Tellus SERIES A DYNAMIC METEOROLOGY AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Randomised multichannel singular spectrum analysis of the 20th century climate data

By TEIJA SEITOLA[1,2]*, JOHAN SILÉN[3] and HEIKKI JÄRVINEN[2], [1]*Climate Service Centre unit, Finnish Meteorological Institute, Helsinki, Finland*; [2]*Department of Physics, University of Helsinki, Helsinki, Finland*; [3]*Climate Research unit, Finnish Meteorological Institute, Helsinki, Finland*

(Manuscript received 24 June 2015; in final form 8 December 2015)

ABSTRACT

In this article, we introduce a new algorithm called randomised multichannel singular spectrum analysis (RMSSA), which is a generalisation of the traditional multichannel singular spectrum analysis (MSSA) into problems of arbitrarily large dimension. RMSSA consists of (1) a dimension reduction of the original data via random projections, (2) the standard MSSA step and (3) a recovery of the MSSA eigenmodes from the reduced space back to the original space. The RMSSA algorithm is presented in detail and additionally we show how to integrate it with a significance test based on a red noise null-hypothesis by Monte-Carlo simulation. Finally, RMSSA is applied to decompose the 20th century global monthly mean near-surface temperature variability into its low-frequency components. The decomposition of a reanalysis data set and two climate model simulations reveals, for instance, that the 2–6 yr variability centred in the Pacific Ocean is captured by all the data sets with some differences in statistical significance and spatial patterns.

Keywords: multichannel singular spectrum analysis, random projection, dimensionality reduction, El Niño – southern oscillation, 20th century reanalysis, HadGEM2-ES, MPI-ESM-MR

1. Introduction

Our motivation to focus on advanced spatio-temporal data analysis is to better understand the decadal climate variability in the Earth system and illuminate the capabilities of prediction tools to capture the associated signals (Meehl et al., 2014). Inter- and intra-decadal climate variability is inherent to the ocean–atmosphere system and is further coupled to other Earth system components, such as sea-ice and land surface (Meehl et al., 2009). The variability appears as complex four-dimensional (or spatio-temporal) structures in Earth system variables, such as wind, temperature and precipitation (Solomon et al., 2011).

These structures are embedded in extremely large-dimensional data sets gathered and generated in reanalysis of atmospheric and oceanic observations, and in massive simulation endeavours using Earth system models worldwide. Applicability of advanced data analysis tools is severely hampered by the very large dimensionality of the climate data.

Many common analysis methods, such as principal component analysis (PCA; Von Storch and Zwiers, 1999), involve eigen-problems, which become impossible to solve with increasing data dimension. Earlier we illustrated the use of random projections (RP) as a tool to tackle high-dimensional problems (Seitola et al., 2014). We demonstrated how PCA can be applied in three-dimensions to problems that are beyond practical computational limits without efficient dimension reduction. PCA is not an ideal tool, however, to extract and illustrate four-dimensional eigen-features in climate data. In this respect, the multichannel singular spectrum analysis (MSSA; Broomhead and King, 1986a, b) is a much more appealing method since the MSSA eigen-problem inherently contains the auto-covariance in the lagged copies of the original data vectors. The computational burden is, however, even larger than in PCA. We overcome this burden by a novel randomised version of MSSA, called RMSSA. To our knowledge, this approach has not been suggested before. We note that Oropeza and Sacchi (2011) incorporate a randomising operator into MSSA for noise attenuation in seismic data, but their algorithm is not aimed directly at large-dimensional problems. In RMSSA, RPs are used essentially to enable analysis of extremely large-dimensional data sets.

*Corresponding author.
email: teija.seitola@fmi.fi

Citation: Tellus A 2015, **67**, 28876, http://dx.doi.org/10.3402/tellusa.v67.28876

In this article, we present the RMSSA algorithm in detail and also include a test for the statistical significance of the results (Monte-Carlo MSSA; Allen and Robertson, 1996) in the algorithm. We demonstrate the use of RMSSA by decomposing the 20th century global monthly mean near-surface temperature variability into its low-frequency components. The data sources are described in Section 2.3.

2. Methods and Data

2.1. *Multichannel singular spectrum analysis*

MSSA was introduced into the study of dynamical systems by Broomhead and King (1986a, b). The method is equivalent to extended empirical orthogonal function (EEOF) analysis (Weare and Nasstrom, 1982), but there are differences in the choice of some important parameters and in the interpretation of the results (Plaut and Vautard, 1994).

In traditional PCA or EOF analysis (e.g. Rinne and Karhila, 1979), spatial correlations (in case of climatic data sets) are used in determining the patterns that explain most of the variability in the data set, but MSSA differs from this traditional method by also taking into account the temporal correlations. In other words, standard PCA decomposes a spatio-temporal field into spatial PC loading patterns (EOFs) and corresponding PC score time series (PCs), whereas MSSA also adds a temporal dimension to EOFs. MSSA PCs and EOFs are often called space-time PCs (ST-PCs) and space-time EOFs (ST-EOFs), and we have adopted this notation here. A more detailed description of MSSA is presented in Ghil et al. (2002) and in Appendix A.1 here.

2.1.1. Choice of the lag window. The idea of MSSA, in brief, is to find the patterns that maximise the lagged covariance of the data set $\mathbf{X}_{N \times L}$ within M lags. In case of a gridded climate data set, N represents the time steps and L is the number of grid points. The columns of the data matrix $\mathbf{X}$ are often called channels. The length of the lag window M is a user choice. For example, Elsner and Tsonis (1996) suggest that the results of MSSA do not change significantly with varying M as long as $N>>M$ and they recommend using $M=N/4$. Vautard and Ghil (1989) recommend to choose M no larger than approximately $N/3$. Clearly, if the number of channels L is large in the beginning, choosing large M would result in a very high-dimensional data matrix with $M \times L$ columns, including lagged copies of each channel in $\mathbf{X}$.

Determining the length of the lag window M is a trade-off between spectral resolution and statistical significance of the obtained components. The larger M is chosen, the more temporal information can be extracted but at the same time the variance is distributed on a larger set of components. If M is small, the statistical significance of the obtained components is enhanced. In this study, we used several values of M in order to test its effects on the results.

2.1.2. Assessing statistical significance with Monte-Carlo MSSA. ST-PCs/ST-EOFs often appear in pairs ('sinusoidal') that explain approximately the same variance and are $\pi/2$ out of phase with each other. These pairs are said to present stationary or propagating oscillatory modes of the data set (Plaut and Vautard, 1994). Modes with period less than or equal to M can be only presented by such pairs. However, existence of such a pair does not guarantee any physical oscillation, and according to Allen and Robertson (1996) such pairs can also be generated by non-oscillatory processes, such as first-order autoregressive noise.

This finding led Allen and Robertson (1996) to formulate a test for the statistical significance of MSSA components. The identified components are tested against a null-hypothesis of the data being generated by independent AR(1) processes (i.e. red noise) with the same variance and lag-1 autocorrelation as the original input time series. This procedure is called Monte-Carlo MSSA (MC-MSSA), and it is described in more detail in the original study of Allen and Robertson (1996) as well as in Appendix A.1 of this article.

2.1.3. Reconstructed components. ST-PCs cannot be compared to the original time series as such; instead, they can be represented in the original coordinate system by their reconstructed components, RCs (Plaut and Vautard, 1994; Ghil et al., 2002). In the reconstruction, the ST-PCs are projected back onto the eigenvectors (ST-EOFs) and each RC is a kind of filtered version of the original multivariate time series. Construction of RCs is illustrated in Fig. 1. Several ST-PCs/ST-EOFs can be used in the reconstructions, and if there is an oscillation that appears as a sinusoidal pair, both of these ST-PCs/ST-EOFs should be included in the reconstruction of that certain oscillatory mode. This is done by summing up the corresponding RCs. No information is lost in the reconstruction, and the original time series is a sum of all individual RCs.

2.2. *Randomised algorithm for MSSA*

As mentioned earlier, the computational burden of MSSA becomes soon prohibitively high if the original data set is high-dimensional and M is chosen to be large. This is typically the situation in studies of low-frequency variability in climate data sets. Traditionally, the dimensionality reduction has been obtained by calculating first a conventional PCA and retaining a set of dominant PCs for the MSSA

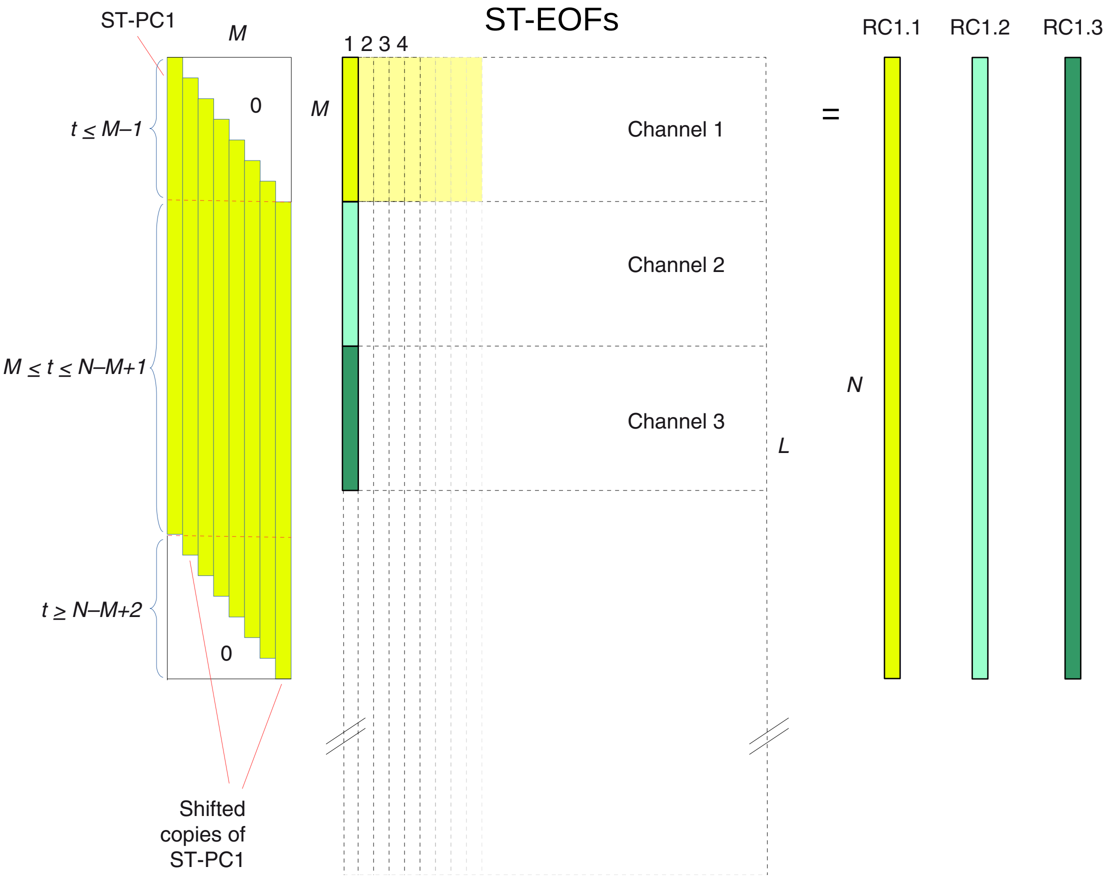


Fig. 1. Example calculation of the reconstructed components (RCs). A matrix of M shifted copies of a ST-PC (ST-PC1 in this example) is constructed to calculate reconstructions of that ST-PC in a time series of each channel (grid point). This matrix is then multiplied with that part of ST-EOF that corresponds to each channel. If $t \leq M$, the elements of RC are divided by t, if $M \leq t \leq N-M+1$, divided by M, and if $t \geq N-M+2$, divided by $N-t+1$.

