Nonlinear Laplacian spectral analysis: application to climate science

Alexis Montavon
Under the supervision of Eniko Szekely & Olivier Verscheure

January 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. This algorithm has the advantage that it does not require any preprocessing of the data.

1 Introduction

Most of the modern techniques applied in climate science for dimension reduction and prediction, such as temperature forecast and 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 (SSA) with the differences that it is nonlinear 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. 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 [5], aims to extract spatiotemporal patterns form 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 δt, 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:

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

$X(t)$ 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 [4].

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:

$$K(X(t_i), X(t_j)) = exp\left(-\frac{\|X(t_i) - X(t_j)\|^2}{\epsilon\|\zeta(t_i)\|\|\zeta(t_j)\|}\right) \tag{1}$$

With ϵ 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 [5]. 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].

$$\widehat{K}_{ij} = \frac{K_{ij}}{Q_i Q_j}, Q_i = \sum_{j=1}^{S} K_{ij} \tag{2}$$

$$P_{ij} = \frac{\widehat{K}_{ij}}{D_i}, D_i = \sum_{j=1}^{S} \widehat{K}_{ij} \tag{3}$$

$$L_{ij} = I - P_{ij} \tag{4}$$

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 [5].

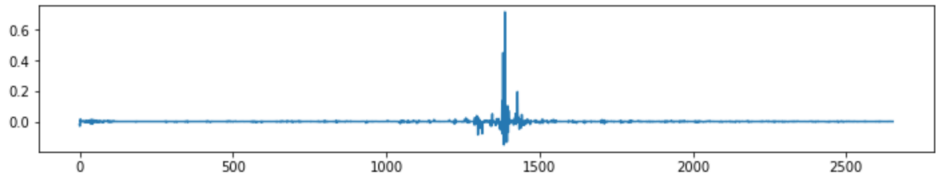

Figure 1: Least significant eigenfunction retrieved with Gaussian kernel on daily data

2.1.3 Empirical set of basis functions

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

$$Lv_i = \lambda_i v_i \tag{5}$$

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 λ_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 λ_i as they will represent the features with slow variation and will reduce noise and parameter sensitivity [5].

We will then use the chosen eigenvectors v_i to retrieve spatiotemporal patterns by reconstructing the data in the delay-coordinate space:

$$\widehat{X}_i = X v_i v_i^t \tag{6}$$

2.2 Gaussian kernel method

To start the analysis and for the sake of comparison we also used a simpler technique using a Gaussian kernel. Step 1 and 3 are the same but in step 2, we lose all the normalizations and apply a simpler Gaussian kernel before building the Laplacian graph. K becomes:

$$K(X(t_i), X(t_j)) = exp\left(-\frac{\|X(t_i) - X(t_j)\|^2}{2 * \sigma^2}\right) \tag{7}$$

With $\sigma^2 = \frac{1}{N}\sum_{i=1,j=1}^{N} \|X(t_i) - X(t_j)\|^2$. But the time series retrieved with this method were either very noisy or very sparse as shown in Figure 1 and we decided to focus only on the Laplace-Beltrami operator.

3 Dataset Description

The dataset we used for this project contained 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 x 144, from 1870 to 2100 (with 95 years of simulated data in a worst case scenario). For the daily data we used an average over 4 different models. Figure 2

