Isolating secular signals in observations and climate model simulations using M-SSA based Wiener filtering

by

Christian Grimm

A Thesis Submitted in
Partial Fulfillment of the
Requirements for the Degree of

Master of Science
in
Mathematics

at

The University of Wisconsin-Milwaukee
May 2018

ABSTRACT

Isolating secular signals in observations and climate model simulations using M-SSA based Wiener filtering

by

Christian Grimm

The University of Wisconsin-Milwaukee, 2018
Under the Supervision of Professors Sergey Kravtsov and Vytaras Brazauskas

In this thesis, Wiener filtering of gridded surface-temperature time series from observations and climate model simulations is performed by using multi-channel singular spectrum analysis (M-SSA) in order to isolate non-stationary climate signals. The contributions to the singular spectrum from shorter-term internal climate variability — treated in this context as noise — are estimated by fitting to the data spatially extended stochastic models, which are subsequently used to produce synthetic ensembles of surface temperature time series and the corresponding synthetic M-SSA spectra. The full spectra are weighted by the signal-to-noise ratios and transformed back to physical space to obtain reconstructions of the non-stationary signal. This methodology was first tested using the twentieth century simulations from the Community Earth System Model Large Ensemble Project, for which the forced climate signal can be reliably estimated by taking the ensemble average over the 40 available climate realizations, then applied to individual model ensembles as well as the overall ensemble from the Coupled Model Intercomparison Project Phase 5 and, finally, to the observational surface-temperature time series from Twentieth Century Reanalysis. The method is shown to successfully recover the low-frequency (decadal or larger time scales) component of the forced signal in model simulations, but fails to isolate shorter-term variability associated with

volcanic eruptions. The secular signals estimated from model simulations and observations exhibit large differences, which indicates the presence, in observations, of a pronounced multi-decadal variability with a distinctive spatiotemporal structure absent in any of the model simulations.

Table of Contents

List of Figures

LIST OF ABBREVIATIONS

SAT Surface Atmospheric Temperature

M-SSA Multi-channel singular spectrum analysis

NMO Northern Hemisphere Multidecadal Oscillation

LENS Large Ensemble Project (data set)

CMIP5 Coupled Model Intercomparison Project Phase 5 (data set)

20CR Twentieth Century Reanalysis (data set)

1 Introduction

Analyzing temporal development and spatial patterns of surface atmospheric temperatures (SAT; measured at a height of 2 meters above the surface), both over ocean and land, is one of the central issues in climate research. Its results have far-reaching relevance: Changing sea surface temperatures can effect atmospheric flows and therefore impact the climate at distant places, too. Also, warming sea surfaces can influence the frequency and power of severe weather conditions such as tropical cyclone activity (Christensen et al., 2013). Changes of land surface temperatures, in contrast, are of particular interest for agricultural and environmental institutions.

An SAT time series at a given spatial location shows an observation of climatic variability for a limited period of time. The main goal of classical time series analysis in climate research (Venegas, 2001) is the decomposition of this climatic variability into a part which is called the *forced climate variability* (e.g. due to solar radiation, anthropogenic warming and essential natural events such as volcanic eruptions or seasonal effects) and a part referred to as the *internal climate variability* (all low-frequency climate variability that is present in the climate system independently of external forcing). In literature, depending on the object of study, forced climate variability is associated with the signal and internal climate variability with noise (Kravtsov and Callicutt, 2017).

Approaches to estimate forced signal and internal noise components from a given time series are numerous in literature. Statistical models using methods such as empirical orthogonal function analysis (Monahan et al., 2009), singular spectrum analysis (Hassani, 2007), multi-

channel singular spectrum analysis (Ghil et al., 2002), multi taper method (Venegas, 2001) or diverse extended and modified versions of these (Venegas, 2001) provide simulated spatial and temporal structures which can then be compared to those in the observed time series. Semi-empirical approaches use partially or fully coupled simulations of the climate and its dynamics combined with observations of the climate system (Kravtsov and Callicutt, 2017). In this thesis, instead of the classical isolation of forced signal and noise, the time series of climatic variability is decomposed into a stationary and a non-stationary part. The non-stationary part of the variability may contain both external forcing (i.e. the signal) and extremely low-frequent internal variability (i.e. the part of the noise that cannot be differentiated from signal due to its highly extended temporal scale). Consequently, the stationary part of variability is defined as the noise except for its ultra-low frequent part.

The main new idea is to use *Wiener Filtering* in the orthogonal subspace associated with dominant spatiotemporal modes of variability associated with the Multi-Channel Singular Spectrum Analysis (M-SSA) (Ghil et al., 2002) to identify the non-stationary part of variability in the surface temperatures. The filter applies under the assumption that the stationary part of variability can be described by a spatially extended linear empirical stochastic model. The linear empirical stochastic model is implemented in a multi-scale formulation comprised of two parts: A classical linear inverse model (Penland, 1996) is used for annual temperature data, and a multilevel nonlinear regression model (Kravtsov et al., 2005) is applied for monthly data. The novel aspect of the analysis here is the blending of the two simulations at annual time scale, which provides a better fit of the simulated M-SSA spectra to the observed M-SSA spectrum for non-leading modes.

For the purpose of analysis, the methodology sketched above will be applied to two different simulated SAT data ensembles and one assimilated data model, which will be introduced in Chapter 2 and includes a more detailed description of the methodology.

Chapter 3 will provide results of the analysis for each of the two data sets and point out

commonalities and differences, particularly in relation to the observed temperature time series. The main research part of the project is the identification of the spatiotemporal structure of the difference between the simulated and the observed signals.

Chapter 4 gives a summary and provides a discussion of the results.

Finally, Chapter 5 highlights potential future work in climate signal detection.

2 Data sets and analysis methodology

This chapter is organized as follows. The analyzed data sets and basic data pre-processing, including removal of the seasonal cycle and empirical orthogonal function (EOF) based data compression, are described in section 2.1. Section 2.2 explains in more detail the empirical stochastic noise models. Finally, section 2.3 describes the Wiener filtering of the non-stationarity part of variability in M-SSA phase space.

Note that parts of the methodology, such as EOF data compression and M-SSA are standard techniques in Atmospheric Sciences. Detailed descriptions of some techniques are therefore moved to Appendices (see notes in this chapter).

2.1 Data sets and pre-processing

2.1.1 Data sets

In this thesis, three different data sets are analyzed: *LENS*, *CMIP5* and *20CR*.

Large Ensemble Project (LENS) data is a publically available data set created by fully-coupled climate model simulations conducted by the *Community Earth System Model* (Kay et al., 2015). It contains 40 ensemble members representing the simulations of the 20th century evolution of surface temperature time series on a *1-degree* latitude-longitude grid covering the globe. Each of the 40 ensemble members experienced the same external forcing (e.g. solar radiation, anthropogenic warming, volcanic eruptions etc.). Since the simulations started at the same atmospheric state back in the year 1850, due to growing errors in the

atmospheric conditions the ensemble members can be seen to be statistically independent from about 1920. The LENS data is available and investigated in the 93 years time range from 1920 to 2012.

The *Coupled Model Intercomparison Project Phase 5 (CMIP5)* data is a data set distributed by the *World Climate Research Programme* (Taylor et al., 2012), which contains, among other things, historical (1880-2005) climate simulations performed by different climate models. Multiple simulations of a given model formed by perturbed initial conditions allow one to estimate the contributions of external forcing and internal variability to the model simulated climates. Following the model selection used by Kravtsov and Callicut (2017), in this thesis 17 different models and the total of 111 simulations are considered.

The *Twentieth Century Reanalysis (20CR)* data (Compo et al., 2011) is the product of a reanalysis project conducted by the *National Oceanic and Atmospheric Administration* that assimilates available observations into a climate model to get the best dynamically consistent fields of climatic variables, including SAT.

