



**HAL**  
open science

## An experimental evaluation of choices of SSA forecasting parameters

Teodor Knapik, Adolphe Ratiarison, Hasina Razafindralambo

► **To cite this version:**

Teodor Knapik, Adolphe Ratiarison, Hasina Razafindralambo. An experimental evaluation of choices of SSA forecasting parameters. 2022. hal-03679576v1

**HAL Id: hal-03679576**

**<https://hal.science/hal-03679576v1>**

Preprint submitted on 26 May 2022 (v1), last revised 22 Mar 2024 (v4)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# An experimental evaluation of choices of SSA forecasting parameters

Teodor Knapik<sup>\*1</sup>, Adolphe Ratiarison<sup>2</sup>, Hasina Razafindralambo<sup>2</sup>

<sup>1</sup>ISEA, Université de la Nouvelle Calédonie

<sup>2</sup>DyACO, Université d'Antananarivo

\*E-mail : [knapik@unc.nc](mailto:knapik@unc.nc)

---

## Abstract

Six time series related to atmospheric phenomena are used as inputs for experiments of forecasting with singular spectrum analysis (SSA). Existing methods for SSA parameters selection are compared throughout their forecasting accuracy. The results confirm that the choices of the window length and of the grouping are essential and that selecting longer windows leads often to poorer predictions.

## Keywords

time series forecasting, singular spectrum analysis, parameter selection

---

## I INTRODUCTION

Many naturally occurring dynamical systems are governed by a huge number of unknown parameters and laws. Most of the time the understanding of such systems seems beyond human reach. Yet, the ability to predict their future trajectory with an acceptable error using past observations is often of paramount importance.

This paper reports an experimental investigation of univariate time series prediction using a singular spectrum analysis (SSA) related method, namely the vector forecasting (see e.g.[15]). SSA involves several computational steps among which the vector forecasting may be seen as a final one. A time series resulting from measurements of an observable of a dynamical system at regular intervals is the input together with a forecast horizon. In its most basic version, SSA has to be provided with an additional information that may require an expert knowledge about the observed phenomenon. Is it possible to reduce such an additional input so that SSA forecasting could be applied to phenomena of unknown dynamics by a user with no knowledge in the underlying field nor in statistics?

A partially positive answer to the latter question comes from the availability of several software packages for SSA, notably SSA-MTM toolkit [5] and the **R** library *Rssa* [14, 17]. Although for an end user, SSA-MTM toolkit may be more suitable due to its graphical interface, the choice of *Rssa* for experiments reported here has been motivated by an intrinsic flexibility of a library. Nevertheless, besides an input time series, both SSA-MTM toolkit and *Rssa* must be provided with a *window length* and a list of *components*

to be grouped together for forecasting.<sup>1</sup> Those two additional inputs greatly impact the forecast quality and several authors discuss methods for inferring them automatically from the input times series. The present paper reports an experimental study of forecast quality obtained using such methods.

In order to recall the importance of the length of the window and the choice of components to group, the next section reviews the essential SSA steps that lead to the vector forecasting further explained in Sect. III. Sect. IV and V review a few methods for selecting SSA parameters. The data sets used in the reported experiments are introduced in Sect. VI and the experiments are described in Sect. VII. Their outcome is discussed in Sect. VIII. Throughout this paper,  $[n]$  stands for  $\{1, \dots, n\}$ .

## II SSA STEPS TOWARDS THE VECTOR FORECASTING

SSA has a history of parallel development on both sides of the iron curtain. As it has been sketched in many papers and several books (e.g. in [15]), it is omitted in the present article.

The first step of SSA is an *embedding* of a real-valued input time series  $\mathbb{X} = (x_1, \dots, x_N)$ ,  $N > 2$ , into a vector space spanned by a sequence of  $K$  *lagged vectors*  $X_1, \dots, X_K \in \mathbb{R}^L$ , with  $X_i := (x_i, x_{i+1}, \dots, x_{i+L-1})^\top$  where  $L \in \mathbb{N}$  is the *window length* and  $K := N - L + 1$ . Those lagged vectors form a *trajectory matrix*  $\mathbf{X} \in \mathbb{R}^{L \times K}$ ,  $\mathbf{X} = [X_1, \dots, X_K]$ ,

$$\mathbf{X} := \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_K \\ x_2 & x_3 & x_4 & \cdots & x_{K+1} \\ x_3 & x_4 & x_5 & \cdots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \cdots & x_N \end{pmatrix}$$

which is a Hankel matrix, viz., the elements of each anti-diagonal are equal. Note that this embedding is a bijection between the set of sequences of length  $N$  and the set of  $L \times K$  Hankel matrices. Consequently, by the *inverse embedding*, from any  $m \times n$  Hankel matrix, one gets the corresponding sequence of length  $m + n - 1$ .

The parameter  $L$  is essential for the whole SSA and is usually chosen so that  $L < N/2$ . This is assumed throughout the paper so as to keep  $L < K$ . Although the rank of  $\mathbf{X}$  is at most  $L$ , in real life time series it is equal to  $L$ , what is also assumed here in order to simplify the presentation.

The above embedding  $\mathbb{X} \mapsto \mathbf{X}$  may be interpreted as a representation of  $\mathbb{X}$  by a trajectory of a hypothetical dynamical system that generated  $\mathbb{X}$ . In the second step of SSA, the singular value decomposition (SVD) of the trajectory matrix is computed:  $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$  where  $\mathbf{U} = [U_1, \dots, U_L] \in \mathbb{R}^{L \times L}$  and  $\mathbf{V} = [V_1, \dots, V_K] \in \mathbb{R}^{K \times K}$  are unitary matrices and  $\mathbf{\Sigma} \in \mathbb{R}^{L \times K}$  is a rectangular diagonal matrix with diagonal  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_L \geq 0$ , viz.,  $\Sigma_{kk} = \sigma_k$ . Every *eigen-triple*  $(U_k, \sigma_k, V_k)$ , for  $i \in [L]$ , where  $U_k$  (resp.  $V_k$ ) is called *left* (resp. *right*) *singular vector* for *singular value*  $\sigma_k$ , yields an *elementary matrix*  $\mathbf{X}_k := \sigma_k U_k V_k^\top$  of rank 1 so that  $\mathbf{X} = \sum_{i=1}^L \mathbf{X}_k$ . Note that all non null elementary matrices in this decomposition are pairwise orthogonal.

---

<sup>1</sup>However, in Rssa, automated grouping is also available.

