An experimental evaluation of choices of SSA forecasting parameters

Teodor Knapik, Adolphe Ratiarison, Hasina Razafindralambo

An experimental evaluation of choices of SSA forecasting parameters

Teodor Knapik[*1], **Adolphe Ratiarison**[2], **Hasina Razafindralambo**[2]

[1]ISEA, Université de la Nouvelle Calédonie
[2]DyACO, Université d'Antananarivo

*E-mail : 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 comparison shows that a widespread practice of selecting longer windows leads often to poorer predictions. It also confirms that the choices of the window length and of the grouping are essential.

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. Such past observations form a time series and the main concern of this paper is the following

TIME SERIES FORECASTING PROBLEM
Input: real-valued time series $x_1, \ldots, x_N$ and a forecast horizon $h \in \mathbb{N}_+$,
Output: estimated future values $\hat{x}_{N+1}, \ldots, \hat{x}_{N+h}$.

Note that N is not fixed here and is a part of the input. As an example, one may think of daily recording of the number of births in a country, starting from, say, 1980 until today. The forecasting consists in predicting the number of births that will take place tomorrow, or, more generally, each of next h days, if the forecast horizon expressed in the number of days ahead is h. In order to assess the quality of a forecasting method, one has to repeat the prediction every day over an extended period of time. One has then to compare formerly predicted values with actually recorded ones. Needless to say, the forecasting is one of the main challenges in science, agriculture, policy making or business administration. Therefore, the development of forecasting methods and easy to use tools is important. Besides recurrent neural networks [25] and observable operator models [18], singular spectrum analysis (SSA) [23] provides one of the most promising forecasting frameworks. Indeed, the experiments of [20] indicate that SSA outperforms standard statistical methods in the field such as ARIMA (see e.g.[28]).

This paper reports an experimental investigation of univariate time series prediction using SSA related method, namely the vector forecasting (see e.g.[23]). SSA involves several computational steps among which the vector forecasting may be seen as a final one. 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 [10] and the **R** library Rssa [21, 26]. 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 the forecasting.[1] 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. The question is whether SSA forecasting can be made fully automated and easy to use. Can those methods be included in decision-support tools which automatically select suitable parameters for SSA and compute the required forecast ? Concerning the window length, the present paper brings a positive answer. The result of authors' investigation is that, among very few existing methods, the one of [15] gives the best accuracy of forecasting, at least for meteorological time series used in this paper. Unfortunately, concerning the grouping, the accuracy of the only known truly automated method (available in Rssa package [21]) is not always satisfactory. Although a suggestion for improvement is given at the end of Sect. VIII, one of the conclusions of the paper is that truly automated grouping algorithms need yet to be developed and implemented.

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 [23]), 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 $K := N - L + 1$.

[1] However, in Rssa, automated grouping is also available.

Those lagged vectors form a *trajectory matrix* $\mathbf{X} \in \mathbb{R}^{L\times K}$, $\mathbf{X} = [X_1, \ldots, X_K]$,

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

which is a Hankel matrix, viz., the elements of each anti-diagonal are equal. The k-th anti-diagonal consists of those element of the matrix that are indexed by A_k defined in Eq. (1) below. 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 $L < K$. As the forecasting problem becomes trivial when the rank of $\mathbf{X}$ is strictly less than L, that special case is not considered here in order to simplify the mathematical treatment. Indeed, when the input time series comes from an intricate dynamical system, as those used in this paper, one cannot expect that the rank of $\mathbf{X}$ is strictly less than L.

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) [1] of the trajectory matrix is computed: $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ where $\mathbf{U} = [U_1, \ldots, U_L] \in \mathbb{R}^{L\times L}$ and $\mathbf{V} = [V_1, \ldots, V_K] \in \mathbb{R}^{K\times K}$ are unitary matrices and, $\mathbf{\Sigma} \in \mathbb{R}^{L\times K}$ is a rectangular diagonal matrix with diagonal $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_L \geq 0$, viz., $\Sigma_{kk} = \sigma_k$. Every *eigentriple* (U_k, σ_k, V_k), for $k \in [L]$, where U_k (resp. V_k) is called *left* (resp. *right*) *singular vector* for *singular value* σ_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.

By averaging over anti-diagonals, from an elementary matrix $\mathbf{X}_k$ one gets its *Hankelization* (see also [27]) $\tilde{\mathbf{X}}_k \in \mathbb{R}^{L\times K}$. More precisely the k-th anti-diagonal of an $L \times K$ matrix has its indexes in

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

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

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

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. $\mathbf{X}_k$). Now, $\tilde{\mathbf{X}}_k$ is a Hankel matrix and 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" $\mathbb{X}_{\overline{I}} := \sum_{k\in[L]\setminus I} \mathbb{X}_k$. 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 $\mathbf{X}_{\overline{I}} := \sum_{k\in[L]\setminus I} \mathbf{X}_k$ or their Hankelizations $\mathbf{X} = \tilde{\mathbf{X}}_I + \tilde{\mathbf{X}}_{\overline{I}}$. The reader should note that I is another additional input for SSA and that the choice of I greatly impacts the quality of forecasting [20].

III VECTOR FORECASTING

The vector forecasting is not a part of SSA *per se*. Like two other forecasting methods described in [23], 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 [23] is to find a homogeneous linear recurrent equation (LRE)

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

that is satisfied by "denoised" time series $\mathbb{X}_I = (x_{I,1}, \ldots, x_{I,N})$. For $\mathbb{X}_I$ to satisfy an LRE means it is generated by a linear dynamical system. This lets $\mathbb{X}_I$ a wide class of behaviours. Although "linearity" may sound restrictive for a computer scientist, within the theory of dynamical systems it refers to the underlying evolution functions not to the behaviour of the system itself. Indeed, it is well known that linear dynamical systems behave like a sum of products of polynomials, exponentials and sinusoids [22]. Finding coefficients

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

of LRE (2) amounts to solving the following system of equations

$$\mathbf{a}\underline{\mathbf{X}}_I = (x_{I,L}, \ldots, x_{I,N}) \tag{3}$$

where $\underline{\mathbf{X}}_I$ denotes matrix $\mathbf{X}_I$ without its last row. As $L < N/2$, this system is overdetermined, and in general, has no exact solution, except when $\mathbb{X}_I$ is actually governed by a linear dynamical system of dimension at most L. However, as dynamical systems considered in this paper are not linear, and this is also the case for all systems of intricate dynamics, only approximate solutions of (3) can be obtained. This can be compared with a linear regression where one fits a line to a set of points in an optimal way. In LRE-based forecasting, such as vector forecasting used in this paper, one fits a linear recurrent sequence to the set of observations. For that, the closest approximate solution of (3) with respect to the Euclidean norm is sought. It is well known that such an approximate solution is given by $\mathbf{a} \approx (x_{I,L}, \ldots, x_{I,N})\underline{\mathbf{X}}_I^\dagger$, where "$\_^\dagger$" stands for the pseudo-inverse of Moore-Penrose of a rectangular matrix [2].

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 $\mathbf{U}_I$. Remember that $\mathbf{U}$ is the matrix of left singular vectors in the SVD of $\mathbf{X}$. 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 $\upsilon^2 := \sum_{i\in I} u_{L,i}^2$. Observe that $\upsilon = \cos\theta$ where θ is the angle between $\mathcal{U}_I$ and $\mathbf{e}_L = (0, \ldots, 0, 1)^\top \in \mathbb{R}^L$. As $\mathbf{e}_L \notin \mathcal{U}_I$, one has $\upsilon \ne 1$ and the following vector is well defined

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

It may be shown that $(x_{I,L}, \ldots, x_{I,N})\underline{\mathbf{X}}_I^\dagger = R^\top$. Consequently R is the closest approximate solution of (3). Now, matrix

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

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 $\mathbf{F}: \mathbb{R}^L \to \mathcal{U}_I$ which extends the orthogonal projection $\Pi\check{\mathbf{z}}$ of $\check{\mathbf{z}}$ with the next term of the recurrent sequence inferred from (2)

