b) can be characterised as being a time series analysis method for high-dimensional problems. It effectively identifies spatially and temporally coherent patterns of a data set by decomposing a lag-covariance data matrix into its eigenvectors and eigenvalues (e.g., Ghil et al., 2002) using singular value decomposition (SVD). The lag window in MSSA is a user choice, recommended typically to be shorter than approximately one third of the length of the time series (Vautard and Ghil, 1989). Long lag window enhances the spectral resolution, i.e., the number of frequencies that can be identified, but distributes the variance on a larger set of components. MSSA eigenvectors are called here space-time EOFs (ST-EOFs), and the projections of the data set onto those ST-EOFs space-time principal components (ST-PCs). Because of the lag window, ST-PCs have a reduced length and they cannot be located into the same index space with the original time series. However, they can be represented in the original coordinate system by the reconstructed components (RC; Plaut and Vautard, 1994).

MSSA is computationally expensive and practical limits are easily exceeded for large data sets and long lag windows. In order to overcome this limitation, a computationally more efficient variant, called Randomised MSSA (RMSSA; Seitola et al., 2015), is applied here. The RMSSA algorithm, in a nutshell, (1) reduces the dimension of the original data set by using so-called random projections (RP; Bingham and Mannila, 2001; Achlioptas, 2003), (2) decomposes the data set by calculating standard MSSA in the low-dimensional space, and (3) reconstructs the components in the original high-dimensional space.

In RP, the original data set is projected onto a matrix of Gaussian distributed random numbers (zero mean and unit variance) in order to construct a lower dimensional representation. In this study, we reduce the data volume to about ~~0.8 to~~ 5 % of the original volume. Since the computational complexity of RP is low, involving only a matrix multiplication, it can be applied to very high-dimensional data sets. Although RP is not a lossless compression, it has the important property that the lower-dimensional data set has essentially the same structure as the original high-dimensional data set. This has been demonstrated for climate model data in Seitola et al. (2014). The RMSSA algorithm is briefly presented in the Appendix **??**.

2.2 Computation of spectra

The ST-PCs represent the different oscillatory modes extracted from the data set. In order to estimate the dominant frequencies associated with each ST-PC, the power spectrum is calculated with the Multitaper spectral analysis method (MTM) (Thomson, 1982; Mann and Lees, 1996). To further compare the ~~spectral properties of~~ variability modes and their intensities in different data sets, the power spectrum of all the ST-PCs of each data set is summed up to obtain so-called total spectrum. The ST-PCs are already weighted by their respective explanatory power, i.e. multiplied by the corresponding eigenvalue. Therefore the components with more explanatory power have also higher spectral densities compared to the ones that explain only a small fraction of the variance. Therefore no extra weighting is needed in this step.

The uncertainty related to the explanatory power of each ST-PC (i.e. the confidence interval of the respective eigenvalue) is estimated using the Norths rule of thumb for sampling errors (North et al., 1982). The sampling error (e_k) is given by $e_k \sim \lambda_k(2/N)$, where λ_k is the eigenvalue associated with the kth ST-PC and N is the length of the time series. Thus, the confidence interval of the total spectrum describes the uncertainties related to the explanatory power of each ST-PC.

2.3 Statistical significance testing

In data sets of dynamical systems, ST-PCs/ST-EOFs of MSSA often appear as quadratic pairs that explain approximately the same variance and are $\pi/2$ out of phase with each other. However, existence of such a pair does not guarantee any physical oscillation in the data set, and it may be due to some non-oscillatory processes, such as first-order autoregressive noise. Allen and Robertson (1996) formulated a test, where the oscillatory modes identified with MSSA are tested against a red noise null-hypothesis through ~~Monte-Carlo~~ Monte Carlo simulation.

Significance testing in MSSA requires solving conventional PCs of the original data set. In case of very high-dimensional problems this easily exceeds practical computational limits. The RMSSA implementation in Seitola et al. (2015) contains the Allen-Robertson test such that the PCs are solved in the dimension-reduced space, and is thus affordable even in very high-dimensional problems. The Appendix **??** also includes a short description of the significance test.

2.4 Data sources

The data consists of the monthly mean near-surface air temperature from the historical 20th Century simulations of 12 different climate models (Table **??**). The selected models originate from different modelling centres, and thus do not have close common ancestor models. Furthermore, the selected models have undergone a long (generally several generations of) history of development, suggesting that the chosen models collectively represent the state-of-the-art. Near-surface temperature was chosen, because many processes must be adequately represented in coupled models to realistically capture the observed temperature distribution (Flato et al., 2013). These include processes in the Earth system component models (atmosphere, ocean, etc.) as well as in their mutual coupling models. Also, for the near-surface temperature, there are corresponding reanalysis data available.

The historical (1901–2005) simulations were extracted from the CMIP5 data archive and they follow the CMIP5 experimental protocol (Taylor et al. 2012). The 20th Century simulations use the historical record of climate forcing factors such as greenhouse gases, aerosols, solar variability, and volcanic eruptions. We used a single ensemble member of each model and the model data sets were interpolated into a common grid of 144×73 points.

As a reference, we used two reanalysis data sets: the 20th Century Reanalysis V2 data (hereafer 20CR) provided by the NOAA/OAR/ESRL PSD (Compo et al., 2011), and ERA-20C data provided by ECMWF (Poli et al., 2013). The data sets are produced using an ensemble of perturbed reanalyses, and the final data set corresponds to the ensemble mean. In 20CR, only surface pressure observations are assimilated, and the observed monthly sea-surface temperature and sea-ice distributions from HadISST1.1 (Rayner et al., 2003) are used as boundary conditions (Compo et al., 2011). In ERA-20C, observations of surface pressure and surface marine winds are assimilated (Poli et al., 2013). Unlike 20CR, it uses a more recent sea-surface temperature and sea ice cover analysis from HadISST2 (Rayner et al., 2006). Both 20CR and ERA-20C are forced by historical record of changes in climate forcing factors (greenhouse gases, volcanic aerosols and solar variations). In order to be consistent with the climate model simulations, the same time period is used (1901–2005, i.e., 1260 monthly mean fields) ~~. 20CR has ∼2.0 degree horizontal resolution and we have used gaussian gridded (192 × 94) data from 3-hour forecast values. The horizontal resolution of ERA-20C is approximately 125 km (T159) in a grid of 360 × 181 points~~and the reanalysis data sets were interpolated into the same grid as the model simulations (144×73 points).

2.5 Data processing

Some pre-processing of the data ~~sets is~~ was needed before applying RMSSA~~and statistical significance testing. The data sets were standardised (i.e. the time series of~~ . At each grid point ~~was mean-centered and divided by its standard deviation) to avoid overweighting the grid points with higher variance. This adds weight on the lower latitude variability, where ENSO-type variability is pronounced. On the other hand, no-scaling would make the higher latitude variability dominant because of larger amplitude variations there. Furthermore, each data set was de-trended, and the dominating~~ the data sets were processed as follows:

- linear trend was fitted and removed,

– annual cycle was estimated using Seasonal-Trend Decomposition (STL; Cleveland et al., 1990) and removed~~from the original time series.~~ ,

– resulting values were mean-centered and divided by the average standard deviation of all the data sets (see Figure **??**). Average standard deviation was obtained after removal of the trend and the annual cycle.

The reanalysis and climate model data sets have different temperature standard deviations, which would impact the temperature variability from inter-annual to multi-decadal timescales (e.g., Thompson et al. 2015). To retain these differences, we have used a common normalization factor (i.e., the average standard deviation of all the data sets). This procedure reduces the weight of grid points with high variance, typically at higher latitudes, and hence adds weight on the lower latitude features. After the pre-processing, the dimension reduction step of RMSSA was applied so that approximately ~~3~~ 5 % of the original ~~dimensions of 20CR data, 0.8 of ERA-20C, and about 5 of climate model data were retained – the different percentages are due to different volumes of the original data~~data dimensions were retained. The lag window in the analysis was 20 yr (240 months). The total spectra ~~was~~ were obtained from this analysis, and are comparable due to normalisation using the common standard deviation of the data sets.

~~In the~~ The statistical significance test ~~, the~~ uses a red noise null hypothesis. In the test we have used data sets that are normalised by their own standard deviations. Using a common normalisation interferes with generating the red noise surrogates corresponding to each data set. The first 50 PCs of each data set were retained as input. Those PCs explain ~~80~~ 79 % of the variability in 20CR, 75 % in ERA-20C, and 70 %–80 % in the climate model data sets. A total of 1000 realisations of red noise surrogate data sets were generated, and confidence ~~intervals (90 and~~ interval (95 %) for the oscillatory modes were estimated. We note that transformation to PCs may ~~intefere~~ interfere with the detection of weak signals, as demonstrated by Groth and Ghil (2015).

