# Nonlinear Laplacian spectral analysis: application to climate science

Alexis Montavon
Under the supervision of Eniko Szekely & Olivier Verscheure

June 2019

**Abstract**

In this report, we present a nonlinear way to extract spatiotemporal patterns from high dimensional data using nonlinear Laplacian spectral analysis (NLSA). A technique that we apply to separate spatiotemporal patterns in order to perform dimension reduction and set the groundwork for later predictions on worldwide temperature data using different time-lags. We also compare this method to the already known Singular spectrum analysis (SSA). This algorithm has the advantage that it does not require any preprocessing of the data.

EPFL

# Contents

# 1 Introduction

Most of the modern techniques applied in climate science for dimension reduction and prediction, such as temperature forecast and the study of climate phenomenons, are using linear algorithms and different types of physical measures. With this paper we aim to demonstrate a nonlinear method using only temperature data in order to find and separate spatiotemporal patterns in the purpose of dimension reduction and prediction. The algorithm used in this study is the nonlinear Laplacian spectral analysis, a technique derived from the singular spectrum analysis with the differences that it is nonlinear, does not require preprocessing and enforces smoothness on the leading Laplace-Beltrami eigenfunctions [3].

We start by presenting the NLSA algorithm in the way we used it for this analysis. Then, we will give a description of the dataset we used and explain the spatiotemporal patterns we found under different parameters and data resolutions. We will also make a comparison with the patterns extracted using SSA on the same dataset. Finally we will conclude with the applications that can be done using the eigenfunctions resulting of this technique in climate science. Although this paper does not contain actual predictions, it sets the foundation for a later study.

# 2 Methodology

## 2.1 NLSA algorithm

The NLSA algorithm presented by D. Giannakis & A. J. Majda [2, 3] and its application to climate science [7], aims to extract spatiotemporal patterns from high dimensional data generated by dynamical systems. It works in three main steps:
1. Construction of an embedded space via the Takens method of delay.
2. Construction of a discrete Laplace-Beltrami operator via kernel methods applied to the embedded space.
3. Construction of an empirical set of basis functions through the eigendecomposition of the Laplace-Beltrami operator that will be used for dimension reduction and pattern extraction.

### 2.1.1 Takens method of delay

Taking a time series $x(t_i) = (x^1(t_i), ..., x^n(t_i))$ in a subspace of $\mathbb{R}^n$ of $s$ samples taken a time $t_i = i\delta t$ with a time interval of $\delta t$, the method of delay helps recover some phase-time information lost by partial observations. We can represent $x(t)$ in the embedded space as the sequences of data over the time-lag window $q\delta t$. Giving:

$$x(t) \mapsto X(t) = (x(t), x(t - \delta t), ..., x(t - (q - 1)\delta t)) \tag{1}$$

$X(t)$ becomes a high dimensional representation of trajectories of length $q\delta t$ in the physical space. For a sufficiently large $q$, this operation recovers the topology lost by partial observation [6]. We can then select the $k$ nearest neighbors for each data point in the embedded space, which reduces the amount of computations and allows us to look only at the most clustered part of the graph. This is one of the main advantages of the Laplacian graph over the SSA algorithm.

### 2.1.2 Discrete Laplace-Beltrami operator

The Laplace-Beltrami operator is based on the construction of a geometrical operator using pairwise measure of similarity decaying exponentially in data space. This specific kernel, $K$, is expressed in the Takens embedded space:

$$K(X(t_i), X(t_j)) = exp\left(-\frac{\|X(t_i) - X(t_j)\|^2}{\epsilon\|\zeta(t_i)\|\|\zeta(t_j)\|}\right) \tag{2}$$

With $\epsilon$ a parameter to control the bandwidth of the kernel and $\zeta(t_i) = X(t_i) - X(t_{i-1})$ a measure of the local space velocity of the data [7]. This pairwise evaluation leads to a symmetric kernel $K$.

After this first evaluation, we construct the discrete Laplace-Beltrami operator, $L$, by performing the following normalizations, proposed in the Diffusion maps algorithms [1].

$$\widehat{K}_{ij} = \frac{K_{ij}}{Q_i Q_j}, Q_i = \sum_{j=1}^{S} K_{ij} \tag{3}$$

$$P_{ij} = \frac{\widehat{K}_{ij}}{D_i}, D_i = \sum_{j=1}^{S} \widehat{K}_{ij} \tag{4}$$

$$L_{ij} = I - P_{ij} \tag{5}$$

With large enough data the discrete Laplace-Beltrami operator converges toward a continuous one. In order to save computational time and by the exponential decay property of the kernel, we will select only the $k_{nn}$ nearest neighbors after the computation of the pairwise distances. We make the assumption that the full attractor of the dynamical systems shows low-dimensional geometric structures associated with climate phenomenons and that the sampling is sufficiently dense to retrieve these structures [7].