(see chapter 2.2.3). However, in this article we apply a different approach to dimensionality reduction. That is, we use RP to reduce the dimensionality of the original data set before performing MSSA.

In Halko et al. (2011), it is stated that randomised methods provide a powerful tool for constructing approximate matrix factorisations. Compared with standard deterministic algorithms, the randomised methods are often faster and more robust. Halko et al. (2011) present also numerical evidence that these algorithms succeed for real computational problems.

2.2.1. Description of RMSSA. In our approach, RP is applied to reduce the dimension of the original data matrix $\mathbf{X}$ after which the traditional MSSA calculation is performed in the lower-dimensional subspace. Finally, we reconstruct the ST-EOFs and RCs in the original space. We call this algorithm randomised multichannel singular spectrum analysis (RMSSA).

In RP, the original data set is projected onto a matrix $\mathbf{R}$ of Gaussian distributed (zero mean and unit variance) random numbers in order to construct a lower-dimensional representation $\mathbf{P}$ of the data set:

$$\mathbf{P}_{N\times k} = \mathbf{X}_{N\times L}\mathbf{R}_{L\times k} \quad (1)$$

In other words, we are projecting our data set onto k random directions determined by the column vectors of $\mathbf{R}$. From these projections a lower-dimensional representation of the original data set can be constructed. Due to the simplicity of RP, involving only matrix multiplication, it can be applied to a wide range of data sets, even very high-dimensional ones.

RP has already been applied to climate data in Seitola et al. (2014) and it has been shown to preserve structures of the original data very well. In that article, the theoretical background of RP is presented in more detail with additional references.

The projected lower-dimensional data set **P** can be processed through MSSA where instead of original L channels we have now only k channels. This implies substantial computational savings (see Appendix A.2, algorithm 1). In the literature, there are some estimates of an appropriate value for k (e.g. Frankl and Maehara, 1988; Dasgupta and Gupta, 2003). However, these theoretical lower bounds for k are the worst case estimates and usually much lower values for k still give good results, retaining most of the information of the original data set (see e.g. Bingham and Mannila, 2001; Seitola et al., 2014). In practice, the value for k is usually chosen adaptively keeping the desired size for lower-dimensional approximation in mind.

A final step of the algorithm is to calculate the reconstructed components. This requires the recovery of the MSSA eigenmodes from the reduced space back to the original space, allowing the reconstruction of the original time series. This means that the eigenvectors (ST-EOFs) should be calculated in the original space instead of the reduced one. This part of the algorithm is also presented in Appendix A.2. Furthermore, in Appendix A.4 we explain how RP preserves the lagged covariance structure of the original data set.

2.2.2. Comparison of RMSSA to previous work. To our knowledge, the proposed RMSSA algorithm is unique. Some published work comes close to our approach but RMSSA has some important differences to the randomised MSSA algorithms used in seismic data processing (Oropeza and Sacchi, 2011; Chiu, 2013). The aim of Chiu (2013) was to introduce a new rank-based-reduction denoising algorithm to perform coherent and random noise filtering concurrently. Chiu (2013) named this algorithm, or rather filter, MSSARD (MSSA in the randomised domain). In MSSARD, the randomising operator randomly rearranges the order of the input data and reorganises the coherent noise into incoherent noise. The most important difference to our algorithm is in the randomising operator: In our case, we are using RP to reduce the dimensionality of the input data whereas Chiu's (2013) approach is to randomly rearrange the input data.

The technique of Oropeza and Sacchi (2011) was to embed a spatial data at a given temporal frequency into a block Hankel matrix after which a randomised singular value decomposition (SVD) was adopted to accelerate the rank reduction stage of the algorithm. Construction of a Hankel matrix corresponds to the construction of an augmented data matrix **A** in our algorithm (see Appendix A.1). Our algorithm is different in the sense that we apply RP on the original input data before construction of the augmented (or Hankel) matrix. This notably reduces the computational burden of MSSA because we are processing a much smaller data set already in the augmentation phase of the algorithm (see algorithm 1 in Appendix A.2).

In addition to these main differences, the above-mentioned seismological applications involve handling a data set where each time/frequency slice of spatial (x-y) data is processed separately through the algorithm. In our case, we are processing the whole time–longitude–latitude data set at once through the RMSSA algorithm.

2.2.3. Enhancing PC-based MSSA. In many studies, where MSSA is used as an analysis method (e.g. Plaut and Vautard, 1994; Moron et al., 2012), the dimension of the original data matrix has been reduced by calculating a conventional PCA of the original data matrix and then limiting MSSA into the dominant PCs. One has to bear in mind that the problem dimension may be prohibitive to contemplate solving even PCA, let alone MSSA. Nevertheless, the number of retained PCs is a somewhat arbitrary choice, but can be estimated by inspecting the eigenvalue spectrum and choosing the PCs that account for the majority of the variance and are separated from the rest of the spectrum. In geophysical datasets, however, the eigenvalue spectrum often decreases monotonically and it is difficult to distinguish the appropriate cut-off point. The aim of the study does also affect the choice of the PCs. For example, if the focus is on large-scale patterns, it might be more convenient to choose the low-frequency PCs for further analysis. Performing the calculations with different number of PCs and comparing the results can also help in finding the appropriate number of PCs. Importantly, RMSSA (Appendix A.2, algorithm 1) does not suffer from this problem because the lower-dimensional data set has essentially the same structure as the original high-dimensional data set.

PCA-based dimensionality reduction is, however, a preferred method if the oscillatory modes identified with MSSA are tested against a red noise null-hypothesis through Monte-Carlo simulation. According to Allen and Robertson (1996) the test is only useful if the channels in the data matrix are orthogonal or at least very nearly so. The PCs fulfil the orthogonality condition exactly. The randomised method can still accelerate – and in the case of a very-high-dimensional data set even enable – the calculation of the PCs (see Appendix A.2, algorithm 2).

This also raises the question as to whether the projected data set [i.e. matrix **P** in eq. (1)] could be used directly in MC-MSSA. Like the PCs, RP is also an orthogonal projection and the columns of **P** are also nearly orthogonal. However, this question is beyond the scope of this study and will not be discussed here any further.

2.3. Data

As an illustration of applying the RMSSA algorithm, we analysed the monthly mean near-surface air temperature field from the 20th Century Reanalysis V2 data, hereafter 20CR, provided by the NOAA/OAR/ESRL PSD (Compo et al., 2011). In addition, we repeated the analysis for the historical 20th century simulations by Hadley Global Environment Model 2 – Earth System HadGEM2-ES (Collins et al., 2011), hereafter HadGEM2, and MPI Earth System Model (ESM) running on a medium resolution grid MPI-ESM-MR (Stevens et al., 2013), hereafter MPI-ESM. We extended the historical simulations (1901–2005) until 2012 using the rcp45 simulations. The historical and rcp45 simulations were extracted from the CMIP5 data archive and they follow the CMIP5 experimental protocol (Taylor et al., 2012). In the 20th century simulations, the historical record of climate forcing factors such as greenhouse gases, aerosols and natural forcings such as solar and volcanic changes is used. Rcp45 simulations follow the RCP4.5 greenhouse gas scenario. We used a single ensemble member of each model: r2i1p1 in case of HadGEM2 and r1i1p1 in case of MPI-ESM.

The 20CR data set is produced using an ensemble of perturbed reanalyses, and the final data set corresponds to the ensemble mean. Only surface pressure observations are assimilated, and the observed monthly sea-surface temperature and sea-ice distributions are used as boundary conditions to generate full three-dimensional estimates of the state of the troposphere (Compo et al., 2011). The 20CR data set is available from 1871 to 2012 but to be consistent with HadGEM2 and MPI-ESM simulations, the time sequence analysed here is 1901–2012 (1344 time steps). 20CR has $\sim$2.0 degree horizontal resolution and we have used Gaussian gridded (192×94) data from 3-hour forecast values. HadGEM2 and MPI-ESM have both a global grid of 144×73 points. Thus, we have original data sets $\mathbf{X}_{N \times L}$ with $N = 1344$, $L = 18048$ (20CR) and $L = 10512$ (HadGEM2 and MPI-ESM).

As an illustrative example of the high-dimensionality of the MSSA problem, let's choose a lag window of $M = 240$ (months). In the case of the 20CR data set, this would result in an augmented matrix with $M \times L = 4331520$ columns. Clearly some kind of dimensionality reduction is needed in order to make the computations more efficient or even make them possible.

3. Results

3.1. Application of RMSSA to climatic data sets

In the previous section, we have introduced the RMSSA algorithm and the data sets to be analysed. Next we will proceed to the applications of the proposed method and discuss the results.

First, the original data sets were mean centred and RMSSA (algorithm 1 in Appendix A.2) was applied with $k = 500$. The first 1–30 ST-PCs of 20CR are shown in Fig. 2. In order to find the most powerful frequencies associated with the ST-PCs, the Multitaper spectral analysis method (Thomson, 1982; Mann and Lees, 1996) was applied. The power spectra of the ST-PCs are shown on the right in Fig. 2. The first pair of ST-PCs is clearly related to the annual cycle and this pair together explains the majority of the variance of the data set (almost 90%). The pairs of ST-PCs 3–4, 7–8 and 12–13 are the subharmonic frequencies of the annual cycle. The periods of ST-PCs 5, 6 and 11 as well as of ST-PCs 14, 17 and 18 fall outside the lag window length M and are the so-called trend components. ST-PC5 may be related to a centennial scale warming trend and ST-PC11 has a multi-decadal scale variability. ST-PCs 22 and 24 have clear spectral peaks on a 5–6 yr period and ST-PCs 29 and 30 are oscillating on a period of 3–4 yr. Those ST-PCs might be related to the El Niño-Southern Oscillation (ENSO) which is a prominent phenomenon on those time scales. ST-PCs 19–21 are related to a decadal scale variability, but the spectra of those components are quite broad on a 10–20 yr time scale.