~~In the following section we will compare the spectral properties of the reanalysis and model data sets. Furthermore, we will test the spectra of each data set against a red noise null hypothesis in order to distinguish signal from noise. Finally, we will compare the spatial patterns of an oscillatory modewith a~~

2.6 Data visualization

We used reconstructed components (RC; see Appendix **??**) for visualisation of the spatial patterns related to ST-PCs. For each grid point time series, we can calculate the RCs corresponding to the ST-PCs (or modes) of interest. These RC values, reflecting the contribution of each grid point to the mode, can be plotted on a map at each time step. We have used these maps to construct videos of the spatio-temporal modes. In Section 3.5~~yr period as represented by different data sets.~~ , we have analysed RCs corresponding to 3–4 yr variability. The result is a time series of the data corresponding to the 3–4 yr mode in each grid point and according to its variance after detrending and removing the annual cycle. In the analysis we have neglected 5 yrs in the beginning and the end of the time series, because the reconstruction procedure may be biased there (see the Appendix, eq. A4). The videos can be found at our Youtube channel (https://www.youtube.com/channel/UCu1zJdwJfLaXvfvTqsKCLHw).

To summarise the animations, we have calculated composite maps of the modes. The compositing procedure 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 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 (RC_l, $RC_l^{'}$) and the vector $(0, 1)$. These phases, in the interval $(0, 2\pi)$, are then classified into eight equally populated categories. Composite maps are constructed from these categories.

3 Results

3.1 ~~Spectral similarity of the two reanalysis data sets~~

~~We first demonstrate the spectral similarity of the two reanalysis data sets. Figure ?? displays, in terms of explained variance, the leading 30~~

3.1 Reanalysis decompositions

The main outcome of the RMSSA method, the space-time principal components (ST-PCs~~for~~) characterise both the spatial and temporal structure of the modes of variability. Sections 3.1 – 3.4 focus on their temporal aspects. The leading 30 ST-PC time series and the corresponding power spectra are displayed in Figure ?? for 20CR and ERA-20C~~and the corresponding power spectra. The decomposition reveals that variance is distributed in a very similar way in 20CR and ERA-20C. The ordering of component pairs is not identical but there is a very clear correspondence of the spectral peaks. For instance, the 1.7 yr peak in the 20CR components 29–30 corresponds to the components 24–25 in ERA-20C. In summary,~~ , ordered according to the explained variance. We can see that

- ~~the trend components with multi-decadal periods (~~components with predominantly multi-decadal periodicity (1~~and~~ , 2~~)~~ ~~explain 6~~ , 5, and 6) explain a total of 7.2 % and ~~5.3~~ 5.9 % of the variance in 20CR and ERA-20C, respectively, with ~~very similar spectra (the length of the time series – 105 years – restricts, of course, the correct identification of multi-decadal oscillations) the spectral peak at 3.5 yr and the broad one around 5 yr are very similar in components 3–6, although the associated 10–20 yr variability in~~ clear similarities in their time series and spectra
- multi-annual components (3, 4, 7, and 8) explain 4.2 % and 3.2 % of the variance in 20CR ~~is separate in components 9–10 in ERA-20C (i. e. ENSO variability has a decadal component in 20CR)~~and ERA-20C, respectively
- there is a broad multi-annual peak centered at 5 yr and a narrower peak at 3.5 yr in both reanalyses; these are clearly separated in ERA-20C at the components 3 and 4 versus 7 and 8. This separation in 20CR is less clear
- there are many spectral peaks in the reanalyses at 2–3 year periods with little explained variance but some are well separated and distinct

~~The similarity of~~ The conclusion based on Figure 2 is that the leading sources of the near-surface air temperature variability at multi-decadal and multi-annual periods are well identifiable in the reanalysis data sets. 20CR and ERA-20C ~~is striking in the total spectra (Fig. ??a). Variability shorter than about 10 years is captured similarly by the two reanalyses . 20CR is somewhat more lively in decadal scale (10–30 yr) and has about 30 higher spectral density there~~ are composed of very similar components explaining the variance in the two data sets. This is of course expected but it is also reassuring from the methodological view point: despite its complexity, the RMSSA decomposition is consistent.

It is noteworthy in Figure **??** that the components 3, 4, 7, and 8 in both reanalyses have become more prominent with time. Since about the 1970's, the amplitudes of these 3.5 yr and 5 yr oscillations have been at a higher level, presumably due to some combination of forced climate change, intrinsic low-frequency climate variability, or changes in global observing network (the rather sudden increase in the amplitude seems to coincide with the onset of the modern era of satellite observations). This finding seems to be in support of e.g. Russell and Gnanadesikan (2014). In this connection it should be noted, however, that apparent low-frequency variations and changes in amplitude may simply arise from random fluctuations of the time series (Wunsch, 1999; Wittenberg, 2009). ~~One explanation for this difference may be the decadal variability of ENSO which is present in~~ Back-projection of these components into the original grid representation (Figure **??**), reveals that the components are indeed associated with the ENSO phenomenon and are geographically similar in 20CR and ERA-20C. In the snapshots from January 1987 and January 1998 (Figure **??**), there is a typical El Niño pattern with positive anomalies in the equatorial Pacific Ocean, South-America, and northwestern North-America. These are associated with synchronous evolution of (i) a dipole structure in the western Antarctica with easterly motion, and (ii) a wave-train type pattern in the northernmost North-America with north-easterly motion. The components 3, 4, 7, and 8 thus represent a global phenomenon, with an increased amplitude in recent decades. These features are nicely depicted in our Youtube channel (https://www.youtube.com/watch?v=vehbT8fOHeM, https://www.youtube.com/watch?v=xG--SiUqqAI).

3.2 Reanalysis total spectra

Figure **??**a shows the total spectrum for the reanalyses constructed from the ST-PCs, and their confidence intervals (dashed lines). As in the ST-PCs, there is most power in the slow modes. At periods of about 3.5 yr and 5 yr, there are the spectral peaks of the components 3, 4, 7, and 8. The dip at 1 yr reflects the removed annual cycle.

As Fig. **??** already suggests, the shape of the two spectra is remarkably similar in all time scales (Fig. ~~**??**, components 5–6)but is missing from~~ **??**a). This leaves hardly any doubt that the data assimilation systems of 20CR and ERA-20C ~~.~~

~~The statistical significance testing of the periodicities~~ extract observed information in a very similar manner. There are some differences, however. The spectral power in ERA-20C is systematically slightly higher than in 20CR. This difference is statistically significant at almost all time scales. This is most likely due to generally higher temperature variance in ERA-20C compared to 20CR, especially in the Southern Ocean and Arctic Ocean. Also, in 20CR~~(Fig. **??**b)~~, the 3.5 yr and 5 yr spectral peaks are relatively more pronounced than in ERA-20C.

Statistical significance tests are presented in Figs. **??**b and **??**c for 20CR and ERA-20C~~(Fig. **??**c) reveals that nearly the same periods rise above the red noise in the two data sets (~~, respectively. The multi-annual periods (less than 7 yr) rising above

the 95 % confidence interval (i.e., the red dots above the ~~area~~ region covered by the vertical bars) are 3.5 yr, 3.6 yr, and 5.7 yr in 20CR and 3.6 yr, 5.2 yr, 5.5 yr, and 5.7 yr in ERA-20C. Thus, nearly the same periodicities rise above the red noise in the two data sets. It is logical that the frequency corresponding to the annual cycle is present in the red noise surrogates while it is ~~removed~~ absent from the data~~sets (~~, and therefore the red dots ~~are far below the bars). The periods rising above the red noise at 5 and 10 significance level are tabulated in Table ??, columns 1 and 2, and are nearly the same~~ fall far below their expected values. Interestingly, the period of 2.9 yr in 20CR and ERA-20C fall below the 95 % confidence interval. Our conclusion is therefore that the ~~low-frequency~~ multi-annual climate variability in the near-surface air temperature is very similar in ~~the two reanalyses~~20CR and ERA-20C.

3.3 ~~Evaluation of the simulated~~ CMIP5 model total spectra

The ~~climate model simulation data was processed exactly the same way as the reanalysis data. The total spectrum of each model is shown in Figure ?? with~~ total spectra for the 12 CMIP5 model are shown in Fig. ?? (solid lines) with their 95 % confidence intervals (dashed envelopes) and the reanalysis spectra ~~on the background~~ as a reference (~~dotted~~ thin lines). ~~The spectra are objective measures of model performance. We evaluate *subjectively* how the models reproduce~~ Statistically significant multi-annual modes (at 5 % level) are denoted by vertical dashed lines. As in the case of reanalyses, these spectra are unique expressions of the low-frequency variability present in the simulation data. A comparison between the simulated and the reanalysis spectra ~~, and adopt the following terminology: a model can be either under-active~~ provides one means to assess the strengths and weaknesses of these models. However, one cannot simply rank the models based on how "far off" the model spectra are from the reference, because this comparison focuses on just one (although important) aspect of model performance and because seemingly good agreement with observations might occasionally result from compensating errors in model processes.