### 2.1.3 Empirical set of basis functions

The eigendecomposition of $L$ results in a set of eigenvectors $v_i$ and their corresponding eigenvalues $\lambda_i$ such that:

$$Lv_i = \lambda_i v_i \tag{6}$$

These eigenvectors can be seen as discretely sampled functions or as time series $v_i(t_j) = v_{ji}$. Their ability to separate different kind of climate phenomenons, such as seasonal cycle, volcanic cycle or El Niño, as well as the smoothness of the retrieved patterns, will depend on the selected time-lag and the kernel parameters.
The eigenvalues $\lambda_i$ have a geometrical interpretation as the gradient of $v_i$. Following this interpretation we will select the leading $v_i$ with the smallest eigenvalues $\lambda_i$ as they will represent the features with slow variation and will reduce noise and parameter sensitivity [7].

One of the most important feature of the eigenvectors is the ability to recover spatiotemporal patterns in the ambient data space. We will then use the chosen eigenvectors $v_i$ to reconstruct these patterns and visualize them:

$$\widehat{X}_i = X v_i v_i^t \tag{7}$$

In order to keep a smaller scale in the reconstructions, we advise to centralize the original data.

## 2.2 SSA algorithm

The singular spectrum analysis is a known method to extract trends from climate data, as demonstrated by Mahdi Haddad in [4]. In general it is a tool used for a number of problems including smoothing, extraction of seasonality, periodicities and trends, finding structure in short time series and change-point detection [5]. In order to recover spatiotemporal patterns, we will use the two decomposition steps of the SSA algorithm, namely the embedding and the singular value decomposition.

### 2.2.1 Embedding

The embedding step is a mapping from a one-dimensional time series $x(t)$ into a multi-dimensional series $X(t)$, which is the same as the one described in 2.1.1.

### 2.2.2 Singular Value Decomposition

The singular value decomposition (SVD) is a factorization of a matrix $M$ of the form:

$$M = U\Sigma V^T \tag{8}$$

Where $U$ and $V^T$ are unitary matrices and their columns (respectively rows) are called left (right)-singular vectors. $\Sigma$ is a diagonal matrix with non-negative entries called singular values. We will look at the left or right-singular vectors $u_i$ or $v_i^t$, as the embedded matrix is symmetric, $u_i = -v_i^t, \forall i$. The reconstruction is done in the same way as (7) using $u_i$.

# 3 Dataset Description

The dataset we used for this project contains three dimensional worldwide images of near-surface air temperatures (in Kelvin), provided by the Max Planck Institute for Meteorology[1]. We used monthly and daily images of size $72 * 144$, from January 1870 to 2100. It is important to specify that the observed data exists only until 2005, the remaining 95 years are simulated data. In order to reduce the internal variability of the simulated data, we computed an average over 4 different models, each with different initial conditions. It also turns out that it was crucial in retrieving smooth eigenvectors. Figure 1 is an example of one of those images. Note that we did not apply any kind of preprocessing, thus keeping the resulting patterns untouched and decreasing the possibility of adding subjective features [7].

# 4 Analysis & Results

## 4.1 A quick word on memory

Due to the nature of the data, the embedding as well as the pairwise distance computations, the algorithm requires to store in memory very large matrices. To give just one example, the pairwise distance matrix for the complete 231 years of daily data has a size of 56GB. For this reason we first started to analyse only a small portion of 30 years, form January 1970 to December 2000, before moving to batch computation on the complete data. The analysis and results found will be presented in that order.

[1] https://www.mpimet.mpg.de/en/mpimet-homepage/

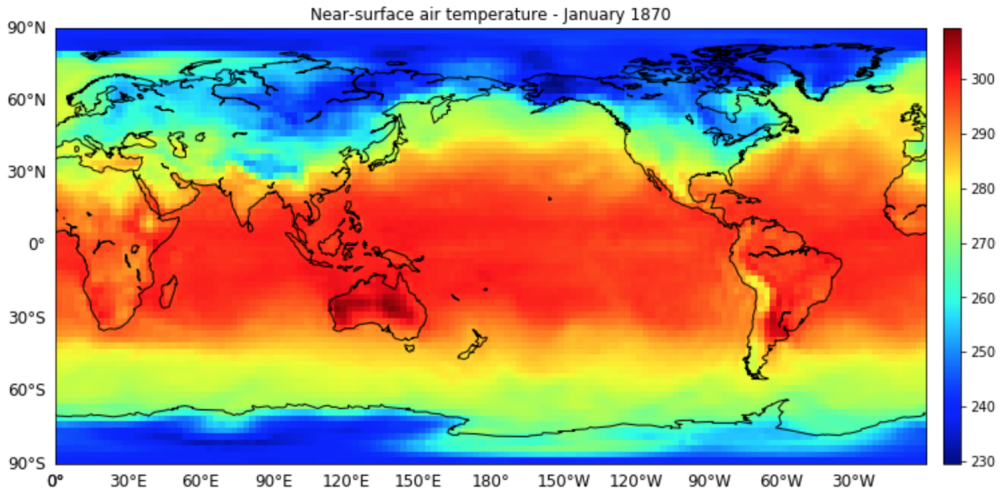


Figure 1: An example of the data used in this project

Memory stops being and issue when the amount of data is drastically reduced, which is the case when using monthly data. We were able to apply the method on the complete 231 years period, but due to the sample's lower temporal resolution, the retrieved eigenvectors are rapidly a bit noisier. The results with monthly images will be presented last.