The above analysis was also performed for the HadGEM2 and MPI-ESM data sets (figures not shown). As the annual cycle is too dominating in each data set, the analysis in the following sections will be repeated without the annual cycle. We also integrate a MC-MSSA step in the RMSSA algorithm (Appendix A.2, algorithm 2) in order to study the statistical significance of the obtained components.

3.1.1. Pre-processing the data for Monte-Carlo MSSA. Some pre-processing of the original data sets was crucial in order to assess the statistical significance of the low-frequency variability using MC-MSSA. First of all, the original data sets were standardised (i.e. the time series of each grid point was mean centred and divided by its standard deviation) in order to avoid overweighting the grid points with higher variance. Furthermore, the annual cycle of the time series of each grid point was estimated by STL (Loess based Seasonal-Trend Decomposition) and removed from the original data set. The STL method is a filtering procedure for decomposing a time series into trend, seasonal and remainder components. It includes some parameter choices controlling, for instance, how rapidly the trend and seasonal components can change. The method is described in detail in Cleveland et al. (1990) and we have followed their guidelines in choosing the related parameters. Without this procedure the annual cycle would dominate the results and starve the lower ranked MSSA components of power when tested against the red noise null-hypothesis. Linear trends were also fitted and removed from the data sets in order to avoid the dominance of the centennial scale trend.

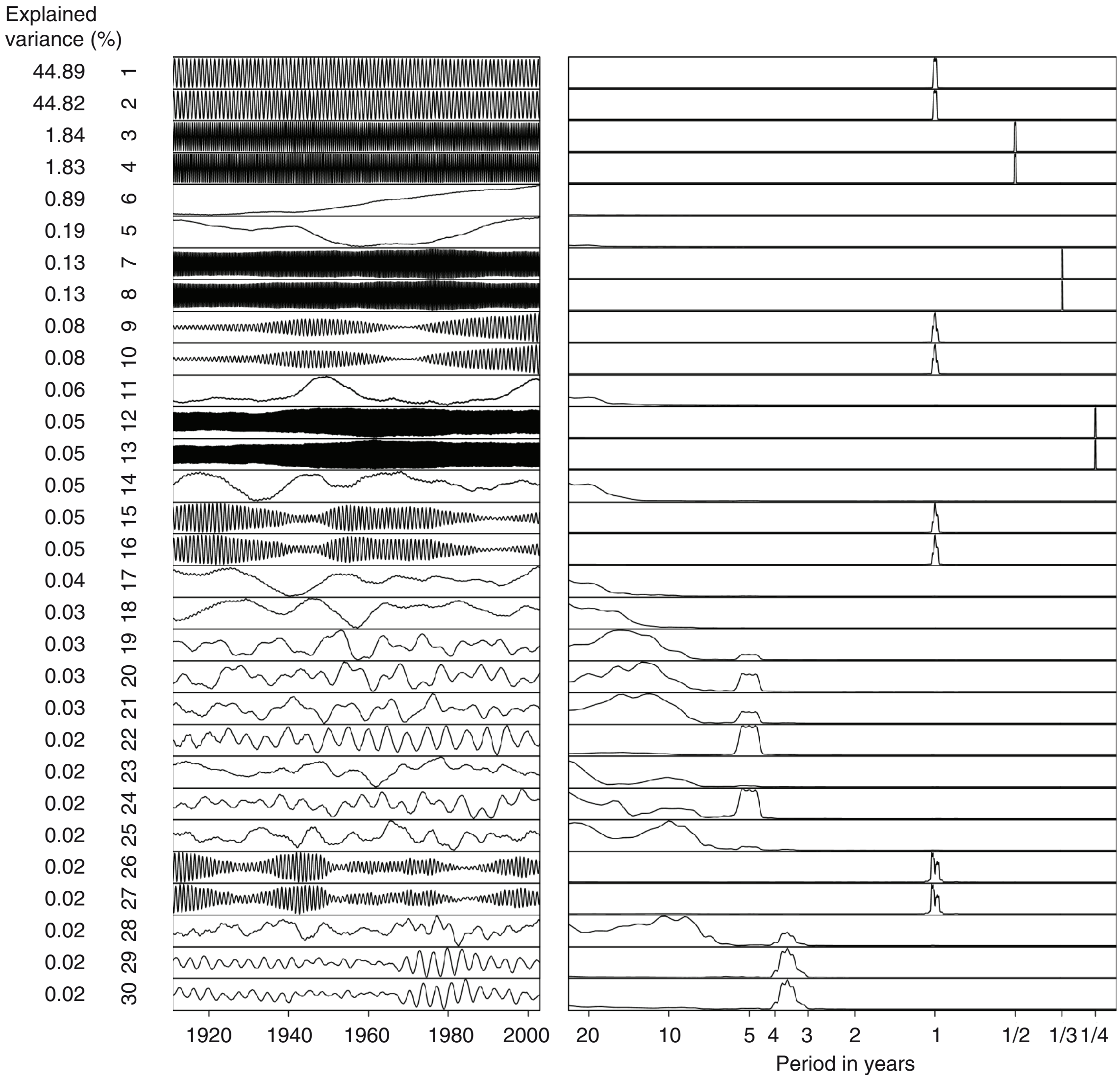


Fig. 2. ST-PCs 1–30 of 20CR monthly near-surface temperature 1901–2012 and their spectra. The lag window length M used in RMSSA is 20 yr (240 months). The data set is centred and algorithm 1 of Appendix A.2 is applied. The proportion of the variance explained by each component is also presented in the figure.

For the sake of comparison, the annual cycle was also estimated by calculating the mean values of each calendar month and those values were subtracted from the data to get monthly anomaly time series. However, determining the base for the anomaly calculation is not that straightforward and the choice of a base period may have severe impacts on the results (Kawale et al., 2011). Furthermore, the average annual cycle is only removed and if the annual cycle varies in the time series, the anomalies still contain a residual annual cycle.

The dimensions of the original data sets were reduced by applying RP with $k = 500$ to have a lower-dimensional approximation $\mathbf{P}_{N\times k}$ of each data set. To be able to perform MC-MSSA, we further calculated SVD of $\mathbf{P}$ and retained 30 first PCs of each data set, explaining approximately 72% (20CR), 67% (HadGEM2) and 64% (MPI-ESM) of the variance. Those 30 PCs were used as input channels in the MC-MSSA-step.

3.1.2. Decomposition of the pre-processed data sets. The ST-PCs 1–30 of each data set and their spectra are presented in Figs. 3–5. These figures show the results after applying the steps 1–8 of algorithm 2 in Appendix A.2 (note that the annual cycle and linear trend were removed from the original data sets). In 20CR (Fig. 3), the ST-PCs 1–2 are so-called trend components explaining together almost 9% of the variability of the data set. Pairs of

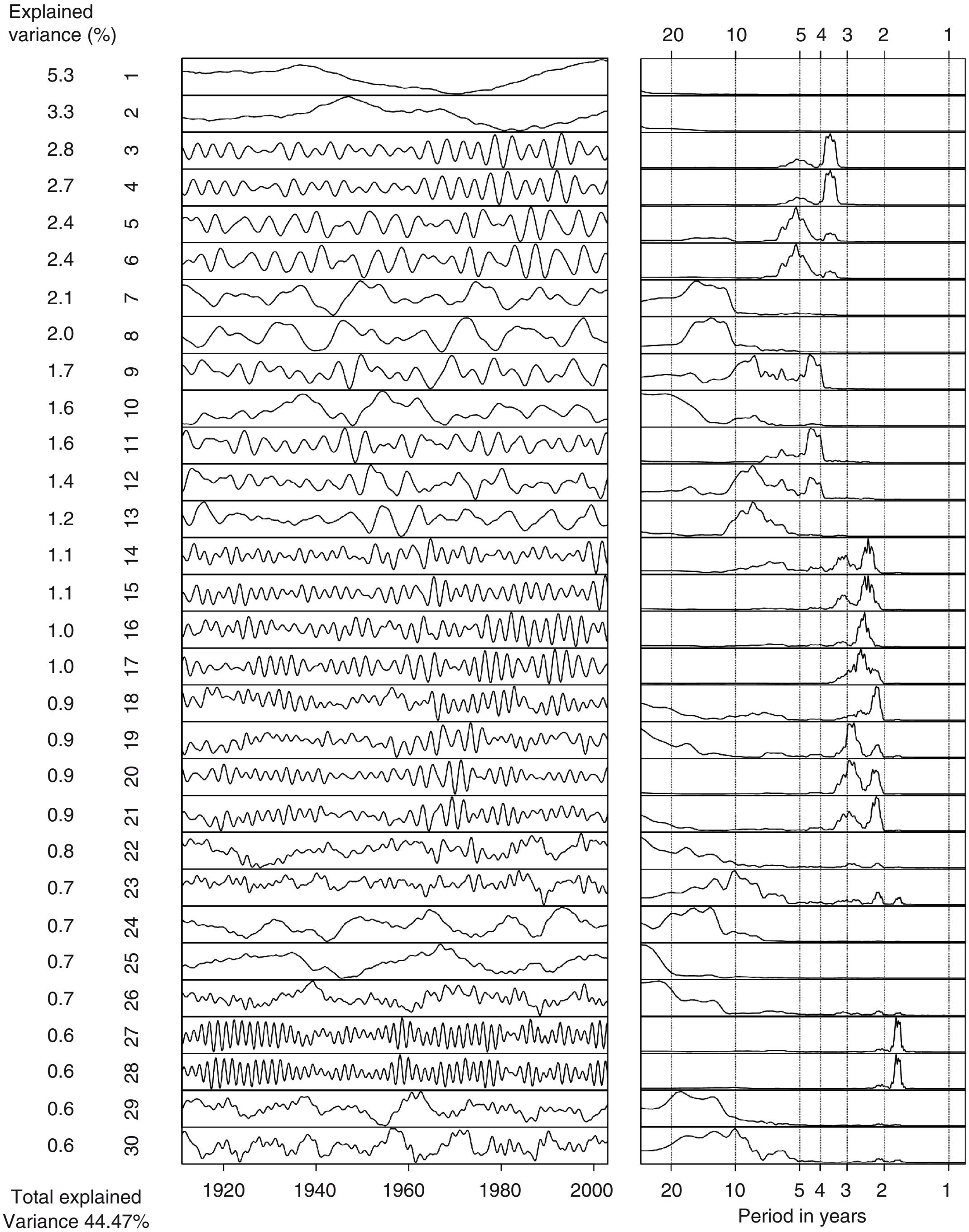


Fig. 3. ST-PCs 1–30 of 20CR monthly near-surface temperature 1901–2012 and their spectra. The lag window length M used in RMSSA is 20 yr (240 months). The annual cycle and linear trend are removed from the original data set and algorithm 2 of Appendix A.2 is applied. The proportion of the remaining variance explained by each component is also presented in the figure.