Here we will concentrate on the multi-annual aspects but note in passing that the level of multi-decadal variability (~~over-active) if the spectral density is lower (higher) than in the reanalysis spectra, or a model is on-target with respect to spectral density. This enables us to make an overall evaluation of the current capabilities of climate models to represent low-frequency variability .~~ > 20 yr) is close to reanalyses in models a, c, d, e, and g. In the rest of the models, the level seems too low. In the decadal scale (∼10 ... 20 yr), the level of variance is close to reanalyses in a, b, c, f, i, j, and l. Subjectively, the shape of the low-frequency end of the spectra appears most realistic in models a and c.

~~The subjective evaluation is summarised in Table ??. In representing multi-decadal variability, three models are on-target while the rest are under-active. In decadal variability, majority of the models are on target, while four are over-active. Only one model has both of these variabilities on-target (a). In ENSO variability, two models (i and g) seem to have both the~~ In multi-annual scales, the model performance varies a lot among the models. There is a group of models (a, b, d, and e) with high spectral density at about 3 – 7 yr periods. The models d and e have a bi-modal spectral structure, as in the reanalyses, while models a and b have a broad unimodal peak. Decompositions (available in the Supplementary material, S1) partly explain the reasons leading to these total spectra.

In model a, for instance, there is a unimodal broad peak at 3.5 ~~and 5 yr periods on-target. Five models have one of these periods on-target, and the rest of the models are either under- or over-active. With respect to the inter-annual variability, four models seem to be over-active, while the rest are on target.~~ – 4 yr periods (Fig. **??**a). The decomposition reveals that there are, in fact, two well separated component pairs at 3.5 yr and 4 yr generating one merged peak to the total spectrum (Fig. S1a in the Supplement). A development hint is thus to investigate these modes which can help to better understand some underlying modelling deficiencies, and to keep monitoring how this aspect of model performance evolves in the future model upgrades. An additional concern in model a is the excessive spectral density at about 2 yr and 7 – 10 yr periods.

~~We are not going to rank the models, but two models (i and g) seem to perform particularly well and have four out of five key periodicities on-target. The IPSL-CM5B-LR model (i) is close to the reanalyses all the way from inter-annual and ENSO variability to the decadal scale of 20 years and only seem to have too little multi-decadal variability . The HadGEM2 model (g~~In model e, there is a bimodal total spectrum (Fig. **??**e), although far too pronounced as compared with the reanalyses. The decomposition (Fig. S1e in the Supplement) reveals that the ST-PC components 1 – 10 (except 7–8) are all multi-annual and peak strongly and well in isolation at 3 yr, 3.5 yr, 4 yr, and 5 yrs, explaining together no less than 13.9 % of the total variance. The development hint for model e is thus to investigate the mechanisms behind the components 1 – 10 and thereby obtain guidance for improving the realism of simulations.

In most other models, the multi-annual variability is less prominent than in the reanalyses. In model c (Fig. **??**c), on ~~the other hand, is reasonably close to the reanalyses across the spectrum, except for the notable over-activity at periods at and around 10 yr. CanESM2 (a)is the only model that closely reproduces both the decadal and~~ one hand, the decomposition points out (Fig. S1c in the Supplement) that there are about 12 ST-PC components with periods between 1.5 – 3 yrs leading to a total spectrum with a broad peak of 2 – 3 yr periods. These components tend to have very regular cycles, remotely resembling a coupled harmonic oscillator and seemingly missing the "offbeats" or true quasi-periodicity of the reanalyses. The task seems to be to find out reasons why model c produces too rapid and regular multi-annual variability. In model g (Fig. **??**g), on the other hand, the leading ST-PC components 1 – 9 are on either decadal or multi-decadal ~~variability~~periods and these overwhelm the total spectrum. It should be important to find out the causes for this accentuated variability, especially on the decadal scale.

~~Overall, the clearest signal here is that the modelsgenerally seem to lack multi-decadal variability , and some models are over-active in representing decadal and inter-annual variability. In ENSO time scales, only two models are on-target~~Finally, Fig. **??** casts light on models' overall level of variability compared to reanalyses. Clearly, this level in model h (Fig. **??**h) is low. Curiously enough, the leading ST-PC component pair in model h explains only 1.4 % of variance and peaks at 3.2 yr. This corresponds to the isolated peak in the total spectrum.

3.4 Significance ~~test by Monte-Carlo MSSA~~of multi-annual modes in CMIP5 models

~~Table ?? contains the periods rising~~ In the reanalyses (Fig. **??**), only a few multi-annual periods rise above the red noise ~~at 5 and 10 significance level. The periods for~~ (three in 20CR and four in ERA-20C~~are in the columns 1 and 2. Both data sets have five periodicities significant at~~). They are at approximately 3.5 yr and 5 ~~level, four of which are common and one~~

~~specific. Additionally, there are three common periodicities significant at 10 level. In terms of statistical significance, 20CR and ERA-20C behave in a very similar manner.~~

~~Table ?? gives an impression that the number of statistically significant periodicities is very large in models, in general. In this respect, the climate model data is dissimilar to the reanalysis data. There is only one model (model k) which has fewer significant periodicities than the reanalyses, and three models (g, h, j) are somewhat close to reanalyses with 11–13 significant periodicities. The rest of the models have many more periodicities (up to 30). Interestingly,~~ yr periods. For the CMIP5 models, the test results are available in the Supplementary material (S2). In Fig. **??**, the multi-annual modes with periods less than 7 yrs at the ~~HadGEM2 model (g) has the 10 yr peak, discussed earlier, significant at~~ 5 % ~~level.~~

~~The number of models that correctly detected either the 20CR or ERA-20C periodicities were seven models for~~ significance level are denoted by dashed vertical lines.

In summary, there are 5 – 15 statistically significant periods in the models, except model k (Fig. **??**k) with three and model g (Fig. **??**g) with zero periods. The large number of significant periods (models d and e, for instance) can be explained, at least partly by the fact that the modes are quasi-periodic and the spectral density therefore appears on a range of frequencies. This manifests as excursion of the ~~1.7 yr period, five for 2.2 yr, four for~~ red-noise threshold on several adjacent frequencies. This is typical for models with large spectral power on certain time scales. In model l (Fig. **??**l), for instance, there are two broad and distinct spectral peaks at about 3.5 yr ~~, seven for 3.6, four for 4.2 yr , and two for 5.2~~ and 6 yr periods, and many significant periods are gathered at these and nearby frequencies. In contrast, models f and h (and to some extent model c) have several significant and distinct periods between 2 yr and 7 yr. In ~~addition to these, there were many "false alarms", as seen in Table ??~~terms of number of significant modes, models a, i, j, and k seem to be closest to the reanalyses.

3.5 Spatial patterns of the ~~3.5~~ 3-4 yr mode

~~The oscillatory mode with a 3.5 yr period was significant in 20CR and ERA-20C and in seven climate models. The~~ ST-PC components can be represented in the original coordinate system as so called reconstructed components (RC) that can be visualised. In this section, some visualisation results are presented and discussed.

In ERA-20C, there is a spectral peak at 3.5 yr ~~has some power on both sides of the peak. Therefore, a 3–4 yr mode would perhaps be a more appropriate name for the peak. Oscillations at these periods are usually attributed to ENSO~~ period, which is significant at 5 % level (Fig. **??**). This peak is due to the ST-PC components 7 and 8 with spectral density closely concentrated on this frequency (Fig. **??**). Figure **??** depicts composite maps of each of the eight phases of the 3.5 yr mode in ERA-20C. Firstly, the mode is global with the largest temperature anomalies in the Pacific and North-America. Secondly, the mode contains tropical Pacific temperature anomalies, like in the ENSO phenomenon (e.g. Kleeman, 2008). ~~We illustrate here how this periodicity appears in different data sets. Since the 3.5 yr period is not reproduced by all of the climate models, we have chosen in these cases a periodicity that is close to 3.5 years.~~

~~We have calculated the reconstructed components (RC) corresponding to the 3.5 yr mode in order to visualise the associated spatial temperature anomaly patterns. The result is a time series of the original (centered)data corresponding to the 3.5 yr mode in each grid point and according to its original variance. We therefore have an animation of~~ The cold (warm) maximum is in

phase 1 (5) with the anomalies extending to the ~~global 3.5 yr mode for each data set for the whole timeperiod (1901–2005).~~ South-American continent. Thirdly, there are traveling temperature anomalies at high latitudes on both hemispheres. These are described next.

~~To synthesise the animations, we have calculated composite maps of the 3.5 yr mode for each data set~~ In phase 1 (Fig. **??**), there is a small warm temperature anomaly in the North-Pacific (lon 160°W, lat 30°N). This pattern slowly moves northeast reaching Alaska in phase 5 and then gradually dissipating over the northernmost North-America in phase 8 (and being visible still in phase 1). There is a very similar evolution of a cold anomaly starting in phase 5. At the same time, there is an oscillating temperature anomaly over the Eurasian continent in opposite phase. In Fig. **??**, there is also a traveling temperature anomaly in the Southern Hemisphere. In phase 8 (Fig. **??**~~). The patterns are composites of eight instances when the mode is in its maximum positive phase in the Niño3.4 region (120~~**??**), there is a cold anomaly over the Southern Ocean (lon 160°~~W–170~~W). This strengthens, moves east, weakens, and crosses the Antarctic Peninsula in phase 4 and remains in the Weddell Sea until phase 7. Similarly, there is a warm anomaly in phase 4 (lon 160°W~~, 5° N–5° S) . Positive events are defined as an average of winter months (November–March~~) with similar evolution as the cold one.