## 4.2 Analysis from 1970 to 2000

The analysis was done using the daily data over a thirty-one years period, from January 1st, 1970 to December 31st, 2000, due to the large computational space requirement.
Following the algorithm presented above we embedded with a 5 years time-lag (1825 days). Each embedded vector was of size $S = 1825 * 72 * 144 \approx 1.9 * 10^7$. The application of NLSA takes on its full meaning, as the extraction of spatiotemporal patterns from this very high dimensional data space is not trivial.
As the results of the algorithm will depend on different parameters, the following results were obtained using a kernel parameter $\epsilon = 2$ and 2000 $k_{nn}$ nearest neighbors ($\approx 20\%$). We settled on these numbers once the resulting patterns looked good enough but there is no guarantee that they are optimal to perform prediction on this dataset. Further in-depth analysis, with higher computational power, would be necessary in order to get the best parameters possible.

We kept our analysis to the first 100 eigenfunctions obtained with the algorithm, a choice based on the eigenvalues plot at Figure 2. As shown in Figure 3 with their frequencies spectrum, the periodic patterns go in pairs, $v_0$ and $v_1$ demonstrate an annual, $v_2$ and $v_3$ a semiannual, while $v_4$ and $v_5$ show a triannual periodic pattern. These periodic patterns are the most characteristic pattern that can be extracted from temperature data.