2.1.2 Pre-processing of data sets

To pre-process SAT data and exclude effects of irrelevant sources of variation to later results, the climatological mean is subtracted, and the first five harmonics of the annual cycle are linearly removed from the data at each grid point to form anomalies (Cryer and Kellet, 1991).

Next, EOF based data compression (Monahan et al., 2009) is applied separately to annual-mean data and to monthly data anomalies with respect to annual mean for each year; 75 annual EOFs and 200 EOFs of monthly anomalies account for more than 95% of the respective data sets' total variance (see the Appendix A for further details on EOF analysis). Note that, due to the prior pre-processing steps, the 75 annual and 200 monthly PCs of SAT-anomalies coming along with EOF analysis represent the *stationary* part of the noise.

To account for meridian convergence (grid box sizes vary for different latitudes and therefore

bias variance analysis), in computing EOFs, the anomalies data is weighted by $\sqrt{cos(\phi)}$ where $-90° \leq \phi \leq 90°$ is the degree of latitude of the considered grid box (Chung and Nigam, 1999).

2.2 Empirical stochastic noise model

The main goal of the analysis methodology explained in this chapter is to filter out the stationary part of the SAT internal variability from the multi-variate time series associated with the SAT principal components (PCs) which are a result of EOF analysis. The internal variability (noise) in this context is defined as the part of variability that can be modeled as a spatially extended stationary stochastic process.

The corresponding empirical stochastic model of stationary noise is constructed in two parts, the annual (implemented by a one-level linear inverse model; see subsection 2.2.1) and the monthly sub-model (implemented by a three-level linear inverse model; see subsection 2.2.2). While integrating the annual sub-model, the monthly model is blended into it in order to obtain a combined set of 100 annual stationary noise realizations (see subsection 2.2.3).

2.2.1 One-level linear inverse stochastic models

Classical linear inverse models were first introduced by Penland (1989, 1996). In our case, the value of the i-th mode of the noise matrix $\boldsymbol{x}$ at time $n+1$ is modeled by a linear function of 15 surrounding modes[1] at time n with residuals $\boldsymbol{r}$ being simulated as spatially correlated Gaussian white noise:

$$\boldsymbol{x}_i^{n+1} = \boldsymbol{B}_i \; \boldsymbol{x}^n + \boldsymbol{r}^{n+1} \tag{2.1}$$

The coefficient estimates $\boldsymbol{B}$ and the covariance matrix $C = \langle \boldsymbol{r} \; \boldsymbol{r}^T \rangle$ are obtained by multiple linear regression from the 75 annual PCs available after EOF data compression of annual anomalies data.

[1]This restriction reduces the effective number of coefficients to be estimated and avoids overfitting.

The noise data for initial time step $n = 1$ is simulated by the spatially correlated Gaussian white noise only.

2.2.2 Three-level linear inverse stochastic models

Multi-level inverse models were described by Kravtsov et al. (2005) including quadratical terms. In this thesis, only a linear version of it is considered. Compared to the one-level version, the multi-level inverse model includes additional model levels in order to account for serial correlations and dependence of residuals $\boldsymbol{r}$ on the modeled process $\boldsymbol{x}$. In the second model level, the first-level residuals $\boldsymbol{r}^{(0),n+1}$ given at time step $n+1$ are expressed as a linear function of the extended matrix $[\boldsymbol{x}^n,\ \boldsymbol{r}^{(0),n}]$ of 15 surrounding modes of the modeled process $\boldsymbol{x}$ and first-level residuals at time n. The third level sets the second-level's residuals $\boldsymbol{r}^{(1),n+1}$ into linear relation with $[\boldsymbol{x}^n,\ \boldsymbol{r}^{(0),n},\ \boldsymbol{r}^{(1),n}]$:[2]

$$\begin{aligned} \boldsymbol{x}_i^{n+1} &= \boldsymbol{B}_i^{(0)}\ \boldsymbol{x}^n + \boldsymbol{r}_i^{(0),n+1} \\ \boldsymbol{r}_i^{(0),n+1} &= \boldsymbol{B}_i^{(1)}\ [\boldsymbol{x}^n\ \boldsymbol{r}^{(0),n}] + \boldsymbol{r}_i^{(1),n+1} \\ \boldsymbol{r}_i^{(1),n+1} &= \boldsymbol{B}_i^{(2)}\ [\boldsymbol{x}^n,\ \boldsymbol{r}^{(0),n},\ \boldsymbol{r}^{(1),n}] + \boldsymbol{r}_i^{(2),n+1} \end{aligned} \tag{2.2}$$

Again, the coefficient estimates $\boldsymbol{B}^{(0)}$, $\boldsymbol{B}^{(1)}$ and $\boldsymbol{B}^{(2)}$ as well as the covariance matrices $C^{(0)} = \langle \boldsymbol{r}^{(0)}\ \boldsymbol{r}^{(0)^T} \rangle$, $C^{(1)} = \langle \boldsymbol{r}^{(1)}\ \boldsymbol{r}^{(1)^T} \rangle$ and $C^{(2)} = \langle \boldsymbol{r}^{(2)}\ \boldsymbol{r}^{(2)^T} \rangle$ are obtained by prior multiple linear regression from the 200 PCs of monthly anomalies data.

The third-level's residuals $\boldsymbol{r}^{(2)}$ as well as the initial time step of the modeled process $\boldsymbol{x}$ are modeled by spatially correlated Gaussian white noise.

[2]Note that the linear three-level inverse model (2.2) fulfills the form of an autoregressive-moving average model (Cryer and Kellet, 1991) also, which can be seen by formulating the three-level model in one equation including time-lagged values of $\boldsymbol{x}_i$, $\boldsymbol{r}^{(0)}$, $\boldsymbol{r}^{(01)}$ and $\boldsymbol{r}^{(2)}$.

2.2.3 Model Integration and Blending

To produce 100 statistical realizations of the stationary noise, the three-level linear inverse model (2.2) is integrated forward 100 times for monthly temperature anomalies. By construction, the monthly noise realizations $\boldsymbol{x}_m$ are given in the phase space of monthly EOFs E_m and need to be projected onto the phase space of the annual EOFs E_a by

$$\begin{aligned} &\boldsymbol{T}_m = \boldsymbol{x}_m \cdot \boldsymbol{E}_m^T, \\ &\boldsymbol{T}_a \text{ is obtained by annual averages of } \boldsymbol{T}_m, \\ &\boldsymbol{x}_{a,(1)} = \boldsymbol{T}_a \cdot \boldsymbol{E}_a^T \cdot \left(\boldsymbol{E}_a^T \cdot \boldsymbol{E}_a\right)^{-1}. \end{aligned} \tag{2.3}$$

where $\boldsymbol{x}_{a,(1)}$ denotes the annualized noise realizations from the monthly model.

While integrating the one-level inverse model (2.1) for annual simulations $\boldsymbol{x}_{a,(2)}$, the annualized monthly simulations $\boldsymbol{x}_{a,(1)}$ are blended into them at each integration time step by

$$\boldsymbol{x}_a^{(n+1)} = c \cdot \boldsymbol{x}_{a,(2)}^{(n+1)} + \boldsymbol{x}_{a,(1)}^{(n+1)} \tag{2.4}$$

The blending factor c is computed as

$$c = \sqrt{1 - \frac{F2}{F1}}$$

where the fraction $\frac{F2}{F1}$ represents the ensemble-average proportion of variance that is carried by the 75 modes of simulated noise and variance that is inherent in the 75 original annual PC modes, i.e.

$$F1 = \sum_{k=1}^{75} Var(P_a^{k-th\ mode}) \qquad \text{and} \qquad F2 = \frac{\sum_{s=1}^{100}\sum_{k=1}^{75} Var(x_{a,(1),s-th\ simul.}^{k-th\ mode})}{100}.$$

In other words, the blending factor c secures that the final version of noise realizations $\boldsymbol{x}_a$

has the same overall variance as the original PC input.