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

one defines a sequence of vectors

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

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

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

obtained by extending $\mathbf{X}_I$ on the right with vectors $Y_{K+1}, \ldots, Y_{N+h}$ resulting from iterating $\mathbf{F}$ on X_K, where $Y_{K+i} = \mathbf{F}^i X_K$. In this context, operator $\mathbf{F}$ can be understood as a recurrence over vectors of $\mathcal{U}_I$ obtained by an appropriate lifting of LRE (2). By Hankelization of $\mathbf{Y}$ and its subsequent inverse embedding, one gets a time series $\mathbb{Y} = (y_1, \ldots, y_{N+h+L-1})$ where the portion $(y_{N+1}, \ldots, y_{N+h})$ 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. Indeed, L should be understood as the chosen dimension for the model, built from SSA, of the observed dynamical system. It determines the order of LRE (2) which is precisely $L-1$. Foundational texts (e.g. [4, 6]) and books [7, 8, 23] 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 of periodic components of the underlying dynamical system. 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 [13] throughout an adaptation to SSA of the minimum description length principle (see e.g. [11, 24]) better known as Kolmogorov complexity. 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 [13] 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 $L \in \{2, \ldots, 301\}$, one has to perform 300 singular value decompositions (SVD). Beyond formal demonstrations, in [17] and [16] the authors of [13]

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*

$$R(\tau) := \frac{1}{\sigma^2} \sum_{i=1}^{N-\tau} (x_{i+\tau} - \mu)(x_i - \mu) \tag{4}$$

where μ (resp. σ) is the empirical mean (resp. empirical standard deviation) of $\mathbb{X}$. In [12] the smallest value of τ where R crosses the confidence interval corresponding to (95% of) the white Gaussian noise (with parameters μ and σ) is used as estimate of L. In [15] and [19] the smallest value of τ such that $R(\tau)R(\tau+1) < 0$ is used as estimate of L.

In the sequel, $L^{[15]}$ stands for the length of the window chosen with the latter method whereas L_{lo} and L_{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. Indeed, as mentioned earlier, SSA should be understood merely as a method for separating the true signal from the noise within the raw signal obtained from observations. Besides theoretical results, the experiments carried in [16] 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 $\mathbb{X}_I = \sum_{i=1}^{k} \mathbb{X}_i$. This shall be called a *prefix grouping* in the sequel.

A common practice for the grouping (see e.g. [23]) 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 $\mathbf{w}$-*correlation matrix*, is defined upon the following *weighted inner product*

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

where $\mathbb{Y}$ and $\mathbb{Z}$ are time series of length N, 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

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

Function grouping.auto.wcor from Rssa implements the latter clustering-based grouping.

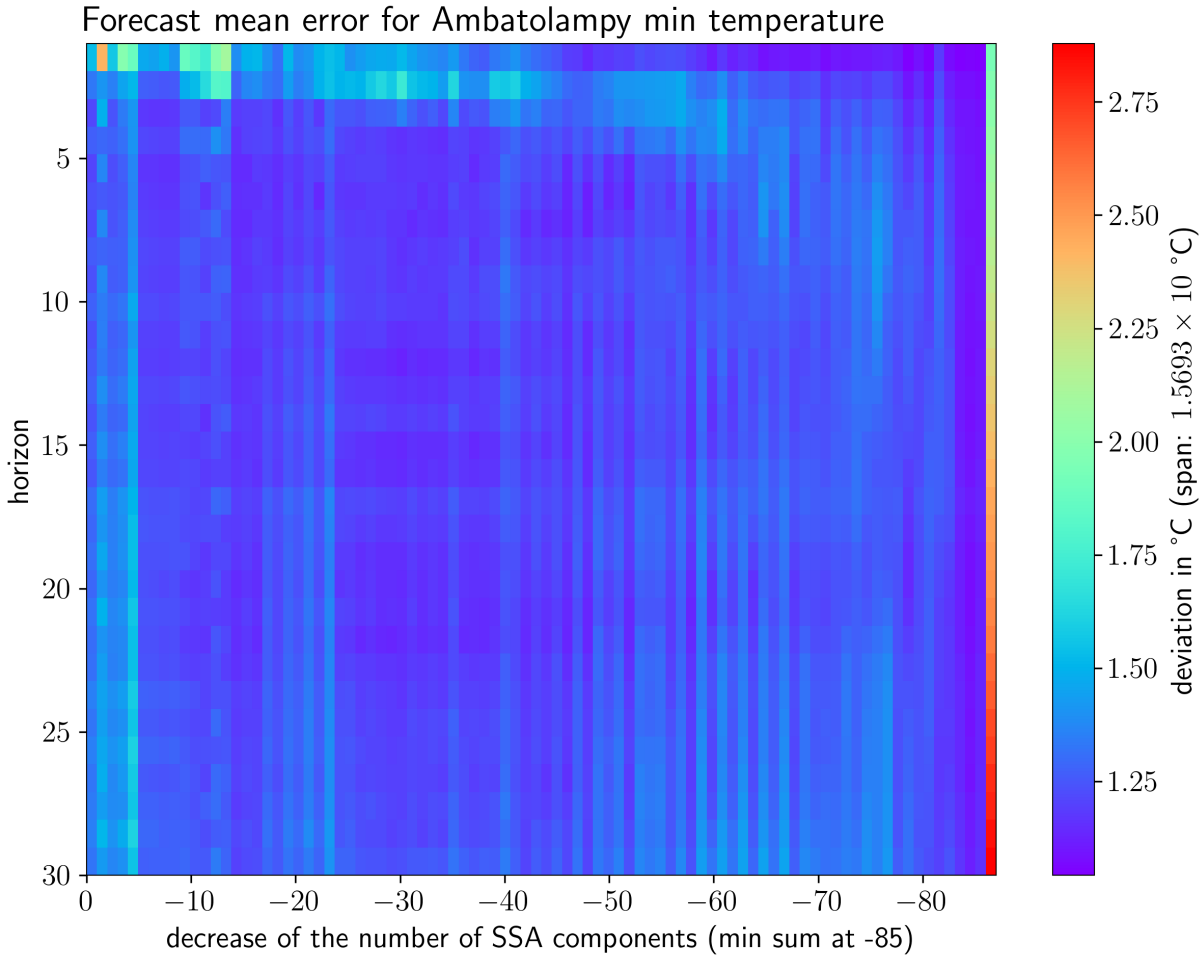


Figure 1: Forecast mean error for all prefix groupings from [89] down to [1]

It can be considered as a fully automated 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 [13] 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 from the ERA-Interim archive of the European Centre for Medium-range Weather Forecasts (ECMWF). The ERA-Interim archives historical forecasts for horizons from 0 to 240

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/m²	Ambovombe	25°10'S 46°05'E	1979-01-01	2018-12-31
ozone	kg/m²	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

hours. These forecasts consist of reanalysis data. The meaning of "reanalysis" for horizon 0 is that the data either come form observations or, if an observation is unavailable at a given location, the corresponding value is interpolated using a meteorological model. Whether a data set comes entirely from observations or has some interpolated parts (due e.g. to a time outage of a recording station) does not matter for this study, in view of a high accuracy of interpolation of those specialised models. On the contrary, the location of each data set matters, as explained in the sequel. It should be noted that all data sets used here are "forecasts" for horizon 0 which means that these are not predicted but rather measured values or, exceptionally, interpolated ones.