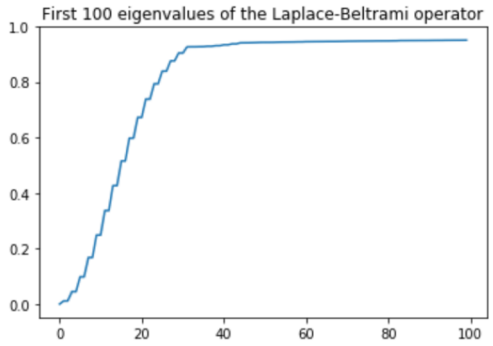


Figure 2: Eigenvalues of the Laplacian

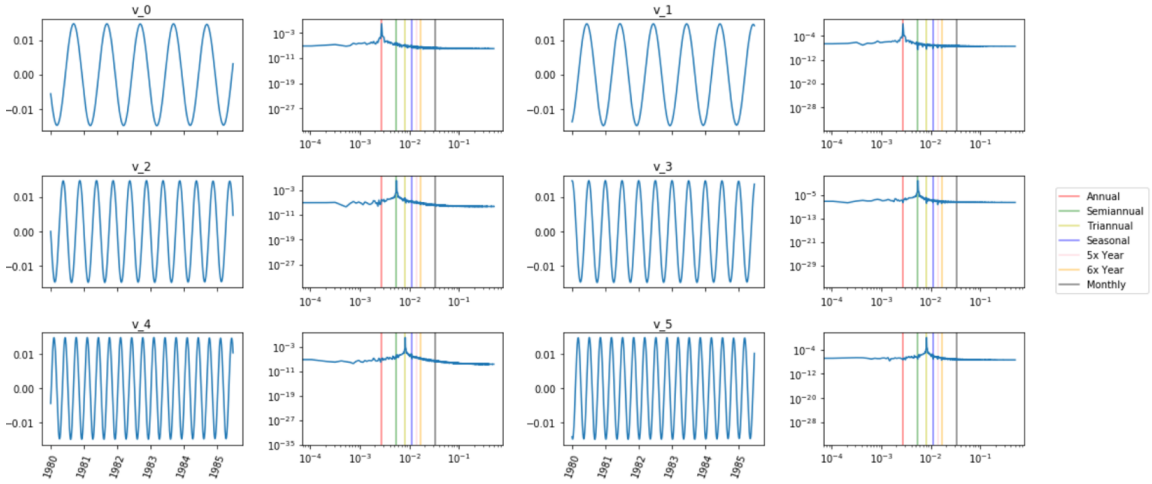


Figure 3: Frequency spectrum of Laplace-Beltrami eigenfunctions

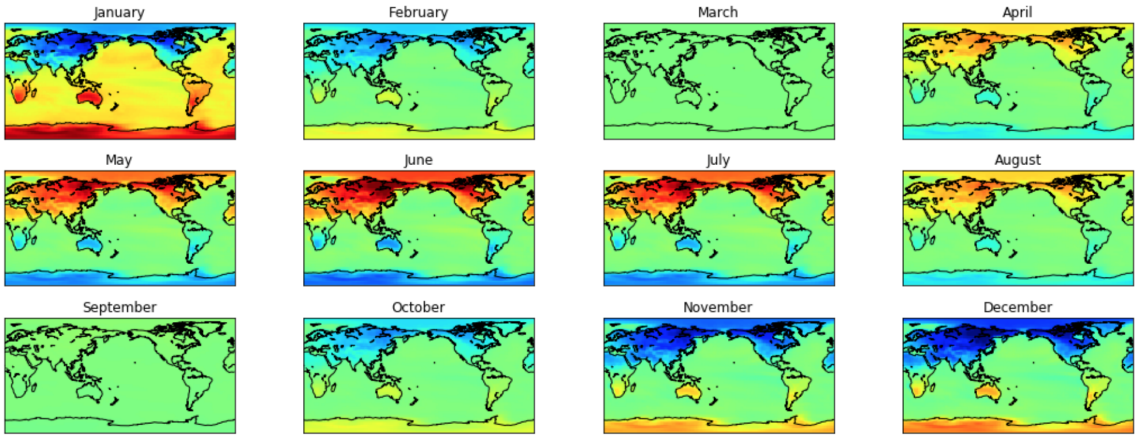


Figure 4: Spatiotemporal reconstruction with eigenfunction $v_1$

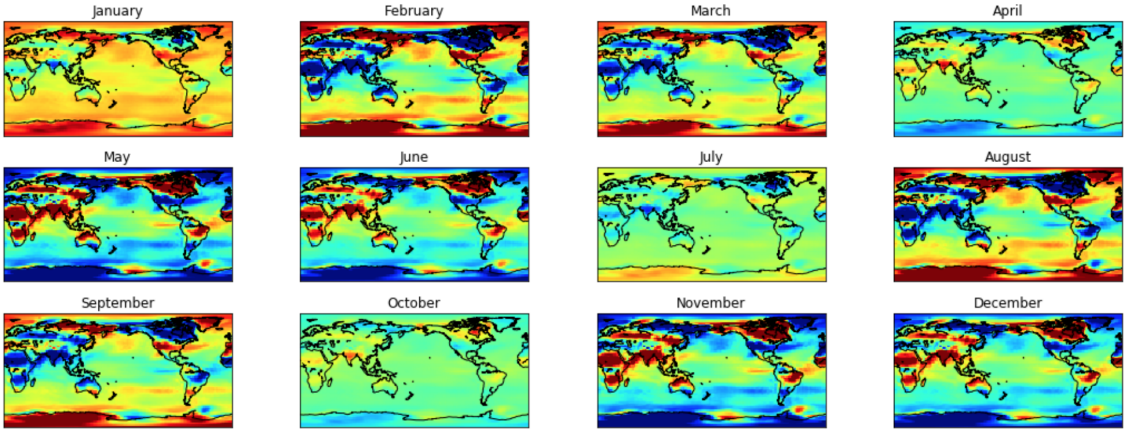


Figure 5: Spatiotemporal reconstruction with eigenfunction $v_2$

#### 4.2.1 Spatiotemporal pattern reconstruction

Once reconstructed these pattern show clear physical phenomenon such as the seasonal cycle or some jet-stream effect in the North America's dipole. The spatiotemporal reconstruction of $v_1$ is shown in Figure 4. We can clearly see the annual pattern as the seasons change, with an emphasis on the northern and southern hemispheres, as well as a greater intensity on the land, while the ocean is less impacted during the transition. This is a very distinctive pattern that was already shown in [7], with data sampled every 3h. The semiannual cycle can be seen in Figure 5, with the same peculiar pattern happening twice in a year. The first 30 eigenfunctions look very similar with an increase of frequency between each pair of functions. At that point it seems that different kinds of patterns start to appear in each function. Another example of the information that comes from the reconstruction of eigenfunctions is the one found with $v_2$, which seems to shows a typical pattern of dipole Pacific - North America, with a warm wave coming from the west coast of Canada and the USA and a polar jet stream bringing cold temperature in the northern parts of Canada (see Figure

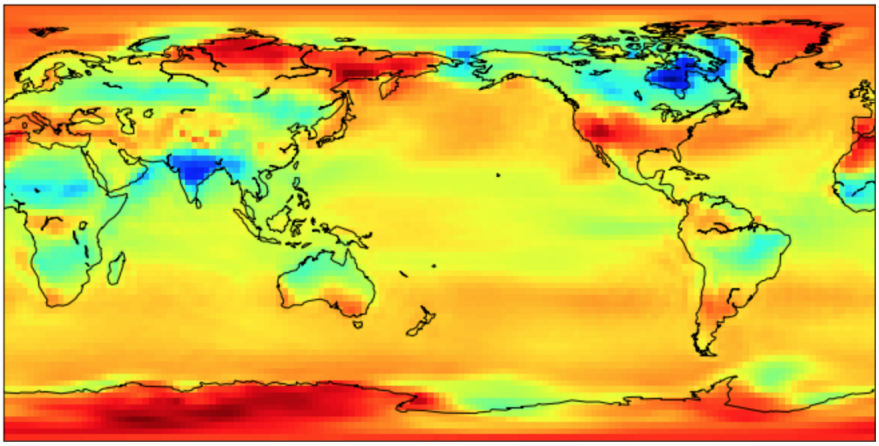

Figure 6: Pacific - North America Dipole

6). Although not all vector are periodic and could carry some other type of valuable information, the 30 years period that we have studied is not long enough to extract this types of phenomenons, such as climate change.

