Singular Spectrum Analysis with Conditional Predictions for Real-Time State Estimation and Forecasting

H. Reed Ogrosky[1], Samuel N. Stechmann[2,3], Nan Chen[2], and Andrew J. Majda[4,5]

[1]Department of Mathematics and Applied Mathematics, Virginia Commonwealth University, Richmond, VA, USA
[2]Department of Mathematics, University of Wisconsin-Madison, Madison, WI, USA
[3]Department of Atmospheric and Oceanic Sciences, University of Wisconsin-Madison, Madison, WI, USA
[4]Department of Mathematics and Center for Atmosphere Ocean Science, Courant Institute of Mathematical Sciences, New York University, New York, New York
[5]Center for Prototype Climate Modeling, NYU Abu Dhabi, Saadiyat Island, Abu Dhabi, United Arab Emirates

Key Points:

- Singular spectrum analysis (SSA) and extended empirical orthogonal function (EEOF) methods suffer from endpoint issues
- SSA with conditional predictions (SSA-CP) is presented as a simple modification to improve real-time estimates near endpoints.
- Forecasts are also possible, including error estimates, and are optimal for Gaussian data and shown to also be skillful for non-Gaussian data.

January 19, 2019

Corresponding author: H. Reed Ogrosky, hrogrosky@vcu.edu

Abstract

Singular spectrum analysis (SSA) or extended empirical orthogonal function (EEOF) methods are powerful, commonly-used data-driven techniques to identify modes of variability in time series and space-time datasets. Due to the time-lag embedding, these methods can provide inaccurate reconstructions of leading modes near the endpoints, which can hinder the use of these methods in real time. A modified version of the traditional SSA algorithm, referred to as SSA with conditional predictions (SSA-CP), is presented to address these issues. It is tested on low-dimensional, approximately Gaussian data, high-dimensional non-Gaussian data, and partially-observed data from a multiscale model. In each case SSA-CP provides a more accurate real-time estimate of the leading modes of variability than the traditional reconstruction. SSA-CP also predictions of the leading modes and is easy to implement. SSA-CP is optimal in the case of Gaussian data, and the uncertainty in real-time estimates of leading modes is easily quantified.

1 Introduction

Singular spectrum analysis (SSA) or extended empirical orthogonal function (EEOF) methods are powerful, commonly-used tools available for identifying modes of variability in time series and space-time datasets. SSA's usefulness has been demonstrated in a variety of fields over the last 3-4 decades, including, e.g., nonlinear dynamics (e.g., Broomhead and King, 1986), geoscience (e.g., Weare and Nasstrom, 1982; Vautard and Ghil, 1989; Keppenne and Ghil, 1990; Vautard *et al.*, 1992; Mo, 2001; Kikuchi and Wang, 2008; Roundy and Schreck, 2009), and economics (e.g., Lisi and Medio, 1997; Hassani *et al.*, 2014). Its popularity is due both to its ease of implementation and to its ability to eliminate noise and extract trends, oscillations, and other signals in both univariate and multivariate time series.

As with some other methods for mode identification in space-time data (e.g., Fourier filtering), SSA suffers from endpoint issues; i.e., estimates of leading modes can be inaccurate in real-time without future information. Therefore, SSA may provide inaccurate initial conditions for real-time forecasts. Despite these challenges, it is sometimes used either as a filtering step prior to generating real-time forecasts (e.g., Mo, 2001; Golyandina *et al.*, 2001; Hassani *et al.*, 2014), or in tests of forecast models (e.g., Kang and Kim, 2010; Kondrashov *et al.*, 2013; Chen and Majda, 2016), due to its effectiveness at mode identification.

This motivates the question: Is there a modified version of SSA that (i) is as straightforward to implement as SSA, but that (ii) provides the most accurate real-time state estimation possible of leading modes of variability?

This question, along with the related question of how to best modify SSA for use on datasets with gaps in the data, has motivated the proposal and study of numerous modified versions of SSA. These methods include schemes for modifying incomplete columns of the lag-embedded matrix by weighting known values (Schoellhamer, 2001), iterative SSA methods (Kondrashov and Ghil, 2006; Kondrashov *et al.*, 2010), methods based on linear recurrent formulae (Golyandina and Osipov, 2007), methods that project smoothed data onto leading SSA modes computed with Fourier filtered data (Roundy and Schreck, 2009), combined recurrent forecasting and hindcasting (Rodrigues and Carvalho, 2013), energy-minimizing reconstructions of principal components (Shen *et al.*, 2014; 2015), and a method utilizing a predicted spatial basis (Chen *et al.*, 2018). Some of these methods will be discussed in Section 5.

Here, we propose and study yet another modification of SSA. This method makes use of conditional mean predictions based on the covariance matrix of the lag-embedded data, and we refer to it as SSA with conditional predictions (SSA-CP). Another appropriate name would be real-time SSA (RT-SSA), though we use SSA-CP here to avoid

confusion with other methods proposed for using SSA in real-time, some of which are discussed further in Section 5.

The results of tests shown here suggest that this method is effective at addressing these endpoint issues in a variety of settings. The datasets used in these tests include both univariate datasets and multivariate datasets with small (2-3) or somewhat large (64) number of spatial dimensions; partially observed systems and datasets with all dynamical variables observed; Gaussian and non-Gaussian data; and synthetic time series and time series generated by observational data.

Given these results, there are at least four reasons for using this method. First, it is simple and easy to implement, requiring only small additional steps during the normal SSA algorithm. Second, it provides both state estimation and prediction of leading modes of variability. Third, it provides an optimal reconstruction if the data is Gaussian using the statistics of the first two moments. Fourth, it outperforms many other proposed methods of SSA state estimation for both Gaussian and non-Gaussian data.

The rest of the paper is organized as follows: Section 2 describes the traditional SSA method and the proposed modification. Section 3 lists datasets and models used in tests of this method. Results are presented in Section 4. Discussion of the methods and results is given in Section 5, including a brief comparison of the results with those of other modified SSA methods. Conclusions are given in Section 6.