ST-PCs 3–4 and 5–6 in 20CR have clear peaks in frequencies corresponding to 3–4 yr and over 5 yr periods. In addition, 2–3 yr periodicities are distributed on several ST-PCs beginning from the 14th one. When the model simulations are compared to the 20CR components, the main differences are the prominent decadal scale components of HadGEM2 (ST-PCs 2–3, 9.3% of explained variance) and the 2–7 yr variability of MPI-ESM that is distributed on a large set of successive components. For more details, the readers are advised to study Figs. 3–5.

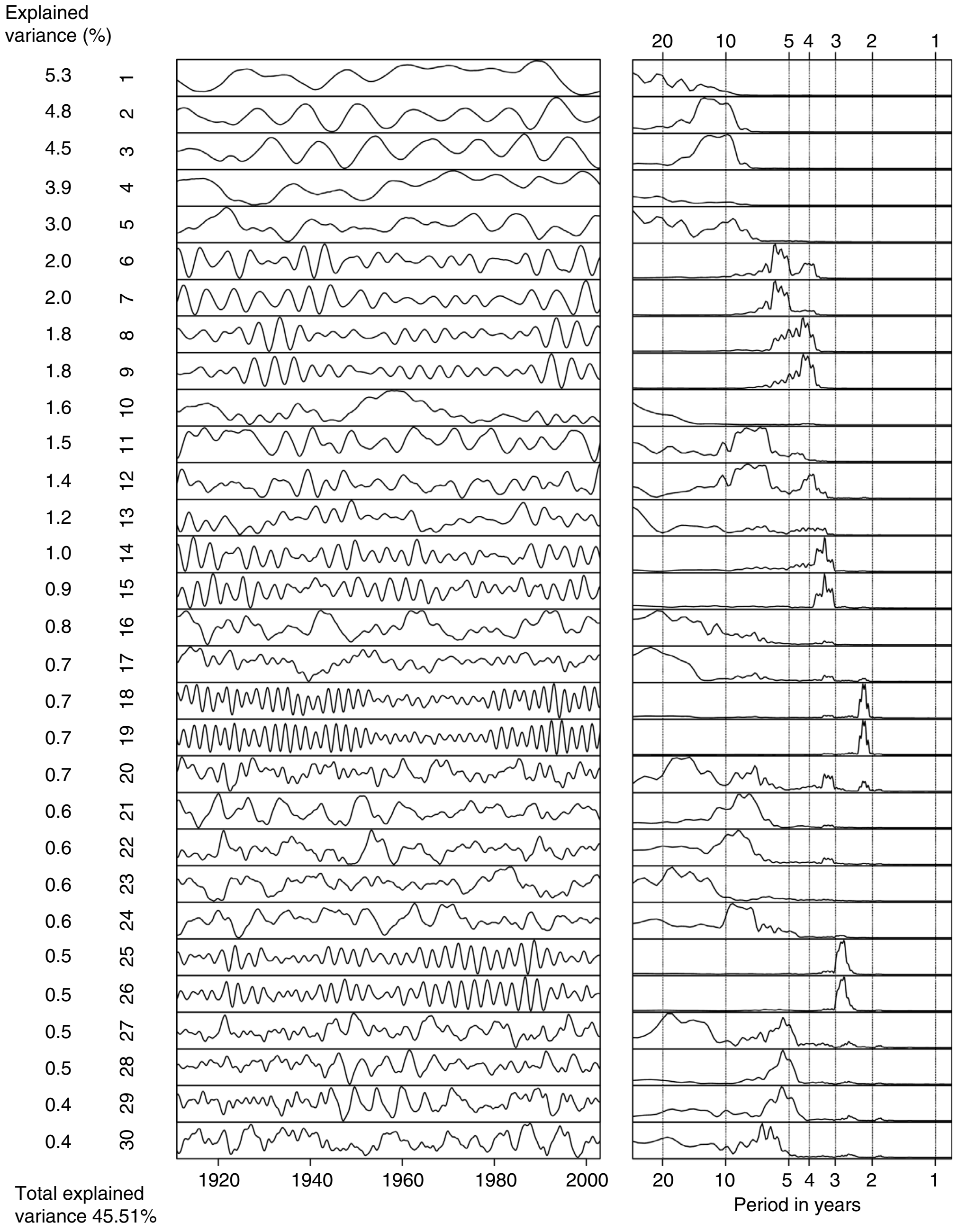


Fig. 4. ST-PCs 1–30 of HadGEM2 monthly near-surface temperature 1901–2012 and their spectra. The lag window length M used in RMSSA is 20 yr (240 months). The annual cycle and linear trend are removed from the original data set and algorithm 2 of Appendix A.2 is applied. The proportion of the remaining variance explained by each component is also presented in the figure.

3.2. *Identifying significant oscillations*

In MC-MSSA step, in total of 1000 realisations of red-noise surrogates were generated and the red-noise basis was used to estimate the 90, 95 and 99% confidence intervals for the eigenvalues generated by the noise model that consists of independent first-order autoregressive processes. Figure 6 shows the results of the Monte-Carlo significance test of 20CR, HadGEM2 and MPI-ESM with a 20 yr lag window ($M=240$ months). In that figure, the data eigenvalues and 2.5th and 97.5th percentiles of the distribution of the surrogate eigenvalues are plotted against the dominant frequencies of the corresponding red-noise basis vectors (noise ST-EOFs). The dominant frequencies are estimated using fast

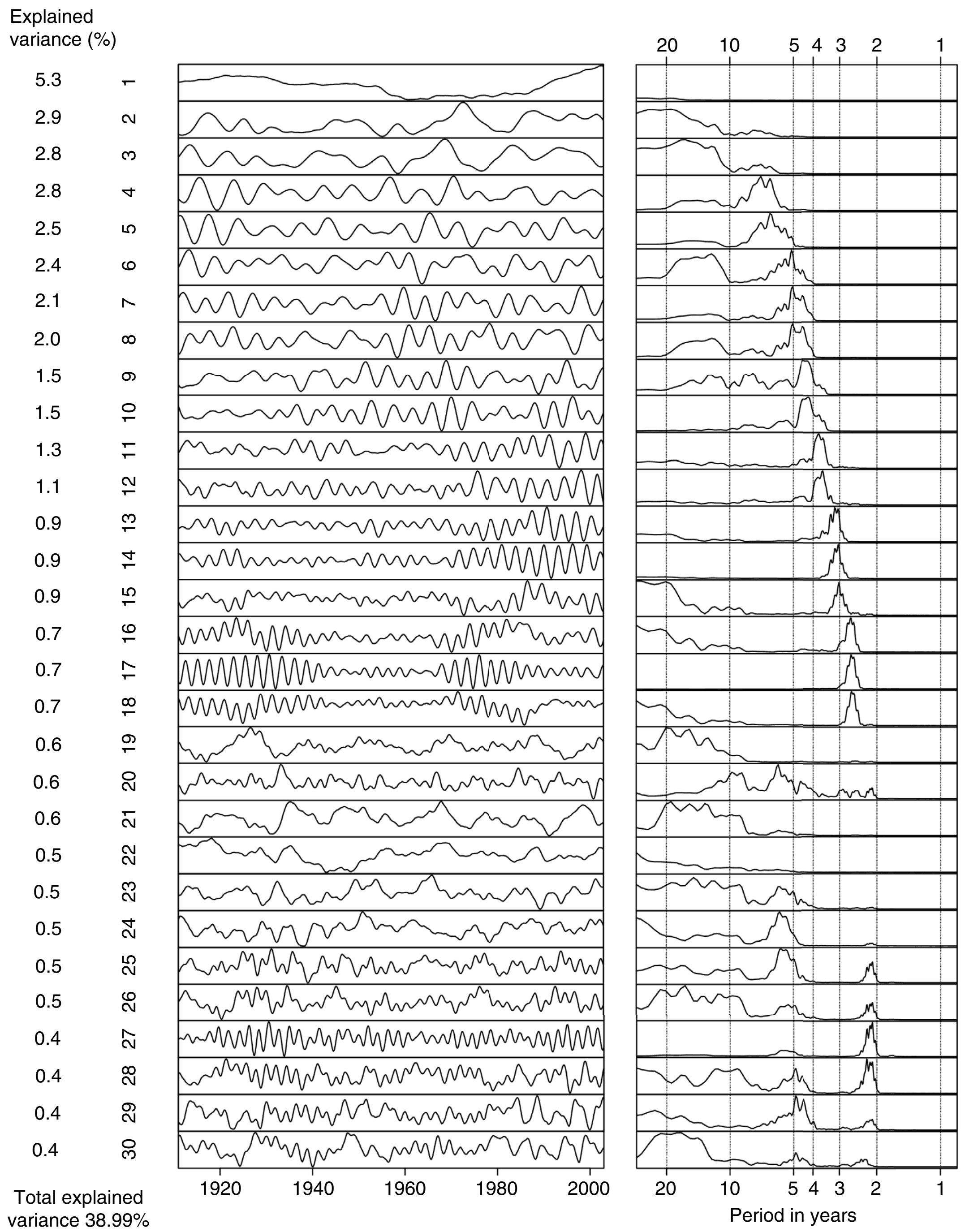


Fig. 5. ST-PCs 1–30 of MPI-ESM monthly near-surface temperature 1901–2012 and their spectra. The lag window length M used in RMSSA is 20 yr (240 months). The annual cycle and linear trend are removed from the original data set and algorithm 2 of Appendix A.2 is applied. The proportion of the remaining variance explained by each component is also presented in the figure.

Fourier transform (FFT). It should be noted, that the estimate of the dominant frequency of the noise ST-EOFs may not be exactly the same as the dominant frequency of the data ST-EOFs which may cause some small inaccuracies in the results.