~~The top row of Fig. **??** displays the patterns for the reanalyses. One can see typical El Niño related temperature anomalies, such as positive anomalies in the equatorial Pacific Ocean and South-America,~~ Next, 20CR and the CMIP5 model behaviour is studied. The 3.5 yr mode is significant in 20CR and ~~northwestern North-America. There is also a cold anomaly over the northern parts of the Eurasian continent, a dipole structure in the western Antarctica, and warm Africa. The patterns are remarkably similar~~ ERA-20C. For the illustration, we have chosen component pairs from the model decompositions (Supplementary material Fig. S1) that have spectral peaks between 3 and 4 years and do not express substantial variability on other time scales. In most climate models, such a corresponding mode exists, except in models g and k. In model c this mode is not significant at 5 % level, but it is illustrated anyway. Supplementary material reveals how these modes are represented in different data sets (Fig. S3–S14). The format is the same as in Fig. **??**. A short summary is presented next.

In 20CR (Fig. S3), the anomalies are weaker compared to ERA-20C (Fig. S4). This is mainly because the 3 – 4 yr mode is distributed on two component pairs in 20CR ~~and~~ whereas in ERA-20C ~~with somewhat larger amplitudes~~ it is concentrated on one pair. Nevertheless, similar although weaker signal is evident in 20CR, ~~except the stronger dipole in ERA-20C. Next, we will try to *subjectively* assess model performance in reproducing the spatial patterns of~~ such as the northeast propagation of the North-Pacific temperature anomaly. (Note that in Fig. **??**, the ~~3.5 yr oscillation.~~

~~All models produce a warm pool~~ combination of components 3, 4, 7, and 8 produce highly similar global patterns for 20CR and ERA-20C.) A prominent feature is also the opposite temperature anomalies in the northern Eurasia versus North-America. All models (Figs. S5–S14) produce a temperature anomaly to the equatorial Pacific Ocean (and South-America). The amplitude is larger and/or the area extends ~~too far~~ further to the west ~~in five~~ than in ERA-20C in six models (a, b, d, e, h~~). Unlike in the reanalyses, the pool extends to the Atlantic in four models (a, d, e, h). The warm anomaly in the South-American land area is too weak in five models (c, f, g, j, k).~~

~~The warm anomaly~~ , l). The anomaly pattern in the northwestern North-America is present in all the models to some extent. In the reanalyses, the anomaly is strictly confined to land areas but in most models, it is either somewhat misplaced or extends to the adjacent sea areas and the Eurasian continent.

~~The cold anomaly over the Eurasian continent is not well represented in the model data. There is a weak cold anomaly in three models (f, i, l), a weak warm anomaly in one model (c), and a mixture of cold and warm areas in the rest.~~

~~The warm pool in the Amundsen sea related to the warm-cold dipole around the western Antarctica is present in nine models (a~~Models c, e, ~~f, g, h, i, j, k, l). The cold pool in the Weddell sea is present in five models (c, d, f, g, k). Three models (f, g, k) represent both pools, and thus the dipole structure in well represented. In Africa, the anomaly is on the positive side in all models.~~

~~In Table ??, there are three over-active models with respect to~~ and f produce the North-American pattern quite similar to reanalyses, and the ~~3.5 yr period (a, d, e). These are associated with large amplitudes in Fig. ??. Additionally, two models also seem to have large amplitudes in Fig. ?? (b, h). Both of these have higher spectral density than the reanalyses (Table ??), but were anyhow assessed as "on-target" in Table ??. Four models were under-active in Table ?? (~~northeast propagation is captured to some extent by models b, c, f, ~~j, k). These correspond to the low-amplitude maps in Fig. ??. In summary, the patterns of Fig. ?? are in support of the subjective analysis of the total spectra of Table ??~~i, and l.

4 Discussion

~~Table ?? shows that there is a much larger number of significant periodicities in the model data than in the reanalyses. This seems to imply that "nature" tends to produce only a few but statistically significant periodicities, and the other potential periodicities are somehow~~ We note that a substantial portion of variance at inter-annual to inter-decadal timescales can be attributed to "~~worn out~~climate noise" ~~via non-liner interactions and feedbacks so that they cannot be distinguished from red noise. We can think of two possible causes for this. First, it may simply be because the reanalysis data represent ensemble mean while the model data are individual simulations. This difference may appear as a difference in the number of significant periodicities. Second, it may be that something in models prohibits the wearing process from occurring, and we can observe the excessive number of periodicities above the noise level – as if the models were missing some non-linear processes. However, MPI-ESM (model k)has fewer significant periodicities than observed, and to our knowledge, this model does not fundamentally differ from the other models. Therefore, it is unclear what exactly lies behind this result.~~

~~We have not discussed the interpretation of the 1.7 yr interannual variability. Based on the related spatial patterns, it may be that it is a harmonic of the ENSO variability (2×1.7 yr $= 3.4$ yr)~~associated with processes with timescales much shorter than the inter-annual scale (Wunsch 1999; Feldstein 2000). If the amplitude of the variability mode exceeds some noise threshold (such as red noise), then the variability mode is also likely driven by some process external to the atmosphere, in addition to the climate noise. For example, large part of the inter-annual atmospheric ENSO pattern is presumably driven by anomalies of tropical diabatic heating associated with sea surface temperature anomalies (Feldstein, 2000). We assume that for this reason the multi-annual patterns related to ENSO clearly exceed the noise threshold in the results of this study.

5 Conclusions

~~In this study , we decomposed~~ The aim of this study is to decompose the 20th ~~Century near-surface temperature~~ century climate variability into its ~~inter-annual to multi-decadal eigenmodes. We used two state-of-the-art~~ multi-annual modes, and to assess how these modes are represented by the contemporary climate models. To this end, two 20th ~~Century~~ century reanalysis data sets and 12 ~~historical climate model simulations extracted from the CMIP5 data archive. The analysis was performed using the randomised multi-channel singular spectrum analysis~~ CMIP5 model simulations for years 1901–2005 of the monthly mean near-surface air temperature have been decomposed using Randomised Multi-Channel Singular Spectrum Analysis (RMSSA)~~, which is particularly well suited to high-dimensional time series analysis~~. The statistical significance of the identified ~~eigenmodes was estimated with Monte-Carlo~~ modes has been estimated with Monte Carlo simulations. The main conclusions are as follows~~:~~.

~~The spectral~~ Spectral properties of the ~~two reanalyses (~~20CR and ERA-20C ~~) are remarkably similar, the only notable difference being the different spectral density in the decadal scale variability (10–30 yr). Also, nearly the same periodicities rise above the red noise at the 5 and 10 significance levels. Majority of the climate models are under-active in representing the multi-decadal climate variability (> 30 yr), some models are over-active in decadal (10–30 yr) or inter-annual (< 2 yr) variability, and only two models are on-target in both the~~ reanalysis data appear remarkably similar. The most prominent forms of variability in both data sets are related to approximately 3.5 yr and 5 ~~yr ENSO variabilities. The IPSL-CM5B-LR and the HadGEM2-ES are the closest to the total spectra of~~ yr modes which are significant at 5 % level. The spectral power in ERA-20C is systematically slightly higher than in 20CR. The 3.5 yr mode is illustrated in more detail. In ERA-20C, the ~~reanalyses.~~ mode is associated with typical ENSO pattern of temperature anomalies in the equatorial Pacific Ocean, South-America, and northwestern North-America. On top of these, the mode also contains a northeast propagating temperature anomaly over the northernmost North-America, and another eastward propagating anomaly in the vicinity of western Antarctica. Since about the 1970's, the amplitude of this 3.5 yr global mode have increased.

~~Relaxation~~ None of the 12 coupled climate models closely reproduce all aspects of the reanalysis spectra, although some models represent many aspects well. For instance, the GFDL-ESM2M model has two nicely separated ENSO -related periods although they are relatively too prominent as compared to the reanalyses. Also, a number of models represent the propagating temperature anomalies at 3 – 4 yr time frame. Some suggestions are provided in the text for potential model development aspects.

There is an extensive Supplement available presenting the results in visual format for each reanalysis and model data set. In the future, relaxation of the uni-variate nature of the present study ~~remains as a subject for future research. It would be very interesting to extend the set of variables~~ would seem a natural extension. This is now possible since the use of random projections ~~allows efficient data~~ allow efficient data structures preserving compression. Of special interest would be to study behaviour of variables directly linked with atmosphere-ocean coupling processes, such as heat, momentum, and moisture fluxes over oceans.