2 SSA algorithms

A brief review of the traditional SSA algorithm is now given, followed by a description of the proposed modification. When used on multivariate time series, SSA is often referred to as Multichannel SSA (MSSA) in the literature; here SSA will be used to refer to either the univariate or multivariate cases. The theory of SSA, which has been developed over the last several decades, is not discussed here; see, e.g., Aubry *et al.* (1991); Ghil *et. al.* (2002); Golyandina *et al.* (2001); Hassani (2007) for discussion of this underlying theory.

2.1 Traditional SSA

We briefly describe the traditional SSA algorithm for a dataset with spatial dimension D; the traditional univariate SSA algorithm can be reproduced by setting $D = 1$ below.

Let $\vec{x}_i$ be a D-dimensional column vector at time i, with $1 \leq i \leq N$. The four steps of SSA are as follows:

Step 1: Create the time-lagged embedding matrix $\mathbf{X}$ of size $(MD)\times(N-M+1)$:

$$\mathbf{X} = \begin{bmatrix} \vec{x}_1 & \vec{x}_2 & \dots & \vec{x}_{N-M+1} \\ \vec{x}_2 & \vec{x}_3 & \dots & \vec{x}_{N-M+2} \\ \vdots & \vdots & & \vdots \\ \vec{x}_{M-1} & \vec{x}_M & \dots & \vec{x}_{N-1} \\ \vec{x}_M & \vec{x}_{M+1} & \dots & \vec{x}_N \end{bmatrix} \tag{1}$$

where M is the length of the embedding window.

Step 2: Find eigenvalues and eigenvectors of the covariance matrix $\mathbf{C} = \mathbf{X}\mathbf{X}^T/(N-M+1)$. Each eigenvector $\vec{v}$ (sometimes referred to as an empirical orthogonal function, or EOF) is an (MD)-dimensional column vector with corresponding eigenvalue λ:

$$\vec{v} = [\vec{v}_1^T, ..., \vec{v}_M^T]^T, \tag{2}$$

where $\vec{v}_s$ is a D-dimensional column vector used to denote the lag-s portion of the eigenvector.

Step 3: Find the principal component (PC) of each mode by projecting the lag-embedded data onto the appropriate eigenvector:

$$\vec{\phi} = \mathbf{X}^T \vec{v}. \tag{3}$$

The entries of each principal component will be denoted $\vec{\phi} = [\phi_1, ..., \phi_{N-M+1}]^T$.

Step 4: Reconstruct the data corresponding to each mode by calculating the reconstructed component (RC) $\vec{z}(t)$:

$$\vec{z}(t) = \frac{1}{M_t} \sum_{i=L_t}^{U_t} \phi_{t-i+1} \vec{v}_i \tag{4}$$

where (M_t, L_t, U_t) are defined by (see, e.g., Ghil et al., 2002)

$$(M_t, L_t, U_t,) = \begin{cases} (t, 1, t)\,, & 1 \le t \le M-1 \\ (M, 1, M)\,, & M \le t \le N-M+1 \\ (N-t+1, t-N+M, M)\,, & N-M+2 \le t \le N \end{cases} \tag{5}$$

so that each reconstructed component $\vec{z}$ is a (possibly multivariate) time series of length N, with each $\vec{z}(t)$ a D-dimensional column vector.

Each reconstructed component entry at time t^* depends directly on one embedding window of principal component entries, and each principal component entry depends on one embedding window of data. As a result, each reconstructed component entry at time t^* is influenced primarily by the values of $\vec{x}_{t^*-M+1}$ through $\vec{x}_{t^*+M-1}$; i.e., two embedding windows worth of data, spanning the window immediately prior to t^* and the window immediately following t^*, contribute directly to the reconstruction at t^*. For $t^* > N-M$, the embedding window's worth of data immediately following t^* is not entirely known. The reconstruction process makes use of the known data by averaging over the available products $\phi_{t-i+1}\vec{v}_i$ in (4), but these final $M-1$ entries of each reconstruction are only estimates of the state of each mode, and can be expected to change as data becomes available at times occurring after the end of the time series. (The same endpoint issues affect the reconstruction for $t^* < M$.)

The reconstruction method in (4) has been shown to be an optimal method, in the sense that, for $D = 1$, e.g., it produces the Hankel matrix that is closest to the matrix $\vec{\phi}\vec{v}^T$ in matrix norm (Golyandina *et al.*, 2001; Hassani, 2007). However, other reconstruction formulas may be considered, including ones that avoid the endpoint issues of the traditional reconstruction in (4). One such method is the 'predicted spatial basis' method of Chen *et al.* (2018), in which a method that shifts future information to the spatial basis (and not the principal components) is tested on a monsoon intraseasonal oscillation index. The method in (4) is used as the primary basis for comparison here due to its optimality with respect to Hankelization and its somewhat standard use.

2.2 SSA with conditional predictions (SSA-CP)

The primary goal of this section is to present a simple method, SSA with conditional predictions (SSA-CP), that improves the estimates of the final $M-1$ entries of each reconstructed component, including in particular the current state estimate. In addition, SSA-CP will provide a prediction of reconstructed components for $t > N$. (The same procedure may be directly applied to the first $M-1$ entries of each reconstruction, but for simplicity of presentation, we focus solely on the last $M-1$ entries.)

The steps of SSA-CP are as follows:

Step 1: Perform steps 1 and 2 of traditional SSA.

Step 2: Construct an extended lag-embedded matrix $\tilde{\mathbf{X}}$ of size $(MD) \times N$. The first N columns of $\tilde{\mathbf{X}}$ are identical to the columns of $\mathbf{X}$. For the final $M-1$ columns, those entries which are known from the time series are filled in. The unknown entries below the diagonal consisting of x_N's are estimated using their conditional mean prediction,

$$\tilde{\mathbf{X}} = \begin{bmatrix} \vec{x}_1 & \dots & \vec{x}_{N-M+1} & \vec{x}_{N-M+2} & \dots & \vec{x}_{N-1} & \vec{x}_N \\ \vec{x}_2 & \dots & \vec{x}_{N-M+2} & \vec{x}_{N-M+3} & \dots & \vec{x}_N & \vec{\mu}_{N+1|N} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots \\ \vec{x}_{M-1} & \dots & \vec{x}_{N-1} & \vec{x}_N & \dots & \vec{\mu}_{N+M-3|N-1,N} & \vec{\mu}_{N+M-2|N} \\ \vec{x}_M & \dots & \vec{x}_N & \vec{\mu}_{N+1|N-M+2,\dots,N} & \dots & \vec{\mu}_{N+M-2|N-1,N} & \vec{\mu}_{N+M-1|N} \end{bmatrix}. \tag{6}$$

The calculation of each $\vec{\mu}_{i|N-l,\dots,N}$ in (6) is as follows.