2.3 Wiener filtering in M-SSA phase space

2.3.1 Multi-channel singular spectrum analysis

M-SSA, the multivariate extension of singular spectrum analysis, is a methodology for decomposing a set of time series at different spatial locations, where each spatial series is referred to as a *channel*, into the sum of independent components including trends, oscillatory behavior and unstructured noise (Hassani, 2007). By selecting a subset of the decomposing components, one can then reconstruct the time series fully (if one selects all components) or in parts (see Appendix B for technical details).

Mainly, M-SSA is a version of the EOF analysis applied to the data time series augmented by its lagged copies. The covariance matrix of so-called trajectory matrices is diagonalized, which provides a set of eigenpairs (λ_k, E^k) where the eigenvectors E^k can be interpreted as EOFs, and corresponding *space-time PCs (hereafter ST-PCs)* A^k. The eigenvalues λ_k are the variances associated with the decomposing modes and therefore provide the variance spectrum (Ghil et al., 2002).

Note that, in this thesis, the trajectory matrices for both original data and noise realizations are constructed with $M = 15$ lags, which allows to capture periodicities as long as 15 years and enables at least $N/M = 126/15 \approx 8$ repetitions of the features at interest (Ghil et al., 2002). The number of channels in the annual phase space is $L = 75$.

2.3.2 M-SSA based Wiener filtering in the analyzed models

Fig. 2.1 presents the variance spectra of the original anomalies input time series (blue crosses) and the ensemble-mean noise (black plus-symbols) shown on the example of the first simulation run of the *CNRM-CM5* model from *CMIP5* data. Apparently, the mean noise spectrum

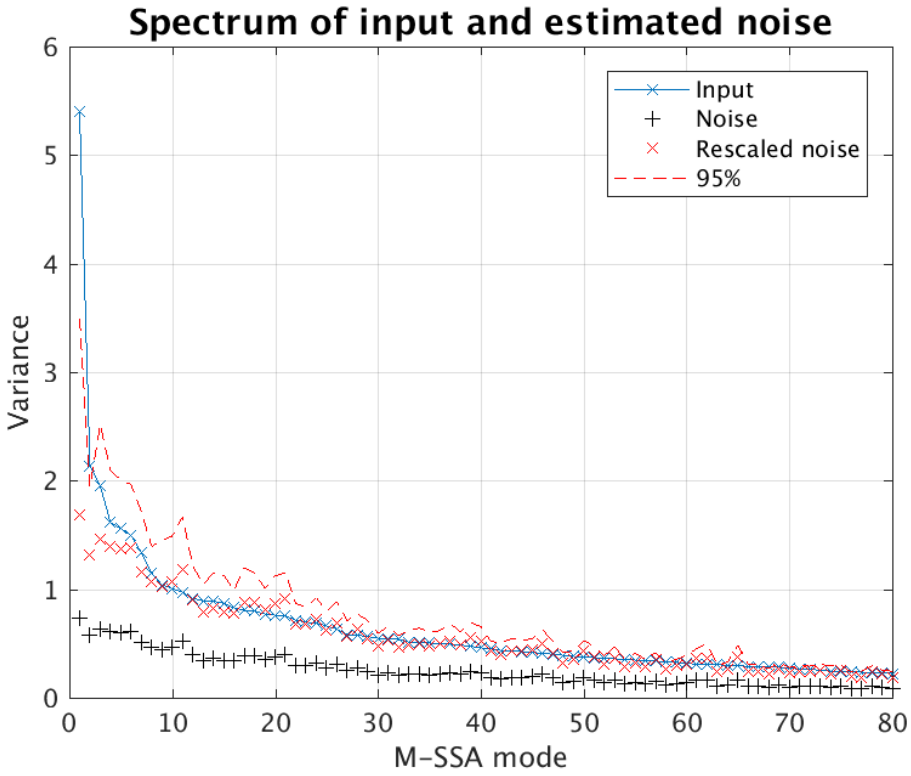

Figure 2.1: M-SSA spectrum of input data (blue crosses) and ensemble-mean noise spectrum (black plus-symbols), as well as the mean (red crosses) and the 95th percentile of rescaled noise spectra (dashed red line; see text for details).

underestimates the input spectrum. We therefore rescale the noise spectra by least-square fitting the ensemble-mean noise spectrum to the spectrum of the input signal for the trailing modes, where the input time series is expected to be dominated by noise. The rescaled ensemble-mean noise spectrum (red crosses) and the 95% percentile of the noise spectra (dashed red line) complete the graph.

Since the noise was modeled by a stochastic model describing the stationary part of climatic variability, one can derive that those M-SSA modes for which the input variance exceeds a high percentile of the stationary noise spectra (in this thesis it was chosen the 95th percentile) are very likely associated with the non-stationary part of variability. For the simulation run depicted in fig. 2.1, the first two of the M-SSA modes were detected to include non-stationary variance.

Now, Wiener filtering (Allen and Smith, 1997; Vaseghi, 2008) is applied in order to select and weight those ST-PC modes that should be used for reconstruction of the secular signal (i.e. the stationary part of climate variability). For the k-th M-SSA mode, the so-called *signal-to-noise ratio*

$$w_k = \frac{Var(Signal_k)}{Var(Signal_k) + Var(Noise_k)} \tag{2.5}$$

provides a measure of how many percent of the total variance is captured by noise. Since $Var(Total_k) = Var(Signal_k) + Var(Noise_k)$, equation 2.5 can be restated as

$$w_k = \frac{Var(Total_k) - Var(Noise_k)}{Var(Total_k)}. \tag{2.6}$$

In our case, $Var(Total_k)$ is given by the k-th value of the input spectrum and $Var(Noise_k)$ is, by assumption, represented by the 95th percentile of the rescaled simulated variance noise spectra. Note that w_k is set equal to 0 if $Var(Noise_k) > Var(Total_k)$, i.e. if the 95th percentile of noise variance exceeds the input spectrum value.

The so-computed weights reduce or eliminate the impact of the ST-PCs in the k-th reconstruction equation

$$R_l^k(t) = \frac{1}{M} \sum_{j=L_t}^{U_t} w_k A^k(t-j+1) E_l^k(j)$$

for certain lower and upper summation boundaries L_t and U_t and channel $l \in \{1, 2, ..., L\}$ (compare to Appendix B). The reconstructions (RC) are then summed over the modes of ST-PCs in order to obtain the portion of variability associated with the signal. The resulting time series is transformed back to physical space by means of the initial EOFs which provides the filtered estimate of the non-stationary signal for the considered model realizations. The reconstruction is repeated for all L channels.

The above-described filtering technique is in its way "optimal" because it minimizes the error and provides, by construction of the filter, the estimate closest to the original signal (Allen

and Smith, 1997).

3 Results

In sections 3.1 to 3.3, the estimation of the secular, non-stationary signal (including both forced component and secular low-frequency component), described in chapter 2, is analyzed for the data sets introduced in subsection 2.1.1. In section 3.4 we then subtract the models from observations and analyze the difference.

3.1 LENS analysis

For *LENS* simulations, the secular signal is estimated by the ensemble-mean of the reconstructed secular signals from 40 individual runs after M-SSA based Wiener filtering. Estimation uncertainty is derived from the spread of the 40 individual estimations in terms of its standard deviation. The "true" signal is defined via the ensemble average of the original SAT data from the 40 runs.