By averaging over anti-diagonals, from an elementary matrix  $\mathbf{X}_k$  one gets its *Hankelization*  $\tilde{\mathbf{X}}_k \in \mathbb{R}^{L \times K}$ . More precisely the  $k$ -th anti-diagonal of an  $L \times K$  matrix has its indexes in

$$A_k := \{(i, j) \in [L] \times [K] : i + j = k + 1\} \quad \text{for } k \in [N] \quad (1)$$

and the Hankelization  $\tilde{\mathbf{X}}_k$  of  $\mathbf{X}_k$  is defined by

$$\tilde{x}_{k,i,j} := \frac{1}{|A_{i+j-1}|} \sum_{p+q=i+j} x_{k,p,q}$$

where  $\tilde{x}_{k,i,j}$  (resp.  $x_{k,p,q}$ ) stands for the element at row  $i$  (resp.  $p$ ) and column  $j$  (resp.  $q$ ) of  $\tilde{\mathbf{X}}_k$  (resp.  $\mathbf{X}_k$ ). As  $\tilde{\mathbf{X}}_k \in \mathbb{R}^{L \times K}$  is a Hankel matrix, by an inverse embedding, one gets an *elementary component* time series  $\mathbb{X}_k$  of  $\mathbb{X} = \sum_{k=1}^L \mathbb{X}_k$ . Some of those components are considered as noise whereas others as carrying valuable information about the underlying dynamical system.

This leads to the step of grouping which aims at removing the noise by choosing a strict subset  $I$  of  $[L]$  to get the “signal”  $\mathbb{X}_I := \sum_{k \in I} \mathbb{X}_k$  separated from the “noise”  $\mathbb{X}_{\bar{I}} := \sum_{k \in [L] \setminus I} \mathbb{X}_k$ . Similarly, one may write  $\mathbf{X}$  as the sum of its relevant parts  $\mathbf{X}_I := \sum_{k \in I} \mathbf{X}_k$  and its noisy part  $\mathbf{X}_{\bar{I}} := \sum_{k \in [L] \setminus I} \mathbf{X}_k$  or their Hankelizations  $\mathbf{X} = \tilde{\mathbf{X}}_I + \tilde{\mathbf{X}}_{\bar{I}}$ . The reader should note that  $I$  is another additional input for SSA and that the choice of  $I$  greatly impacts the quality of forecasting.

### III VECTOR FORECASTING

The vector forecasting is not a part of SSA *per se*. Like two other forecasting methods described in [15], the vector forecasting uses SSA for extracting a low rank approximation of a subspace of a hypothetical dynamical system that generated  $\mathbb{X}$ . The common idea of the three forecasting methods presented in [15] is to find a linear recurrent equation (LRE)

$$x_{I,i+L-1} = \sum_{k=1}^{L-1} a_k x_{I,i+L-1-k} \quad \text{for } 1 \leq i \leq K \quad (2)$$

that is satisfied by “denoised” time series  $\mathbb{X}_I = (x_{I,1}, \dots, x_{I,N})$ . Finding coefficients

$$\mathbf{a} := (a_{L-1}, \dots, a_1)$$

of LRE (2) amounts to solving the following system of equations  $\mathbf{a}\underline{\mathbf{X}}_I = (x_{I,L}, \dots, x_{I,N})$  where  $\underline{\mathbf{X}}_I$  denotes matrix  $\mathbf{X}_I$  without its last row. As  $L < N/2$ , this system is over-determined, and in general, has no exact solution. The closest approximate solution with respect to the Euclidean norm is therefore used. It is well known that such an approximate solution is given by  $\mathbf{a} \approx (x_{I,L}, \dots, x_{I,N})\underline{\mathbf{X}}_I^\dagger$ , where “ $\dagger$ ” stands for the pseudo-inverse of Moore-Penrose of a rectangular matrix.

Let  $\mathbf{U}_I := [U_i : i \in I]$  be the matrix formed with columns of  $\mathbf{U}$  with indexes in  $I$  and let  $\mathcal{U}_I$  be the subspace of  $\mathbb{R}^L$  spanned by columns of  $\mathbf{U}_I$ . Let  $(u_{L,i} : i \in I)$  be the last row of  $\mathbf{U}_I$  and let  $\underline{\mathbf{U}}_I = [\underline{U}_i : i \in I]$  stand for  $\mathbf{U}_I$  with its last row removed. Similarly, let  $\underline{\mathcal{U}}_I$  be the subspace of  $\mathbb{R}^{L-1}$  spanned by columns of  $\underline{\mathbf{U}}_I$ . Note that  $(u_{L,i} : i \in I)$  can be expressed as a linear combination of rows of  $\underline{\mathbf{U}}_I$  because  $\underline{\mathbf{U}}_I$  is of rank  $|I| < L$ . Let  $v^2 := \sum_{i \in I} u_{L,i}^2$ .

Note also that  $v = \cos \theta$  where  $\theta$  is the angle between  $\mathcal{U}_I$  and  $\mathbf{e}_L = (0, \dots, 0, 1)^\top \in \mathbb{R}^L$ . As  $\mathbf{e}_L \notin \mathcal{U}_I$ , one has  $v \neq 1$  and the following vector is well defined

$$R := \frac{1}{1-v^2} \sum_{i \in I} u_{L,i} \underline{U}_i .$$

It may be shown that  $(x_{I,L}, \dots, x_{I,N}) \underline{\mathbf{X}}_I^\dagger = R^\top$ . Consequently  $R$  is the closest approximate solution of (2). Now, matrix

$$\Pi := \underline{\mathbf{U}}_I \underline{\mathbf{U}}_I^\top + (1-v^2) R R^\top$$

defines the orthogonal projection of  $\mathbb{R}^{L-1}$  onto  $\underline{\mathcal{U}}_I$ . Let  $\check{\mathbf{z}}$  be a vector  $\mathbf{z} \in \mathbb{R}^L$  without its first coordinate. Using a linear operator  $\mathbf{F}: \mathbb{R}^L \rightarrow \mathcal{U}_I$  which extends the orthogonal projection  $\Pi \check{\mathbf{z}}$  of  $\check{\mathbf{z}}$  with the next term inferred from (2)

$$\mathbf{F} \mathbf{z} := \begin{pmatrix} \Pi \check{\mathbf{z}} \\ R^\top \check{\mathbf{z}} \end{pmatrix}$$

one defines a sequence of vectors

$$Y_i := \begin{cases} X_{I,i} & \text{for } i \in [K], \\ \mathbf{F} Y_{i-1} & \text{for } i \in \{K+1, \dots, N+h\}, \end{cases}$$

where  $[X_{I,1}, \dots, X_{I,K}] = \mathbf{X}_I$  and  $h \in \mathbb{N}$  is a forecast horizon. This leads to matrix

$$\mathbf{Y} := [Y_1, \dots, Y_{N+h}] .$$

By Hankelization of  $\mathbf{Y}$  and its subsequent inverse embedding, one gets a time series  $\mathbb{Y} = (y_1, \dots, y_{N+h+L-1})$  where the portion  $(y_{N+1}, \dots, y_{N+h})$  is the forecast up to horizon  $h$  obtained by the vector forecasting method.