Let $\vec{y}$ refer to the k-th column of $\tilde{\mathbf{X}}$, with $N+1 \leq k \leq N+M-1$, and let $\vec{y}_1$, $\vec{y}_2$ refer to the known and unknown portions of $\vec{y} = [\vec{y}_1^T, \vec{y}_2^T]^T$, respectively. If $\vec{y}$ is a Gaussian random variable with mean $\vec{\mu} = 0$ and covariance matrix $\mathbf{C}$, then $\vec{y}_2$ has a conditional distribution that is Gaussian with mean

$$\vec{\mu}_{2|1} = \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\vec{y}_1. \tag{7}$$

where $\mathbf{C}$ can be written as

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} \tag{8}$$

with $\mathbf{C}_{11}$ describing the covariance of the known values with themselves, etc. (Kaipio and Somersalo, 2005). The unknown entries $\vec{y}_2$ are then filled in with the appropriate entries of $\vec{\mu}_{2|1}$, where $\vec{\mu}_{N+j|k-M+1,\dots,N}$ in (6) denotes a D-dimensional column vector, i.e. the j-th set of D entries of the vector $\vec{\mu}_{2|1}$, calculated for the k-th column of $\tilde{\mathbf{X}}$ (with $N+1 \leq k \leq N+M-1$). If necessary, a small amount of noise may be added to the covariance matrix in order to evaluate $\mathbf{C}_{11}^{-1}$ in (7).

Step 3: Modify step 3 of traditional SSA by replacing $\mathbf{X}$ with $\tilde{\mathbf{X}}$; this change results in extended principal components $\tilde{\vec{\phi}} = \tilde{\mathbf{X}}^T\vec{v}$; each extended principal component is a column vector of length N.

Step 4: Modify step 4 of traditional SSA by replacing ϕ with $\tilde{\phi}$ to construct an extended RC:

$$\tilde{\vec{z}}(t) = \frac{1}{\tilde{M}_t}\sum_{i=\tilde{L}_t}^{\tilde{U}_t}\tilde{\phi}_{t-i+1}\vec{v}_i \tag{9}$$

where $(\tilde{M}_t, \tilde{L}_t, \tilde{U}_t)$ are defined by

$$(\tilde{M}_t, \tilde{L}_t, \tilde{U}_t,) = \begin{cases} \left(\frac{1}{t}, 1, t\right), & 1 \leq t \leq M-1 \\ \left(\frac{1}{M}, 1, M\right), & M \leq t \leq N \\ \left(\frac{1}{N-t+M}, t-N+1, M\right), & N+1 \leq t \leq N+M-1 \end{cases} \tag{10}$$

so that each extended reconstructed component $\tilde{\vec{z}}$ is a (possibly multivariate) time series of length $N+M-1$, with the last $M-1$ entries corresponding to predictions of the future state of the mode.

In the case that the dataset has a Gaussian distribution, the conditional mean provides an optimal estimate of the missing data (Kaipio and Somersalo, 2005).

177 # 3 Data and Methods

178 The SSA-CP method will be tested on several datasets and compared to the tra-
179 ditional SSA reconstruction.

180 ## 3.1 Data

181 The first test uses a fifteen year portion of the daily Real-time Multivariate MJO
182 (RMM) indices (Wheeler and Hendon, 2004) from 1 January 1999 through 31 Decem-
183 ber 2013. The RMM indices have a distribution that is approximately normal with mean
184 and variance approximately 0 and 1, respectively (Chen and Majda, 2015). For this 2-
185 dimensional dataset, $D = 2$ and $N_{tot} = 5479$, with N_{tot} referring to the number of days.

186 GPCP daily precipitation data (Huffman *et. al.*, 2012) are used for the second test.
187 This dataset has a spatial resolution of $1^\circ \times 1^\circ$; the portion from 1 January 1997 through
188 31 December 2013 is used. Prior to applying SSA, the following steps were taken: (i) a
189 meridional mode truncation to move from 2D(x, y) to 1D(x), (ii) removal of annual mean
190 and seasonal cycle, and (iii) interpolation to 64 equally-spaced zonal gridpoints. The merid-
191 ional mode truncation step is a projection of the data onto the leading meridional mode
192 proportional to $e^{-y^2/2}$ where y is proportional to latitude; this step is identical to that
193 used in, e.g., Stechmann and Majda (2015); Stechmann and Ogrosky (2014). Steps (i)
194 and (iii) reduce the number of dimensions to $D = 64$, and the number of times is $N_{tot} =$
195 6209. Note that these anomalies have a non-Gaussian distribution at each longitude; see
196 the SI for the statistics of these anomalies.

197 A simulation of a multiscale model (Majda and Harlim, 2012) is used for the third
198 test. The model equations are

$$du_1 = (-\gamma_1 u_1 + F(t))\, dt + \sigma_1 dW_1, \tag{11a}$$

$$du_2 = (-\gamma_2 + i\omega_0/\epsilon + ia_0 u_1)\, u_2 dt + \sigma_2 dW_2, \tag{11b}$$

199 where $\gamma_1 = \gamma_2 = 0.2$, $\sigma_1 = \sigma_2 = 0.5$, $\omega_0 = a_0 = 1$, $\epsilon = 0.5$, and $F(t) = \sin(t/5)$. An
200 approximate solution was calculated numerically with the Euler-Maruyama method us-
201 ing $dt = 0.005$ and $t_{end} = 2000$. The real part of u_2 was then sampled every 0.5 time
202 units to create a dataset with $D = 1$ and $N_{tot} = 4000$. A portion of this signal can
203 be seen in Figure S2 in the SI.