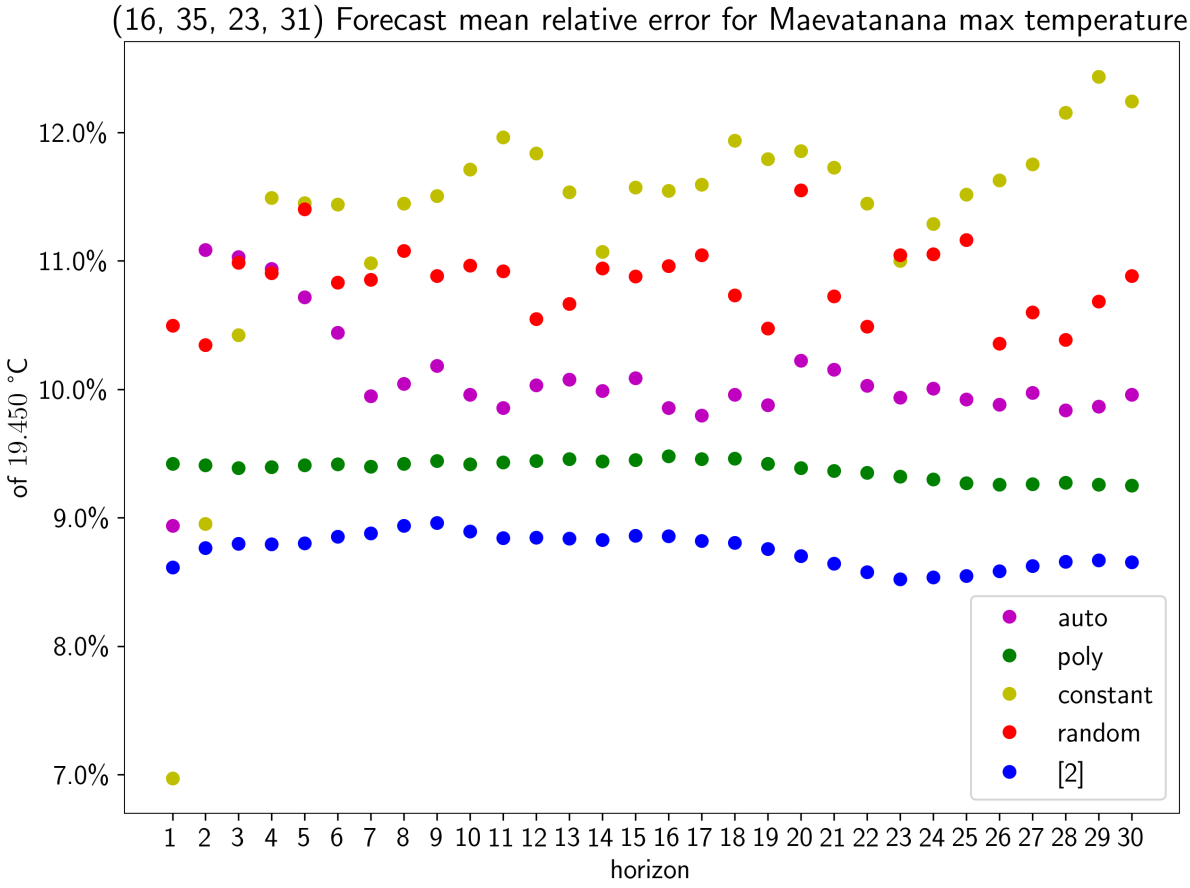


Figure 2: Forecast mean relative error for Maevatanana maximum temperature

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.

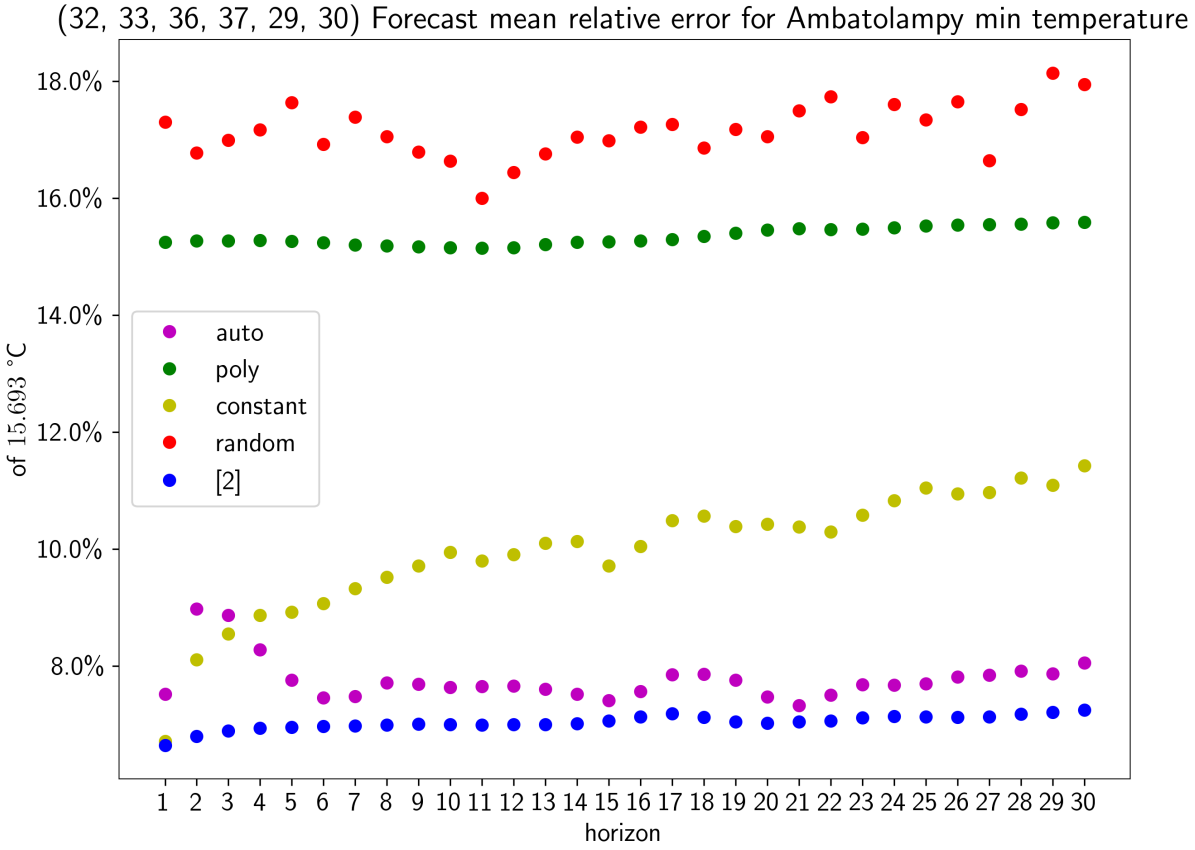


Figure 3: Forecast mean relative error for Ambatolampy minimum temperature

All recordings are daily and start from 1979-01-01. Both water vapor and ozone are expressed in kg/m^2 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 asses the quality of SSA forecasting from user's point of view for a short duration with horizon $h \in [30]$. Here "short duration" is relative to the length of the time series. In fact, meteorologist speak of medium-range when $h \in [10]$ and long-range when the horizon exceeds 7 days although these limits are not strict. Specialised meteorological models have excellent accuracy of forecasts up to 5 days. The choice for $h \in [30]$ is motivated by a potential future comparative study of forecasting accuracies of specialised meteorological models vs. general-purpose time series forecasting methods such as those from SSA.

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 $\mathbb{X}_{\leq j} := (x_1, \ldots, x_j)$ 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 corresponding forecasts. Thus, for a fixed method of the window length estimation, the most computationally expensive part, namely SVD, was computed only once for each forecasting day.