#### IV CHOICE OF THE WINDOW LENGTH

From the presentation of SSA method including the vector forecasting, it follows that the length of the window,  $L$ , is a crucial parameter. Foundational texts (e.g. [1, 2]) and books [3, 4, 15] give no general estimation methods of this parameter. The prevailing opinion is that choosing  $L$  only slightly less than  $N/2$  allows capturing all significant frequencies. Choosing  $L$  equal to the longest oscillation period or a multiple of that period not exceeding  $N/2$  is also often suggested. Unfortunately, if the data comes from a poorly understood dynamical system, such a period is unknown. Therefore, providing methods and, more importantly, efficient algorithms for estimating  $L$  from the input time series is essential.

An appealing formal approach for estimating an adequate window length is developed in [8] throughout an adaptation to SSA of the minimum description length principle (see e.g. [6, 16]). The method consist in a cross-optimisation of two functions, say  $f(L, M)$  and  $g(L, M)$  wrt.  $L$  and  $M$ . This yields an estimation of  $L$  and also of the number  $M$  of the most significant components of  $\mathbb{X}$  to be considered as signal. Unfortunately, for each evaluation step of  $f(L, M)$  or  $g(L, M)$ , singular values of  $\mathbf{X}$  have to be computed because  $\mathbf{X}$  depends on  $L$ . As a practical rule, the authors of [8] suggest to take  $(\log N)^c$  with  $c \in (1.5, 2.5)$  as an upper bound for  $L$ . Although  $(\log N)^c \in o(N)$  and therefore  $(\log N)^c \ll N/2$ , for  $N$  sufficiently large, with  $c = 2.5$  the method is still computationally

demanding when  $\mathbb{X}$  is a time series with daily samples over, say, 50 years, because the maximum value of  $L$  then equals 301. Indeed, in case of an exhaustive search over  $L \in \{2, \dots, 301\}$ , one has to perform 300 singular value decompositions (SVD). Beyond formal demonstrations, in [11] and [10] the authors of [8] provide an experimental evaluation of their method on real world data sets which confirms that choosing  $L$  much smaller than  $N/2$  significantly improves the quality of forecasting.

Several authors use the *autocorrelation function*

$$R(\tau) := \frac{1}{\sigma^2} \sum_{i=1}^{N-\tau} (x_{i+\tau} - \mu)(x_i - \mu)$$

where  $\mu$  (resp.  $\sigma$ ) is the mean (resp. standard deviation) of  $\mathbb{X}$ . In [7] the smallest value of  $\tau$  where  $R$  crosses the confidence interval corresponding to (95% of) the white Gaussian noise (with parameters  $\mu$  and  $\sigma$ ) is used as estimate of  $L$ . In [9] and [12] the smallest value of  $\tau$  such that  $R(\tau)R(\tau + 1) < 0$  is used as estimate of  $L$ .

In the sequel,  $L^{[9]}$  stands for the length of the window chosen with the latter method whereas  $L_{\text{lo}}$  and  $L_{\text{hi}}$  denote two extreme values for the maximum window length in  $\{(\log N)^c : c \in (1.5, 2.5)\}$  discussed formerly.

## V CHOICE OF THE GROUPING

The choice of index set  $I$  of components that are used as signal in forecasting is as essential as the choice of the window length. Besides theoretical results, the experiments carried in [10] show that both the grouping and the window length selection have a tremendous impact on forecast accuracy. This is not a surprise as both affect directly LRE (2).

On the contrary to the window length where the search space is linear in  $N$  (yet brute force methods are limited by the computationally costly step of SVD), the search space for grouping is in  $O(2^L)$ . Several authors only consider groupings such that  $I = [M]$  with  $M \in [L - 1]$  which lets reducing the search space into  $O(L)$ . In other words, the signal is selected as the first  $M$  elementary components of  $\mathbb{X}$ , viz.,  $I := [M]$  and  $\mathbb{X}_I = \sum_{i=1}^M \mathbb{X}_i$ . This shall be called a *prefix grouping* in the sequel.

A common practice for the grouping (see e.g. [15]) is to rely on visual examination of scatter plots and recurrence plots which involves subjective assessment of parameters. Although, pattern recognition techniques can be used within this approach, those also require some parameters.

The **R** package `Rssa` implements two methods for the grouping. The first method uses frequency analysis via discrete Fourier transform. Again, as it requires additional parameters, it cannot be qualified as automated grouping. The second method runs a clustering algorithm using a similarity matrix between time series' elementary components. That similarity matrix,  $(s_{i,j}) \in \mathbb{R}^{L \times L}$ , which is more precisely a *w-correlation matrix*, is defined upon the following *weighted inner product*

$$(\mathbb{Y}, \mathbb{Z})_{\mathbf{w}} := \sum_{i=1}^N |A_i| y_i z_i$$

and the corresponding *weighted norm*  $\|\mathbb{X}\|_{\mathbf{w}} := \sqrt{(\mathbb{X}, \mathbb{X})_{\mathbf{w}}}$ . Remember that  $A_i$ , defined in Sect. II Eq. (1) is the set of indexes of the  $i$ -th anti-diagonal of an  $L \times K$  matrix. The  $\mathbf{w}$ -correlation matrix is defined as follows

$$s_{i,j} := \frac{(\mathbb{X}_i, \mathbb{X}_j)_{\mathbf{w}}}{\|\mathbb{X}_i\|_{\mathbf{w}} \|\mathbb{X}_j\|_{\mathbf{w}}}$$

where  $\mathbb{X}$ ,  $\mathbb{Y}$  and  $\mathbb{Z}$  are time series of length  $N$ . Function `grouping.auto.wcor` from `Rssa` implements the latter clustering-based grouping. It can be considered as a fully automated