#### 4.2.2 SSA Comparison

In order to make sure that the patterns we extracted from our data would also come out if we were using a linear method, we decided to compare our results with the SSA algorithm (2.2), commonly used in climate science.

We could confirm that the periodic patterns we found were also detected using a linear technique such as SSA, see Figure 7. These basis functions demonstrated the same patterns once projected onto the original data, as the ones extracted with NLSA. The biggest difference that we observed can be seen when comparing Figure 2 with Figure 8, which demonstrate that fewer vectors extracted with SSA contain meaningful information than when we used NLSA.

### 4.3 Analysis using the complete data

To be able to detect changes in climate phenomenons we needed to apply the algorithm on a larger dataset. We decided to use the complete 231 years of data we had available, processing them in batches due to the memory necessary for the computations. But with this very large amount of data we were only able to select as much as 10% of nearest neighbors and even though new patterns appeared, as well as the same periodic ones that we previously extracted (see Figure 9), the lack of information due to the small number of neighbors created glitches in the resulting eigenvectors, demonstrated in Figure 10 and 11.

The lack of smoothness in those vectors makes it impossible to interpret meaningful physical events and would make it very hard to get accurate predictions. To get around that problem we decided to used monthly data, this would both reduce the amount of data and let us work on the 231 years scope with the trade-off of using a lower temporal resolution.