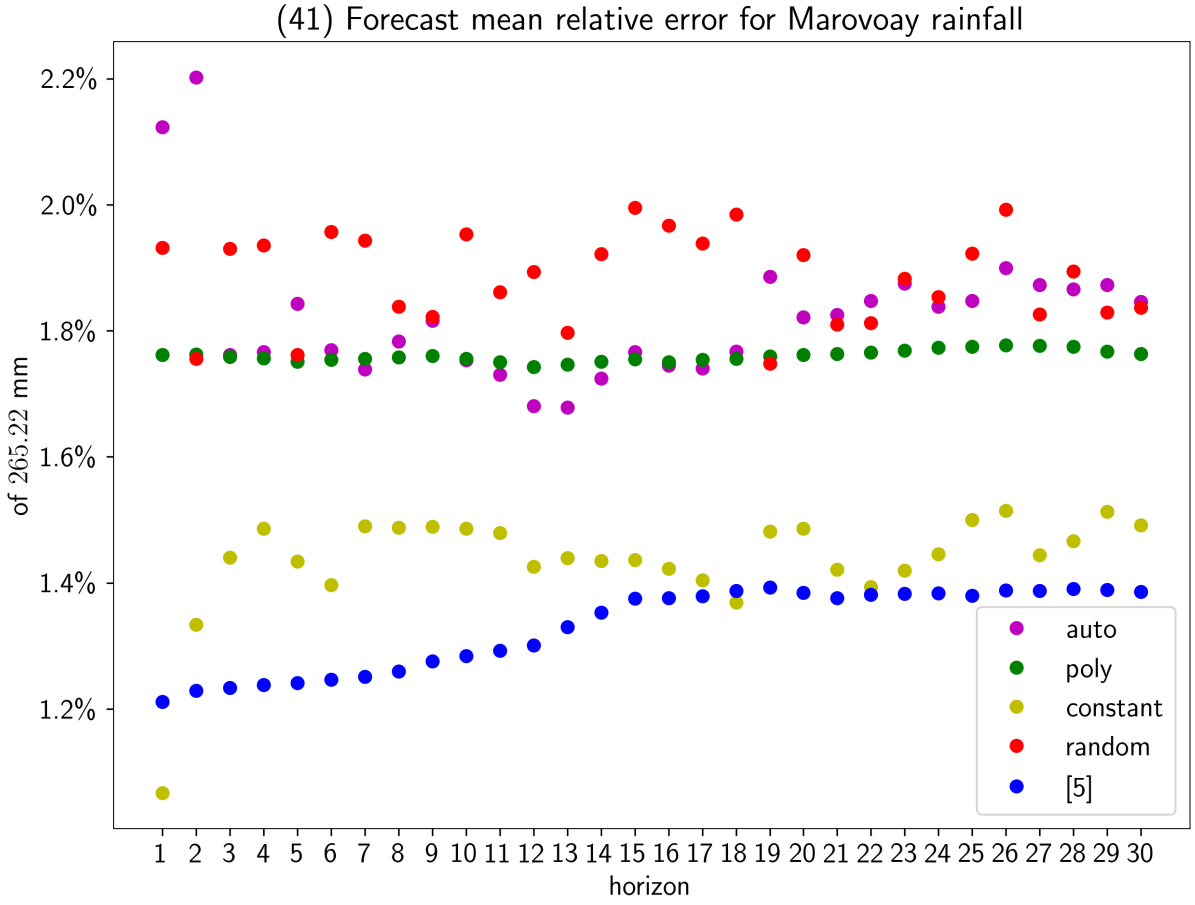


Figure 4: Forecast mean relative error for Marovoay rainfall

Assuming the methods for the window length and the grouping are fixed, by repeating steps 1–4 for all forecasting days, one gets vectors $(L_j : j \in D_1)$ and $(I_j : j \in D_1)$ of window lengths and groupings, and, for every horizon h, a vector of forecasted values

$$\mathbf{y}_h = (y_{h,j} : j \in F_h),$$

where $y_{h,j}$ is the value for day $j+h$ forecasted from $\mathbb{X}_{\leq j}$ (as if it were done on day j). 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 $\mathrm{mean}(|\boldsymbol{\xi}_h|)$ (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

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

Although when it comes to forecasting, the mean squared error is mostly used, the

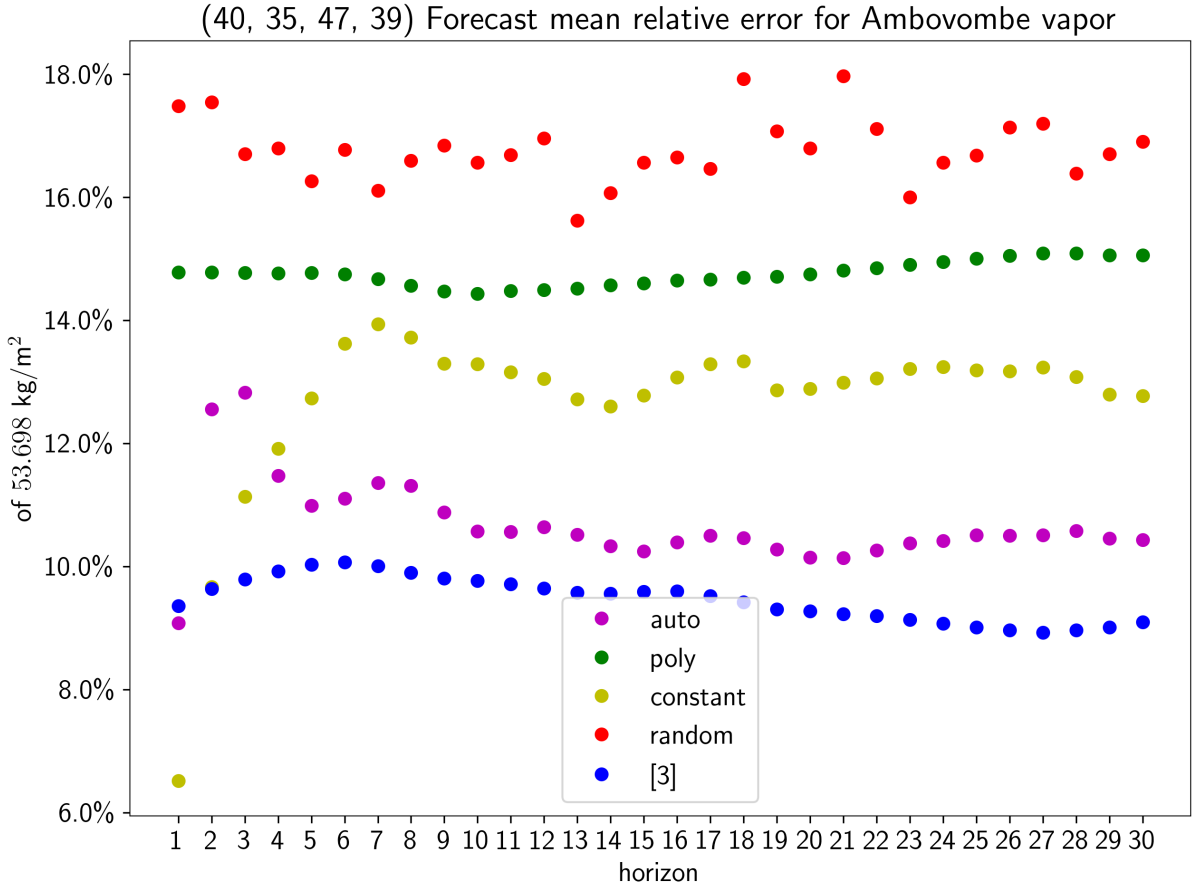


Figure 5: Forecast mean relative error for Ambovombe vapor

authors believe that the arithmetic mean has a clear intuitive meaning for an average user. By the way, the maximum error can also be important in many forecasting tasks. It can for instance bring some insight about the ability to forecast extreme events. This 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 $j \in D_h$:

- random forecast – the forecasted value is sampled from the distribution inferred from $\mathbb{X}_{\leq j}$,
- constant forecast – the forecasted value equals x_j,

- regression based forecast – uses polynomial regression (with polynomials of degree 4) from $\mathbb{X}_{\leq j}$ to extrapolate the value used as the forecast.

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

	$\min_{j\in D_1} L_{\mathrm{lo},j}$	$\max_{j\in D_1} L_{\mathrm{lo},j}$	$\min_{j\in D_1} L_j^{[15]}$	$\max_{j\in D_1} L_j^{[15]}$	$\min_{j\in D_1} L_{\mathrm{hi},j}$	$\max_{j\in D_1} L_{\mathrm{hi},j}$
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 $L_{\mathrm{big},j}$ taken as the largest multiple of a mean year duration 365.25 smaller than $j/2$, together with prefix grouping $I_{\mathrm{big},j} = [M_{\mathrm{big},j}]$ for $M_{\mathrm{big},j} = L_{\mathrm{big},j} - 1$. Another part was done for $L_{\mathrm{hi},j}$ together