The significant signals (at 5% significance level) in Fig. 6 are those whose data eigenvalues lie above the 97.5th percentiles of the surrogate eigenvalues. According to the test these signals have more variance than would be expected to have from a noise process. According to Plaut and

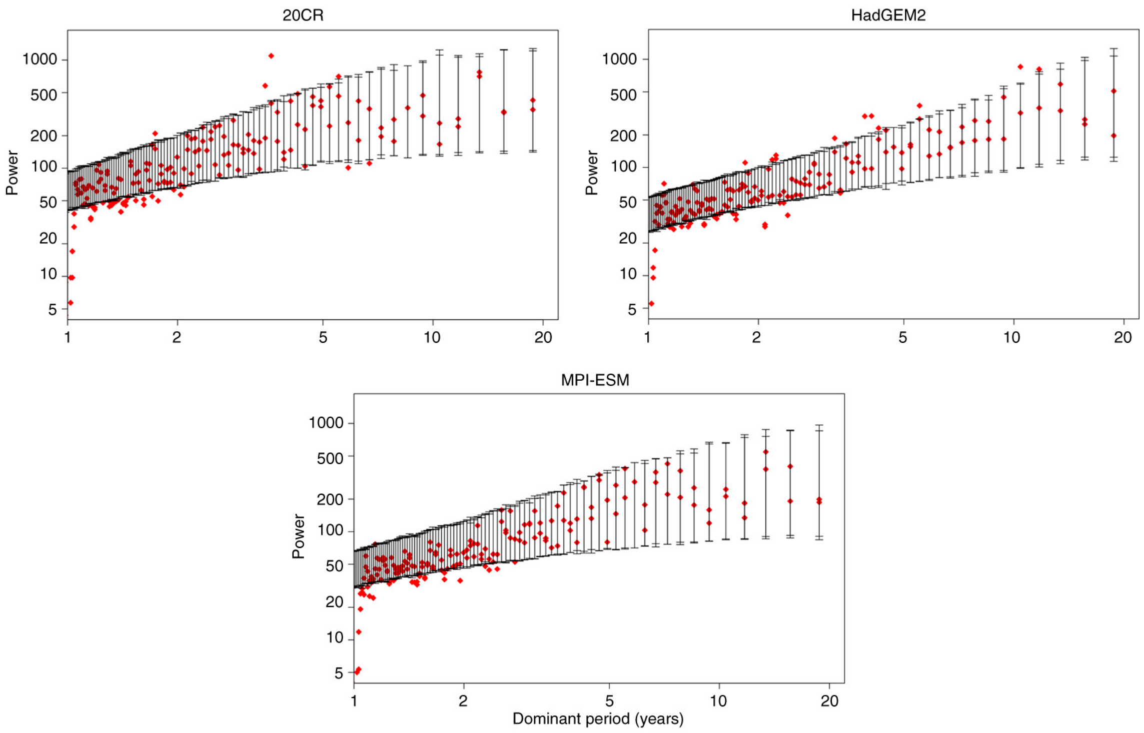


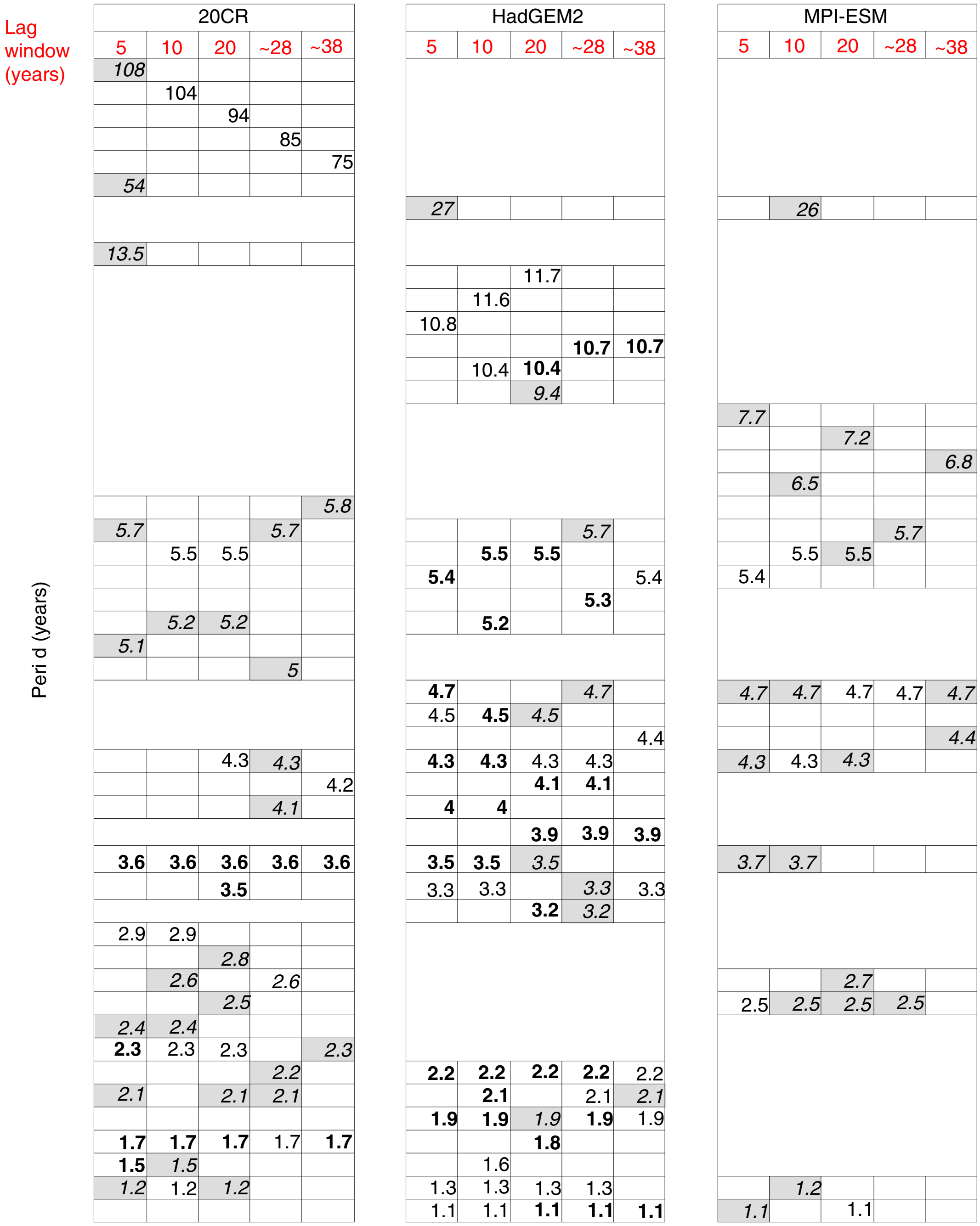
Fig. 6. MC-MSSA test of the monthly near-surface temperature variability in 20CR, HadGEM2 and MPI-ESM data sets 1901–2012. PCs 1–30 of RP + PCA (see Appendix A.2, algorithm 2) are used as input channels in the analysis and the lag window length M is 20 yr (240 months). In MC-MSSA, the red-noise basis is used. Red squares show the data eigenvalues plotted against the dominant frequency of the ST-PC corresponding to each eigenvalue. The vertical bars show the 2.5th and 97.5th percentiles of the eigenvalue distribution calculated from 1000 realisations of the red-noise surrogates. The ST-PCs that correspond to eigenvalues rising above the 97.5th percentiles are considered significant at the 5% level. Note the missing power at 1 yr due to the removal of the annual cycle.

Vautard (1994) the use of a lag window length M typically allows the distinction of oscillations with periods in the range [$M/5$, M]. Therefore we only show the significance test of the periodicities that are covered by the 20 yr lag window used in this example. From Fig. 6, we can see that in 20CR data set there are some significant periodicities (at 5% level) between 1.7 and 5.5 yr. HadGEM2 has somewhat more significant periodicities compared to 20CR, especially on 10 yr time scales, but MPI-ESM has hardly any eigenvalues lying above the 97.5th percentile.

3.2.1. Results with different lag window lengths. As noted earlier, the Monte-Carlo simulations were performed with varying lag window M to estimate its effect on the statistical significance of the oscillations. Spectral resolution increases with lag window length and oscillatory pairs with longer periodicity can be identified. However, at the same time the statistical significance of the identified oscillations may decline. We used the following values of M: 5 yr ($M = 60$ months), 10 yr ($M = 120$), 20 yr ($M = 240$), approx. 28 yr [$M = 340 \approx N/4$, following the recommendation of Elsner and Tsonis (1996)] and approx. 38 yr [$M = 450 \approx N/3$, following Vautard and Ghil (1989)].

The identified periodicities and their significance levels with increasing lag window are presented in Fig. 7. The numbers in Fig. 7 show the dominant periods associated with the oscillations. These dominant periods are estimated using FFT. From Fig. 7 we can see that in 20CR the significant periodicities are consistently found at 3.6, 2.3 and 1.7 yr, depending to some extent on M. Those periods are more or less visible in HadGEM2 and to a lesser extent in MPI-ESM. Significant 5–6 yr oscillations are identified in all the data sets and especially a ~5.5 yr variability is found consistently.

2–6 yr oscillations are usually attributed to ENSO which is a globally dominating form of variability on annual to decadal time scales (e.g. Kleeman, 2008). It is a broadband phenomenon with several spectral peaks and the highest peak is around 4 yr. This can also be seen in our analysis of 20CR, HadGEM2 and MPI-ESM data sets because most of the significant oscillations are concentrated on 2–6 yr time scales. However, the spectra of MPI-ESM (Fig. 5) differs distinctly from the spectra of the other two data sets: the


Lag window (years)
Peri d (years)

20CR 5	10	20	~28	~38
108				
	104			
		94		
			85	
				75
54				
13.5				
				5.8
5.7			*5.7*	
	5.5	5.5		
	5.2	*5.2*		
5.1				
			5	
		4.3	*4.3*	
				4.2
			4.1	
3.6	**3.6**	**3.6**	**3.6**	**3.6**
		3.5		
2.9	2.9			
		2.8		
	2.6		2.6	
		2.5		
2.4	*2.4*			
2.3	2.3	2.3		*2.3*
			2.2	
2.1		*2.1*	*2.1*	
1.7	**1.7**	**1.7**	1.7	**1.7**
1.5	*1.5*			
1.2	1.2	*1.2*		

HadGEM2 5	10	20	~28	~38
27				
		11.7		
	11.6			
10.8				
			10.7	**10.7**
	10.4	**10.4**		
		9.4		
			5.7	
	5.5	**5.5**		
5.4				5.4
			5.3	
	5.2			
4.7			*4.7*	
4.5	**4.5**	*4.5*		
				4.4
4.3	**4.3**	4.3	4.3	
		4.1	**4.1**	
4	**4**			
		3.9	**3.9**	**3.9**
3.5	**3.5**	*3.5*		
3.3	3.3		*3.3*	3.3
		3.2	*3.2*	
2.2	**2.2**	**2.2**	**2.2**	2.2
	2.1		2.1	*2.1*
1.9	**1.9**	*1.9*	**1.9**	1.9
		1.8		
	1.6			
1.3	1.3	1.3	1.3	
1.1	1.1	**1.1**	**1.1**	**1.1**