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

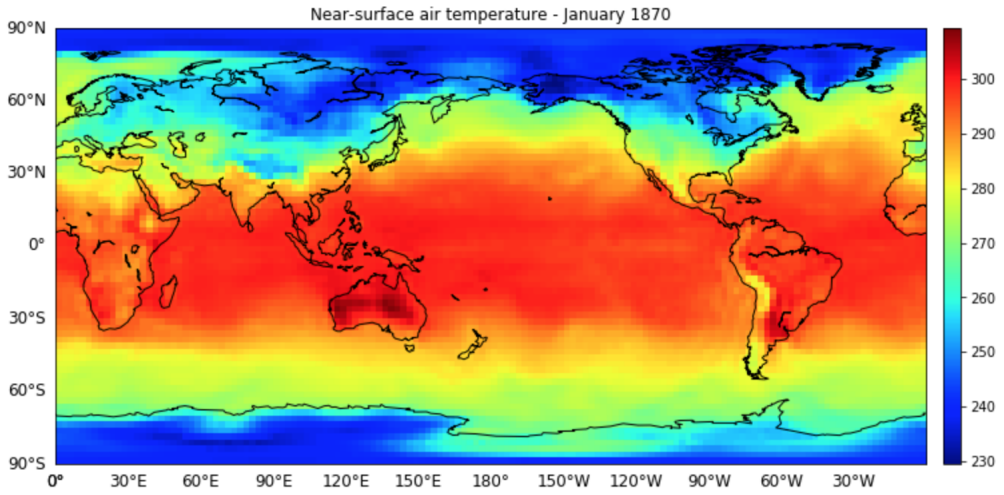

Figure 2: An example of the data used in this project

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 [5]. 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. The same method was applied to the monthly data, on the complete 231 years period, but the results were too noisy due to the sample's low temporal resolution. The retrieved eigenvectors only showed a mix of a lot of spatial patterns.

4 Analysis & Results

Following the algorithm presented above and the trade-off between space requirement and the right amount of past data, 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 50 eigenfunctions obtained with the algorithm, a choice based on the time at our disposal, there could be other interesting patterns left past this point. Figure 3 shows a plot of the eigenvalues. As shown in Figure 4, the most characteristic 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. The spatiotemporal reconstruction of v_1 is shown in Figure 5. 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 [5], with data sampled every 3h. The semiannual cycle can be seen in Figure 6, with the same peculiar pattern happening twice in a year. The first 30 eigenfunctions look