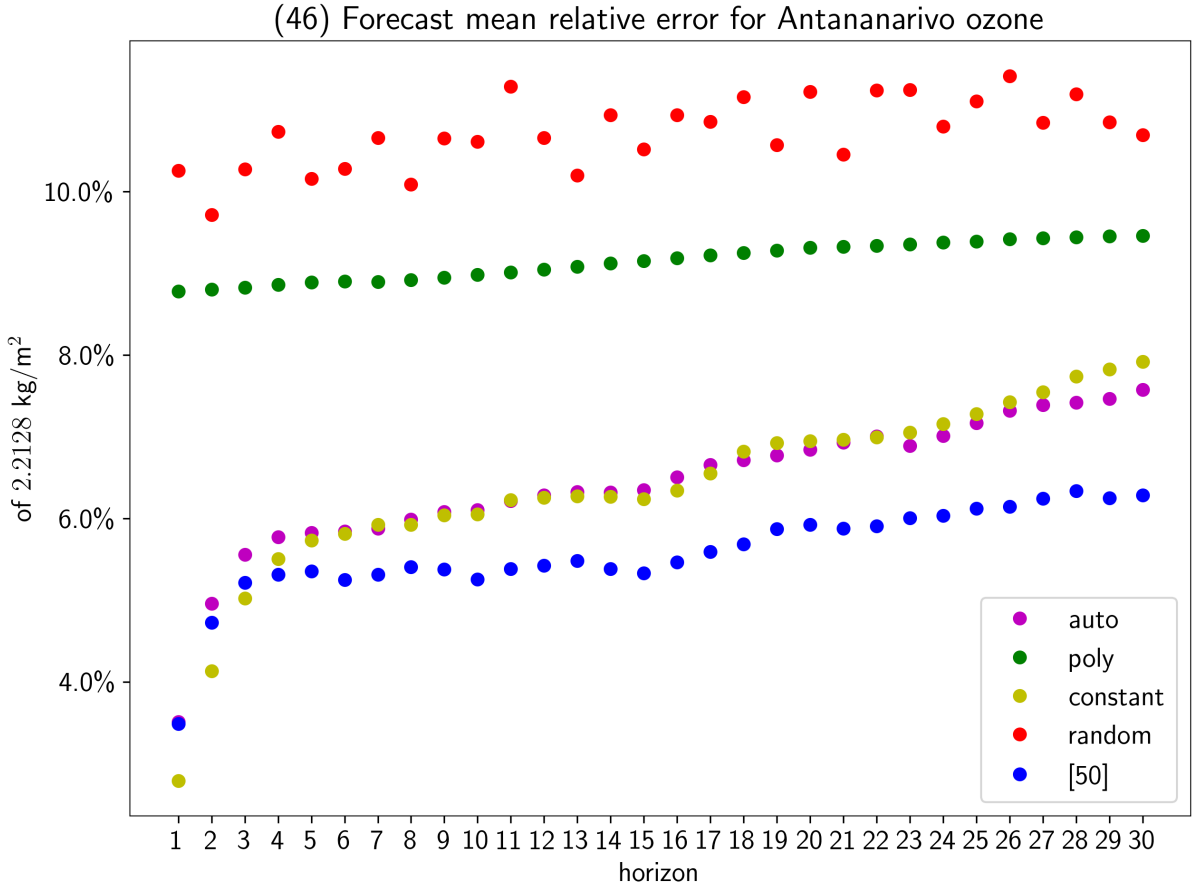


Figure 6: Forecast mean relative error for Antananarivo ozone

with prefix grouping $I_{\mathrm{hi},j} = [M_{\mathrm{hi},j}]$ for $M_{\mathrm{hi},j} = L_{\mathrm{hi},j} - 1$. Remember that $L_{\mathrm{lo},j} = \lceil (\log j)^{1.5} \rceil$

and $L_{\mathrm{hi},j} = \lfloor (\log j)^{2.5} \rfloor$ are extreme integer values of possible window sizes suggested in [13].

A deeper evaluation relied essentially on the method of [15] with estimating the window length for $\mathbb{X}_{\leq j}$ written $L_j^{[15]}$. In a more systematic way, for each data set, and every forecasting day, $L_j^{[15]}$, $L_{\mathrm{lo},j}$ and $L_{\mathrm{hi},j}$ have been computed but only $L_j^{[15]}$ and $L_{\mathrm{lo},j}$ have been used for forecasting with various groupings. More precisely, prefix groupings have been performed for all $M_j \in [L_j^{[15]}]$ (resp. $M_j \in [L_{\mathrm{lo},j}]$). 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 $I_{\mathrm{mean}} = [M_{\mathrm{mean}}]$ (resp. $I_{\max} = [M_{\max}]$) with respect to the mean (resp. maximum) forecast error has been selected for comparing with automated groupings $I_{\mathrm{auto},j}$ computed by grouping.auto.wcor. Also the closest neighbourhood of I_{mean} (resp. $I_{\max}$) has been examined. This neighbourhood consists of all index sets that differ

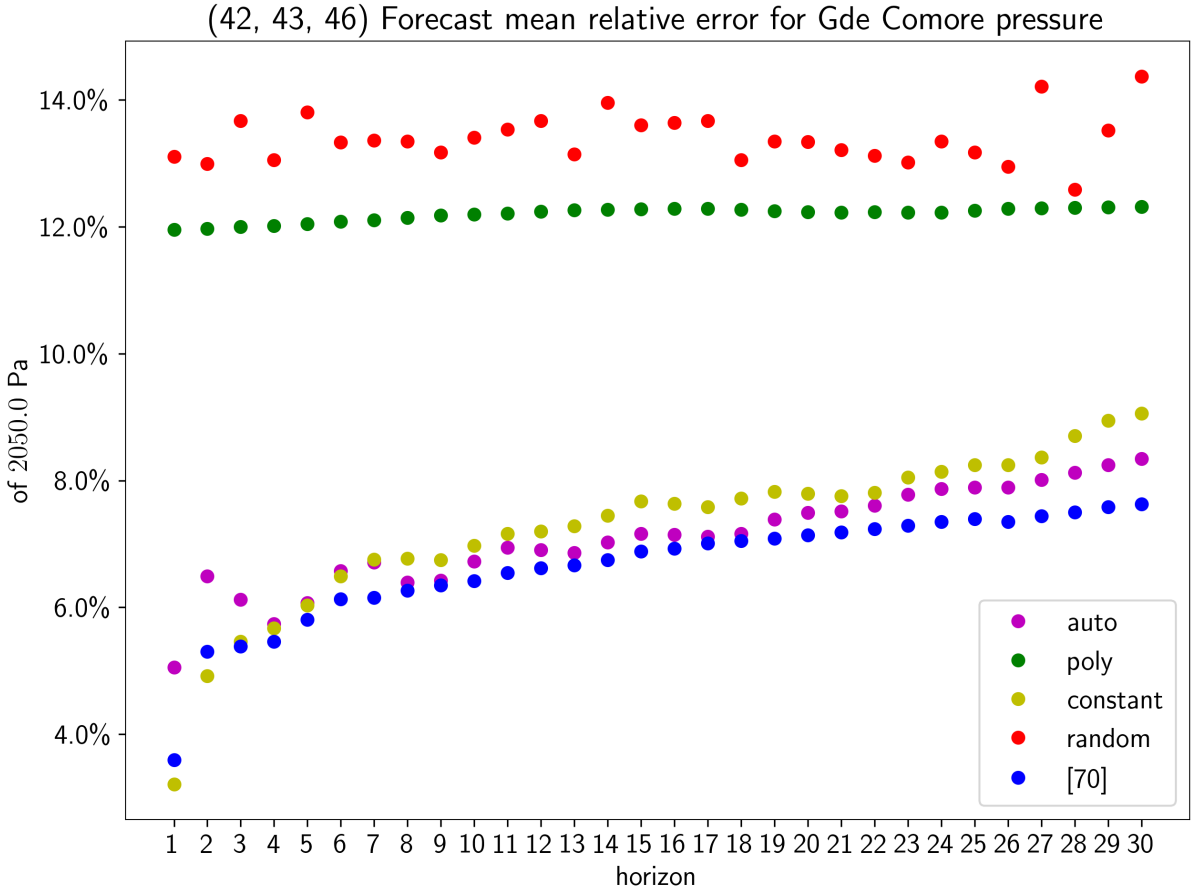


Figure 7: Forecast mean relative error for Grande Comore atmospheric pressure

from I_{mean} (resp. $I_{\max}$) by one element only:

$$\begin{aligned} \mathcal{V}_{\mathrm{mean}} :=& \big\{ [M_{\mathrm{mean}}] \setminus \{k\} : k \in [M_{\mathrm{mean}}] \big\} \cup \\ & \big\{ [M_{\mathrm{mean}}] \cup \{k\} : k \in [\min_{j \in D_1} L_j^{[15]}] \setminus [M_{\mathrm{mean}}] \big\} \\ \big(\text{resp. } \mathcal{V}_{\max} :=& \big\{ [M_{\max}] \setminus \{k\} : k \in [M_{\max}] \big\} \cup \\ & \big\{ [M_{\max}] \cup \{k\} : k \in [\min_{j \in D_1} L_j^{[15]}] \setminus [M_{\max}] \big\} \big) \end{aligned} \tag{5}$$

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

VIII RESULTS AND DISCUSSION

As a preamble to this chapter, it is worth mentioning that SSA and the corresponding forecasting methods are not considered as learning algorithms. Nevertheless, some analogies with machine learning or their lack are worth highlighting.

1. Undertraining happens exactly as in machine learning when the input time series is too short to capture all essential behaviour of the observed dynamical system. In other words, there is not enough of observations.
2. Overfitting results from the choice of index set I including too many inessential components (obtained using SVD). This choice, called "grouping" is discussed in Sect. V. Since those inessential components are considered as noise, a model (an LRE in the case of SSA forecasting) capturing such noise is overfitted.
3. Underfitting is, as usual, the opposite of overfitting. It occurs when index set I does not include enough of essential components. The reader should keep in mind that $I \subseteq [L]$. Thus, when L is too small, the underfitting cannot be compensated by a good choice of I.
4. The choice of window length L does not seem to have a straightforward machine learning counterpart. It can be understood as the choice of the dimension of the model. This is for instance similar to the choice of the number of states of a hidden Markov model (see e.g. [5]) required by spectral learning algorithms [14] or by the classical Baum-Welch algorithm [3, 9]. The choice of L can be also compared to the choice of the degree of the polynomial in polynomial regression. Choosing this degree too big leads typically to an overfitting. The potential overfitting when L is too big can be however compensated to some extent by the choice of I.