The right upper plot in fig. 3.2 depicts the estimated forced signal (black line) together with its estimation uncertainty (dashed black lines; showing the range of +/- 1 standard deviation) next to the "true" forced signal (red line) and the 40 individual runs (grey lines) for the Northern Hemispheric Multidecadal Oscillation (NMO) which is defined as the multidecadal component of internal Northern Hemisphere mean temperature variance (Steinman et al., 2015).

The ensemble-mean reconstruction of the secular signal is fairly close to the model forced signal. One can therefore conclude that M-SSA based Wiener Filtering is able to identify the forced signal in *LENS* simulations. Furthermore, the smooth form indicates that the

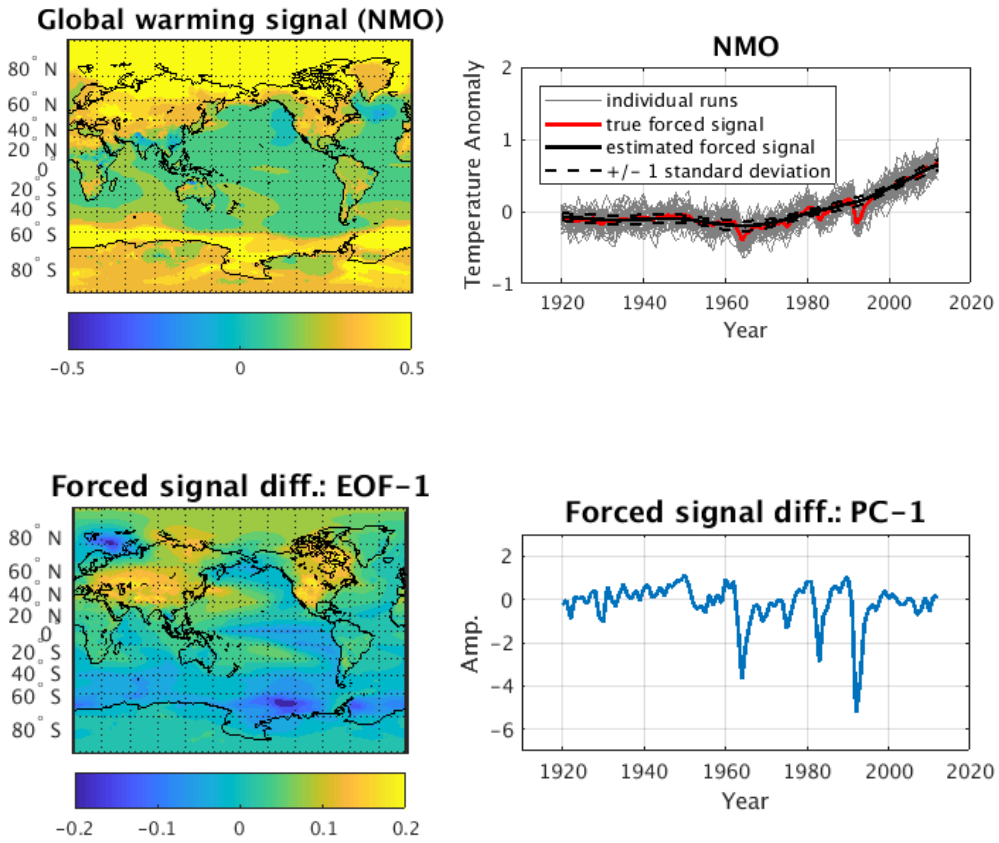

Figure 3.2: **On top:** Spatial global warming map (left) and estimated forced signal (black line), estimation uncertainty (dashed black lines), "true" forced signal (red line) and the 40 individual runs (grey lines) of *LENS* data for the NMO index (right). **Below:** Leading EOF (left) and PC (right) of the difference between non-stationary signal estimate and true forced signal for *LENS* data.

signal in this model is primarily associated with low-frequency forced variability, whereas the intermittent interannual forced signal associated with volcanic eruptions (recognisable by pronounced troughs in grey and red curves) is not captured. Secular internal variability appears to be rather small in *LENS* simulations.

Also, we observe a fairly narrow ensemble spread coming along with the secular signal estimation using *LENS* data. The narrow band of standard deviation can be explained by the relatively large number of ensemble members belonging to a consistent climate model (Kravtsov and Callicutt, 2017).

The top left plot depicts corresponding spatial global warming tendencies. Obviously, *LENS* data predicts fastest warming in the two polar regions (so-called polar intensification).
The two bottom plots of fig. 3.2 show the leading EOF and PC after an EOF analysis of the difference between the non-stationary estimate and the true forced signal. Apparently, the time points of pronounced troughs in the leading PC curve, which captures about 24.8% of the total variance, coincide with the dates of volcanic activity seen in the original runs. This suggests that the difference between non-stationary estimate and true forced signal primarily stems from volcanic forcing. The remaining 75% of the total variance distribute among the subsequent modes, with 12.7% to the second and 10.8% to the third PC, and are explained by internal variability and biases in the estimation of the forced signal.

3.2 CMIP5 analysis

3.2.1 Individual model estimation

Signal estimation by an individual *CMIP5* model is exemplarily illustrated for the *CNRM-CM5* model. Note that analysis is representative for any other choice of the set of 17 models. Analogously to section 3.1, estimated and true signal are computed by the ensemble average over the 10 simulation runs contained by *CNRM-CM5*.
Fig. 3.3 (top right) shows that the implemented filtering methodology provides an equally good isolation of the forced signal for the *CNRM-CM5* model than it did for the *LENS* ensemble. It similarly fails in capturing interannual volcanic forcing. The observations therefore corroborate conclusions made in section 3.1.
The secular signal estimates in the *CNRM-CM5* model show a larger spread. Probably this is because the CNRM has a larger internal variability at low frequencies. Also the uncertainty of the ensemble mean goes down with the increasing ensemble size, which however does not affect the spread of the individual simulations (or their filtered analogs).
Also, *CNRM-CM5* seems to predict a similar polar intensification in global warming (see