6 Data and code availability

All data used in this study was downloaded from open sources. The RMSSA algorithm and the statistical significance testing are implemented using GNU licensed free software from the R Project for Statistical Computing (www.r-project.org). Our implementation is available on request. The animations of the ~~3.5 yr periodicity~~ 3–4 yr mode are available for all data sets at https://www.youtube.com/channel/UCu1zJdwJfLaXvfvTqsKCLHw.

Appendix A: Randomised multi-channel singular spectrum analysis (RMSSA)

The RMSSA algorithm and the significance test is briefly presented here. The original data matrix is $\mathbf{X}_{N\times L}$, where the columns are called *channels*. In case of gridded data set, N represents the time steps and L is the number of grid points. It is useful to think N as the time steps when the *sample* of dimension L is collected. The dimension reduction is a projection $\mathbf{X}_{N\times L} \rightarrow \mathbf{P}_{N\times k}$, where $L \gg k$. In other words, we preserve all samples but reduce the sample dimension from L to k. The dimension reduction is performed in two steps: (1) generate a random matrix $\mathbf{R}_{L\times k}$, where the matrix elements are $r_{ij} \sim N(0,1)$ and column vectors of $\mathbf{R}$ are normalised to unit length, and (2) project $\mathbf{X}$ onto $\mathbf{R}$:

$$\mathbf{P}_{N\times k} = \mathbf{X}_{N\times L}\mathbf{R}_{L\times k}. \tag{A1}$$

The next step is to construct an augmented data matrix $\mathbf{A}$, which contains M lagged copies of each channel in $\mathbf{P}$. In RMSSA, M represents the lag window. $\mathbf{A}$ now has Mk columns and $N' = N - M + 1$ rows. The singular value decomposition of $\mathbf{A}$ is

$$\mathbf{A} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{V}^T \tag{A2}$$

The vectors of $\mathbf{U}$ are the eigenvectors of $\mathbf{Z} = \frac{1}{Mk}\mathbf{A}\mathbf{A}^T$ and $\mathbf{V}^T$ contains the eigenvectors of $\mathbf{C} = \frac{1}{N'}\mathbf{A}^T\mathbf{A}$. These vectors are orthogonal and often called space-time principal components (ST-PCs) and space-time empirical orthogonal functions (ST-EOFs), respectively. Note that the ST-EOFs are now in reduced space k. Diagonal elements of $\mathbf{D}$ are the eigenvalues of $\mathbf{C}$ or $\mathbf{Z}$. Finally, the eigenvectors (ST-EOFs) are calculated in the original L-dimensional space by

$$\mathbf{V} \approx \mathbf{A}_o^T\mathbf{U}(\mathbf{D}^{1/2})^{-1}, \tag{A3}$$

where $\mathbf{A}_o$ is the augmented matrix of the original data matrix $\mathbf{X}$. Note that the calculation of ST-EOFs in Eq. (**??**) can be limited only to the eigenmodes of interest.

The ST-PCs can be represented in the original coordinate system by the reconstructed components, RCs (Plaut and Vautard, 1994; Ghil et al., 2002). This transformation is given by

$$rc_{le}(n) = \frac{1}{M_n}\sum_{m=I_n}^{J_n} u_e(n-m+1)v_{le}(m), \tag{A4}$$

where u_e are the ST-PCs and v_{le} are the ST-EOFs calculated in Eq. (**??**) (the part of ST-EOF corresponding to channel l). e is the index of the eigenmode that is calculated. The normalisation factor M_n and the summation bounds I_n and J_n are given in Ghil et al. (2002) and for the central part of the time series $(M \leq n \leq N - M + 1)$ they are $(M, 1, M)$, respectively.

RMSSA with significance testing is briefly presented in the following. Testing the MSSA components against a red-noise null-hypothesis requires orthogonal input vectors, which are obtained by calculating first a conventional PCA and retaining a set of dominant PCs. Therefore some additional calculation steps are included in the RMSSA-algorithm:

SVD of lower dimensional matrix $\mathbf{P}$ is calculated to obtain the principal components (PCs, calculated as $\mathbf{UD}^{1/2}$). PCs fullfil the orthogonality constraint exactly. PCs, that explain large part of the variance of the data set (e.g. 50 first), are retained to obtain matrix $\mathbf{T}$, where the columns are the PCs. Next, the augmented matrix $\mathbf{A}_{PC}$ is constructed from $\mathbf{T}$ and SVD is calculated as in Eq. (**??**).

Finally, a large number of red-noise processes (i.e. surrogate data sets) are generated, and the confidence limits for the MSSA eigenmodes are determined. This signicance test (~~Monte-Carlo~~ Monte Carlo MSSA) is described in detail in Allen and Robertson (1996).

Author contributions. HJ suggested the study and mostly wrote the article. TS implemented the methods, performed all computations, and wrote data and method descriptions, JS supported the method development, and JR the climate model data analysis.

Acknowledgements. This research has been funded by the Academy of Finland (project number 140771), Academy of Finland Centre of Excellence Programme (project number 272041), and the Fortum Foundation (grant number 201500127).

References

Achlioptas, D.: Database-friendly random projections: Johnson-Lindenstrauss with binary coins, Journal of computer and System Sciences, 66(4), 671–687, 2003.

Allen, M. R. and Robertson, A. W.: Distinguishing modulated oscillations from coloured noise in multivariate datasets, Climate Dynamics, 12(11), 775–784, 1996.

Ba, J., Keenlyside, N. S., Latif, M., Park, W., Ding, H., Lohmann, K., Mignot, J., Menary, M., Otterå, O. H., Wouters, B., Salas y Melia, D., Oka, A., Bellucci, A. and Volodin, E.: A multi-model comparison of Atlantic multidecadal variability, Climate Dynamics, 9, 2333–2348, 2014.

Bellenger, H., Guilyardi, E., Leloup, J., Lengaigne, M. and Vialard, J.: ENSO representation in climate models: from CMIP3 to CMIP5, Climate Dynamics, 42, 1999–2018, 2014.

Bingham, E. and Mannila, H.: Random projection in dimensionality reduction: applications to image and text data, in: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '01, ACM, New York, 245–250, 2001.

Broomhead, D. S., and King, G. P.: Extracting qualitative dynamics from experimental data, Physica D, 20, 217–236, 1986a.

Broomhead, D. S., and King, G. P.: On the qualitative analysis of experimental dynamical systems, in: Nonlinear Phenomena and Chaos, S. Sarkar (Ed.), Adam Hilger, Bristol, England, 113–144, 1986b.

Clement, A., Bellomo, K., Murphy, L. N., Cane, M. A., Mauritsen, T., Rädel, G. and Stevens B.: The Atlantic Multidecadal Oscillation without a role for ocean circulation, Science, 350, 320–324, doi: 10.1126/science.aab3980, 2015.

Cleveland, R.b., Cleveland, W.S., McRae, J.E. and Terpenning, I.: STL: A Seasonal-Trend Decomposition Procedure Based on Loess, Journal of Official Statistics, 6, 3–73, 1990.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, Ø., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D. and Worley, S. J.: The Twentieth Century Reanalysis Project, Quarterly J. Roy. Meteorol. Soc., 137, 1–28, 2011.

Feldstein, S. B.: The timescale, power spectra, and climate noise properties of teleconnection patterns, J. Climate, 13, 4430-4440, 2000.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C. and Rummukainen, M: Evaluation of Climate Models, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, [Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V. and Midgley, P.M. (Eds.)], Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Fredriksen, H.-B. and Rypdal, K.: Spectral Characteristics of Instrumental and Climate Model Surface Temperatures, J. Climate, 29, 1253–1268, doi: http://dx.doi.org/10.1175/JCLI-D-15-0457.1, 2016.

Ghil, M.: Natural climate variability, in: Encyclopedia of Global Environmental Change, T. Munn (Ed.), Vol. 1, J. Wiley & Sons, Chichester/New York, pp. 544–549, ISBN 0-471-97796-9, 2002.

Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., Robertson, A.W., Saunders, A., Tian, Y., Varadi, F. and Yiou, P.: Advanced spectral methods for climatic time series, Reviews of geophysics, 40(1), 1–41, 2002.

Groth, A. and Ghil, M.: Monte Carlo Singular Spectrum Analysis (SSA) Revisited: Detecting Oscillator Clusters in Multivariate Datasets, Journal of Climate, 28(19), 7873–7893, 2015.

Kleeman, R.: Stochastic theories for the irregularity of ENSO, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 366(1875), 2509–2524, 2008.

Knutson, T. R., Zeng, F. and Wittenberg, A. T.: Multimodel Assessment of Regional Surface Temperature Trends: CMIP3 and CMIP5 Twentieth-Century Simulations, J. Climate, 26, 870–8743, 2013.

Mann, M. E. and Lees J. M.: Robust estimation of background noise and signal detection in climatic time series, Clim. Change, 33, 409–445, 1996.