204 ## 3.2 Methods

205 The results of each real-time reconstruction method (SSA-CP and traditional) will
206 be compared with the traditional reconstruction that has knowledge of future data. This
207 is done in two steps.

208 First, both the traditional SSA and SSA-CP methods were applied to each dataset
209 after removing the final $2M-2$ time entries from the dataset; e.g., using an embedding
210 window of $M = 51$ days for the RMM indices, the methods were applied to the first
211 $N = N_{tot} - 2M + 2 = 5379$ days. The embedding window was chosen to be large
212 enough to be consistent with the intraseasonal timescale of the indices and is similar to
213 that used in Chen and Majda (2015); other choices of this parameter value will be dis-
214 cussed in Section 5. The standard reconstruction $z(t)$ for each mode therefore has $N =$
215 5379 entries, while the SSA-CP reconstruction $\tilde{z}(t)$ has $N + M - 1 = 5429$ entries.
216 Note that the first $N-M+1 = 5329$ entries for each reconstruction method are iden-
217 tical to one another; i.e. $z(t) = \tilde{z}(t)$ for $1 \le t \le N - M + 1$. Next, the traditional re-
218 construction method was used again, this time on the full $N_{tot} = 5479$ entries, result-
219 ing in a reconstruction $u(t)$ with $N_{tot} = 5479$ entries. The entries of $u(t)$ up to $N_{tot}-$
220 $M+1 = 5429$ are taken to be 'truth', and each of the methods applied to the shorter
221 time series are compared with this truth.

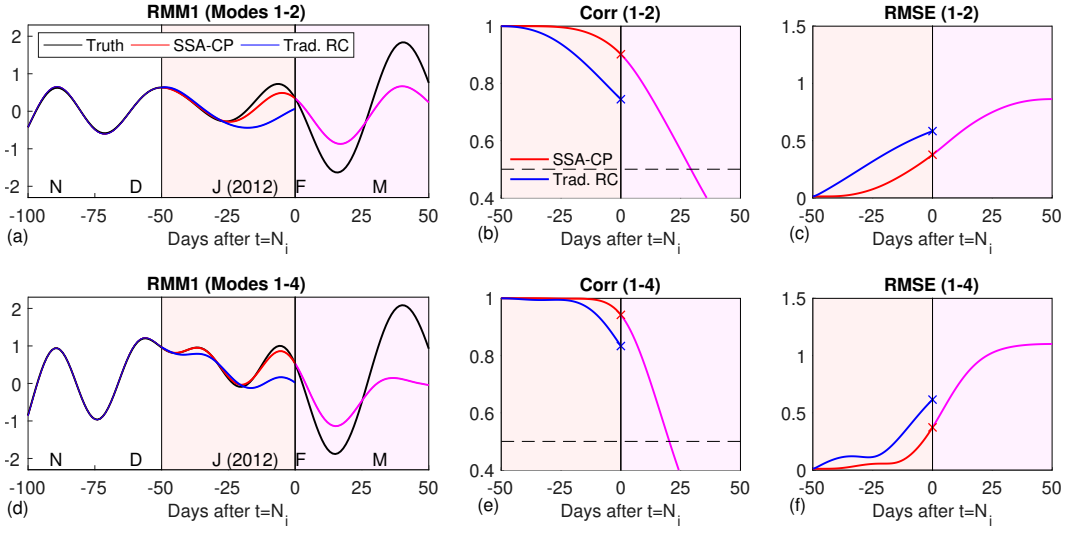

Figure 1. (a) Reconstructed RMM1 using components 1-2 with $t = N_{601} = 4779$ (31-Jan-2012) using (blue) traditional reconstruction, (red/magenta) SSA-CP, and (black) reconstruction using future information. (b,c) Bivariate pattern correlation and RMSE of the (blue) traditional reconstruction and truth as a function of days prior to/after N_i, and (red/magenta) SSA-CP reconstruction and truth using modes 1-2. (d-f) Same as (a-c) but using components 1-4.

Second, these tests are repeated for each dataset with decreasing N_{tot}; i.e., define $N_{tot,i} = N_{tot} - i + 1$, and repeat the test described above but using only the first $N_{tot} = N_{tot,i}$ entries of the dataset, so that $N = N_i := N_{tot,i} - 2M + 2$. For the RMM indices and multiscale model, $i \in I = [1, ..., 1001]$; for the GPCP data, $i \in I = [1, 6, 11, ..., 1001]$. The pattern correlation and root mean square error (RMSE) are then calculated as a function of days before or after N_i; see the SI for details.

4 Results

We next show results for three tests.

4.1 RMM index

How well does the method perform on low-dimensional data that is nearly Gaussian?

Figure 1(a,d) shows the results of using the SSA-CP or traditional reconstruction methods on the RMM indices with an embedding window $M = 51$ days. For times away from the endpoints of the data i.e. $t < N_i - M + 1$, both methods are in agreement with the truth. For past times near the endpoints, i.e. $N_i - M + 1 < t < N_i$ (light orange shaded region), SSA-CP captures both the phase and amplitude of the RMM1 index better than the traditional reconstruction. For future times $t > N_i$, SSA-CP is able to make predictions, with good agreement in phase and an underestimate of the amplitude of the true reconstruction. This underestimate of amplitude is due to using conditional mean predictions which tend to zero as $t \to \infty$.