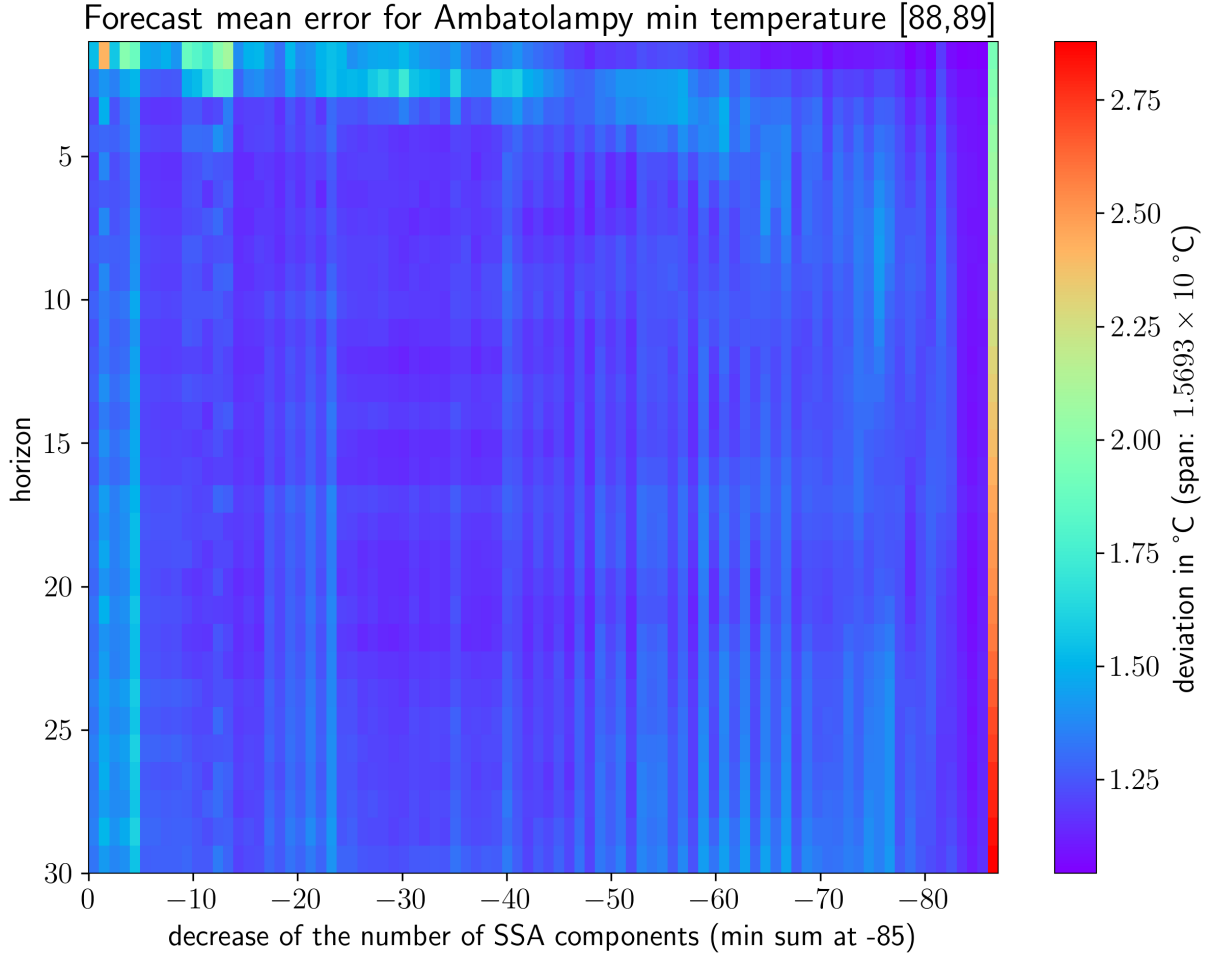


Figure 1

grouping method.

There is no *a priori* restriction on the form of index set  $I$  of “signal” resulting from clustering by `grouping.auto.wcor`. On the contrary, the method of [8] based on minimum description length yields  $M \in [L - 1]$  to be used as prefix grouping [M]. Unfortunately, the method is difficult to implement. No algorithm is clearly stated. Implementations, if any, do not seem available in the public domain.

## VI DATA SETS

Several methods discussed above have been evaluated as a part of the present work using real word data summarised in Table 1. The data sets used here have been downloaded

| kind                    | unit              | location      | coordinates     | begins on  | ends on    |
|-------------------------|-------------------|---------------|-----------------|------------|------------|
| maximum temperature 24h | °C                | Maevatanana   | 16°57'S 46°50'E | 1979-01-01 | 2017-12-31 |
| minimum temperature 24h | °C                | Ambatolampy   | 19°23'S 47°25'E | 1979-01-01 | 2018-12-31 |
| rainfall 24h            | mm                | Marovoay      | 16°6'S 46°38'E  | 1979-01-01 | 2017-12-31 |
| water vapor             | kg/m <sup>2</sup> | Ambovombe     | 25°10'S 46°05'E | 1979-01-01 | 2018-12-31 |
| ozone                   | kg/m <sup>2</sup> | Antananarivo  | 18°56'S 47°31'E | 1979-01-01 | 2018-12-31 |
| mean pressure           | Pa                | Grande Comore | 11°55'S 43°25'E | 1979-01-01 | 2016-12-31 |

Table 1

from the ERA-Interim archive of the European Center for Medium-range Weather Forecasts (ECMWF). The ERA-Interim archives historical forecasts for horizons from 0 to 240 hours. These forecasts consist of reanalysis data. The meaning of “reanalysis” for horizon 0 is that the data either come from observations or, if an observation is unavailable at a given location, the corresponding value is extrapolated/interpolated using a meteorological model. The data sets chosen for this paper concern atmospheric and oceanic phenomena at the northern extremity of the Mozambique Channel (“Grande Comore”) and at several locations in Madagascar. This region of the Western Indian Ocean has a tropical climate experiencing different micro-climates from part to part. In addition to the concern to compare different climatic parameters, each location also has particularities. Maevatanana, (resp. Ambatolampy, Ambovombe, Antananarivo), is the place reputed to be the hottest (resp. the coldest, the driest, the most polluted) on the island. Marovoay is an area with high agricultural potential. The study of rainfall is therefore as interesting as it is essential. The study of the atmospheric pressure in Grande Comore is mainly motivated by forecasting the trajectories of cyclones.

All recordings are daily and start from 1979-01-01. Both water vapor and ozone are expressed in kg/m<sup>2</sup> representing their total amount in a column extending from the surface of the Earth to the top of the atmosphere.