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., Corti, S., Danabasoglu, G., Doblas-Reyes, F., Hawkins, E., Karspeck, A., Kimoto, M., Kumar, A., Matei, D., Mignot, J., Msadek, R., Navarra, A., Pohlmann, H., Rienecker, M., Rosati, T., Schneider, E., Smith, D., Sutton, R., Teng, H., van Oldenborgh, G., Vecchi, G. and Yeager, S.: Decadal Climate Prediction: An Update from the Trenches, Bull. Amer. Meteor. Soc., 95, 243–267, 2014.

North, G. R., Bell, T. L., Cahalan, R. F. and Moeng, F. J.: Sampling errors in the estimation of empirical orthogonal functions, Monthly Weather Review, 110(7), 699–706, 1982.

Plaut, G. and Vautard, R.: Spells of low-frequency oscillations and weather regimes in the Northern Hemisphere, Journal of the atmospheric sciences, 51(2), 210–236, 1994.

Poli, P., Hersbach, H., Tan, D., Dee, D., Thepaut, J. N., Simmons, A., Peubey, C., Laloyaux, P., Komori, T., Berrisford, P., Dragani, R., Trémolet, Y., Hólm, E. V., Bonavita, M., Isaksen, L. and Fisher, M.: The data assimilation system and initial performance evaluation of the ECMWF pilot reanalysis of the 20th-century assimilating surface observations only (ERA-20C), ERA Report Series, ECMWF, 2013.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, Journal of Geophysical Research: Atmospheres, 108(D14), 446–469, 2003.

Rayner, N. A., Brohan, P., Parker, D. E., Folland, C. K., Kennedy, J. J., Vanicek, M., Ansell, T. J., and Tett, S. F. B.: Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: The HadSST2 dataset, Journal of Climate, 19(3), 446–469, 2006.

Russell, A. M. and Gnanadesikan, A.: Understanding Multidecadal Variability in ENSO Amplitude, Journal of Climate, 27, 4037–4051, doi:10.1175/JCLI-D-13-00147.1, 2014.

Seitola, T., Mikkola, V., Silen, J. and Järvinen, H.: Random projections in reducing the dimensionality of climate simulation data, Tellus A, 66, doi:http://dx.doi.org/10.3402/tellusa.v66.25274, 2014.

Seitola, T., Silén, J. and Järvinen, H.: Randomised multichannel singular spectrum analysis of the 20th century climate data, Tellus A, doi:http://dx.doi.org/10.3402/tellusa.v67.28876, 2015.

Slingo, J. and Palmer, T.: Uncertainty in weather and climate prediction, Phil. Trans. R. Soc. A., 369, 4751–4767, 2011.

Stevenson, S., Fox-Kemper, B., Jochum, M., Rajagopalan, B. and Yeager, S. G.: ENSO model validation using wavelet probability analysis. Journal of Climate, 23(20), 5540–5547, 2010.

Taylor, K. E., Stouffer, R. J. and Meehl, G. A.: An Overview of CMIP5 and the experiment design, Bull. Amer. Meteor. Soc., 93, 485–498, 2012.

Thompson, D. W. J., Barnes, E. A., Deser, C., Foust, W. E., and Phillips, A. S.: Quantifying the role of internal climate variability in future climate trends. J. Climate, 28, 6443-6456, 2015.

Thomson, D. J.: Spectrum estimation and harmonic analysis, Proc. IEEE, 70, 1055–1096, 1982.

Vautard, R. and Ghil, M.: Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series, Physica D: Nonlinear Phenomena, 35(3), 395–424, 1989.

Wittenberg, A. T.: Are historical records sufficient to constrain ENSO simulations?, Geophys. Res. Lett., 36, L12702, 2009.

Wunsch, C.: The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations, Bulletin of the American Meteorological Society, 80, 245-255, 1999.

Table 1. Climate models used in the study. For more details of the models, see Table 9.1. in Flato et al. (2013).

Model ID	Model name	Modeling center	Country
a	CanESM2	CCCMA	Canada
b	CESM1(CAM5)	NSF-DOE-NCAR	USA
c	CNRM-CM5-2	CNRM-CERFACS	France
d	CSIRO-Mk3.6.0	CSIRO-QCCCE	Australia
e	GFDL-ESM2M	NOAA GFDL	USA
f	GISS-E2-R	NASA GISS	USA
g	HadGEM2-ES	MOHC	UK
h	INM-CM4	INM	Russia
i	IPSL-CM5B-LR	IPSL	France
j	MIROC-ESM	JAMSTEC/AORI/NIES	Japan
k	MPI-ESM-MR	MPI-M	Germany
l	MRI-CGCM3	MRI/JMA	Japan

~~Evaluation summary. Under-active On-target Over-active Multi-decadal (> 30 yr) b, c, d, e, f a, g, j h, i, k, l Decadal (10–30 yr) a, b, c, e, f, d, g, j, k h, i, l ENSO at ∼5 yr a, f, h, j, l c, g, i, k b, d, e ENSO at ∼3.5 yr c, f, j, k b, g, h, i, l a, d, e Inter-annual (<~~

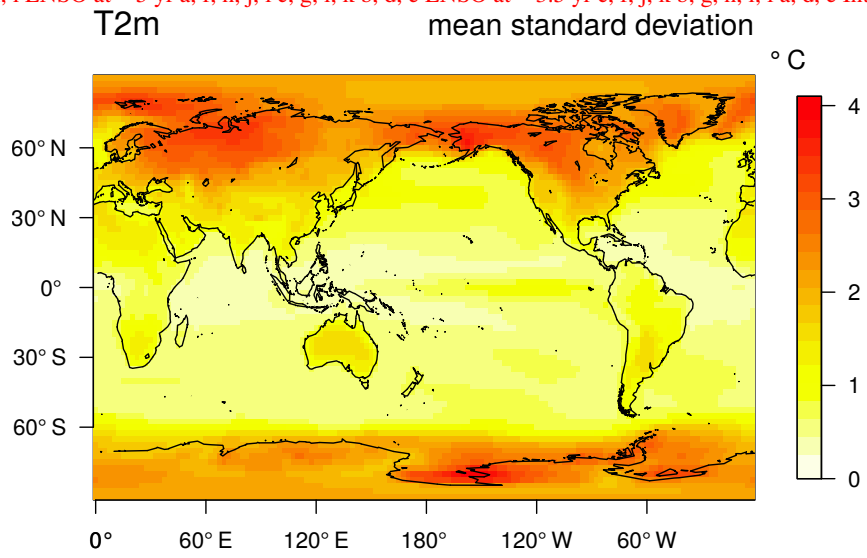


Figure 1. Map of the common normalisation factor. Shown is the mean standard deviation of 2 ~~yr) d, f, g, h, i, a, b, c, e j, k, l~~ metre temperature (degC) across all the data sets.

~~Approximate periodicities (in years) detected by the Monte-Carlo significance test with a 20 year lag window length. Similar periodicities among the data sets are aligned. Numbers in the table are in bold if the significance level is at 5 and in italics if at 10 . Dominant periodicities of the oscillations are estimated with MTM.~~