While varying day j over D_1, computed window lengths $L_{\mathrm{lo},j}$, $L_{\mathrm{hi},j}$. and $L_j^{[15]}$ remained stable (see Table 2). Interestingly, for every considered data set and each $j \in D_1$, one has

$$L_{\mathrm{lo},j} < L_j^{[15]} < L_{\mathrm{hi},j} \ .$$

This is surprising as $L_{\mathrm{lo},j}$ and $L_{\mathrm{hi},j}$ are the extreme integer values within the real interval $[(\log j)^{1.5}, (\log j)^{2.5}]$ which depends only on length j of $\mathbb{X}_{\leq j}$. On the other hand, $L_j^{[15]}$ depends not only on j but also on the autocorrelation function (see Eq. 4). Consequently, on the contrary to $L_{\mathrm{lo},j}$ and $L_{\mathrm{hi},j}$, $L_j^{[15]}$ depends on the values of $\mathbb{X}_{\leq j}$. The above inequality is not a general rule and 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 $(L_{\mathrm{big},j}, I_{\mathrm{big},j})$ and $(L_{\mathrm{hi},j}, I_{\mathrm{hi},j})$ confirms the findings of [13, 16, 17] and [20] that using longer windows do not improve forecast accuracy. Indeed, the forecasts obtained within the present experimentation with parameters $(L_{\mathrm{big},j}, I_{\mathrm{big},j})$ were not better than random ones. Even with $(L_{\mathrm{hi},j}, I_{\mathrm{hi},j})$ the accuracy was only slightly better than using random forecast. The only convincing results come with $L_j^{[15]}$ and $L_{\mathrm{lo},j}$. These are depicted in Appendix B. Their comparison lets conclude that the results obtained with $L_j^{[15]}$ are better than with $L_{\mathrm{lo},j}$. Consequently the method of [15] based on the autocorrelation function (see Eq. (4)) is favoured by the authors as it is easy to implement and gives satisfactory results.

A systematic evaluation of prefix grouping for window length $L_j^{[15]}$ 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 all horizons from 1 to 30.

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 which show the mean error per horizon. Similar plots for maximum error appear in Appendix A and use the same colour codes. 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 $\mathbb{X}_{\leq j}$ 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_{lo} confirms the above observations about automated grouping. It also lets comparing forecast accuracy for window lengths $L^{[15]}$ and L_{lo} (see figures in Appendix B). As far as mean errors are used for comparison, the automated (resp. optimal) grouping with L_{lo} underperforms the automated (resp. optimal) grouping with $L^{[15]}$ in all examples studied, except for Marovoay rainfall. However, when one uses maximum errors (right column) instead of mean errors (left column), no winner can be clearly declared. Moreover, the plots of maximum errors for $L^{[15]}$ (see Appendix A) 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%. This is perhaps a general drawback of all general-purpose forecasting methods for time series, as the maximum error criterion seems to be deliberately avoided in the corresponding literature.

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 $\mathcal{V}_{\max}$ (see Eq. (5)) 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. In any case, the latter observation leads to the conclusion that a viable strategy for improving the automated grouping would be to start with the value yield by grouping.auto.wcor and find a nearest local optimum for prefix grouping.

IX CONCLUSION

The experiments reported in this paper confirm that the choice of the window length and of the grouping are essential for the accuracy of SSA forecasting. The window length selection method of [15] 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. Consequently, this is where the research on SSA should focus in order to make SSA forecasting ready to be included in decision-support tools. This conclusion is completed with the result of the comparison of the window length selection method of [15] – it outperforms other methods evaluated in reported experiments.

When the maximum error matters, SSA forecasting seems rather unsuitable for short horizon prediction, at least for atmospheric/oceanic phenomena illustrated by the time series used in this study. The lack of literature with the maximum error criterion does not let suggest an alternative.

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 form 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] C. Eckart and G. Young. "The approximation of one matrix by another of lower rank". In: *Psychometrika* 1 (1936), pages 211–218.

[2] R. Penrose. "On best approximate solutions of linear matrix equations". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 52.1 (1956), pages 17–19.

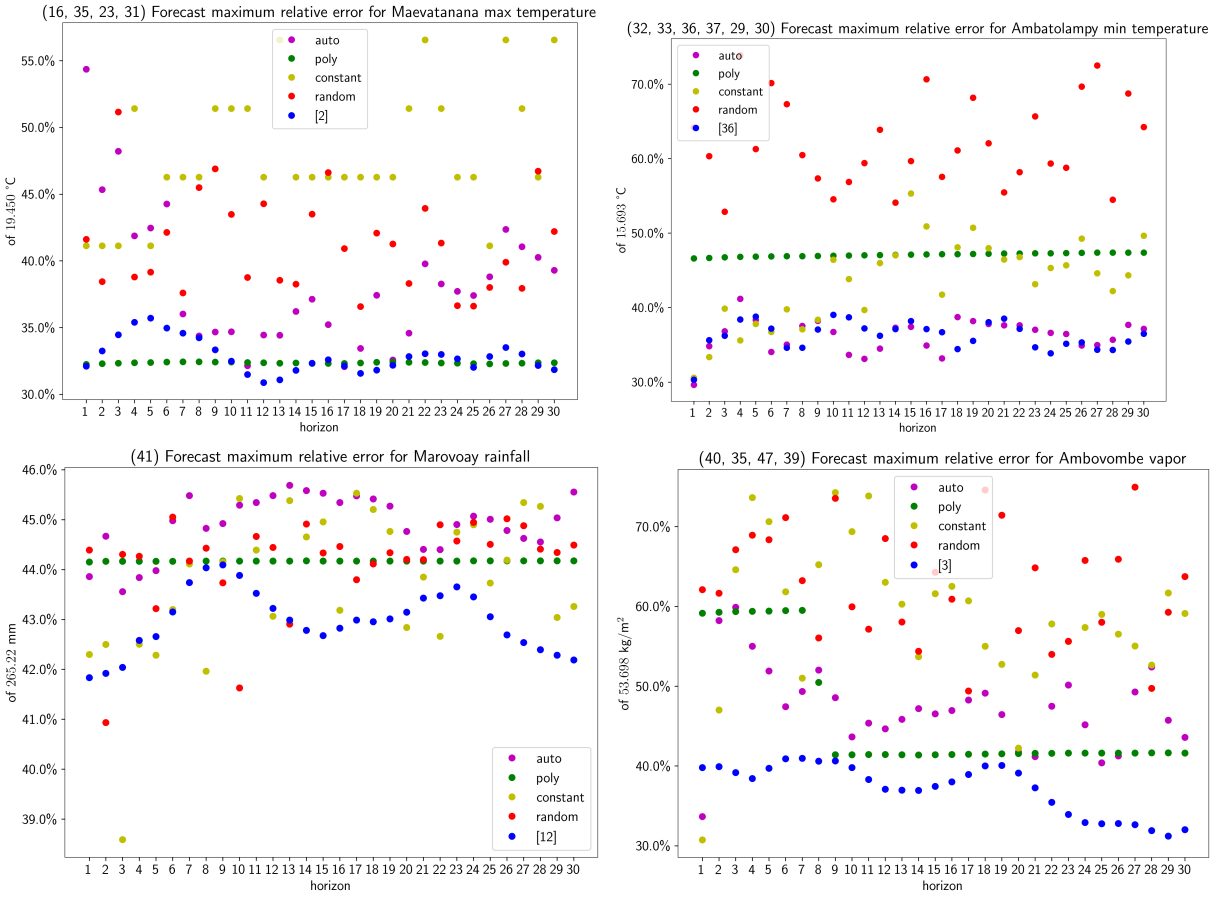
[3] L. E. Baum. "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of a Markov Process". In: *Inequalities* 3 (1972), pages 1–8.