## VII EXPERIMENTS

The aim of numerical experiments reported in this paper was to assess the quality of SSA

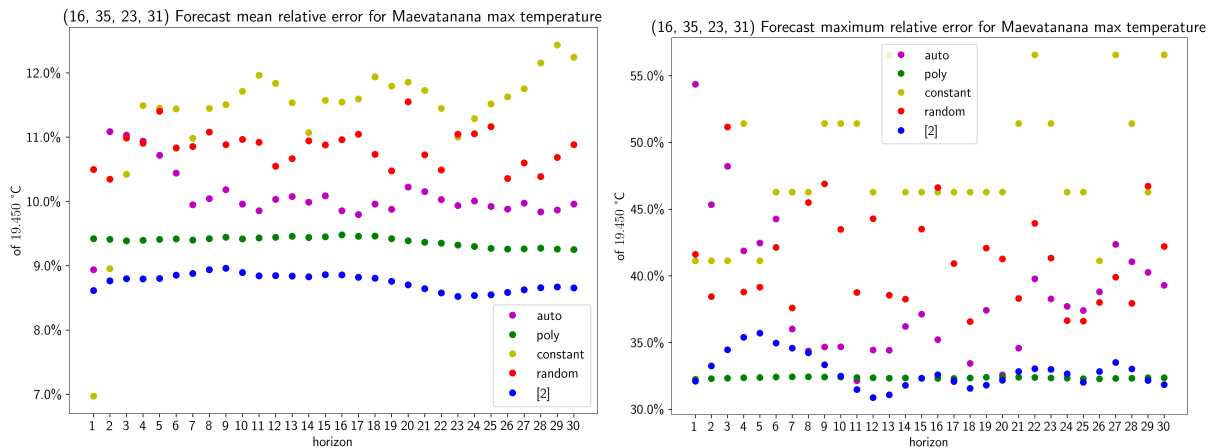


Figure 2

forecasting from user’s point of view for a short duration with horizon  $h \in [30]$ . For each data set, the forecast has been computed on every day of the last year of data, except



on December 31. For a given horizon  $h$ , this resulted in  $365 - h$  forecasting days, except  $366 - h$  forecasting days for leap year 2016 (Grande Comore time series only). Let  $D_h$  (resp.  $F_h$ ) denote the set of forecasting (resp. forecasted) days for horizon  $h$ . For every forecasting day  $j \in D_h$  computing a forecast consisted in taking  $(x_1, \dots, x_j)$  as input time series for

1. estimating the window length (see Sect. IV),
2. embedding and decomposition (see Sect. II),
3. grouping (see Sect. V),
4. vector forecasting (see Sect. III),

where the two latter steps were repeated using various grouping choices in order to collect

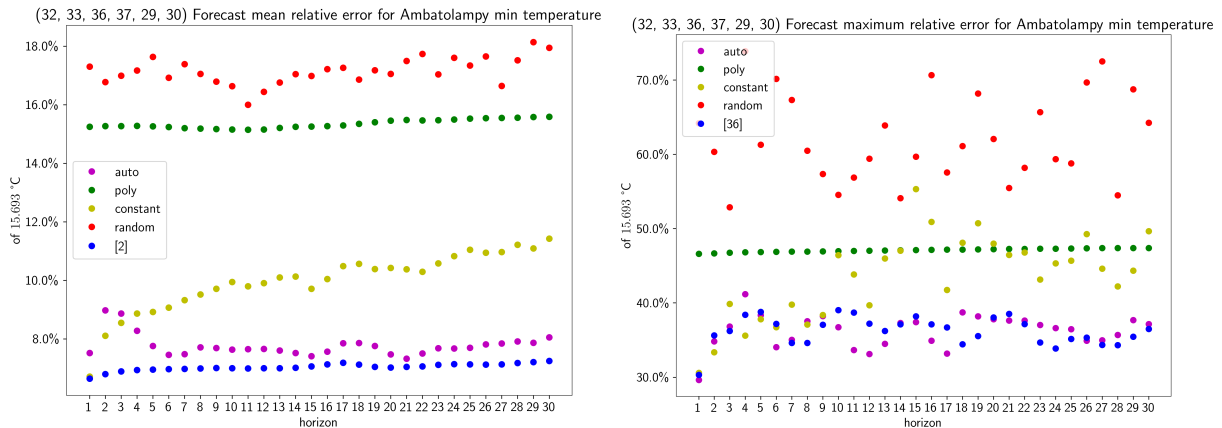


Figure 3

corresponding forecasts. Thus, for a fixed method of the window length estimation, the most computationally expensive part of decomposition was computed only once for each forecasting day.

Assuming the methods for the window length and the grouping are fixed, by repeating steps 1–4 for all forecasting days, one gets vectors  $(L_j : j \in D_1)$ ,  $(I_j : j \in D_1)$  of window lengths and groupings, and, for every horizon  $h$ , a vector of forecasted values  $\mathbf{y}_h = (y_{h,j} : j \in F_h)$ . The *forecasting error vector for  $h$* , is therefore  $\boldsymbol{\xi}_h := \mathbf{y}_h - \mathbf{x}_h$ , where  $\mathbf{x}_h = (x_j : j \in F_h)$  is the vector of the corresponding actual values, viz., the corresponding terminal portion of  $\mathbb{X}$ . The *mean* (resp. *maximum*) *error for  $h$*  is  $\text{mean}(|\boldsymbol{\xi}_h|)$  (resp.  $\max(|\boldsymbol{\xi}_h|)$ ) where “mean” stands for the arithmetic mean. These absolute errors have their relative variants, each one defined as the ratio of the corresponding absolute error divided by span of the data, namely

$$\frac{\text{mean}(|\boldsymbol{\xi}_h|)}{\max(\mathbb{X}) - \min(\mathbb{X})} \quad \text{and} \quad \frac{\max(|\boldsymbol{\xi}_h|)}{\max(\mathbb{X}) - \min(\mathbb{X})} .$$

Although when it comes to forecasting, the mean squared error is mostly used, the authors believe that the arithmetic mean has a clear intuitive meaning for an average user of forecasting. Moreover, she or he can be also interested in the worst-case error. Furthermore, the maximum error can bring some insight about the ability to forecast extreme events which is particularly important for phenomena expressed by data sets used in this study.

For comparison with SSA vector forecasting, the following “naive” forecasting methods have been used as benchmark at every forecasting day  $j \in D_h$ :