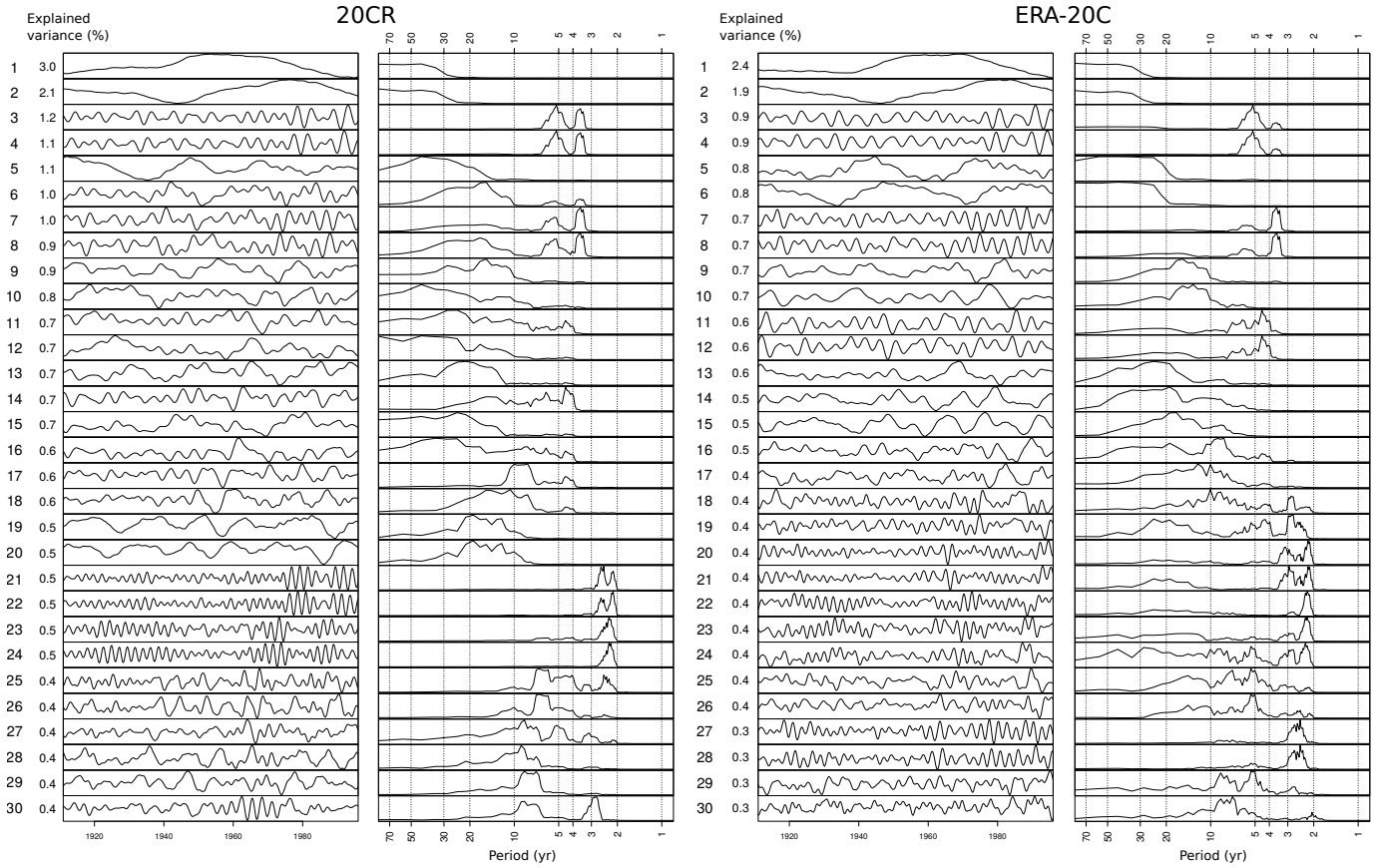


Figure 2. ~~ST-PCs 1–30 of 20CR~~ Reanalysis ST-PC time series (columns 1 and ~~ERA-20C~~ 3) of monthly near-surface temperature 1901–2005 and their spectra ~~. The lag window length M used in RMSSA is 20 years~~ (~~240 months~~columns 2 and 4) for 20CR and ERA-20C. The ~~annual cycle and linear trend~~ components are ~~removed from~~ ordered according to the ~~original data set before applying RMSSA~~explained variance (%).~~The proportion~~

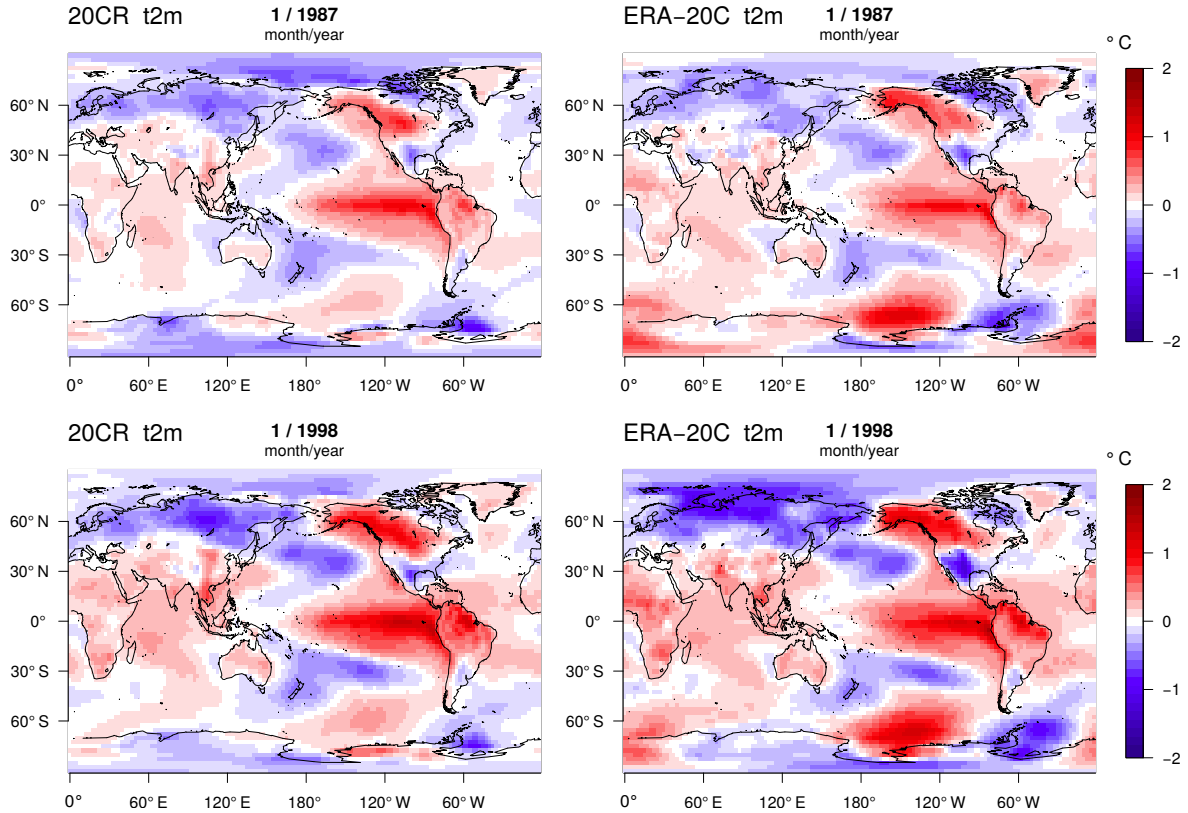


Figure 3. Global patterns of 2 metre temperature for the ~~remaining variance explained by each component is also presented~~ components 3, 4, 7 and 8 in ~~the figure~~20CR (left column) and ERA-20C (right column). Snapshots are taken from Jan 1987 (top row) and Jan 1998 (bottom row). Unit degC.

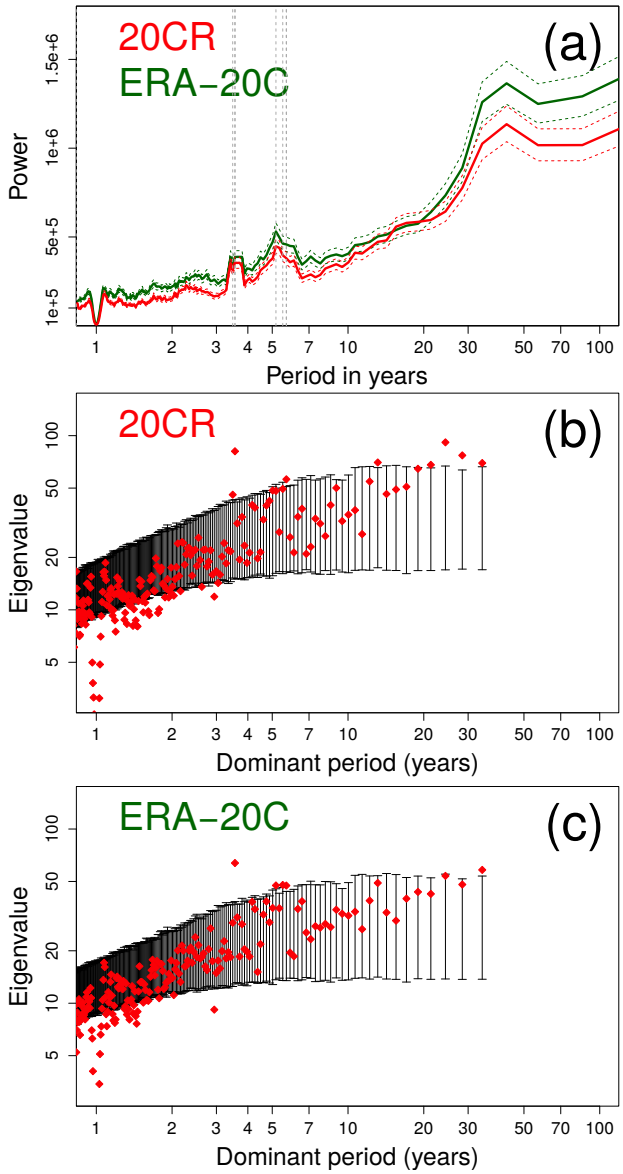


Figure 4. (a) Total spectrum of ~~the reanalysis data sets,~~ 20CR (red line) and ERA-20C ~~. The total spectrum is a sum of the spectral density of all components~~ (~~ST-PCs~~green line) ~~related to each data set~~with their min-max confidence intervals. The unit of the ~~spectra~~ spectral density is arbitrary. (b) Significance ~~test~~ of the ~~monthly near-surface temperature variability in the~~ 20CR ~~data set 1901–2005. PCs 1–50 are used as input channels and the lag window length M is 20 years (240 months). In the test the~~ periodicities against red-noise ~~basis is used~~null-hypothesis. ~~Red squares show~~ Shown are the data eigenvalues ~~plotted against the dominant frequency of the ST-PC corresponding to each eigenvalue. The vertical bars show~~ (red squares) and the 2.5^{th} and 97.5^{th} percentiles of the eigenvalue distribution ~~calculated from 1000 realisations~~ of the red-noise surrogates ~~. The ST-PCs that correspond to eigenvalues rising above the 97.5^{th} percentiles are considered significant at the 5 level. There is missing power at 1 yr due to the removal of the annual cycle~~(vertical bars). (c) Same as (b), but for the ERA-20C data set.