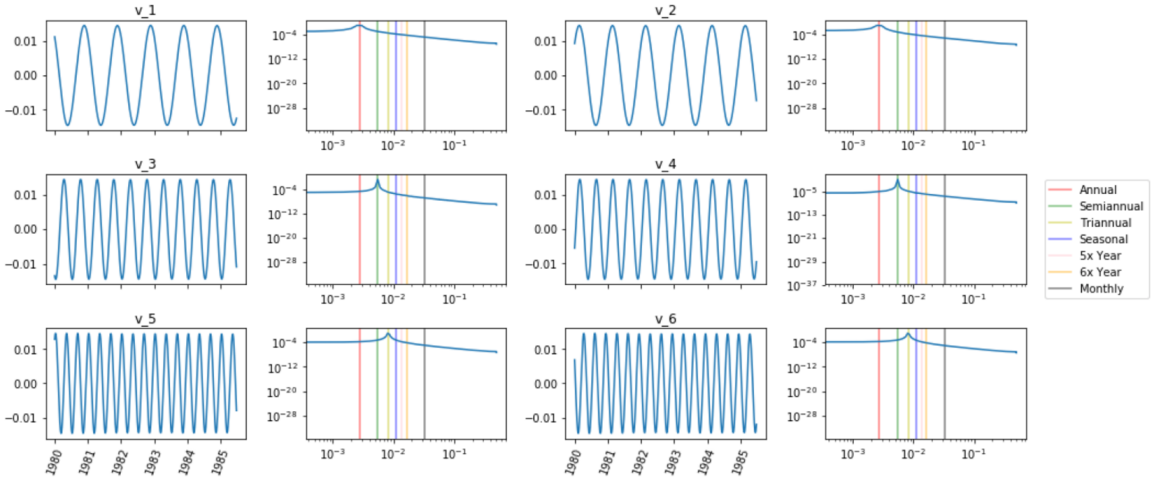


Figure 7: Frequency spectrum of the first left singular vectors

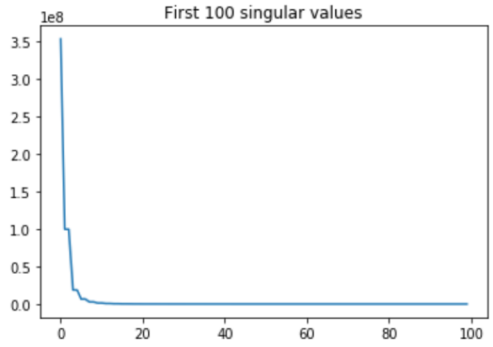


Figure 8: Singular values of SSA

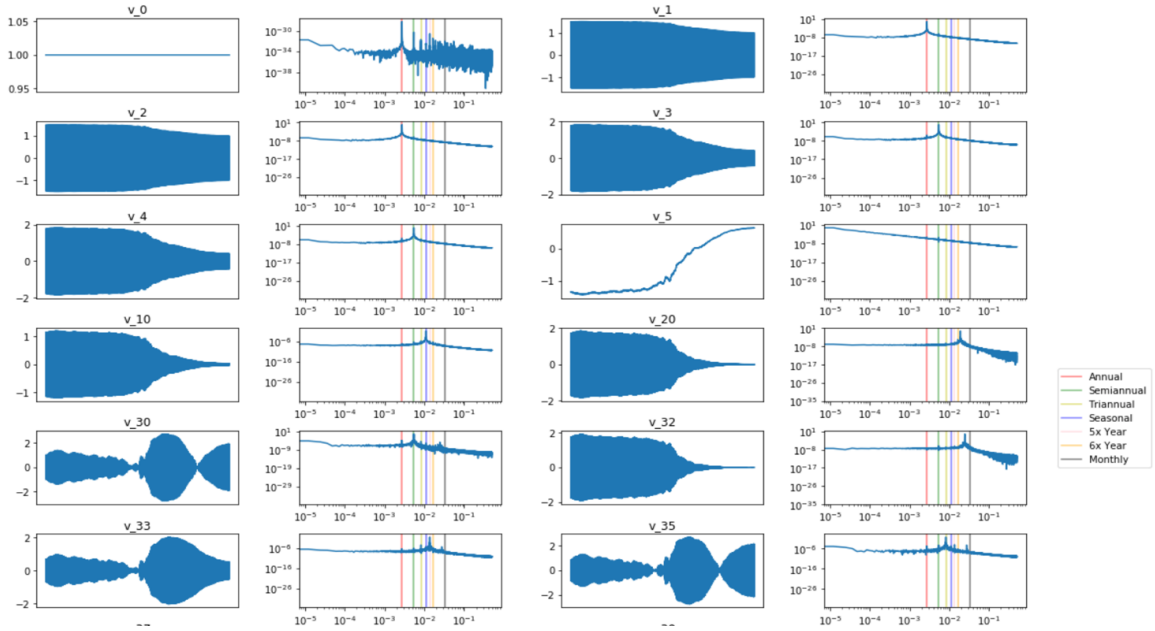


Figure 9: Frequency spectrum of the Laplace Beltrami Operator with 231 years of data