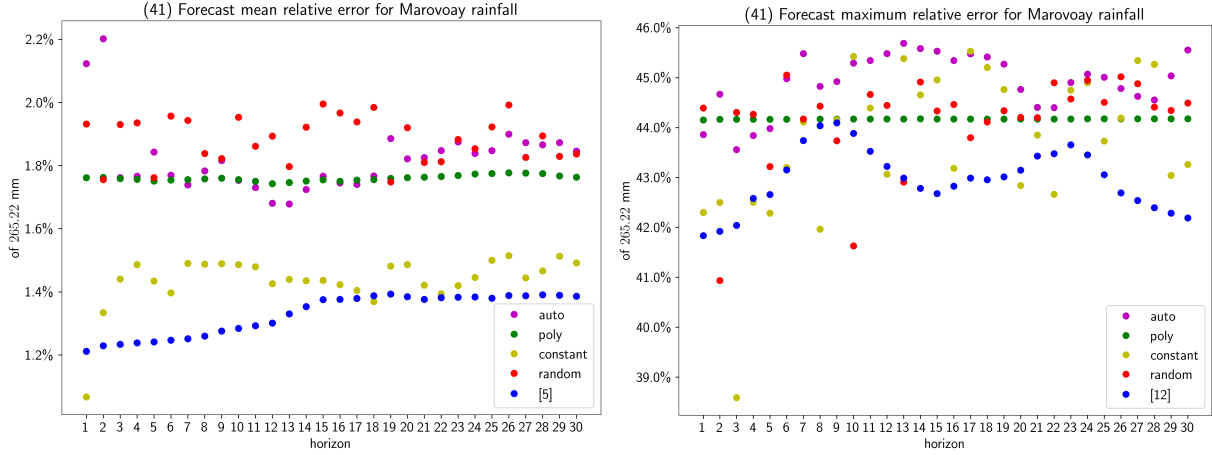


Figure 4

- random forecast – the forecasted value is sampled from the distribution inferred from  $(x_1, \dots, x_j)$ ,
- constant forecast – the forecasted value equals  $x_j$ ,
- regression based forecast – uses polynomial regression (with polynomials of degree 4) from  $(x_1, \dots, x_j)$  to extrapolate the value used as the forecast.

Only rough evaluation of forecast quality with varying window lengths (see Table 2) has

|               | $\min_{j \in D_1} L_{lo,j}$ | $\max_{j \in D_1} L_{lo,j}$ | $\min_{j \in D_1} L_j^{[9]}$ | $\max_{j \in D_1} L_j^{[9]}$ | $\min_{j \in D_1} L_{hi,j}$ | $\max_{j \in D_1} L_{hi,j}$ |
|---------------|-----------------------------|-----------------------------|------------------------------|------------------------------|-----------------------------|-----------------------------|
| Maevatanana   | 29                          | 30                          | 52                           | 52                           | 281                         | 283                         |
| Ambatolampy   | 30                          | 30                          | 90                           | 91                           | 283                         | 285                         |
| Marovoay      | 29                          | 30                          | 78                           | 80                           | 281                         | 283                         |
| Ambovombe     | 30                          | 30                          | 90                           | 91                           | 283                         | 285                         |
| Antananarivo  | 30                          | 30                          | 95                           | 97                           | 283                         | 285                         |
| Grande Comore | 29                          | 29                          | 90                           | 90                           | 279                         | 281                         |

Table 2: Window lengths computed using three methods

been conducted because of an important computational cost of the decomposition step. On an *ad hoc* basis, a part of this evaluation was done for window length  $L_{big,j}$  taken as the largest multiple of a mean year duration 365.25 smaller than  $j/2$ , together with prefix grouping  $I_{big,j} = [M_{big,j}]$  for  $M_{big,j} = L_{big,j} - 1$ . Another part was done for  $L_{hi,j}$  together with prefix grouping  $I_{hi,j} = [M_{hi,j}]$  for  $M_{hi,j} = L_{hi,j} - 1$ .

A deeper evaluation relied essentially on the method of [9] for estimating the window length. In a more systematic way, for each data set, and every forecasting day,  $L_j^{[9]}$ ,  $L_{lo,j}$  and  $L_{hi,j}$  have been computed but only  $L_j^{[9]}$  and  $L_{lo,j}$  have been used for forecasting with various groupings. More precisely, prefix groupings have been performed for all  $M_j \in [L_j^{[9]}]$  (resp.  $M_j \in [L_{lo,j}]$ ). Fig. 1 displays a result of such an exhaustive evaluation. By averaging over the last year of the time series, the best *a posteriori* prefix grouping  $I_{max} = [M_{max}]$  (resp.  $I_{mean} = [M_{mean}]$ ) with respect to the maximum (resp. mean) forecast error has been

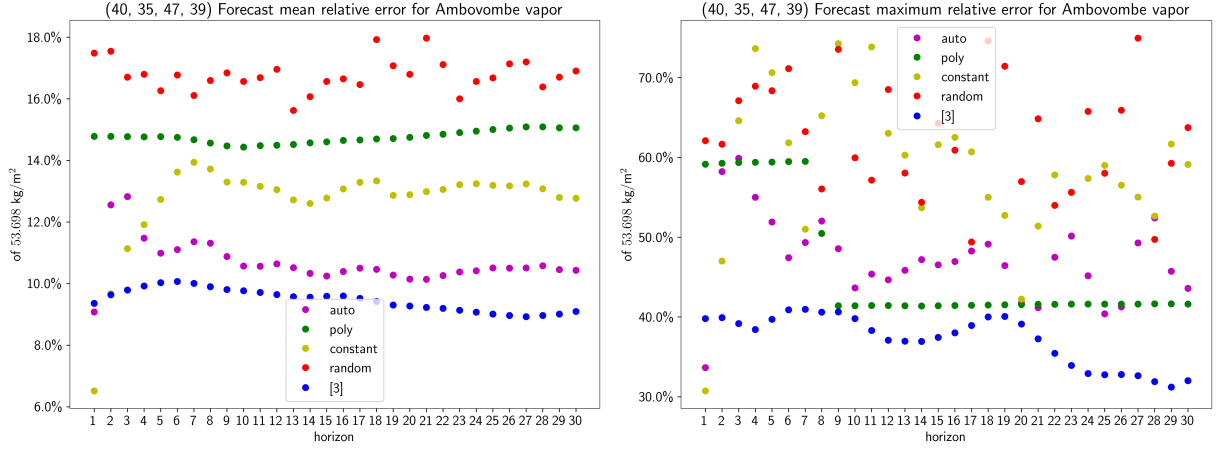


Figure 5

selected for comparing with automated groupings  $I_{\text{auto},j}$  computed by `grouping.auto.wcor`. Also the closest neighbourhood of  $I_{\text{max}}$  (resp.  $I_{\text{mean}}$ ) has been examined. This neighbourhood consists of all index sets that differ from  $I_{\text{max}}$  (resp.  $I_{\text{mean}}$ ) by one element only:

$$\begin{aligned}
 \mathcal{V}_{\text{max}} := & \{ [M_{\text{max}}] \setminus \{k\} : k \in [M_{\text{max}}] \} \cup \\
 & \{ [M_{\text{max}}] \cup \{k\} : k \in [\min_{j \in D_1} L_j^{[9]}] \setminus [M_{\text{max}}] \} \\
 (\text{resp. } \mathcal{V}_{\text{mean}} := & \{ [M_{\text{mean}}] \setminus \{k\} : k \in [M_{\text{mean}}] \} \cup \\
 & \{ [M_{\text{mean}}] \cup \{k\} : k \in [\min_{j \in D_1} L_j^{[9]}] \setminus [M_{\text{mean}}] \} )
 \end{aligned} \tag{3}$$

All numerical experiments were programmed in Python and R. The programs are available upon request from the corresponding author.