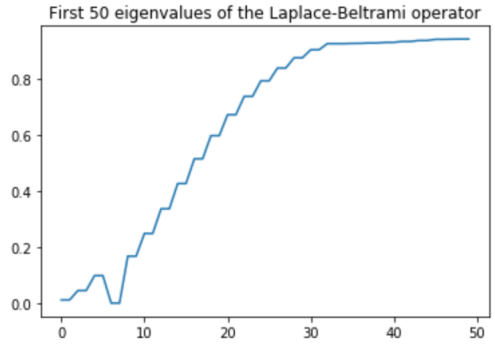

Figure 3: Eigenvalues of the Laplacian

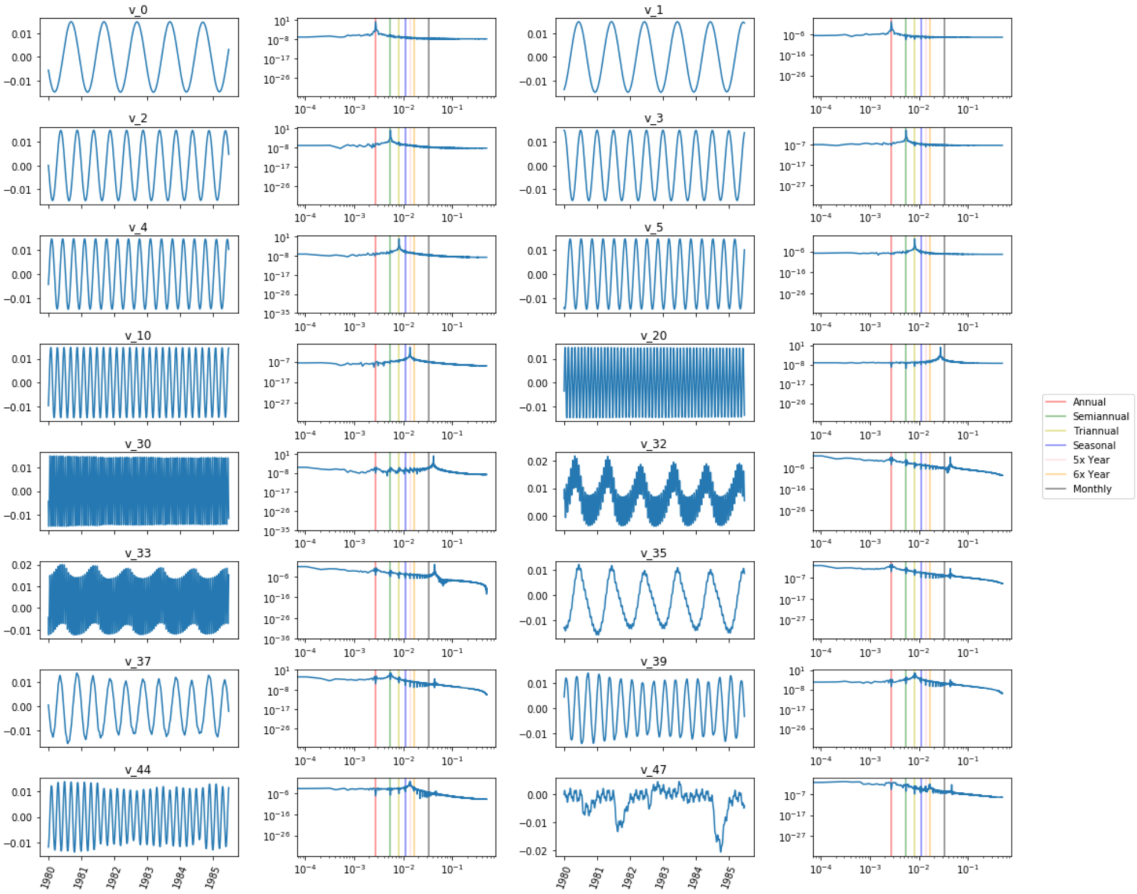

Figure 4: Frequency spectrum of Laplace-Beltrami eigenfunctions

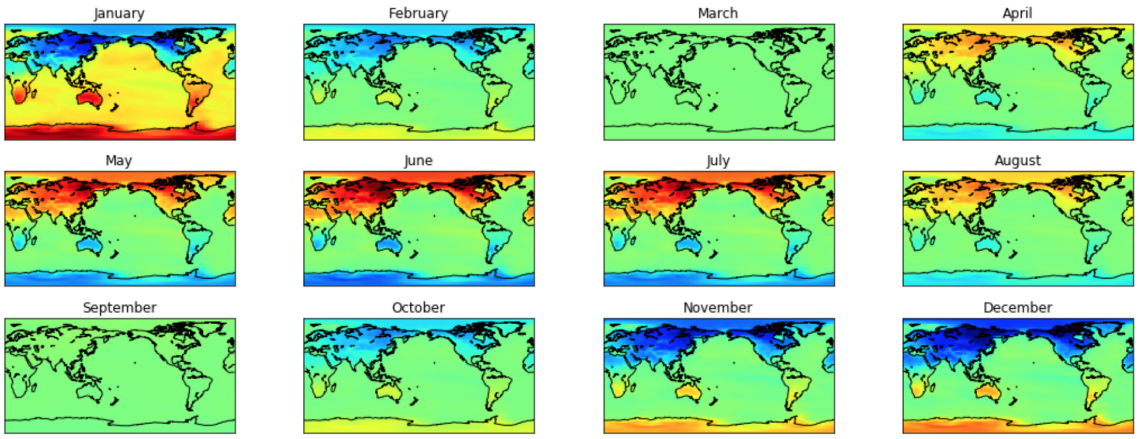

Figure 5: Spatiotemporal reconstruction with eigenfunction v_1

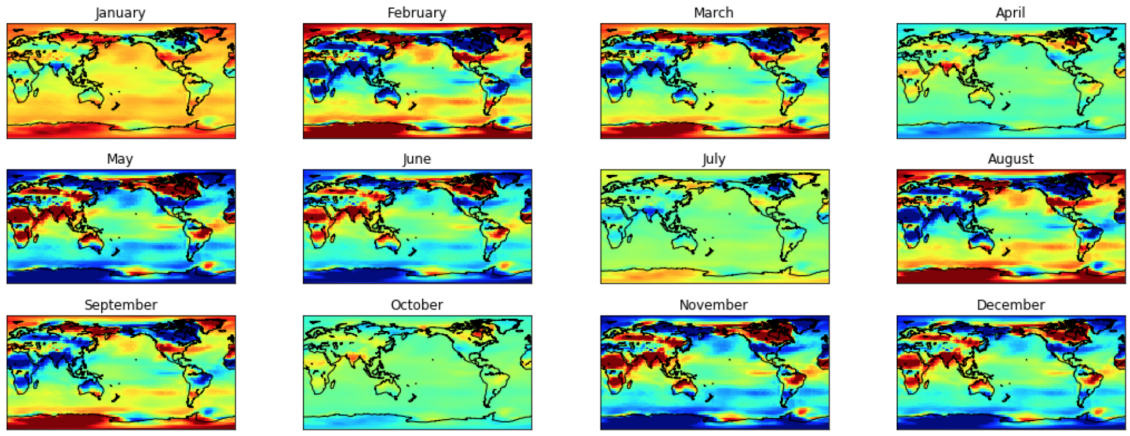

Figure 6: Spatiotemporal reconstruction with eigenfunction v_2

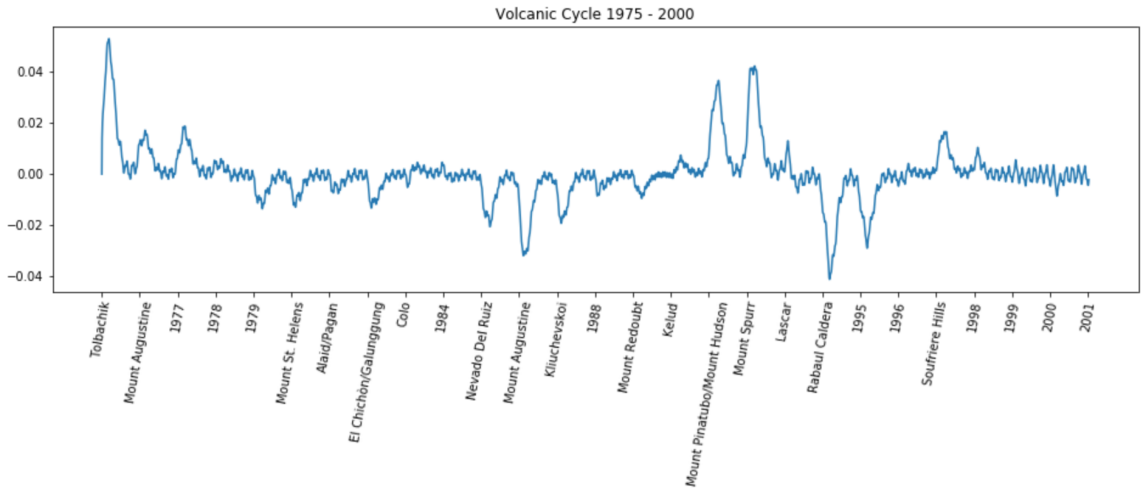

Figure 7: Possible volcanic cycle pattern

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 one function, as in the frequency plot of v_{32} to v_{47}. Nonetheless, the lack of a periodic pattern does not necessarily means that the eigenfuction does not contain any important information. Such an example is v_{47}, plotted completely on Figure 7. This vector seems to be representing a possible volcanic cycle, as shown with the list of large volcanic eruptions of the 20th century [2], starting in 1975 (because of the 5 year time-lag, when reconstructing $S = N - q + 1$ hence the starting date is $1970 - 5 + 1day =$ January 2nd 1975). The other theory we have is that v_{47} could be a representation of ENSO (El Niño and Southern Oscillation), we still need to discuss our results with a climatologist to get a confirmation.
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 8). The eigenvector v_{37}, shown in its full size in Figure 9, could well represents the solar cycle when looking at years 1987 to 1998.