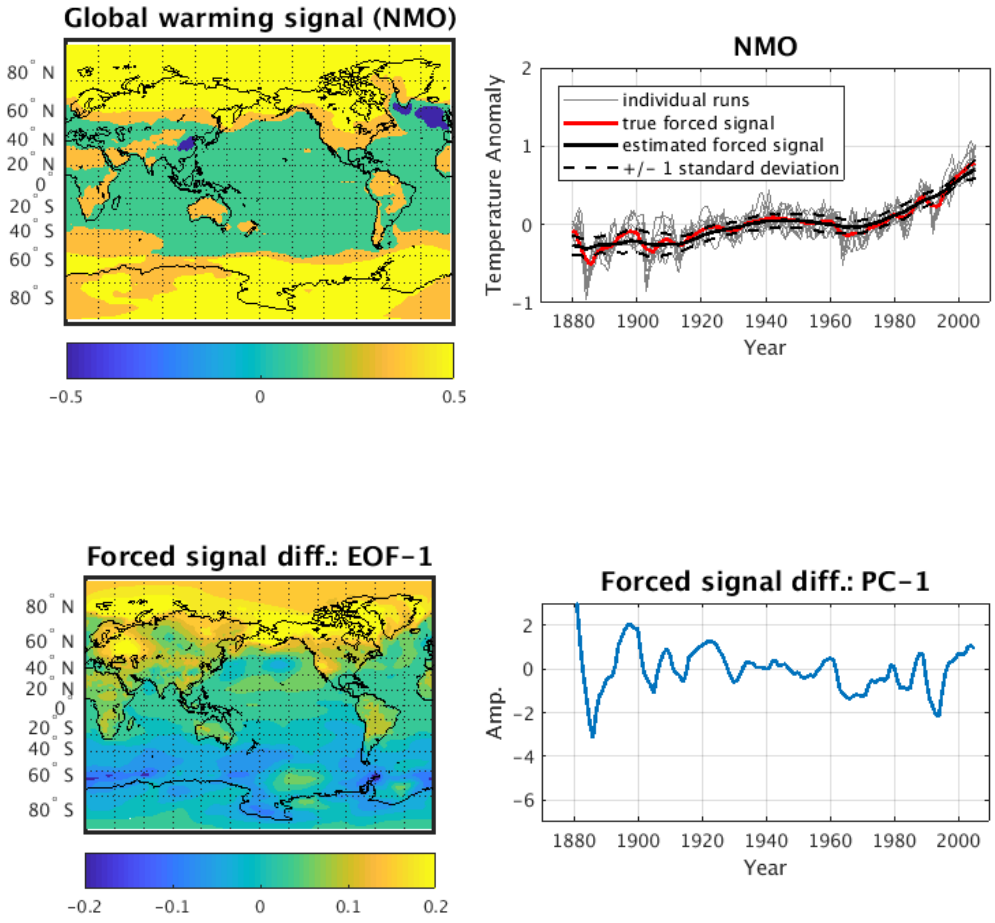

Figure 3.3: **On top:** Spatial global warming map (left) and estimated forced signal (black line), estimation uncertainty (dashed black lines), "true" forced signal (red line) and the 10 individual runs (grey lines) of the *CNRM-CM5* model in *CMIP5* data for the NMO index (right). **Below:** Leading EOF (left) and PC (right) of the difference between non-stationary signal estimate and true forced signal for *CNRM-CM5* model.

top right plot). EOF analysis of the difference between the estimated secular and the "true" signal provides three leading modes capturing 35.9%, 15.6% and 7.3% of the variance. The leading PC (see lower plots in fig. 3.3) show similar, however not as pronounced troughs at time of volcanic eruptions in external forcing.

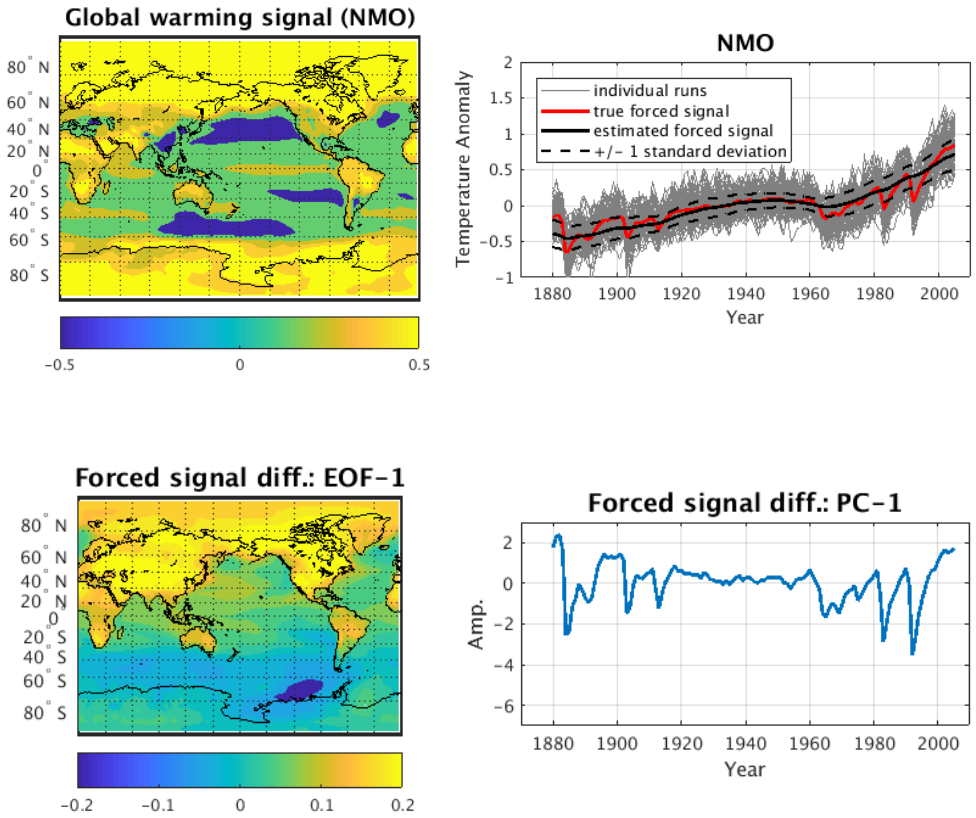

Figure 3.4: **On top:** Spatial global warming map (left) and estimated forced signal (black line), estimation uncertainty (dashed black lines), "true" forced signal (red line) and the 111 individual runs (grey lines) of the overall *CMIP5* model ensemble for the NMO index (right). **Below:** Leading EOF (left) and PC (right) of the difference between non-stationary signal estimate and true forced signal for *CMIP5* model ensemble.

3.2.2 Overall ensemble estimation

Now, the estimated and true signal are computed by the overall *CMIP5* models ensemble data consisting of 111 simulation runs. This provides a larger ensemble size but involves 17 models characterized by different sets of external forcing.

While fig. 3.4 suggests a similar signal estimation quality and spatial global warming structure and therefore confirms prior analysis and conclusions, the spread of the secular signal estimates is clearly larger here, since internal variability is dominated by the *model uncertainty* here. We define model uncertainty as the uncertainty which is the uncertainty arising

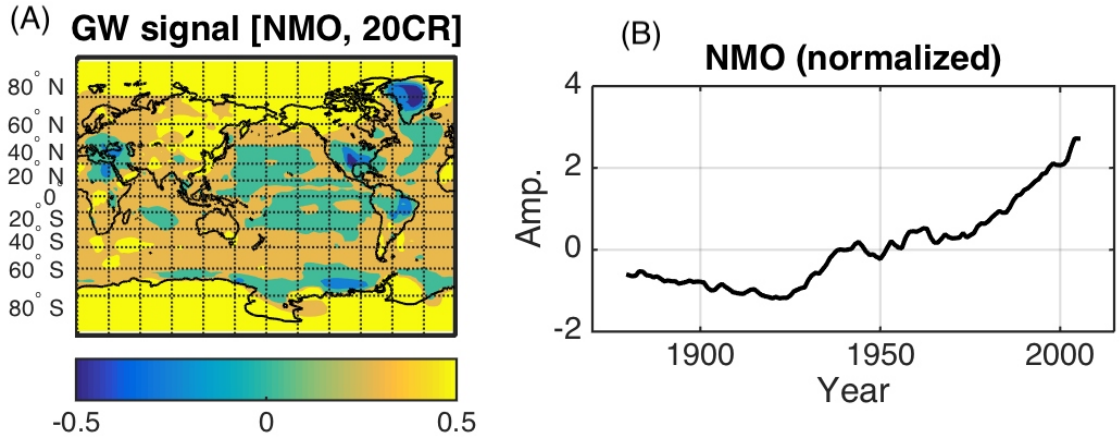

Figure 3.5: Signal Spatial global warming map (left) and estimated forced signal (black line) of the assimilated observational *20CR* data for the NMO index (right).

from different forcing assumptions and physical parameterizations used in the 17 considered *CMIP5* sub-models (Kravtsov and Callicutt, 2017).

The variability of differences between secular signal estimation and true signals is clearly dominated by the leading PC with 69.8%; the second and third PC capture another 5.1% and 4.1% of the total variability. The dominating PC shows pronounced troughs at the dates of volcanic activity.

3.3 20CR analysis

For the one-run *20CR* observational data, we only show the filtered M-SSA based signal reconstruction. The estimated signal is obtained by its filtered M-SSA based reconstruction. Fig. 3.5 suggests that, compared to model estimations, the M-SSA based Wiener filtering of the observed signal leads to a richer structure of non-stationary variability since it is described by a larger number of significant M-SSA modes (right plot).