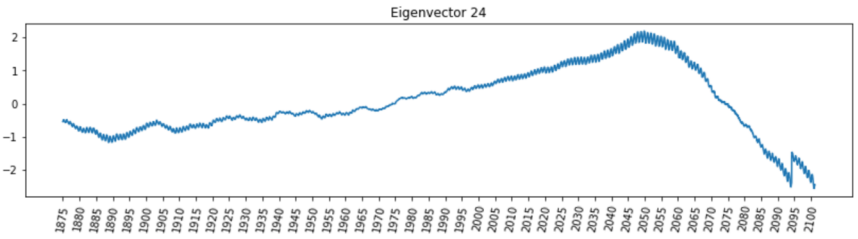


Figure 10: Vector 24 extracted with 231 years of data

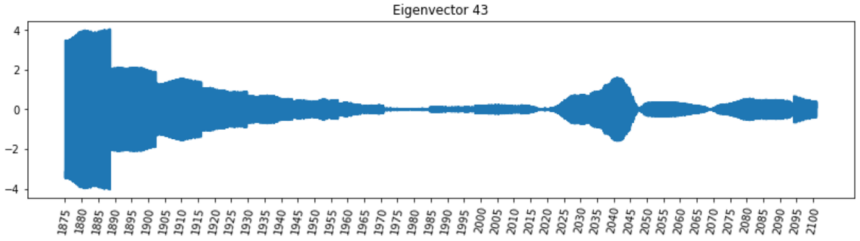


Figure 11: Vector 43 extracted with 231 years of data

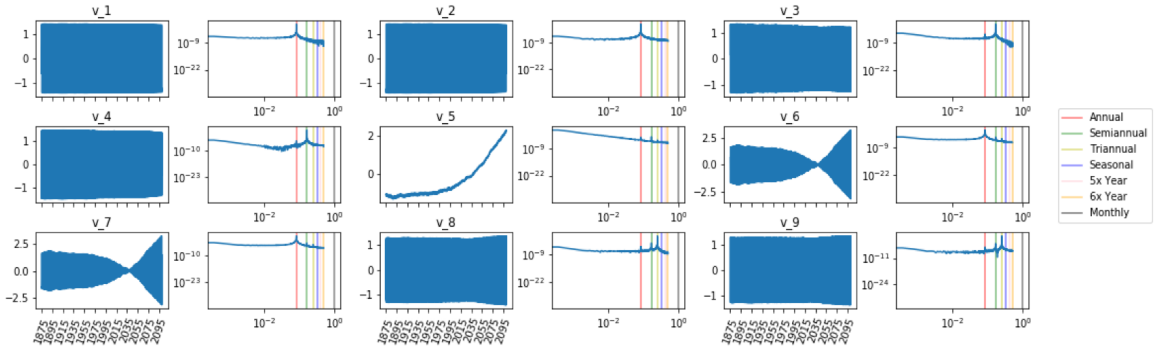


Figure 12: Frequency spectrum of the Laplace Beltrami Operator with monthly data

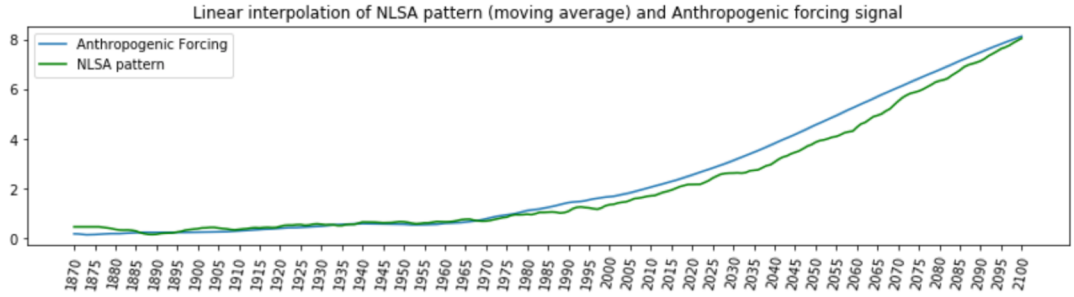


Figure 13: Linear correlation with RCP 8.5

## 4.4 Analysis with monthly images

As monthly image take a lot less space we were able to apply the algorithm without the use of batches and using around 50% of nearest neighbors, which fixed the glitches issue that we had. The extracted vectors were much smoother and we were able to detect signals such as the anthropogenic forcing (see Figure 12). In order to estimate how close this signal was from the real Representative Concentration Pathway (RCP), we averaged the extracted function and computed the Pearson correlation coefficient using the RCP 8.5 and RCP 4.5 simulations. We obtained a very strong correlation in both cases with a coefficient of respectively 0.9948 and 0.9943, see Figure 13 and 14.

The reconstruction of the anthropogenic forcing (see Figure 15) showed a shift in temperature during the end of the 20th century and a fast increase in the coming decades, with a focus on the North Pole, which matches the predictions done by climatologists.
Lastly we applied NLSA on a standardized set of monthly data and recovered again the anthropogenic forcing, but this time the reconstruction focuses on the tropics and mark a clear separation between tropics and poles (see Figure 16), which was the desired outcome.