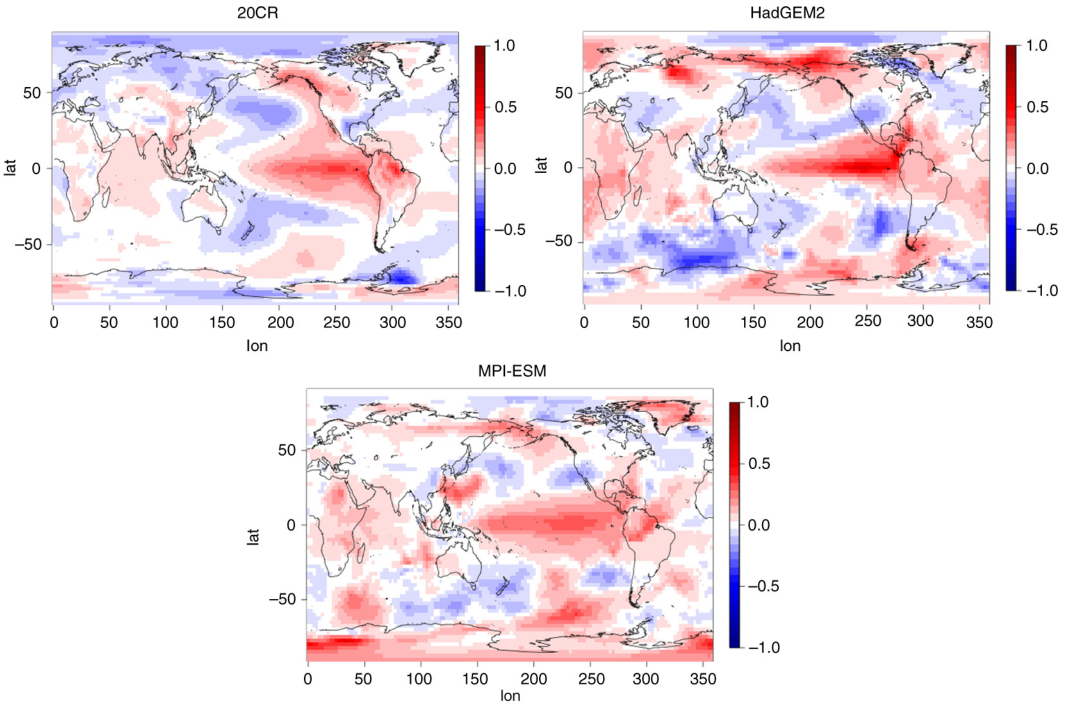
MPI-ESM 5	10	20	~28	~38
	26			
7.7				
		7.2		
				6.8
	6.5			
			5.7	
	5.5	*5.5*		
5.4				
4.7	*4.7*	4.7	4.7	*4.7*
				4.4
4.3	4.3	*4.3*		
3.7	*3.7*			
		2.7		
2.5	*2.5*	*2.5*	*2.5*	
	1.2			
1.1		1.1		

Fig. 7. Periodicities (years) detected by RMSSA/MC-MSSA with varying lag window length (years) for each data set (20CR, HadGEM2 and MPI-ESM). The similar periodicities among the data sets are aligned. Numbers in the figure are in bold if the significance level of a periodicity is 1%, and with grey background if 10%. Otherwise the significance level is 5%. Dominant frequencies of the oscillations are estimated using fast Fourier transform (FFT).

power on 2–8 yr time scales is distributed on a large set of components (especially ST-PCs 4–18) which also decreases the statistical significance of oscillations on those time scales.

In HadGEM2, significant decadal scale oscillations are identified with all lag window lengths. Dominant peak on the decadal time scales has been noted by Collins et al. (2008) and one of the possible reasons for this is in deficiencies of simulation of the ENSO phase-changing process in HadGEM2 (Martin et al., 2010).

There are also significant multi-decadal components in 20CR data set, but their period decreases with increasing lag window M. The time series to be analysed become


20CR

HadGEM2

MPI-ESM

Fig. 8. Global patterns of ~5.5 yr oscillation of the near-surface temperature anomaly (°C) in 20CR, HadGEM2 and MPI-ESM data sets 1901–2012. The patterns are calculated as composites of eight cases, when the oscillation is in its maximum positive phase in the equatorial Pacific. Those positive events are defined as an average of winter months (Nov–Mar). See the text for more details on the reconstruction procedure. The identified patterns have similarities to El Niño -phenomenon.

shorter with increasing M and this may have an effect on the identified period length. We did not find significant multi-decadal components in HadGEM2 and MPI-ESM data sets, although 27 yr and 26 yr periods are identified on 10% significance levels, but only with a single lag window length. However, the use of a lag window M typically allows only the distinction of oscillations with periods $\leq M$ and thus the interpretation of those multi-decadal components remains uncertain.

3.3. *Reconstruction of the significant oscillations*

The final step of our analysis is to reconstruct the decomposed signals in the original space. As an illustration, we have chosen to reconstruct the signal corresponding to approximately 5.5 yr variability, which was identified in all the data sets.

In order to see the time evolution of the ~5.5 yr variability, we have reconstructed the time series in each gridpoint of the original data set with the ST-PCs corresponding to the signal of interest. I.e., in the reconstruction we have projected the original (centred) data set onto ST-PCs (calculated in the reduced space) to obtain ST-EOFs in the original space and then projected the ST-PCs onto those ST-EOFs (see Appendices A.2 and A.4 for more details). In order to see the global effects of the ~5.5 yr cycle, the time series of each grid point has its original variance. The above calculations were completed for each data set using their own ~5.5 yr patterns. ST-PCs 5 and 6 of the 20CR data set (Fig. 3), ST-PCs 6 and 7 of HadGEM2 (Fig. 4) and ST-PCs 4 and 5 of MPI-ESM (Fig. 5) were used in the reconstruction.

Once we have reconstructed the time series in each gridpoint we can plot the anomalies related to the signal month by month. These plots are presented as animations of each data set (20CR, HadGEM2 and MPI-ESM) for a time period of 1901–2012 (the animations are available at www.youtube.com/channel/UCRjwc6cI-TzbvtShONYZ7cg). In Fig. 8, we also show the global patterns of the ~5.5 yr variability of near-surface temperature anomaly. The patterns are composites of eight cases, when the oscillation is in its positive phase in the equatorial Pacific. Positive events are defined as an average of winter months (November–March).

The temperature anomalies of 20CR have many similarities to global El Niño effects, such as above average temperatures in the central and eastern equatorial Pacific Ocean, in the western and northern parts of North-America and South-America as well as in South-East Asia, Australia and southern Africa. Below average anomalies are found in the south-east parts of North-America, in the north-west and south-west Pacific as well as in northern parts of Eurasia.

In 20CR, a typical north-south wave train is also seen, but the east-west patterns are weaker, except for the anomalies at the Amazonas.

HadGEM2 and MPI-ESM show similarities to 20CR, but differences can be seen, for example, in the Pacific forcing patterns. Especially in MPI-ESM the centre of the forcing pattern seems to be more western. In the model simulations, the negative anomaly near the west-coast of North-America extends to the continent, which is not detected in 20CR. The positive anomalies in HadGEM2 and MPI-ESM also extend into the northern Eurasia and there are anomaly patterns in the southern Indian Ocean which are absent in 20CR. MPI-ESM has a stronger positive anomaly in the coast of South-East Asia compared to the other two data sets. In addition, there is a strong anomaly near the Antarctic Peninsula in the Weddel Sea in the 20CR data set which is not detected in the model simulations. The anomaly patterns in the Atlantic Ocean are also weaker in 20CR compared to simulations.

The animations of the ~5.5 yr pattern (available at www.youtube.com/channel/UCRjwc6cI-TzbvtShONYZ7cg) show some more features in addition to the ones seen in Fig. 8. For instance, in 20CR animation there is a quite strong anomaly pattern to the west of Ural Mountains. This pattern is not usually associated with ENSO, and its maximum negative and positive phases seem to occur at different times compared to the ENSO-related anomaly patterns in the Pacific. However, this pattern to the west of Ural might also reflect some other phenomenon, mixed with the ENSO patterns.

The animations also show that the variability has a more propagating character in 20CR data set whereas the anomaly patterns in the model simulations are more stationary. In the northern and southern Pacific Ocean, for example, the anomalies seem to propagate eastward in the 20CR animation.

Compared to 20CR, HadGEM2 and MPI-ESM show a richer structure in Fig. 8 and in the animations. One has to remember that the reanalysis data set is an ensemble mean whereas the analysis of the climate model simulations is conducted on a single ensemble member of each model. This may also contribute to the structure seen in the model simulations. Different, more or less real, phenomena may also be mixed in the variability patterns of the simulations.

4. Summary and Discussion

We have introduced an RMSSA algorithm, which allows the calculation of MSSA of extremely high-dimensional problems. The RMSSA algorithm first reduces the dimension of the original data set by RP, then decomposes the data set into components of different frequencies by calculating MSSA in a reduced space, and finally reconstructs the components in the original high-dimensional space.

We have applied the RMSSA algorithm to decompose the monthly mean near-surface air temperature of the 20th century reanalysis and the historical 20th century simulations of HadGEM2-ES and MPI-ESM-MR extracted from the CMIP5 data archives. We have also performed Monte-Carlo simulations in order to estimate the significance of the identified low-frequency components. Our analysis shows that 2–6 yr oscillations are present in all the data sets. Their statistical significance is highest in HadGEM2 while in MPI-ESM the power on those timescales is distributed on a large set of components decreasing their statistical significance.

2–6 yr oscillations are usually attributed to ENSO which is a globally dominating form of variability on annual to decadal time scales. Our global monthly animations of 5–6 yr near-surface temperature cycle match quite well with the known temperature anomalies related to ENSO. The reanalysis and the historical simulations have similar anomaly patterns in the central and eastern Pacific Ocean, around the northern part of Indian Ocean as well as in the north-west North-America, but also some notable differences in several areas, such as Eurasia. Also, our animations of the 5–6 yr cycle reveal a propagating structure in the near-surface temperature anomalies of 20CR, while the variability in HadGEM2 and MPI-ESM data sets is more stationary. The focus of this study was to introduce the RMSSA algorithm and the discussion on the possible causes for the differences in oscillatory patterns of the data sets is limited. However, this would be a subject for a further study with a larger set of climate model data sets included.

RMSSA algorithm is a powerful tool when the dimensions of the data sets become prohibitively large. It allows a computationally efficient way of decomposing a data set into its spatio-temporal patterns. Several climatic state variables can be incorporated in the RMSSA at the same time in order to find the co-varying signals and illustrate their propagation. RMSSA can also be used in studying the oscillations in three dimensions including data from several atmospheric levels in the analysis.

5. Acknowledgements

We thank Petteri Uotila and Mikko Alestalo from Finnish Meteorological Institute for interesting discussions at earlier phases of the work. We also thank Jouni Räisänen from University of Helsinki for preparing the HadGEM2-ES and MPI-ESM-MR data sets used in this study. In addition, we thank the reviewer for her/his valuable comments and suggestions that helped in improving the manuscript. This research has been funded by the Academy of Finland (project number 140771).