The spatial analysis of global warming supports conclusions from models. Since *20CR* consists of only one run, it is an unseparable mixture of internal variability and forced signal and the difference between signal reconstruction and original time series would be noise by construction. Therefore, no estimation spread and EOF analysis of the forced signal differences

is provided.

3.4 Comparing CMIP5-based and observed signal

The main research part of this thesis is to identify the spatio-temporal structure of the difference between the observed and simulated signals.

Given that our estimated secular signals in CMIP5 simulations primarily reflect the forced response of CMIP5 models, it makes sense to linearly subtract them from the observed secular signal at each grid point using linear regression. The regression residuals describe the part of the observed secular variability unaccounted for in CMIP5 simulations (Kravtsov and Callicutt, 2017; Steinman et al., 2015). To analyze and visualize the model-data differences, we again apply the M-SSA analysis to the 111 multi-variate difference time series obtained as above. By reconstructing the variability associated with statistically-significant leading modes, one obtains spatially extended time series of the observed multidecadal variability that is unexplained by the climate models considered.

We can also subtract the individual model ensemble-mean secular signals from all of this models simulations to define an internal component of the secular signal in each simulation. As a third component of analysis, the difference time series from step 2, denoted by dT, are projected onto the observed EOFs E_{obs} of M-SSA analysis, i.e.

$$P_{obs,proj} = dT \cdot E_{obs}.$$

Using the 20 leading PC's, a trajectory matrix D with maximal lag 65 is built and then the variance estimated by

$$\frac{\sum (D \cdot E)^2}{L}$$

where the squared in the numerator is meant element-wise and the scalar L is the row dimension of D minus 1.

Fig. 3.6 (A) depicts the observed (black line) and mean CMIP5 (blue line) data-model difference spectrum including errorbars which are representing the spread across the 111 estimated spectra. The 99% percentile of the projected spectrum completes the graph.

The M-SSA analysis of data-model differences identifies a pronounced pair of M-SSA modes altogether absent from model simulations. The black line presents the internal variability estimate in observations which is clearly larger than the one associated with the model-mean difference. One can conclude that the data-model differences are dominated by a pronounced multidecadal signal that is absent from the model simulations. Observational space-time patterns are only barely represented in the models.

The reconstruction of this pair of modes for different regional climate indices (see part (B)) identifies a multidecadal oscillation propagating across the climate index network (see part (C)) – a so-called stadium wave (Wyatt et al., 2012), which we will refer to as the Global Stadium Wave Multidecadal Oscillation.

The order of indices in the sequence of Fig. 3.6, part (C), (except for GMO) is chosen based on the visual analysis of the SAT anomaly propagation over a time period between 1921 and 1963, which roughly spans half of the oscillation period (Fig. 3.7). In year 1921, the oscillation is in its cold phase (cf. Fig. 3.6, part (C)), with the exception of four major positive SAT anomaly spots: west of Weddell Sea, in eastern equatorial Pacific, as well as over central US and Greenland. The development of an oscillation starts with emergence of the positive SST anomaly in the North Atlantic (1921-1930), which subsequently expands and growth along with SST anomalies in the North and Southwestern Pacific (1933-1942), then in the Southern Ocean and Antarctica (1941-1957) and, finally, over the Arctic region (1960-1963), at which point the oscillation arrives at its positive phase throughout the world (with the exception of four major negative SAT anomaly regions roughly at the same locations as their positive analogs 40 years ago).

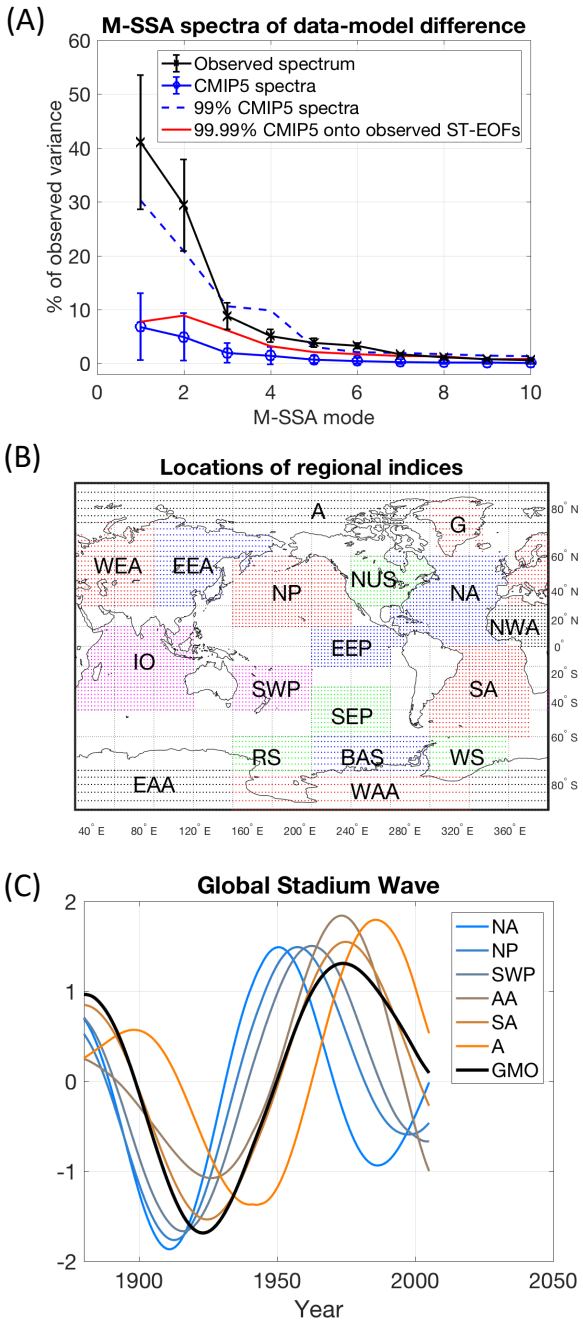

Figure 3.6: **(A)** M-SSA spectra of data-models secular difference (black) and model signals' deviations from individual model ensemble means (blue), both including uncertainty computed over 111 estimates (errorbars), and the 99th percentile of variances obtained by projecting the simulated signals onto the observed ST-EOFs of M-SSA analysis (red). **(B)** Locations of regional SAT indices. **(C)** Reconstructed time series associated with the leading M-SSA pair in selected regional indices. GMO (Global Multidecadal Oscillation) time series represents the reconstruction of the global-mean temperature. All time series are dimensionless.