The example in Figure 1 is a particularly challenging test as it is a case of MJO onset. More specifically, the period being predicted in Figure 1(a,d), namely 01-Feb-2012 through 21-Mar-2012, exhibits a growing amplitude of the RMM1 index (black line), corresponding to the onset of the MJO event sometimes referred to as MJO4 during the 2011-

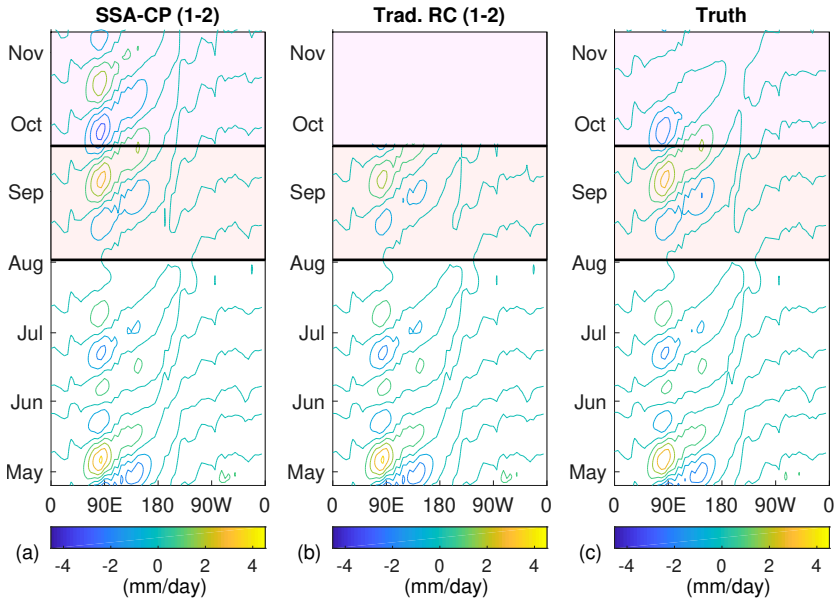

Figure 2. (a) Reconstructed precipitation during 2013 using SSA-CP modes 1-2 with $t_N = 6109$, corresponding to 22-Sep-2013. (b) Same as (a) but using traditional reconstruction. (c) Reconstructed modes 1-2 using future information.

2012 CINDY/DYNAMO field campaign (Yoneyama *et al.*, 2013). This MJO has been considered a 'primary' event, in that there are no clear signals connecting this MJO to the previous MJOs that occurred in October through December 2012. In contrast to the traditional reconstruction, SSA-CP more naturally captures oscillations, and changes in frequency and amplitude, near the endpoint.

Figure 1(b,c,e,f) shows that when these tests are repeated, SSA-CP has significantly improved pattern correlation and reduced error compared to the traditional reconstruction. As a current state estimation, at $t = N_i$ SSA-CP improves the pattern correlation from 0.74 to 0.90 (0.83 to 0.94) for the 2 (4) leading modes. Likewise, SSA-CP reduces the error at $t = N_i$ from 0.58 to 0.38 (0.62 to 0.37). For future times $t > N_i$, SSA-CP is able to make meaningful predictions for an extended period of time, with pattern correlations exceeding 0.5 out to approximately 29 (20) days when 2 (4) leading modes are used.

4.2 Precipitation data

How well does the method perform on large-dimensional, possibly non-Gaussian data?

Figure 2 shows reconstructed precipitation anomalies using the 2 leading modes with an embedding window of 51 days. Both methods produce identical reconstructions prior to Aug. 2, 2013. For 3-Aug-2013 through 22-Sep-2013, SSA-CP produces a reconstruction with amplitude in much better agreement with the non-real-time reconstruction (truth) than the traditional reconstruction. It also provides a prediction with decaying amplitude throughout October, qualitatively similar to the truth but with slower decay.

Repeating these tests for various N_i produces the pattern correlation and RMSE shown in Figure 3. For the recent past in time interval $N_i - M + 1 < t < N_i$, SSA-CP

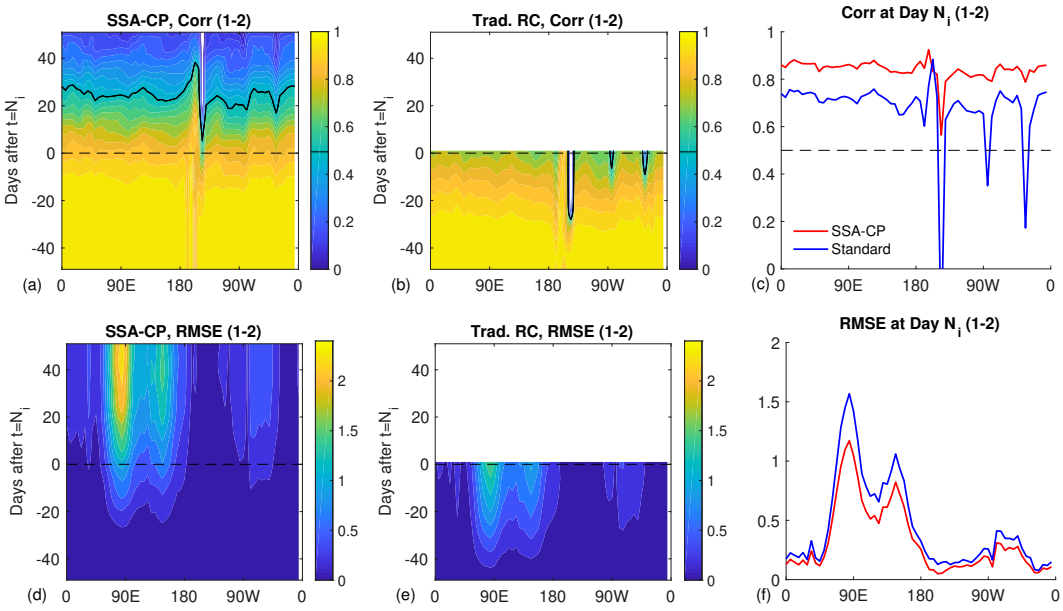

Figure 3. (a) Pattern correlation, using 200 runs of SSA, of reconstructed precipitation components (1-2) using SSA-CP. (b) Same as (a) but for traditional reconstruction. (c) Pattern correlation at Day $t = N_i$ for each method. (d-f) Same as (a-c) but showing RMSE.

produces higher pattern correlation and lower RMSE than the standard reconstruction method. For state estimation at $t = N_i$, the pattern correlation is 0.1-0.2 higher at almost all longitudes when using SSA-CP than when using the standard method. Likewise, the RMSE is lower using SSA-CP than the traditional reconstruction at all longitudes. Note that low pattern correlation values for each method at longitudes like 150W are due to small anomalies in the leading modes.