Appendix A

A.1. MSSA and Monte-Carlo MSSA

A.1.1. Multichannel singular spectrum analysis (MSSA)

The aim of MSSA is to identify spatially and temporally coherent patterns in a multivariate data set. In MSSA terminology, the columns of the original data matrix $\mathbf{X}_{N\times L}$ are called channels. In case of gridded data set, N represents the time steps and L is the number of grid points:

$$\mathbf{X}=\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,L} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,L} \end{bmatrix} \tag{A1}$$

The next step is to construct an augmented data matrix $\mathbf{A}$, which contains M lagged copies of each channel in $\mathbf{X}$:

$$\mathbf{Y_i}=\begin{bmatrix} x_{1,i} & x_{2,i} & \cdots & x_{M,i} \\ x_{2,i} & x_{3,i} & \cdots & x_{M+1,i} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N\prime,i} & x_{N\prime+1,i} & \cdots & x_{N,i} \end{bmatrix}, i=1...L \tag{A2}$$

and

$$\mathbf{A}=[\mathbf{Y_1} \quad \mathbf{Y_2} \quad \cdots \quad \mathbf{Y_L}] \tag{3}$$

In MSSA, M represents the lag window. $\mathbf{A}$ has now ML columns and $N'=N-M+1$ rows. The singular value decomposition (SVD) of $\mathbf{A}$ can now be calculated:

$$\mathbf{A}=\mathbf{U}_A\mathbf{D}_A^{1/2}\mathbf{V}_A^T, \tag{4}$$

The vectors of $\mathbf{U}_A$ are the eigenvectors of $\mathbf{Z}_A=\frac{1}{ML}\mathbf{A}\mathbf{A}^T$ and $\mathbf{V}_A^T$ contains the eigenvectors of $\mathbf{C}_A=\frac{1}{N\prime}\mathbf{A}^T\mathbf{A}$. These vectors are orthogonal and often called space-time principal components (ST-PCs) and space-time empirical orthogonal functions (ST-EOFs), respectively. Diagonal elements of $\mathbf{D}_A$ are the eigenvalues of $\mathbf{C}_A$ or $\mathbf{Z}_A$.

Optionally the lag-covariance matrix $\mathbf{C}_A$ (or $\mathbf{Z}_A$) and its eigendecomposition can be calculated to yield eigenvectors $\mathbf{V}_A^T$ (or $\mathbf{U}_A$) and eigenvalues (diagonal elements of matrix $\mathbf{D}_A=\mathbf{V}_A^T\mathbf{C}_A\mathbf{V}_A$or $\mathbf{D}_A=\mathbf{U}_A^T\mathbf{Z}_A\mathbf{U}_A$). Matrix $\mathbf{U}_A$ (or $\mathbf{V}_A^T$) can be obtained by projecting $\mathbf{A}$ onto $\mathbf{V}_A^T$(or $\mathbf{U}_A$). If $N'>ML$ (or $ML>N'$), it is more convenient to estimate $\mathbf{C}_A$ (or $\mathbf{Z}_A$) because it is smaller. See Allen and Robertson (1996) for details.

A.1.2. Monte-Carlo MSSA

The components obtained by MSSA can be tested against a null-hypothesis of the data being generated by independent AR(1) processes (i.e. red noise). The red noise model has the form:

$$u_{t+1,s}=\gamma_s u_{t,s}+\alpha_s w_{t,s}, \tag{A5}$$

where γ_s is the lag-1 autocorrelation of channel s (in the original data set), $\alpha_s=\sqrt{c_s(1-\gamma_s^2)}$ ($c\ _s$ is the variance of channel s) and $W_{t,s}$ is Gaussian white noise. The data set generated by the model in (A5) is called the surrogate data set and it is subjected to MSSA in the same way as the original data set. Large number of surrogates are generated in order to estimate the confidence limits for the MSSA results of the original data set.

In the test of Allen and Robertson (1996), the lag-covariance matrices of the original data set and the red-noise surrogates are projected either onto the data-adaptive basis (i.e. $\mathbf{U}_A$ or $\mathbf{V}_A^T$) or the null-hypothesis basis. The null-hypothesis basis can be calculated from the expected lag-covariance matrix $\mathbf{C}_N$ of the red-noise surrogates. $\mathbf{C}_N$ can be estimated analytically by

$$[\mathbf{C}_N]=\frac{1}{ML}\sum_{s=1}^{ML}c_s\gamma_s^{|ii-jj|} \tag{A6}$$

Projection onto the red-noise basis is considered more reliable because the use of the data-adaptive basis assumes the existence of an oscillation even in a case where it is uncertain whether or not the oscillation is significant.

According to Allen and Robertson (1996), the input channels should be uncorrelated (or at least nearly) at zero lag for the test to be useful. In the case of a gridded data set, where all the grid point time series are used as input channels, the decorrelation condition is not valid. The test might still be useful if we are using grid points sufficiently far from each other as the input channels for the test (Ghil et al., 2002).

A.2. Randomised algorithms for MSSA

1: Original MSSA algorithm enhanced by RP

(1) construct the original data matrix $\mathbf{X}_{N\times L}$
(2) (pre-processing of $\mathbf{X}$, if needed)
(3) generate k L-dimensional vectors of Gaussian distributed random numbers to get matrix $\mathbf{R}$ (and optionally orthogonalise the random vectors)
(4) project the original data matrix onto random vectors: $\mathbf{P}_{N\times k}=\frac{1}{\sqrt{k}}\mathbf{X}_{N\times L}\mathbf{R}_{L\times k}$
(5) generate augmented matrix $\mathbf{A}_{RP}$ of $\mathbf{P}$
(6) calculate SVD: $\mathbf{A}_{RP}=\mathbf{U}_{RP}\mathbf{D}_{RP}^{1/2}\mathbf{V}_{RP}^T$ (or covariance matrix $\mathbf{C}_{RP}$ or $\mathbf{Z}_{RP}$ and its eigendecomposition)
(7) calculate ST-EOFs in the original space: $\mathbf{V}_A\approx\mathbf{A}^T\mathbf{U}_{RP}(\mathbf{D}_{RP}^{1/2})^{-1}$ (see Appendix A.4 for an explanation)
(8) calculate RCs using ST-EOFs of step 7.

2: PC-based MSSA algorithm enhanced by RP

(1) construct the original data matrix $\mathbf{X}_{N\times L}$
(2) (pre-processing of **X**, if needed)
(3) generate k L-dimensional vectors of Gaussian distributed random numbers to get matrix **R** (and optionally orthogonalise the random vectors)
(4) project the original data matrix onto random vectors: $\mathbf{P}_{N\times k} = \frac{1}{\sqrt{k}}\mathbf{X}_{N\times L}\mathbf{R}_{L\times k}$
(5) calculate SVD of **P** (see Appendix A.3 for an explanation of how the covariance is preserved in RP + SVD)
(6) retain e.g. 30 first PCs of **P** to obtain reduced matrix **T**
(7) generate augmented matrix $\mathbf{A}_{PC}$ of **T**
(8) calculate SVD: $\mathbf{A}_{PC} = \mathbf{U}_{PC}\mathbf{D}_{PC}^{1/2}\mathbf{V}_{PC}^T$ (or covariance matrix $\mathbf{C}_{PC}$ or $\mathbf{Z}_{PC}$ and its eigendecomposition)
(9) (MC-MSSA step)
(10) calculate ST-EOFs in the original space: $\mathbf{V}_A \approx \mathbf{A}^T\mathbf{U}_{PC}(\mathbf{D}_{PC}^{1/2})^{-1}$ (see Appendix A.4 for an explanation)
(11) calculate RCs using ST-EOFs of step 10.

A.3. RP and SVD

The method to back-project from the reduced space to the original space in the case of RP + SVD is explained in Seitola et al. (2014) (Appendix A.1) but we also present it briefly here:

The SVD of the original data matrix $\mathbf{X}_{N\times L}$ is:

$$\mathbf{X}_{N\times L} = \mathbf{U}_{N\times N}\mathbf{D}_{N\times L}\mathbf{V}_{L\times L}^T \tag{A7}$$

U contains the eigenvectors of $\mathbf{Z} = \mathbf{X}\mathbf{X}^T$.

Random projection (RP) of **X** is $\mathbf{P} = \mathbf{X}\mathbf{R}$, where $\mathbf{R}_{L\times k}$ is the projection matrix and the row vectors of **R** are scaled to have unit length. Thus, we can write:

$$\mathbf{Z}_{RP} = \mathbf{X}\mathbf{R}(\mathbf{X}\mathbf{R})^T = \mathbf{X}\mathbf{R}\mathbf{R}^T\mathbf{X}^T \approx \mathbf{X}\mathbf{X}^T = \mathbf{Z} \tag{A8}$$

In the previous, we have assumed that $\mathbf{R}\mathbf{R}^T \approx \mathbf{I}$ because the row vectors of **R** are nearly orthonormal. It is also possible to make the vectors of **R** strictly orthonormal, in which case $\mathbf{R}\mathbf{R}^T = \mathbf{I}$. However, orthogonalisation is often not necessary, because the difference between the orthogonalised and non-orthogonalised random vectors is very small, especially in high-dimensions.

Let's rewrite (A7) as $\mathbf{X}_{N\times L} = \mathbf{U}_{N\times r}\mathbf{D}_{r\times r}\mathbf{V}_{r\times L}^T$, where $r = \text{rank}(\mathbf{X})$. Now we can manipulate (A7):

$$\begin{array}{ll} \mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T & (\mathbf{U}^T\mathbf{U} = \mathbf{I}) \\ \mathbf{U}^T\mathbf{X} = \mathbf{D}\mathbf{V}^T & (\mathbf{D}^{-1}\mathbf{D} = \mathbf{I}) \\ \mathbf{V}^T = \mathbf{D}^{-1}\mathbf{U}^T\mathbf{X} & \text{transpose of both sides} \\ \mathbf{V} = \mathbf{X}^T\mathbf{U}(\mathbf{D}^{-1})^T = \mathbf{X}^T\mathbf{U}\mathbf{D}^{-1} & \end{array} \tag{A9}$$

Because $\mathbf{Z} \approx \mathbf{Z}_{RP}$ we can approximate

$$\begin{array}{l} \mathbf{U} \approx \mathbf{U}_{RP}, \\ \mathbf{D} \approx \mathbf{D}_{RP} \quad \text{and} \\ \mathbf{V} \approx \mathbf{X}^T\mathbf{U}_{RP}\mathbf{D}_{RP}^{-1} \end{array} \tag{A10}$$

In the previous, we have defined $\mathbf{U}_{RP}$ as $N \times k$ and $\mathbf{D}_{RP}$ as a $k \times k$ matrix, where k is the rank of matrix $\mathrm{P}_{N\times k}$.

A.4. RP and MSSA

In this appendix, we will explain how to get from the reduced space back to the original space in the case of RP + MSSA.

Let's write the original data matrix $\mathbf{X}_{N\times L}$ as

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,L} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,L} \end{bmatrix} = \begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \vdots \\ \mathbf{x_N} \end{bmatrix}, \tag{A11}$$

where $\mathbf{x_i}$ are the row vectors of **X**.