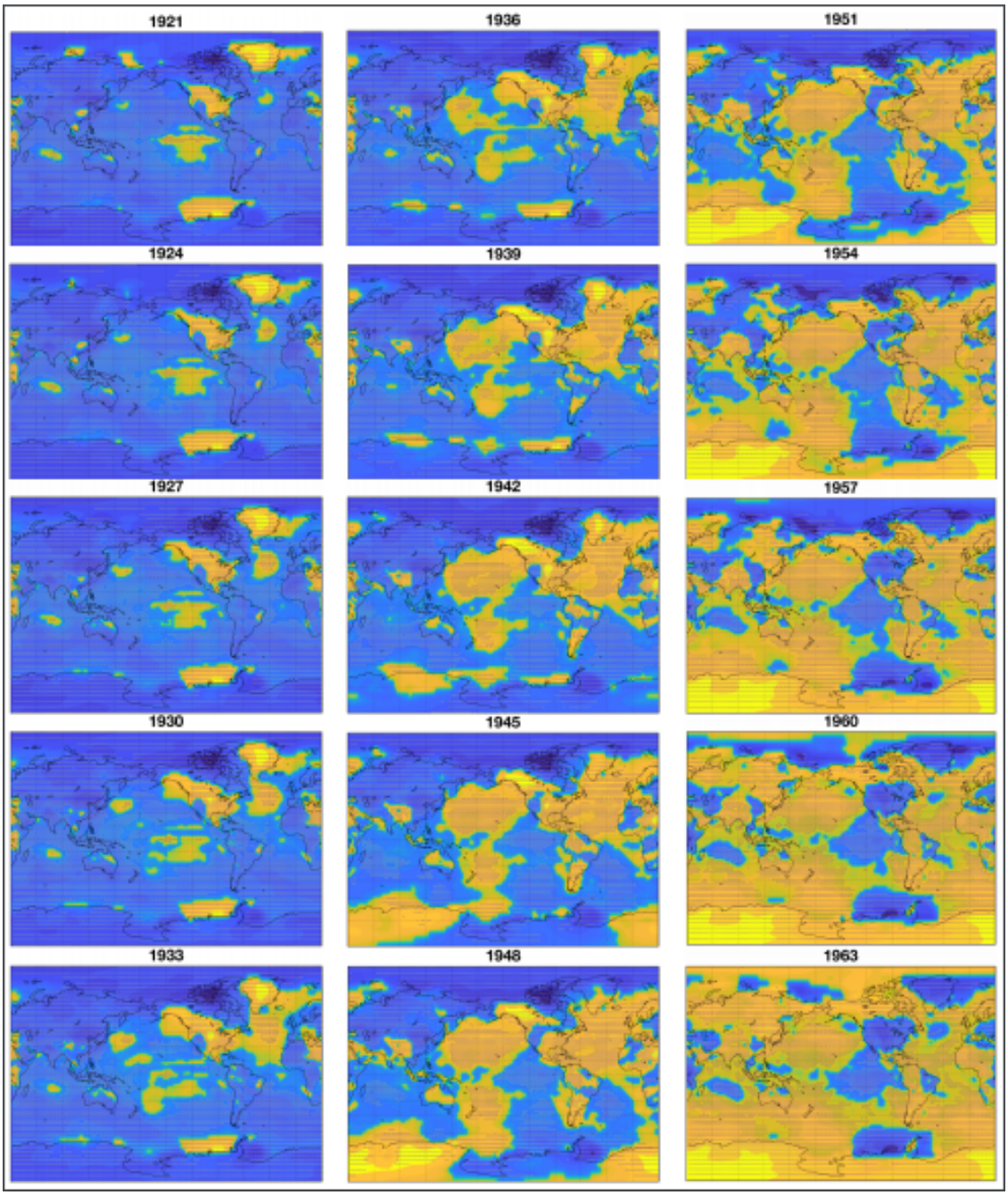

Figure 3.7: A 1921-1963 segment of the global stadium wave; shown are reconstructed SAT anomalies raised to the power of 1/7, which alleviates differences between SAT anomalies over ocean and over land to concentrate on the anomaly patterns and their propagation. Color axis is from -1.5 (saturated blue) to 1.5 (saturated yellow)

4 Summary and Discussion

4.1 Summary

M-SSA based Wiener filtering was implemented in order to filter out the stationary part of the SAT internal variability. In estimating the related secular non-stationary signal which contains both a forced and a secular low-frequent component, we find that this signal is dominated by the forced component. We then subtract the models from observations and see a large difference, which is presumably attributable to the internal variability present in observations but absent from the models.

4.2 Discussion of results

The key result of this study is the identification of a pair of pronounced global-scale modes of multi-decadal climate variability in the twentieth century which is not captured by any state-of-the-art climate model, here shown on the example of *LENS* and *CMIP5* models. Such a mode was previously proposed to explain a major fraction of variability in a network of oceanic and atmospheric climate indices over the Northern Hemisphere and termed the stadium wave (Kravtsov and Callicutt, 2017; Kravtsov et al., 2014; Wyatt et al., 2012). Here we show that the leading of the two modes has a global significance and provide a description of its worldwide evolution throughout the twentieth century.

The global stadium wave presented here was defined in terms of deviations of the observed SATs from the secular trends identified in CMIP5 models, which can be interpreted to be

approximations for the observed forced signal. In principle, it is still possible that these deviations are pronounced in part because of the potential biases in the CMIP5 derived forced signals. However, the oscillatory behavior of the global stadium wave and the spatial pattern of its delayed teleconnections strongly suggest that this mode reflects internal climate variability, perhaps associated with that of global oceanic conveyor-belt circulation.

5 Future Work

Future climate modeling efforts should strive to alleviate discrepancies between the observed and simulated multi-decadal climate variability. Analyzing the signal of data-model differences should help model development efforts and eventually improve the understanding of the physics behind the observed climate change.

In particular, signal estimation should be designed in a way such that it manages to capture the intermittent interannual forced signal associated with volcanic eruptions. A possible approach to do so could be using wavelet transforms which can be seen as a generalized form of (windowed) Fourier transform (Lau and Weng, 1995). Lau and Weng found in 1995, that wavelet transforms provide an improved time-frequency information compared to classical or windowed Fourier transform.

Also, including further fields such as sea level pressure can improve analysis, since multi-decadal variability might correlate between the two fields and therefore provides additional information. For instance, a coupled field analysis of sea surface temperatures and sea level pressure was conducted in Delworth and Mann (2000).

Acknowledgement

Thanks to Professors Kravtsov, Brazauskas and Evans, who helped me as members of the Defense Committee.

Special thanks to Professor Kravtsov for the outstanding support since starting work in the early autumn of 2017. Your detailed and patient feedback has helped me a lot in the preparation of this work.

I also thank the Talanx Foundation (Germany), which has given me financial support for a total of 3 years during my studies.

In general I thank my family and my wonderful friends who have always given me motivation and inspiration for my work.

Bibliography

Allen, M. R. and Smith, L. A. (1997). Optimal filtering in singular spectrum analysis. *Physics letters A*, 234(6):419–428.

Broomhead, D. S. and King, G. P. (1986). Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena*, 20(2-3):217–236.

Christensen, J. H., Kanikicharla, K. K., Marshall, G., and Turner, J. (2013). Climate phenomena and their relevance for future regional climate change.

Chung, C. and Nigam, S. (1999). Weighting of geophysical data in principal component analysis. *Journal of Geophysical Research: Atmospheres*, 104(D14):16925–16928.

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., et al. (2011). The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, 137(654):1–28.

Cryer, J. D. and Kellet, N. (1991). *Time series analysis*. Springer.

Delworth, T. L. and Mann, M. E. (2000). Observed and simulated multidecadal variability in the northern hemisphere. *Climate Dynamics*, 16(9):661–676.

Ghil, M., Allen, M., Dettinger, M., Ide, K., Kondrashov, D., Mann, M., Robertson, A. W., Saunders, A., Tian, Y., Varadi, F., et al. (2002). Advanced spectral methods for climatic time series. *Reviews of geophysics*, 40(1).

Hassani, H. (2007). Singular spectrum analysis: methodology and comparison.

Kay, J., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J., Bates, S., Danabasoglu, G., Edwards, J., et al. (2015). The community earth system model (cesm) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, 96(8):1333–1349.

Kravtsov, S. and Callicutt, D. (2017). On semi-empirical decomposition of multidecadal climate variability into forced and internally generated components. *International Journal of Climatology*, 37(12):4417–4433.

Kravtsov, S., Kondrashov, D., and Ghil, M. (2005). Multilevel regression modeling of nonlinear processes: Derivation and applications to climatic variability. *Journal of Climate*, 18(21):4404–4424.

Kravtsov, S., Wyatt, M. G., Curry, J. A., and Tsonis, A. A. (2014). Two contrasting views of multidecadal climate variability in the twentieth century. *Geophysical Research Letters*, 41(19):6881–6888.