4.3 Partially-observed multiscale model

How well does the method perform on partially-observed data?

Figure S3 in the SI shows the pattern correlation and RMSE for both methods applied to the multiscale model (11). For $N_i - M + 1 < t < N_i$, SSA-CP has significantly higher pattern correlation and lower error than the traditional reconstruction. At $t = N_i$, using SSA-CP improves the pattern correlation from 0.54 to 0.75 for 2 leading modes, and lowers the error from 0.12 to 0.06. For $t > N_i$, predictions using SSA-CP have a pattern correlations of 0.5 or higher out to approximately 23 days when 2 leading modes are used.

5 Discussion

SSA-CP has been proposed as a method that supplements the mode-identification ability of SSA with improved estimates of mode reconstructions near the ends of time series. We note that it is not at all necessarily the best possible data-driven, model-free prediction method that could be designed. Its effectiveness at identifying modes of variability in real-time is of course also limited to cases where SSA is effective at identifying modes of interest.

How sensitive are the results to changes in the embedding window? As a first step towards addressing this question, the RMM tests from the previous section were rerun

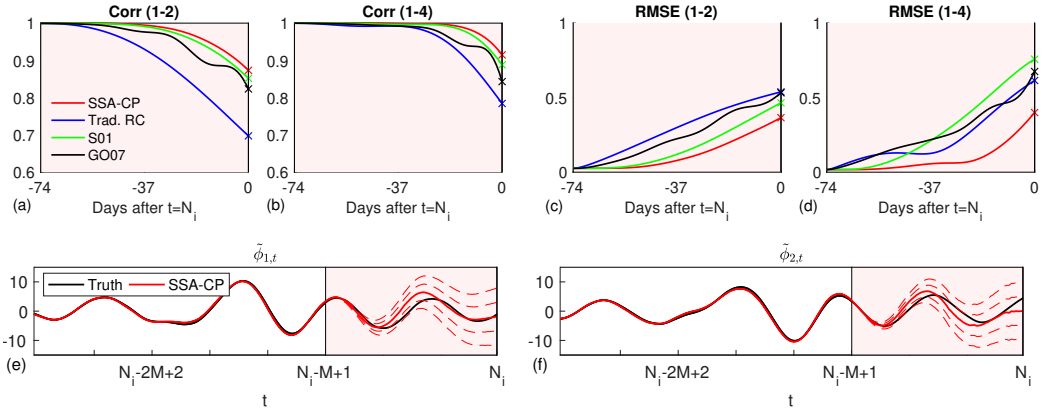

Figure 4. (a-b) Bivariate pattern correlation and RMSE of reconstructed RMM indices using the (blue) traditional reconstruction, (red) SSA-CP reconstruction, (green) weighted reconstruction of Schoellhamer (2001), and (black) Π-projector/simultaneous filling in method of Golyandina and Osipov (2007) as a function of days prior to/after N_i, using modes 1-2; here M = 75, N_i = 4779. (c-d) Same as (a-b) but for modes 1-4. (e-f) Leading two principal components of SSA-CP; dashed red lines indicate $\pm 1, 2$ standard deviations.

with an embedding window of $M = 75$ days. Figure 4(a-d) shows that while both SSA-CP and the traditional reconstruction produce slightly lower pattern correlation at $t = N_i$ than in the previous test with $M = 51$, SSA-CP again results in significantly higher PC and lower RMSE than the traditional reconstruction. For $t > N_i$, the pattern correlation stays higher than 0.5 for 35 (24) days when the leading 2 (4) modes are used (not shown). Results of additional tests using M = 61, 71, 81, and 101, are shown in Figure S4 in the SI; increasing the embedding window provides some small improvement in the pattern correlation and RMSE of predictions, which is likely due to the increased timescales present in the modes identified with a larger embedding window. Further tests indicate that, in addition, the method performs equally well when the leading modes are found using a period of training data that is not near the endpoint; see Figure S5 in the SI for further details.

How does SSA-CP compare with other methods in the literature that have been proposed for either (i) improving state estimation of reconstructed components near the endpoints of time series, or (ii) using SSA on datasets with gaps? We briefly examine this through a comparison of the results of SSA-CP with methods from Schoellhamer (2001) and Golyandina and Osipov (2007) for the first test from Section 4. Figure 4(a-d) shows the pattern correlation and RMSE of these two methods along with the traditional reconstruction and SSA-CP. All of the modified versions of SSA produce higher pattern correlation than the traditional reconstruction, with SSA-CP having the highest. For the leading two modes, all methods produce lower RMSE than the traditional reconstruction, but when the leading four modes are used, only SSA-CP outperforms the traditional reconstruction over each of the final $M - 1$ days. Other methods for using SSA (or other mode-identification methods) in real-time have been proposed, and it would be interesting to investigate further comparisons in a future study. For example, other methods include a predicted-spatial-basis method (Chen *et al.*, 2018), kernel analog forecasting (e.g., Comeau *et al.*, 2017), methods based on linear recurrent formulae (Golyandina *et al.*, 2001), methods that project smoothed data onto leading SSA modes computed with Fourier filtered data (Roundy and Schreck, 2009), and energy-minimizing reconstructions of principal components (Shen *et al.*, 2015).

Many additional tests were conducted beyond the three examples described in detail above. Other tests were conducted using datasets generated by stochastic processes (complex-valued Ornstein-Uhlenbeck process), deterministic dynamical systems (Lorenz 63 model, multiple examples from Golyandina *et al.* (2001)), other observational data (Kelvin wave calculated using NCEP/NCAR reanalysis data (Kalnay *et al.*, 1996) and the methods of Ogrosky and Stechmann, 2015; 2016), and numerous synthetic test signals both with and without noise. SSA-CP significantly outperformed traditional SSA in almost all of these tests. In cases of deterministic signals of Golyandina *et al.* (2001), both methods produced excellent reconstructions of the leading modes near the endpoints. In cases like this, the standard reconstruction may be just as desirable as SSA-CP or any other modification, as the additional effort of implementing SSA-CP, though minimal, may not be necessary to provide reasonable initial conditions for a forecast. In addition, one benefit of the standard reconstruction is its invertibility; if all modes are reconstructed and summed together, the original dataset is recovered. This invertibility is not shared by SSA-CP.