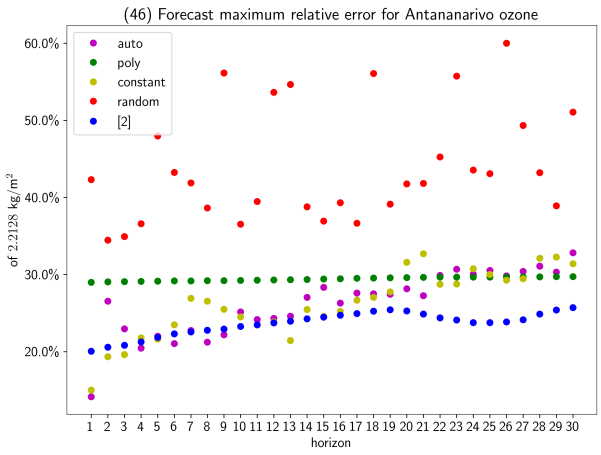
[4] D. S. Broomhead and G. P. King. “Extracting Qualitative Dynamics From Experimental Data”. In: *Physica D: Nonlinear Phenomena* 20 (1986), pages 217–236.
[5] L. R. Rabiner. “A tutorial on Hidden Markov Models and selected applications in speech”. In: *Proceedings of the IEEE* 77.2 (1989), pages 257–286.
[6] 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.
[7] J. R. Elsner and A. A. Tsonis. *Singular Spectrum Analysis: a New Tool in Time Series Analysis*. Springer, 1996.
[8] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky. *Analysis of Time Series Structure: SSA and Related Mehtods*. Chapman and Hall, 2001.
[9] L. R. Welch. “Hidden Markov models and the Baum-Welch Algorithm”. In: *IEEE Information Theory Society Newsletter* 53.4 (2003), pages 1, 10–13.
[10] 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.
[11] P. Grünwald. *The minimum description length principle*. The MIT Press, 2007.
[12] 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.
[13] 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.
[14] D. Hsu, S. M. Kakade, and T. Zhang. “A spectral algorithm for learning Hidden Markov Models”. In: *Journal of Computer and System Sciences* 78.5 (2012), pages 1460–1480.
[15] 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.
[16] 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.
[17] 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.
[18] M. R. Thon and H. Jaeger. “Links between multiplicity automata, observable operator models and predictive state representations: a unified learning framework”. In: *Journal of Machine Learning Research* 16 (2015), pages 103–147.
[19] 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.
[20] 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.
[21] N. Golyandina, A. Korobeynikov, and A. Zhigljavsky. *Singular Spectrum Analysis with R*. Springer, 2018.
[22] J. P. Hespanha. *Linear Systems Theory*. Princeton University Press, 2018.

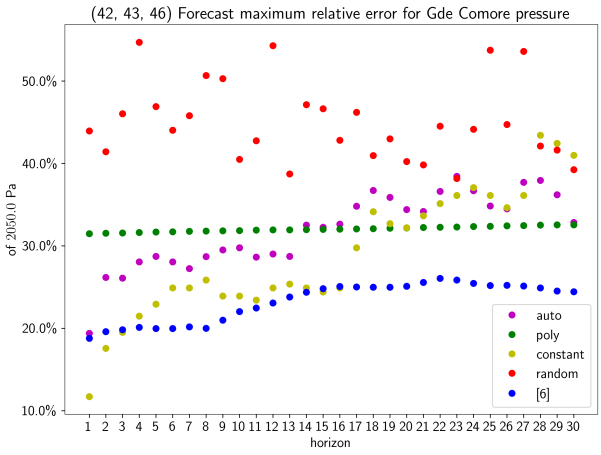
[23] N. Golyandina and A. Zhigljavsky. *Singular Spectrum Analysis for Time Series*. Springer, 2020.
[24] P. Grünwald and T. Roos. "Minimum description length revisited". In: *International Journal of Mathematics for Industry* 11.1 (2020), 22 pp.
[25] H. Hewamalage, C. Bergmeir, and K. Bandara. "Recurrent Neural Networks for Time Series Forecasting: Current status and future directions". In: *International Journal of Forecasting* 37.1 (2021), pages 388–427.
[26] A. Korobeynikov, A. Shlemov, K. Usevich, and N. Golyandina. *Rssa: a collection of methods for singular spectrum analysus*. The Comprehensive R Archive Network. 2021.
[27] F. Sedighin and A. Cichocki. "Image Completion in Embedded Space Using Multi-stage Tensor Ring Decomposition". In: *Frontiers in Artificial Intelligence* 4 (2021).
[28] R. J. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. 2023. URL: https://otexts.com/fpp3/.

APPENDICES

A MAXIMUM ERROR PLOTS FOR $L^{[15]}$

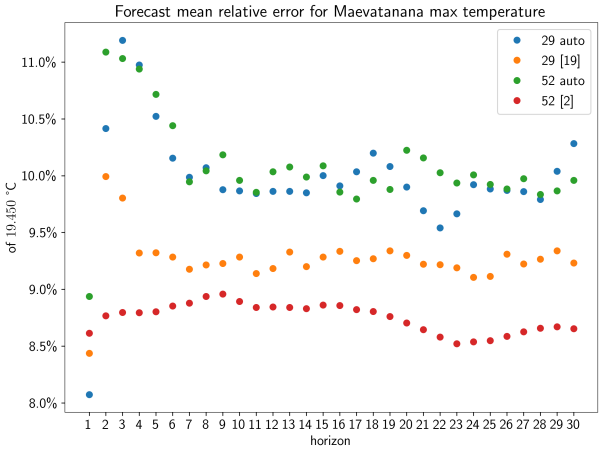


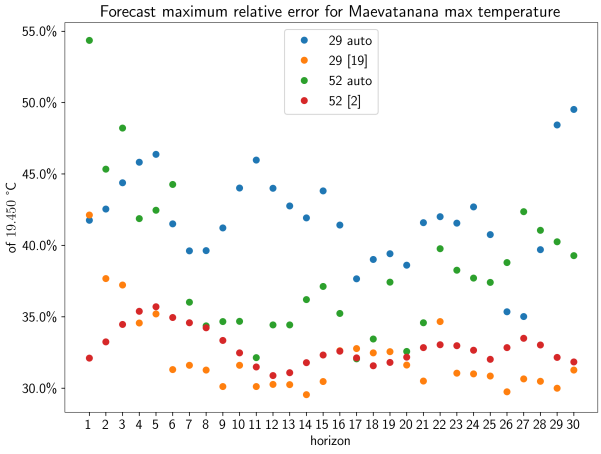


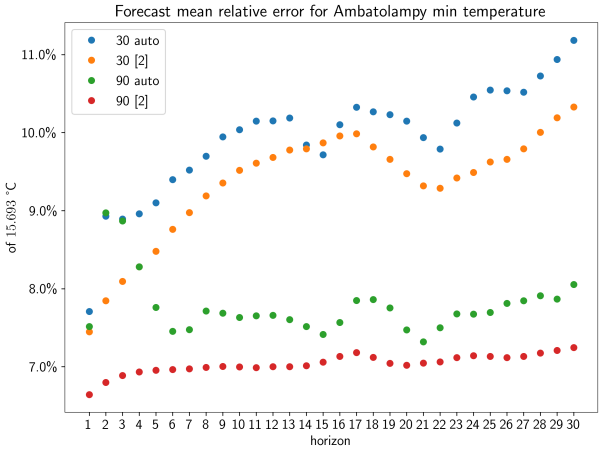


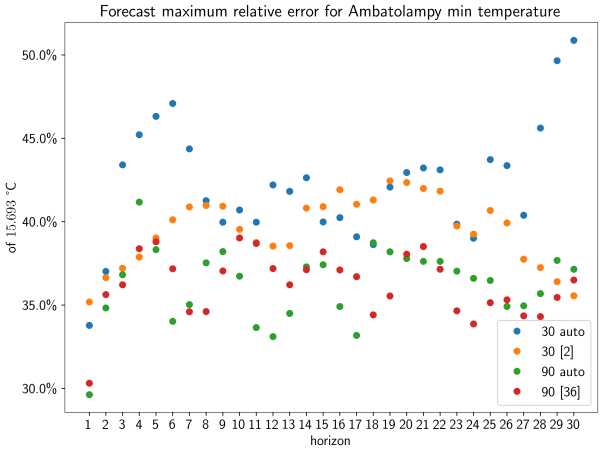
B COMPARATIVE PLOTS FOR TWO WINDOW LENGTHS

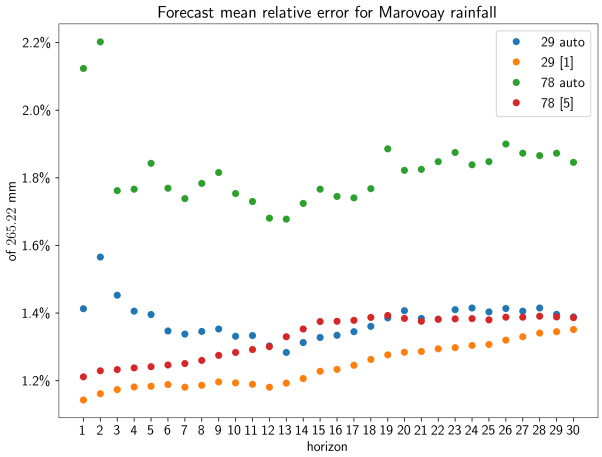
The following plots compare the accuracy of vector forecasting for window lengths $L^{[15]}$ and L_{lo}. Every label of a legend gives the window size followed by either "auto" for automated grouping using grouping.auto.wcor or the optimal prefix grouping $[M]$. For L_{lo} (resp. $L^{[15]}$) the accuracy is plotted in blue (resp. green) for automated grouping and in orange (resp. red) for optimal grouping. The reader may check Table 2 to avoid confusion between $L^{[15]}$ and L_{lo}. Note that the accuracy obtained with L_{hi} do not appear on the following plots as it is significantly worse than with L_{lo} and $L^{[15]}$.