Lau, K.-M. and Weng, H. (1995). Climate signal detection using wavelet transform: How to make a time series sing. *Bulletin of the American meteorological society*, 76(12):2391–2402.

Monahan, A. H., Fyfe, J. C., Ambaum, M. H., Stephenson, D. B., and North, G. R. (2009). Empirical orthogonal functions: The medium is the message. *Journal of Climate*, 22(24):6501–6514.

Penland, C. (1989). Random forcing and forecasting using principal oscillation pattern analysis. *Monthly Weather Review*, 117(10):2165–2185.

Penland, C. (1996). A stochastic model of indopacific sea surface temperature anomalies. *Physica D: Nonlinear Phenomena*, 98(2-4):534–558.

Steinman, B. A., Mann, M. E., and Miller, S. K. (2015). Atlantic and pacific multidecadal oscillations and northern hemisphere temperatures. *Science*, 347(6225):988–991.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A. (2012). An overview of cmip5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4):485–498.

Vaseghi, S. V. (2008). *Advanced digital signal processing and noise reduction*. John Wiley & Sons.

Venegas, S. A. (2001). Statistical methods for signal detection in climate. *Danish Center for Earth System Science Rep*, 2:96.

Wyatt, M. G., Kravtsov, S., and Tsonis, A. A. (2012). Atlantic multidecadal oscillation and northern hemispheres climate variability. *Climate Dynamics*, 38(5-6):929–949.

Appendix A (EOF analysis)

The following explanations and methodologies follow Venegas (2001). The main goal of EOF-based data compression is to reduce the relatively large SAT data set T into a smaller set of independent pieces of information which is, however, accounting for the most part of the time series' variance.
Assume we have N consecutive "observations" of temperature at each of M spatial grid points, forming a SAT data matrix T with dimensions NxM. Using *Singular Value Decomposition (SVD)*, T is decomposed into the product of the three matrices U (NxN), Σ (NxN) and E (MxN) such that

$$T = U \cdot \Sigma \cdot E^T$$

where V^T is the transposed of V and U has orthogonal columns.
By denoting $P := U \cdot \Sigma$ one obtains the decomposition

$$T = P \cdot E^T$$

where we call matrix P the *principal components (PCs)* of the data and matrix E the *empirical orthogonal functions (EOFs)*. In fact one can show that the columns of E solve an eigenvalue problem, and the same decomposition results, i.e. PCs and EOFs, could equally be obtained by a covariance matrix approach which is also presented by Venegas (2001).
The columns of P are time series of the original time length N whereas the columns of E are spatial patterns with length M. Together, P and E encode the entire spatiotemporal information of the time series.
By construction, the columns of P are ordered by descending variances of the associated time series. Also, since P is obtained from U and Σ by a linear mapping, P has orthogonal columns. Consequently, its covariance matrix has diagonal form

$$Cov(P) = \frac{P^T \cdot P}{N-1} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N \end{bmatrix}$$

where the diagonal elements λ_1, λ_2, ..., λ_N are the variances associated with the N PC modes (columns of P) with $\lambda_1 > \lambda_2 > ... > \lambda_N$.
For each of these N modes of the decomposition, $100 \cdot \frac{\lambda_i}{\sum_{j=1}^{N} \lambda_j}$ denotes the percentage of variance that is accounted for by mode $i \in \{1, 2, ..., N\}$. For a particular percentile $\alpha \in (0, 1)$ one can find the minimum number $N_\alpha \in \mathbb{N}$ so that $\frac{\sum_{i=1}^{N_\alpha} \lambda_i}{\sum_{j=1}^{N} \lambda_j} \geq \alpha$, i.e. the first N_α PC modes

account for more than $100 \cdot \alpha$ % of the climatic variability in the original time series. Truncating both PC and EOF matrices to the first N_α modes leads to matrices

$$P_\alpha = \begin{bmatrix} P_{1,1} & P_{1,2} & ... & P_{1,N_\alpha} \\ P_{2,1} & P_{2,2} & ... & P_{2,N_\alpha} \\ \vdots & \vdots & \vdots & \vdots \\ P_{N,1} & P_{N,2} & ... & P_{N,N_\alpha} \end{bmatrix} \quad \text{and} \quad E_\alpha = \begin{bmatrix} E_{1,1} & E_{1,2} & ... & E_{1,N_\alpha} \\ E_{2,1} & E_{2,2} & ... & E_{2,N_\alpha} \\ \vdots & \vdots & \vdots & \vdots \\ E_{M,1} & E_{M,2} & ... & E_{M,N_\alpha} \end{bmatrix}$$

which decompose an approximation of the original time series

$$T_\alpha = P_\alpha \cdot E_\alpha^T.$$

Since P_α and E_α are both only truncated in their column dimension, T_α has still dimension NxM.

Appendix B (M-SSA)

The following explanations are taking from Ghil et al. (2002). In the following, each time series specified at a certain spatial location is referred to as a *channel*.
Given spatial dimension L and using the methodology described by Broomhead and King (1986), for each channel $l \in \{1, 2, ..., L\}$ a trajectory matrix

$$\tilde{\boldsymbol{X}}_m = \begin{bmatrix} X_l(1) & X_l(2) & \dots & X_l(M) \\ X_l(2) & X_l(3) & \dots & X_l(M+1) \\ \vdots & \vdots & \vdots & \vdots \\ X_l(N') & X_l(N'+1) & \dots & X_l(N) \end{bmatrix}$$

with $N' = N - M + 1$ can be constructed.
The L individual trajectory matrices are then combined to a $N' \times (L \cdot M)$ - dimensional joint trajectory matrix

$$\tilde{\boldsymbol{X}} = \left(\ \tilde{\boldsymbol{X}}_1, \ \tilde{\boldsymbol{X}}_2, \ \dots, \ \tilde{\boldsymbol{X}}_M \ \right).$$

The covariance matrix of $\tilde{\boldsymbol{X}}$ is a $(LM \times LM)$ - dimensional blog matrix of the form

$$\tilde{\boldsymbol{C}}_{\tilde{\boldsymbol{X}}} = \frac{1}{N'} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} = \begin{bmatrix} C_{1,1} & C_{1,2} & \dots & C_{1,L} \\ C_{2,1} & C_{2,2} & \dots & C_{2,L} \\ \vdots & \vdots & C_{l,l'} & \vdots \\ C_{L,1} & C_{L,2} & \dots & C_{L,L} \end{bmatrix}$$

with the blog

$$\tilde{\boldsymbol{C}}_{l,l'} = \frac{1}{N'} \tilde{\boldsymbol{X}}_l^T \tilde{\boldsymbol{X}}_{l'}$$

representing the covariance matrix for the individual trajectory matrices of channels l and l'.
Diagonalizing $\tilde{\boldsymbol{C}}_{\tilde{\boldsymbol{X}}}$ provides $L \cdot M$ eigenpairs (λ_k, E^k), where the eigenvectors E^k contain L consecutive M-long segments denoted by E_l^k. Since here diagonalizing is only a special case of a singular value decomposition, the E^k really are EOFs, which explains the affinity of M-SSA and EOF analysis. By projecting $\tilde{\boldsymbol{X}}$ onto these EOFs one obtains the corresponding space-time PCs

$$A^k(t) = \sum_{j=1}^{M} \sum_{l=1}^{L} X_l(t+j-1) E_l^k(j)$$

with $1 \leq t \leq N'$.
The k-th reconstruction for channel l at time t is then given by

$$R_l^k = \frac{1}{M} \sum_{j=L_t}^{U_t} A^k(t-j+1) E_l^k(j)$$

where L_t and U_t are certain lower and upper boundaries of the summation depending on time step t.