~~Total spectrum of each climate model. The total spectrum is the sum of the spectral density of all components (ST-PCs) related to each data set. For comparison, the total spectra of the reanalysis data sets are plotted in each subfigure. The unit of the spectra is arbitrary.~~

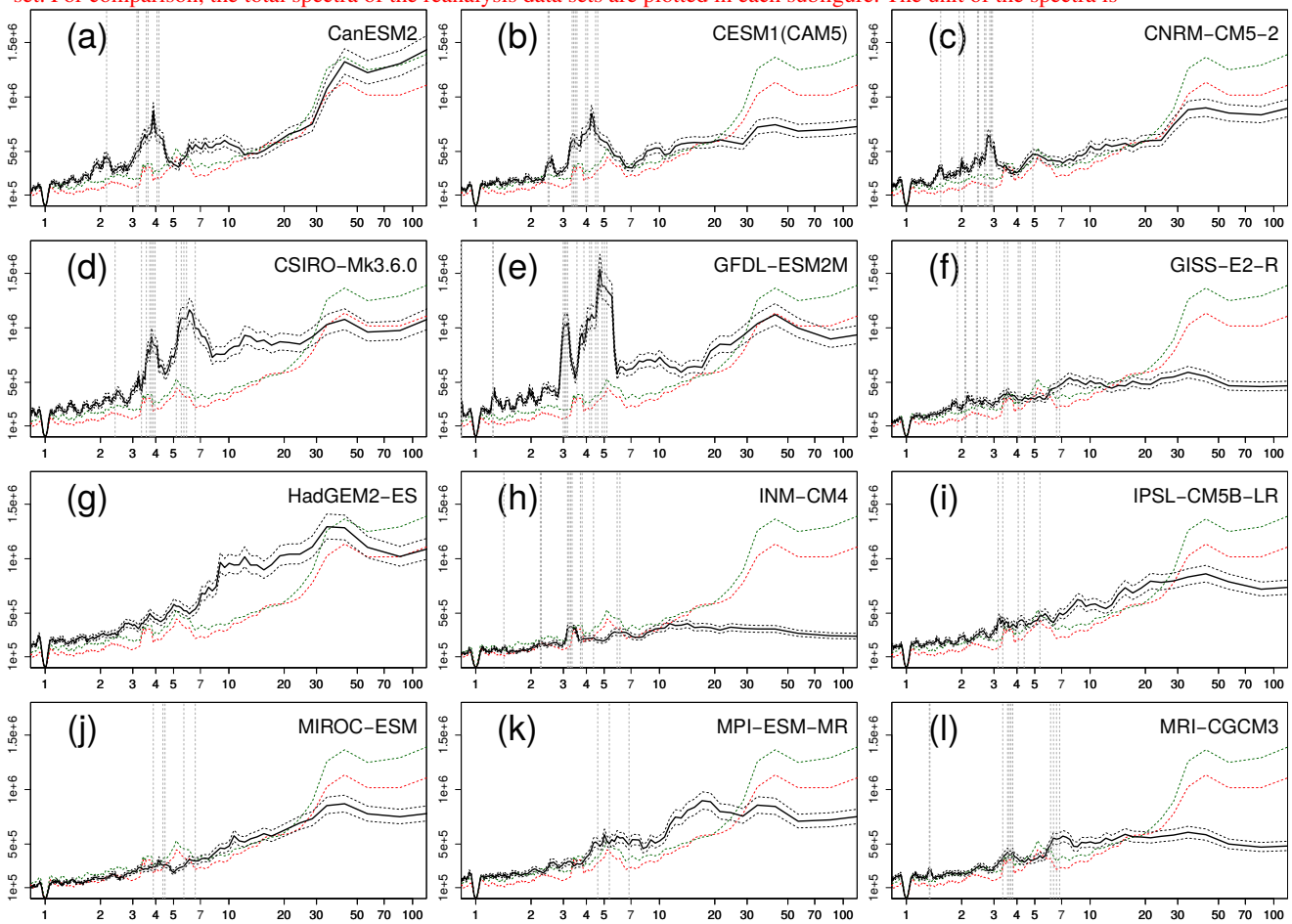
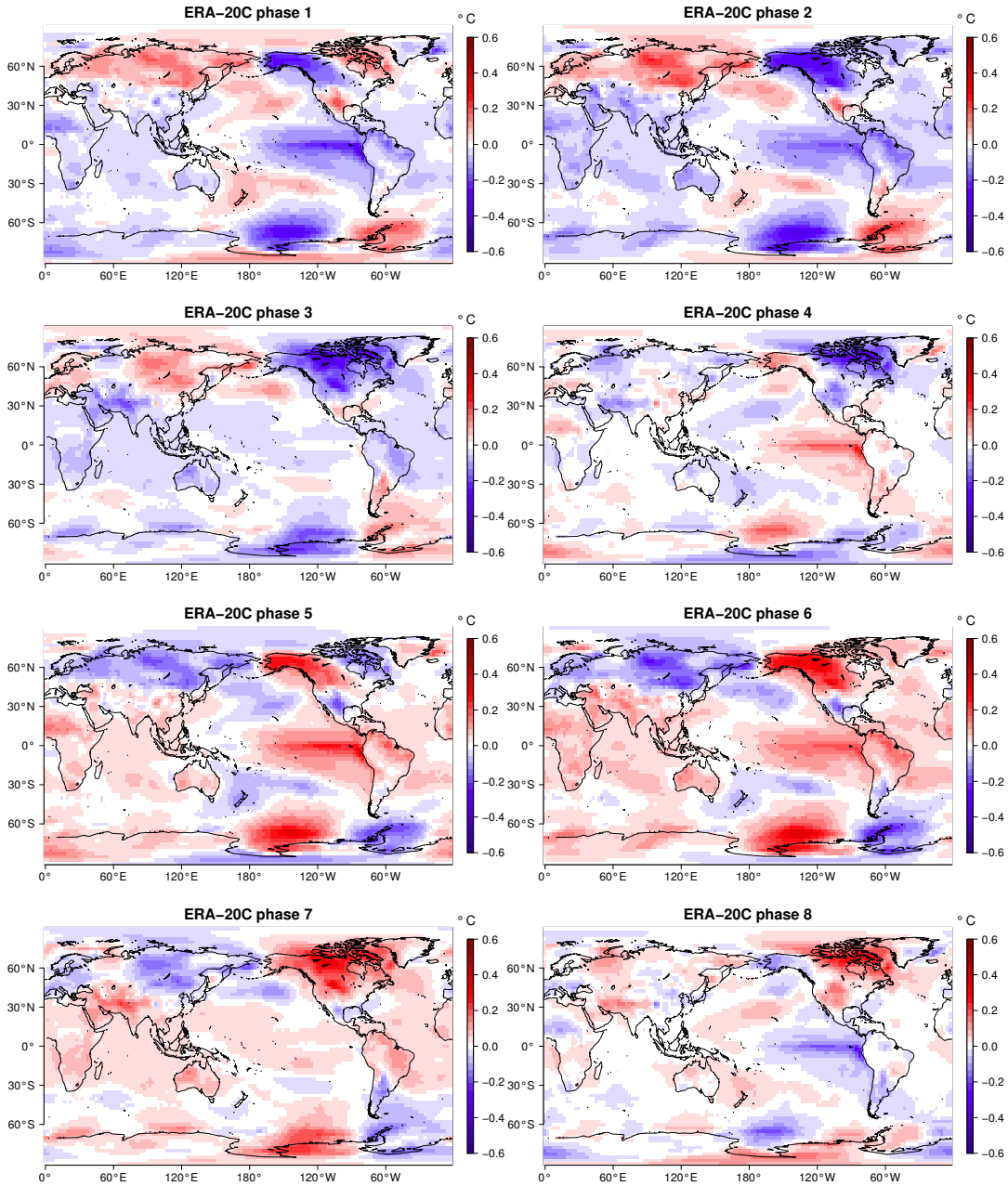


Figure 5. As Figure **??**a but now for each climate model (black line). The reanalysis spectra are shown as a reference (dashed green and red lines). The dashed vertical lines indicate the climate model multi-annual periods significant at 5 % level.

~~Global patterns of 3–4 yr oscillation of the near-surface temperature anomaly (°C) in the reanalysis and climate model data sets 1901–2005. The patterns are composites of 8 cases, when the oscillation is in its maximum positive phase in the Niño3.4 region (120° W–170° W, 5° N–5° S). The positive events are defined as an average of November-March. The identified patterns have similarities to the El Niño~~



~~-phenomenon.~~

Figure 6. ERA-20C phase (1–8) composites of the 3–4 yr variability mode. Unit degC.