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 [5] (that looked only at the tropics), but on worldwide data, something that to our knowledge, was not done before. 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 and an application of dimension reduction and prediction. A research on the influence of the annual cycle might be needed as well as a comparison with SSA.

[2] https://en.wikipedia.org/wiki/List_of_large_volcanic_eruptions_of_the_20th_century

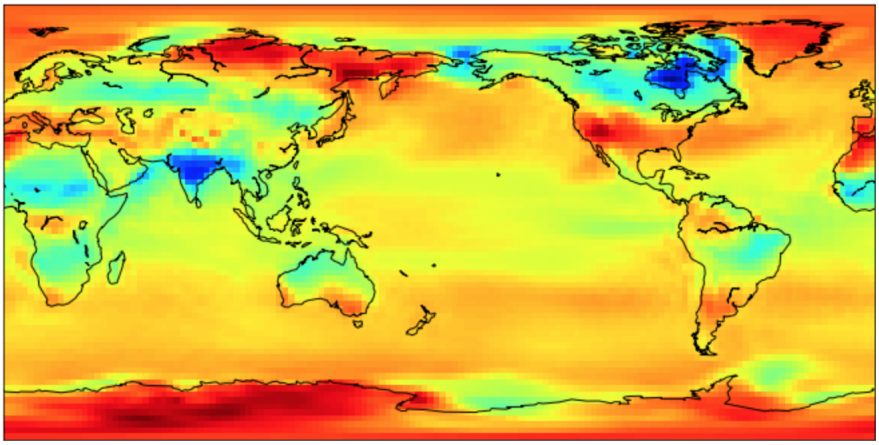

Figure 8: Pacific - North America Dipole

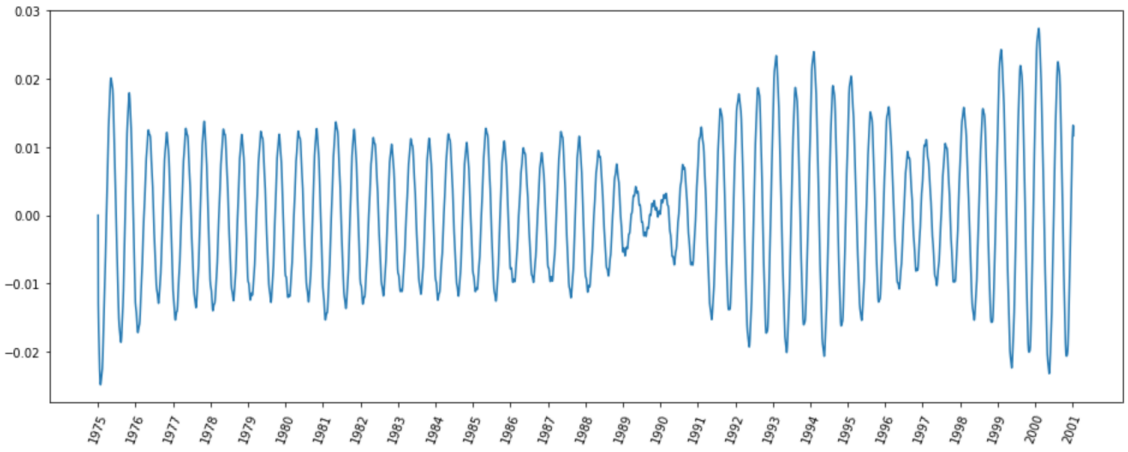

Figure 9: Possible solar cycle representation from v_{37}

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

[5] 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.