The augmented matrix **A** of **X** is already defined in Appendix A.1. Now let's calculate $\mathbf{A}\mathbf{A}^T$.

$$\begin{aligned} \mathbf{A}\mathbf{A}^T &= \begin{bmatrix} \mathbf{Y_1} & \mathbf{Y_2} & \cdots & \mathbf{Y_L} \end{bmatrix} \begin{bmatrix} \mathbf{Y_1^T} \\ \mathbf{Y_2^T} \\ \vdots \\ \mathbf{Y_L^T} \end{bmatrix} \\ &= \left[\mathbf{Y_1}\mathbf{Y_1^T} + \mathbf{Y_2}\mathbf{Y_2^T} + \cdots + \mathbf{Y_L}\mathbf{Y_L^T}\right] \end{aligned} \tag{A12}$$

After some algebra we get

$$\mathbf{A}\mathbf{A}^T = \begin{bmatrix} \mathbf{x}_1\mathbf{x}_1^T + \mathbf{x}_2\mathbf{x}_2^T + \cdots + \mathbf{x}_M\mathbf{x}_M^T & \mathbf{x}_1\mathbf{x}_2^T + \mathbf{x}_2\mathbf{x}_3^T + \cdots + \mathbf{x}_M\mathbf{x}_{M+1}^T & \cdots & \mathbf{x}_1\mathbf{x}_{N\prime}^T + \mathbf{x}_2\mathbf{x}_{N\prime+1}^T + \cdots + \mathbf{x}_M\mathbf{x}_N^T \\ \mathbf{x}_2\mathbf{x}_1^T + \mathbf{x}_3\mathbf{x}_2^T + \cdots + \mathbf{x}_{M+1}\mathbf{x}_M^T & \mathbf{x}_2\mathbf{x}_2^T + \mathbf{x}_3\mathbf{x}_3^T + \cdots + \mathbf{x}_{M+1}\mathbf{x}_{M+1}^T & \cdots & \mathbf{x}_2\mathbf{x}_{N\prime}^T + \mathbf{x}_3\mathbf{x}_{N\prime+1}^T + \cdots + \mathbf{x}_{M+1}\mathbf{x}_N^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{N\prime}\mathbf{x}_1^T + \mathbf{x}_{N\prime+1}\mathbf{x}_2^T + \cdots + \mathbf{x}_N\mathbf{x}_M^T & \mathbf{x}_{N\prime}\mathbf{x}_2^T + \mathbf{x}_{N\prime+1}\mathbf{x}_3^T + \cdots + \mathbf{x}_N\mathbf{x}_{M+1}^T & \cdots & \mathbf{x}_{N\prime}\mathbf{x}_{N\prime}^T + \mathbf{x}_{N\prime+1}\mathbf{x}_{N\prime+1}^T + \cdots + \mathbf{x}_N\mathbf{x}_N^T \end{bmatrix} \tag{A13}$$

Now let's calculate RP of $\mathbf{X}$:

$$\mathbf{XR} = \begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \vdots \\ \mathbf{x_N} \end{bmatrix} \mathbf{R} = \begin{bmatrix} \mathbf{x_1R} \\ \mathbf{x_2R} \\ \vdots \\ \mathbf{x_NR} \end{bmatrix} \tag{A14}$$

The augmented matrix of is $\mathbf{A}_{RP}$:

$$\mathbf{A}_{RP} = \begin{bmatrix} \mathbf{x_1R} & \mathbf{x_2R} & \cdots & \mathbf{x_MR} \\ \mathbf{x_2R} & \mathbf{x_3R} & \cdots & \mathbf{x_{M+1}R} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x_{N\prime}R} & \mathbf{x_{N\prime+1}R} & \cdots & \mathbf{x_NR} \end{bmatrix} \tag{A15}$$

Let's calculate $\mathbf{A}_{RP}\mathbf{A}_{RP}^T$:

$$\mathbf{A}_{RP}\mathbf{A}_{RP}^T = \begin{bmatrix} \mathbf{x_1R} & \mathbf{x_2R} & \cdots & \mathbf{x_MR} \\ \mathbf{x_2R} & \mathbf{x_3R} & \cdots & \mathbf{x_{M+1}R} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x_{N\prime}R} & \mathbf{x_{N\prime+1}R} & \cdots & \mathbf{x_NR} \end{bmatrix} \times \begin{bmatrix} \mathbf{R^Tx_1^T} & \mathbf{R^Tx_2^T} & \cdots & \mathbf{R^Tx_{N\prime}^T} \\ \mathbf{R^Tx_2^T} & \mathbf{R^Tx_3^T} & \cdots & \mathbf{R^Tx_{N\prime+1}^T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R^Tx_M^T} & \mathbf{R^Tx_{M+1}^T} & \cdots & \mathbf{R^Tx_N^T} \end{bmatrix} \tag{A16}$$

Because $\mathbf{RR}^T \approx \mathbf{I}$, the first element of $\mathbf{A}_{RP}\mathbf{A}_{RP}^T$ can be written as $\mathbf{x_1RR^Tx_1^T} + \mathbf{x_2RR^Tx_2^T} + \cdots + \mathbf{x_MRR^Tx_M^T} \approx \mathbf{x_1x_1^T} + \mathbf{x_2x_2^T} + \cdots + \mathbf{x_Mx_M^T}$

After calculating all the elements of $\mathbf{A}_{RP}\mathbf{A}_{RP}^T$ as above, we see that $\mathbf{AA}^T \approx \mathbf{A}_{RP}\mathbf{A}_{RP}^T$. Therefore, as in Appendix A.3, we can approximate

$$\begin{aligned} &\mathbf{U}_A \approx \mathbf{U}_{RP}, \\ &\mathbf{D}_A \approx \mathbf{D}_{RP} \qquad \text{and} \\ &\mathbf{V}_A \approx \mathbf{A}^T\mathbf{U}_{RP}\mathbf{D}_{RP}^{-1} \end{aligned} \tag{A17}$$

Same kind of reasoning applies also when the PCs of the data set are used as channels in MSSA. We can write PCs as $\mathbf{U}_{N\times r}\mathbf{D}_{r\times r} = \mathbf{X}_{N\times L}\mathbf{V}_{L\times r}$, where $r = \text{rank}(\mathbf{X})$. Vectors of $\mathbf{V}$ are orthonormal, so in the above calculations we can replace $\mathbf{R}$ with $\mathbf{V}$.

References

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

Bingham, E. and Mannila, H. 2001. 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. New York, ACM, 245–250.

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

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

Chiu, S. K. 2013. Coherent and random noise attenuation via multichannel singular spectrum analysis in the randomized domain. *Geophys. Prospect.* **61**, 1–9.

Cleveland, R. B., Cleveland, W. S., McRae, J. E. and Terpenning, I. 1990. STL: a seasonal-trend decomposition procedure based on loess. *J Off. Stat.* **6**, 373.

Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P. and co-authors. 2011. Development and evaluation of an Earth-System model HadGEM2. *Geosci. Model Dev.* **4**, 1051–1075.

Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Hinton, T. and co-authors. 2008. Evaluation of the HadGEM2 model. Met Office Hadley Centre Technical Note no. HCTN 74, available from Met Office, FitzRoy Road, Exeter EX1 3PB. Online at: http://www.metoffice.gov.uk/publications/HCTN/index.html

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J. and co-authors. 2011. The twentieth century reanalysis project. *Quart. J. Roy. Meteorol. Soc.* **137**, 1–28.

Dasgupta, S. and Gupta, A. 2003. An elementary proof of a theorem of Johnson and Lindenstrauss. *Rand. Struct. Algo.* **22**, 60–65.

Elsner, J. B. and Tsonis, A. A. (eds.). 1996. *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. Springer Science & Business Media, New York, NY, USA.

Frankl, P. and Maehara, H. 1988. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *J. Combin. Theory Ser. B* **44**, 355–362.

Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D. and co-authors. 2002. Advanced spectral methods for climatic time series. *Rev. Geophys.* **40**(1), 1–41.

Halko, N., Martinsson, P. G. and Tropp, J. A. 2011. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **53**(2), 217–288.

Kawale, J., Chatterjee, S., Kumar, A., Liess, S., Steinbach, M. and co-authors. 2011. Anomaly construction in climate data: issues and challenges. In: *Proceedings of the 2011 Conference on Intelligent Data Understanding*, CIDU, Mountain View, CA, USA, pp. 189–203.

Kleeman, R. 2008. Stochastic theories for the irregularity of ENSO. *Philos. Trans. Roy. Soc. A Math. Phys. Eng. Sci.* **366**(1875), 2509–2524.

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

Martin, G. M., Milton, S. F., Senior, C. A., Brooks, M. E., Ineson, S. and co-authors. 2010. Analysis and reduction of systematic errors through a seamless approach to modeling weather and climate. *J. Clim.* **23**(22), 5933–5957.

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G. and co-authors. 2014. Decadal climate prediction: an update from the trenches. *Bull. Amer. Meteor. Soc.* **95**, 243–267.

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G. and co-authors. 2009. Decadal prediction: can it be skillful? *Bull. Amer. Meteor. Soc.* **90**, 1467–1485.

Moron, V., Robertson, A. W. and Ghil, M. 2012. Impact of the modulated annual cycle and intraseasonal oscillation on daily-to-interannual rainfall variability across monsoonal India. *Clim. Dyn.* **38**, 2409–2435.

Oropeza, V. and Sacchi, M. 2011. Simultaneous seismic data denoising and reconstruction via multichannel singular spectrum analysis. *Geophysics* **76**, V25–V32.

Plaut, G. and Vautard, R. 1994. Spells of low-frequency oscillations and weather regimes in the Northern Hemisphere. *J. Atmos. Sci.* **51**(2), 210–236.

Rinne, J. and Karhila, V. 1979. Empirical orthogonal functions of 500 mb height in the northern hemisphere determined from a large data sample. *Quart. J. Roy. Meteorol. Soc.* **105**, 873–884.

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

Solomon, A., Goddard, L., Kumar, A., Carton, J., Deser, C. and co-authors. 2011. Distinguishing the roles of natural and anthropogenically forced decadal climate variability. *Bull. Amer. Meteor. Soc.* **92**, 141–156.

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T. and co-authors. 2013. Atmospheric component of the MPIM Earth System Model: ECHAM6. *J. Adv. Model. Earth Syst.* **5**(2), 146–172.

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

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

Vautard, R. and Ghil, M. 1989. Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Phys. D Nonlin. Phenom.* **35**(3), 395–424.

Von Storch, H. and Zwiers, F. W. 1999. *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge.

Weare, B. C. and Nasstrom, J. S. 1982. Examples of extended empirical orthogonal function analysis. *Mon. Weath. Rev.* **110**, 481–485.