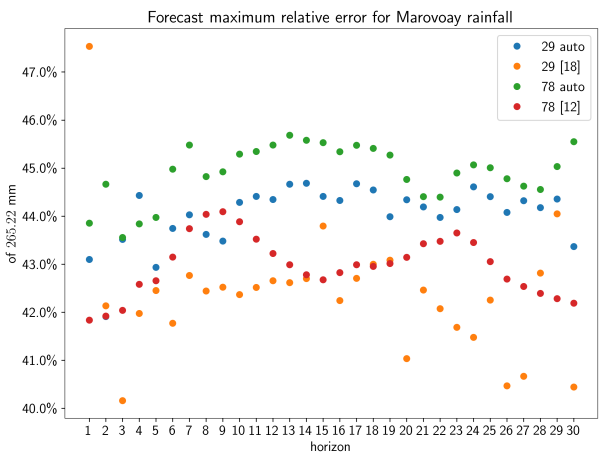


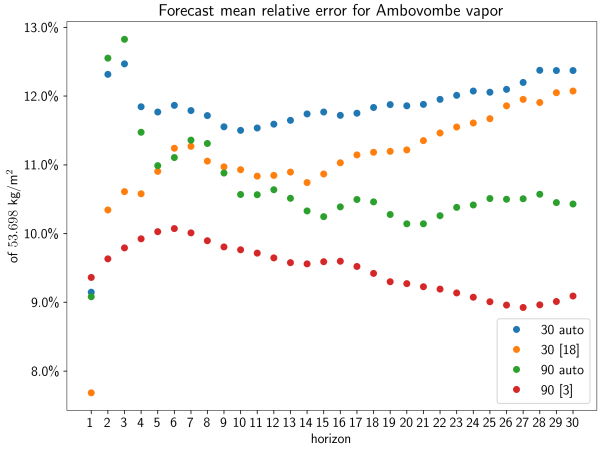


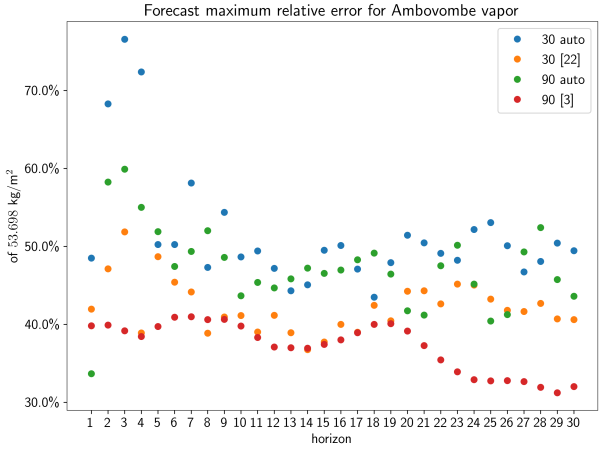


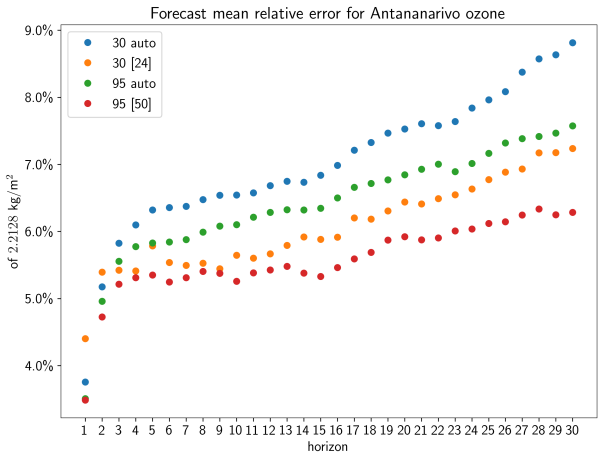


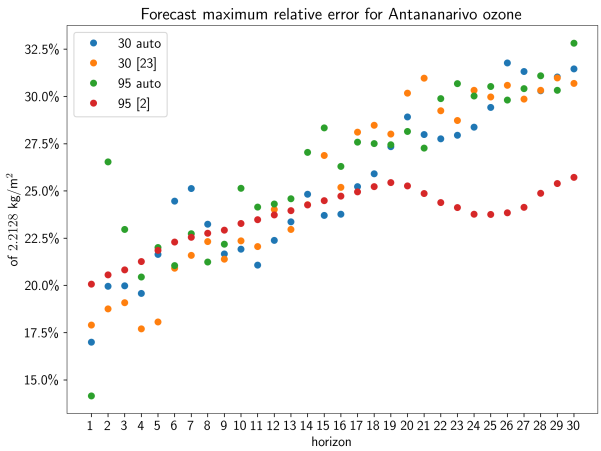

Forecast mean relative error for Marovoay rainfall
29 auto
29 [1]
78 auto
78 [5]
of 265.22 mm
horizon

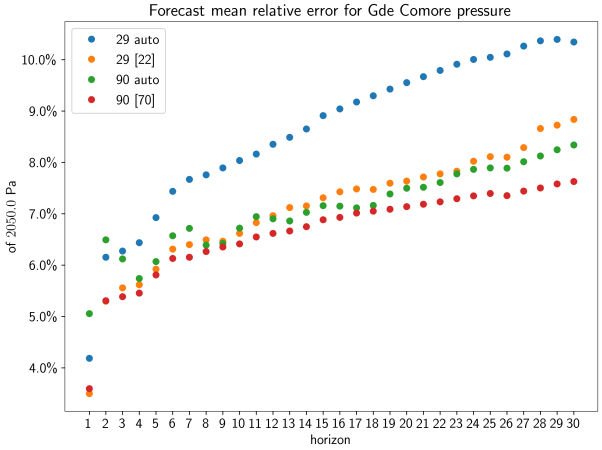

Forecast maximum relative error for Marovoay rainfall
29 auto
29 [18]
78 auto
78 [12]
of 265.22 mm
horizon

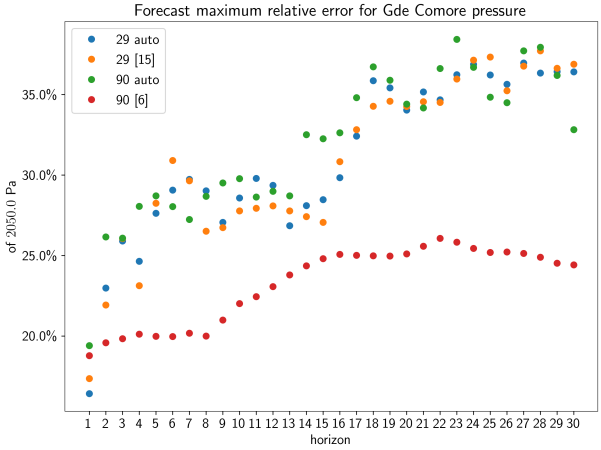

Forecast mean relative error for Ambovombe vapor
30 auto
30 [18]
90 auto
90 [3]
of 53.698 kg/m²
horizon


Forecast maximum relative error for Ambovombe vapor
30 auto
30 [22]
90 auto
90 [3]
of 53.698 kg/m²
horizon


Forecast mean relative error for Antananarivo ozone
30 auto
30 [24]
95 auto
95 [50]
of 2.2128 kg/m²
horizon


Forecast maximum relative error for Antananarivo ozone
30 auto
30 [23]
95 auto
95 [2]
of 2.2128 kg/m²
horizon


Forecast mean relative error for Gde Comore pressure
29 auto
29 [22]
90 auto
90 [70]
of 2050.0 Pa
horizon


Forecast maximum relative error for Gde Comore pressure
29 auto
29 [15]
90 auto
90 [6]
of 2050.0 Pa
horizon