There are several compelling reasons for using SSA-CP rather than the traditional reconstruction, however. First, it is nearly as simple to use as traditional SSA. Second, it is optimal for Gaussian data and is based on well-known theory. Third, it is straightforward to quantify the uncertainty in the extended principal components or reconstruction. For example, the variance of $\tilde{\phi}_{N-l+1}$, where $1 \leq l \leq M-1$, is given by

$$\mathrm{Var}(\tilde{\phi}_{N-l+1}) = \left[\vec{v}_{l+1}^T, ..., \vec{v}_M^T\right] \mathbf{C}_{21} \left[\vec{v}_{l+1}^T, ..., \vec{v}_M^T\right]^T \tag{12}$$

Figure 4(e,f) shows the two leading principal components of the RMM indices calculated using SSA-CP with $N_i = 4779$ and $M = 75$. One and two standard deviations from the extended principal component entries are shown, with the standard deviation calculated using (12).

Finally, since non-Gaussianity leads to a lack of independence between modes in linear methods like empirical orthogonal functions (EOFs), there is no guarantee that the method will work well on data with strong non-Gaussianity (Monehan *et al.*, 2009). However, the method works well on the non-Gaussian data used here, perhaps owing to the somewhat mild deviations from Gaussianity. The method could potentially be extended to non-Gaussian frameworks with conditional Gaussian or Gaussian mixture structures (see, e.g. Chen and Majda (2018); Majda (2016)).

We note that SSA is just one of many data analysis tools capable of identifying modes of variability in spatiotemporal datasets (see Crommelin and Majda, 2004, for a discussion of other linear methods for mode identification). SSA was chosen to be the focus of the current study due to its linearity, simplicity, and popularity, combined with the linearity of the proposed modifications. Other mode identification methods, including nonlinear methods like Nonlinear Laplacian Spectral Analysis (NLSA), have been shown to be effective at capturing modes of variability that SSA has difficulty capturing, like modes with pronounced intermittent behavior (Giannakis and Majda, 2012a,b), and theory supporting both such methods and forecasting techniques of relevance has been developed in recent years (Comeau *et al.*, 2017; Zhao and Giannakis, 2016). Including conditional predictions into such methods is certainly possible; methods like NLSA use a reconstruction approach similar to that of (4), and incorporating conditional predictions into this method could potentially be done in a straightforward manner. It is not clear, however, that such a method would be optimal in the same way that conditional predictions used here are optimal when combined with Gaussian data and with a linear method like SSA.

6 Conclusions

In summary, a modified SSA algorithm, SSA-CP, has been presented and tested. This modification is proposed to address endpoint issues that arise when using SSA. When compared with the traditional reconstruction method, SSA-CP results in significantly improved real-time estimates of leading modes of variability when applied to a variety of datasets.

This method was shown to be useful for providing improved initial conditions for forecasts. It is derived from well-known theory using Gaussian statistics, and provides optimal predictions for Gaussian data, but also performs well in tests with non-Gaussian data. The uncertainty in the real-time estimates may be quantified using the covariance matrix that is inherently part of the method.

While the current study has been primarily focused on applying the method to atmospheric science data, this method may prove useful in application areas outside of atmospheric science. In addition, it is possible that the ideas used here may be adapted for other methods of mode identification. These subjects are left for future work.

Acknowledgments

The GPCP data for this article are available from NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their web site at http://www.esrl.noaa.gov/psd/. The RMM indices can be obtained online at http://www.bom.gov.au/climate/mjo/. Other data used are in the figures.

The research of S.N.S. is partially supported by a Sloan Research Fellowship from the Alfred P. Sloan Foundation and a Vilas Associates Award from the University of Wisconsin-Madison. The research of N.C. is supported by the Office of Vice Chancellor for Research and Graduate Education (VCRGE) at University of Wisconsin-Madison.

References

Aubry, N., Guyonnet, R., & Lima, R. (1991). Spatiotemporal analysis of complex signals: theory and applications. *Journal of Statistical Physics*, *64*, 683–739.

Broomhead, D. S., & King, G. P. (1986). Extracting qualitative dynamics from experimental data. *Physica D*, *20*, 217–236.

Chen, N., & Majda, A. J. (2016). Filtering the stochastic skeleton model for the Madden-Julian oscillation. *Monthly Weather Review*, *144*, 501–527.

Chen, N., & Majda, A. J. (2015). Predicting the real-time multivariate Madden-Julian oscillation index through a low-order nonlinear stochastic model. *Monthly Weather Review*, *143*, 2148–2169.

Chen, N., Majda, A. J., Sabeerali, C. T., & Ravindran, A. J. (2018). Predicting Monsoon Intraseasonal Precipitation using a Low-Order Stochastic Model. *Journal of Climate*, *31*, 4403–4427.

Chen, N., & Majda, A. J. (2018). Conditional Gaussian Systems for Multiscale Nonlinear Stochastic Systems: Prediction, State Estimation and Uncertainty Quantification. *Entropy*, *20*, 509.

Comeau, D., Giannakis, D., Zhao, Z., & Majda, A. J. (2018). Predicting regional and pan-Arctic sea ice anomalies with kernel analog forecasting. *Climate Dynamics*. https://doi.org/10.1007/s00382-018-4459-x.

Crommelin, D. T., & Majda, A. J. (2004). Strategies for model reduction: Comparing different optimal bases. *Journal of the Atmospheric Sciences*, *61*, 2206–2217.

Ghil, M., Allen, R. M., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., Robertson, A., Saunders, A., Tian, Y., Varadi, F., & Yiou, P. (2002). Advanced spectral methods for climatic time series, *Reviews of Geophysics*, *40*,

pp. 3.1–3.41.

Giannakis, D. & Majda, A. J. (2012). Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 2222–2227.