## VIII RESULTS AND DISCUSSION

While varying day  $j$  over  $D_1$ , computed window lengths  $L_{\text{lo},j}$ ,  $L_{\text{hi},j}$ , and  $L_j^{[9]}$  remained stable (see Table 2). Interestingly, for every considered data set, one has

$$L_{\text{lo},j} < L_j^{[9]} < L_{\text{hi},j} .$$

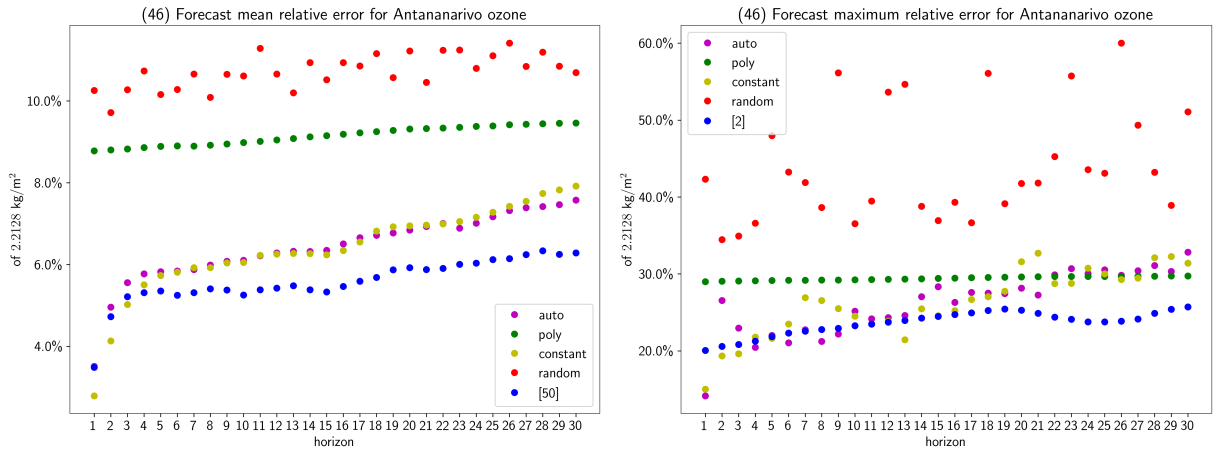


Figure 6

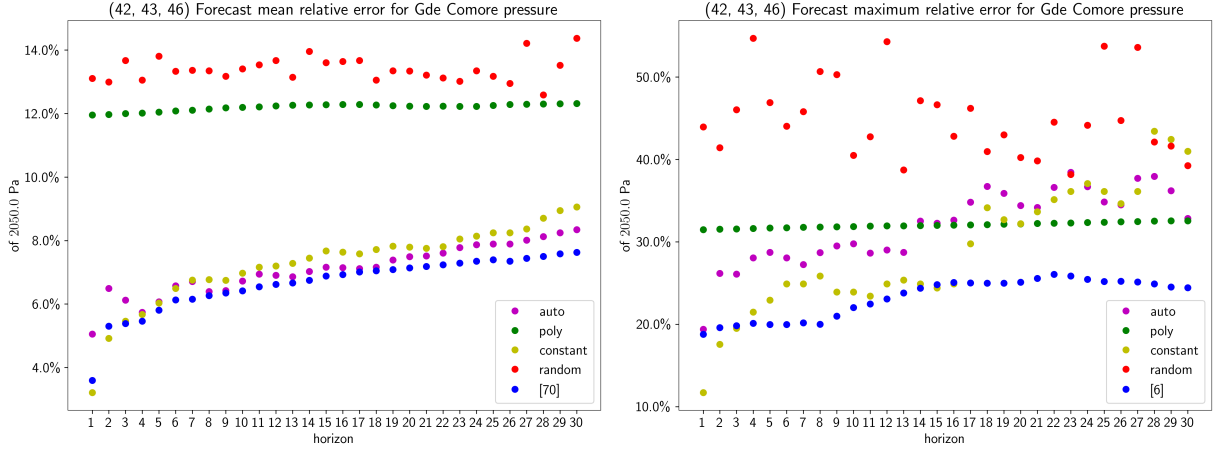


Figure 7

This not a general rule and the fact could be attributed to a specific nature of considered data sets, all resulting from atmospheric and oceanic phenomena.

A rough evaluation of forecasting with parameters  $(L_{\text{big},j}, I_{\text{big},j})$  and  $(L_{\text{hi},j}, I_{\text{hi},j})$  confirms the findings of [8, 10, 11] and [13] that using longer windows do not improve forecast accuracy. Indeed, the forecasts obtained within the present experimentation with parameters  $(L_{\text{big},j}, I_{\text{big},j})$  were not better than random ones. Even with  $(L_{\text{hi},j}, I_{\text{hi},j})$  the accuracy was only slightly better than using random forecast.

A systematic evaluation of prefix grouping for window length  $L_j^{[9]}$  shows that the accuracy of forecasting is very sensitive to it. One could think that including more components would better capture the dynamics via an LRE but often the opposite is true. Fig. 1 shows how, in prefix grouping, decreasing the number of components affects the forecast. The groupings considered in this example range from [88] down to [1]. One can observe that the mean error does not show any regular pattern. The best accuracy is obtained with  $I = [2]$  when comparing mean errors averaged over horizons.

In all time series studied here, the optimal prefix grouping is compared with with clustering-based grouping computed by function `grouping.auto.wcor`. This is displayed on Fig. 2 to 7. Depending on the figure, these show the mean or the maximum error per horizon. The errors are plotted in blue for optimum prefix grouping, in purple for the automated grouping, in green for polynomial regression based forecast, in yellow for constant forecast and in red for random forecast. All errors are given relatively to the span  $\max(\mathbb{X}) - \min(\mathbb{X})$  of the data set as precised on the left edge of each plot. A surprising observation common to all data sets discussed here is that `grouping.auto.wcor` always returned a prefix grouping. This is not a general rule. Indeed, for other time series, `grouping.auto.wcor` may return a cluster of non prefix groupings. Every plot on Fig. 2 to 7 has on top the list of groupings, in parentheses, computed by `grouping.auto.wcor` when varying  $j \in D_1$ . More precisely, an integer  $k$  appearing on the list, means that for some  $j \in D_1$ , `grouping.auto.wcor` returned index set  $[k]$  after taking  $(x_1, \dots, x_j)$  as the input time series. It should be noted here that across  $D_1$ , very few different grouping are obtained and that their variation is non-monotonic. As for all plotted forecast errors for horizon  $h$ , the one resulting from automated grouping is obtained by averaging over  $j \in D_h$ . When the mean error is concerned (Fig. 2 to 7) the forecasts using automated groupings computed by `grouping.auto.wcor` clearly appear as sub-optimal. For Marovoay rainfall (Fig. 4), that

forecast is even close to random forecast and for Maevatanana maximum temperature (Fig. 2) a polynomial regression predicts more accurately.