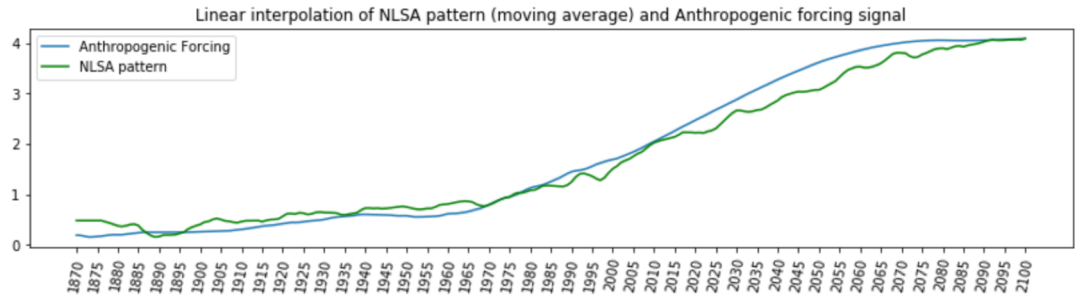


Figure 14: Linear correlation with RCP 4.5

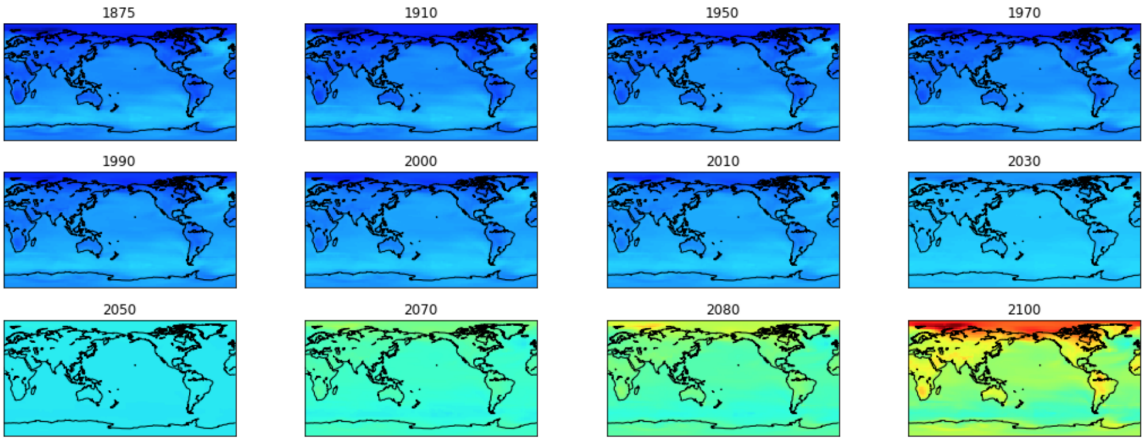


Figure 15: Reconstruction of the Anthropogenic Forcing extracted with NLSA

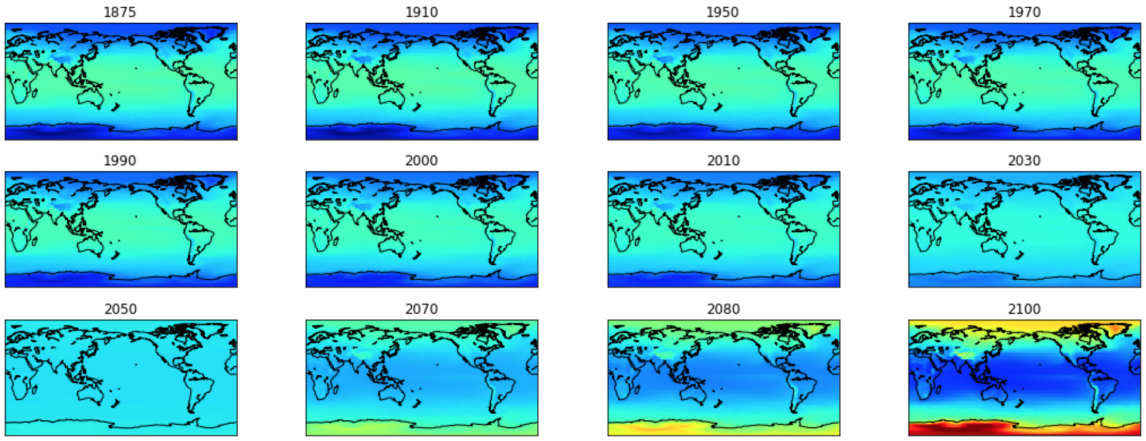


Figure 16: Reconstruction of the Anthropogenic Forcing extracted with NLSA with standardized data

# 5 Conclusion

In this paper we introduced the NLSA algorithm [2, 3] and showed its application to climate science. We were able to retrieve meaningful patterns both in the eigenfunctions and in their reconstructions, reproducing results found in a previous study [7] (that looked only at the tropics), on worldwide data, something that to our knowledge, was not done before. Even though the application on the complete daily data failed due to the memory capacity, we successfully applied the algorithm to batches of data. We showed the robustness of the algorithm under different data resolutions and different set of parameters. We could compare the pattern we extracted with a method used in climate science today and observed that we retrieve the same basis functions. Finally we were able to match the anthropogenic forcing signal with a very high correlation coefficient in two different simulations, RCP 8.5 and 4.5.
Even though we did not get to do predictions, we laid the groundwork for a subsequent study. Further work on the subject should include an in-depth search and validation of the parameters used to construct the kernel on the complete daily data and an application of dimension reduction and prediction.

# 6 Addendum

This project is open source and can be found at:
https://github.com/montalex/NLSA_Application_To_Climate_Science.

## 6.1 Libraries used

Python: https://www.python.org
Numpy: http://www.numpy.org
Pandas: https://pandas.pydata.org
Scipy: https://www.scipy.org
netCDF4: http://unidata.github.io/netcdf4-python/
Basemap: https://matplotlib.org/basemap/

# References


[1] R. R. Coifman and S. Lafon, *Diffusion maps*, Appl Comput Harmon Anal 21(1), (2006), p. 5–30.

[2] D. Giannakis and A. J. Majda, *Nonlinear laplacian spectral analysis for time series with intermittency and low-frequency variability*, Proc Natl Acad Sci 109(7), (2012), p. 2222–2227.

[3] ——, *Nonlinear laplacian spectral analysis: Capturing intermittent and low-frequency spatiotemporal patterns in high-dimensional data*, Stat Anal Data Min 6(3), (2013), p. 180–194.

[4] M. Haddad, *Using multivariate singular spectrum analysis for estimating trends in global climate change indicators*, BioSciencesWorld 2017, (2017).

[5] H. Hassani, *Singular spectrum analysis: Methodology and comparison*, Journal of Data Science 5, (2007), pp. 239–257.

[6] T. Sauer, J. A. Yorke, and M. Casdagli, *Embedology*, Journal of Statistical Physics 65(3-4), (1991), p. 579–616.

[7] E. Székely, D. Giannakis, and A. J. Majda, *Extraction and predictability of coherent intraseasonal signals in infrared brightness temperature data*, Climate Dynamics 46(5-6), (2016), p. 1473–1502.