Giannakis, D., & Majda, A. J. (2012). Comparing low-frequency and intermittent variability in comprehensive climate models through nonlinear Laplacian spectral analysis. *Geophysical Research Letters*, *39*, L10710.

Golyandina, N., Nekrutkin, V., Zhigljavsky, A. A. (2001). *Analysis of Time Series Structure: SSA and Related Techniques*, Chapman and Hall, CRC.

Golyandina, N., & Osipov, E. (2007). The "Caterpillar"-SSA method for analysis of time series with missing values. *Journal of Statistical Planning and Inference*, *137*, 2642–2653.

Hassani, H. (2007). Singular Spectrum Analysis: Methodology and Comparison. *Journal of Data Science*, *5*, 239–257.

Hassani, H., Webster, A., Silva, E.S., & Heravi, S. (2014). Forecasting U.S. tourist arrivals using optimal singular spectrum analysis. *Tourism Management*, *46*, 322–335.

Huffman, G.J., Bolvin, D.T. & Adler, R.F. (2012). GPCP Version 2.2 SG Combined Precipitation Data Set. WDC-A, NCDC, Asheville, NC. Data set accessed 12 February 2014 at http://www.ncdc.noaa.gov/oa/wmo/wdcamet-ncdc.html.

Kaipio, J., & Somersalo, E. (2005). *Statistical and Computational Inverse Problems*, Springer, New York.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R. and Joseph, D., 1996: The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, *77*, 437–471.

Kang, I.-S., & Kim, H.-M. (2010). Assessment of MJO predictability for boreal winter with various statistical and dynamical models. *Journal of Climate 23*, 2368–2378.

Keppenne, C. L., & Ghil, M. (1990). Adaptive filtering and prediction of the Southern Oscillation Index. *Journal of Geophysical Research: Atmospheres, 97*, 20,449–20,454.

Kikuchi, K., & Wang, B. (2008). Diurnal Precipitation Regimes in the Global Tropics. *Journal of Climate*, *21*, 2680–2696.

Kondrashov, D., & Ghil, M. (2006). Spatio-temporal filling of missing points in geophysical data sets. *Nonlinear Processes in Geophysics*, *13*, 151–159.

Kondrashov, D., Shprits, Y., & Ghil, M. (2010). Gap filling of solar wind data by singular spectrum analysis. *Geophysical Research Letters*, *37*, L15101.

Kondrashov, D., Chekroun, M. D., Robertson, A. W., & Ghil, M. (2013). Low-order stochastic model and "past-noise forecasting" of the Madden-Julian Oscillation. *Geophysical Research Letters*, *40*, 5305–5310.

Lisi, F., & Medio, A. (1997). Is a random walk the best exchange rate predictor? *International Journal of Forecasting*, *13*, 255–267.

Majda, A. J. (2016). *Introduction to Turbulent Dynamical Systems in Complex Systems*, Springer.

Majda, A. J., & Harlim, J. (2012). *Filtering Complex Turbulent Systems*, Cambridge University Press.

Mo, K. C. (2001): Adaptive filtering and prediction of intraseasonal oscillations. *Monthly Weather Review*, *129*, 802–817.

Monehan, A. H., Frye, J. C., Ambaum, M. H. P., Stephenson, D. B., & North, G. R. (2009). Empirical Orthogonal Functions: The Medium is the Message. *Journal of Climate*, *22*, 6501–6514.

Ogrosky, H. R., & Stechmann, S. N. (2015). Assessing the equatorial long-wave approximation: asymptotics and observational data analysis. *Journal of the*

Atmospheric Sciences, *72*, 4821–4843.

Ogrosky, H. R., & Stechmann, S. N. (2016). Identifying convectively coupled equatorial waves using theoretical wave eigenvectors. *Monthly Weather Review*, *144*, 2235–2264.

Rodrigues, P. C., & de Carvalho, M. (2013). Spectral modeling of time series with missing data. *Applied Mathematical Modelling*, *37*, 4676–4684.

Roundy, P. E., & Schreck III, C. J. (2009). A combined wave-number–frequency and time-extended EOF approach for tracking the progress of modes of large-scale organized tropical convection. *Quarterly Journal of the Royal Meteorological Society*, *135*, 161–173.

Schoellhamer, D. H. (2001). Singular spectrum analysis for time series with missing data. *Geophysical Research Letters*, *28*, 3187–3190.

Shen, Y., Li, W., Xu, G., & Li, B. (2014). Spatiotemporal filtering of regional GNSS network's position time series with missing data using principal component analysis. *Journal of Geodesy*, *88*, 1–12.

Shen, Y., Peng, F., & Li, B. (2015). Improved singular spectrum analysis for time series with missing data. *Nonlinear Processes in Geophysics*, *22*, 371–376.

Stechmann, S. N., & Majda, A. J. (2015). Identifying the skeleton of the Madden-Julian oscillation in observational data. *Monthly Weather Review*, *143*, 395–416.

Stechmann, S. N., & Ogrosky, H. R. (2014). The Walker circulation, diabatic heating, and outgoing longwave radiation. *Geophysical Research Letters*, *41*, 9097–9105.

Vautard, R., & Ghil, M. (1989). Singular spectrum analysis in non-linear dynamics, with applications to paleoclimatic time series. *Physica D*, *35*, 395–424.

Vautard, R., Yiou, P., & Ghil, M. (1992). Singular spectrum analysis: A toolkit for short noisy chaotic signals. *Physica D*, *58*, 95–126.

Weare, B. C., & Nasstrom, J. S. (1982). Examples of Extended Empirical Orthogonal Function Analyses. *Monthly Weather Review*, *110*, 481–485.

Wheeler, M. C., & Hendon, H. H. (2004). An all-season real-time multivariate MJO index: development of an index for monitoring and prediction. *Monthly Weather Review*, *132*, 1917–1932.

Yoneyama, K., Zhang, C., & Long, C. N. (2013) Tracking pulses of the Madden-Julian oscillation. *Bull. Amer. Meteor. Soc.*, 1871–1891.

Zhao, Z., & Giannakis, D. (2016). Analog forecasting with dynamics-adapted kernels. *Nonlinearity*, *29*, 2888–2939.