A systematic examination of prefix groupings for  $L_{10}$  confirms the above observations about automated grouping. It also lets comparing forecast accuracy for window lengths  $L^{[9]}$  and  $L_{10}$  (see figures in the appendix). As far as mean errors are used for comparison, the automated (resp. optimal) grouping with  $L_{10}$  underperforms the automated (resp. optimal) grouping with  $L^{[9]}$  in all example studied, except for Marovoay rainfall. However, when one uses maximum errors instead of mean errors, no winner can be clearly declared. Moreover, the plots of maximum errors for  $L^{[9]}$  (Fig. 2 to 7) seem to show that SSA is unsuitable for forecasting when maximum errors are the main concern. Indeed, even with optimal prefix grouping, the maximum error is mostly beyond 30%.

As all groupings discussed in this section are prefix ones, one may ask if, for a given window size, the optimal prefix grouping remains optimal among all groupings. The authors do not know the answer although they observed that the values of the errors in the closest neighbourhoods  $\mathcal{V}_{\text{mean}}$  or  $\mathcal{V}_{\text{max}}$  (see Equation (3)) of each optimal prefix groupings (plotted in blue on all figures except on Fig. 1) always exceed those of the latter. Therefore, each optimum prefix grouping for data sets considered here form a local minimum. The authors do not know whether this observation could be turned into a theorem nor if such local minima are also global ones.

## IX CONCLUSION

The experiments reported in this paper confirm that the choice of the window length and of the grouping is essential for the accuracy of SSA forecasting. The window length selection method of [9] together with an adequate grouping enables forecasting with an accuracy significantly better than constant or random forecasting, provided that the mean error is considered. However, the reader should keep in mind that each adequate (optimal) grouping has been selected via an *a posteriori* evaluation. The only widely available method for an automated *a priori* grouping, namely function `grouping.auto.wcor` from `Rssa` package, appears as sub-optimal in the analysed examples.

When the maximum error matters, the SSA forecasting seem rather unsuitable for short horizon prediction, at least for atmospheric/oceanic phenomena illustrated by the time series used in this study.

When examining plots of forecasting errors using various methods, one could ask, what could be learned from those about the dynamics of the underlying phenomena. Somewhat intriguing is the fact that the relative position of the plots differs substantially from one time series to another. How to explain that among “naive” forecasting methods the constant one is the worst for Ambatolampy minimum temperature (see Fig 3) but the best one for Maevatanana maximum temperature (see Fig 2) ?

## ACKNOWLEDGEMENTS

The authors wish to thank Juan Bógalo for general explanation about circulant SSA and Hong-Guang Ma for clarification about his method for estimating the window length, and both, for providing their Matlab codes.

## REFERENCES

### Publications

- [1] D. S. Broomhead and G. P. King. “Extracting Qualitative Dynamics From Experimental Data”. In: *Physica D: Nonlinear Phenomena* 20 (1986), pages 217–236.
- [2] R. Vautard and M. Ghil. “Singular Spectrum Analysis in Nonlinear Dynamics with Applications to Paleoclimatic Time Series”. In: *Physica D: Nonlinear Phenomena* 35.3 (1989), pages 395–424.
- [3] J. R. Elsner and A. A. Tsonis. *Singular Spectrum Analysis: a New Tool in Time Series Analysis*. Springer, 1996.
- [4] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky. *Analysis of Time Series Structure: SSA and Related Methods*. Chapman and Hall, 2001.
- [5] M. Allen, M. Dettinger, K. Ide, D. Kondrashov, M. Ghil, M. Mann, A. W. Robertson, A. Saunders, F. Varadi, Y. Tian, and P. Yiou. *The Singular Spectrum Analysis - MultiTaper Method (SSA-MTM) Toolkit*. UCLA Theoretical Climate Dynamics group. 2007.
- [6] P. Grünwald. *The minimum description length principle*. The MIT Press, 2007.
- [7] G. Tzagkarakis, M. Papadopouli, and P. Tsakalides. “Trend forecasting based on Singular Spectrum Analysis of traffic workload in a large-scale wireless LAN”. In: *Performance Evaluation* 66 (2009), pages 173–190.
- [8] M. A. R. Khan and D. S. Poskitt. *Description Length Based Signal Detection in Singular Spectrum Analysis*. Technical report Working paper 13/10. Monash University, Department of Econometrics and Business Statistics, May 2010.
- [9] H.-G. Ma, R. Lei, X.-Y. Kong, Z.-Q. Liu, and Q.-B. Jiang. “Determine a proper window length for singular spectrum analysis”. In: *IET International Conference on Radar Systems (Radar 2012)*. 2012, pages 1–6.
- [10] M. A. R. Khan and D. S. Poskitt. “A note on window length selection in singular spectrum analysis”. In: *Australian & New Zealand Journal of Statistics* 55.2 (2013), pages 87–108.
- [11] M. A. R. Khan and D. S. Poskitt. “Moment tests for window length selection in singular spectrum analysis of short- and long-memory processes”. In: *Journal of Time Series Analysis* 34.2 (2013), pages 141–155.
- [12] R. Wang, H.-G. Ma, G.-Q. Liu, and D.-G. Zuo. “Selection of window length for singular spectrum analysis”. In: *Journal of the Franklin Institute* 352.4 (2015), pages 1541–1560.
- [13] M. A. R. Khan and D. S. Poskitt. “Forecasting stochastic processes using singular spectrum analysis: Aspects of the theory and application”. In: *International Journal of Forecasting* 33.1 (2017), pages 199–213.
- [14] N. Golyandina, A. Korobeynikov, and A. Zhigljavsky. *Singular Spectrum Analysis with R*. Springer, 2018.
- [15] N. Golyandina and A. Zhigljavsky. *Singular Spectrum Analysis for Time Series*. Springer, 2020.
- [16] P. Grünwald and T. Roos. “Minimum description length revisited”. In: *International Journal of Mathematics for Industry* 11.1 (2020), 22 pp.
- [17] A. Korobeynikov, A. Shlemov, K. Usevich, and N. Golyandina. *Rssa: a collection of methods for singular spectrum analysis*. The Comprehensive R Archive Network. 2021.

## APPENDIX: COMPARATIVE PLOTS FOR TWO WINDOW LENGTHS

The following plots compare the accuracy of vector forecasting for window lengths  $L^{[9]}$  and  $L_{10}$ . Every label of a legend gives the window size followed by either “auto” for automated grouping using `grouping.auto.wcor` or the optimal prefix grouping  $[M]$ .





