Journal of Intelligent Systems and Internet of Things (JISIoT)
Vol. 11, No. 01, PP. 29-43, 2024
DOI: https://doi.org/10.54216/JISIoT.110104

Prediction of Rainfall Trends Using Forecasting Approaches Based on Singular Spectrum Analysis

Kismiantini[1]*, Shazlyn Milleana Shaharudin[2], Ezra Putranda Setiawan[1], Dhoriva Urwatul Wutsqa[1], Muhamad Afdal Ahmad Basri[3], Hairulnizam Mahdin[4], Salama A. Mostafa[4]

[1]Study Program of Statistics, Universitas Negeri Yogyakarta, Indonesia, [2]Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris (UPSI), Malaysia, [3]Universiti Pendidikan Sultan Idris, Malaysia, [4]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor, Malaysia
Emails: kismi@uny.ac.id; shazlyn@fsmt.upsi.edu.my; ezra.ps@uny.ac.id; dhoriva_uw@uny.ac.id; muhamadafdal181@gmail.com; hairuln@uthm.edu.my; salama@uthm.edu.my

Abstract

Advanced technologies such as the Internet of Things provide an integrated platform for weather focusing, including rainfall and flood prediction. Large rainfall data frequently contain noise, which can be difficult to analyze using a standard time series model due to violated assumptions. Singular spectrum analysis (SSA) is a model-free time series analysis method that is widely used. This study aims to predict the rainfall trends in the Special Region of Yogyakarta, Indonesia, using the Recurrent SSA (SSA-R) and Vector SSA (SSA-V). The SSA-R forecasts using the recurrent continuation directly with the linear recurrent formula, while the SSA-V is a modified recurrent method. This study used 50 years of monthly rainfall data (1970-2019) from 25 stations in the special region of Yogyakarta, Indonesia. The SSA steps for forecasting rainfall data include decomposition (embedding and singular value decomposition), reconstruction (grouping and diagonal averaging), and evaluating the SSA model using *w*-correlation (if *w*-correlation is close to zero, returning to the decomposition stage; otherwise, continue the process), forecasting, evaluating the forecast results using root mean square error (RMSE), mean absolute error, *r*, and mean forecast error, and finally selecting the best model (either the SSA-R or SSA-V model). The results showed that the SSA-R performed better than SSA-V due to the smallest RMSE in the dry, rainy, and inter-monsoon seasons. The SSA-R model's forecast results revealed faint, constant patterns for the dry, and rainy seasons and an increasing pattern for the inter-monsoon season. The novelty of this study is to compare the performance of the SSA-R and SSA-V models in the large rainfall data in the special region of Yogyakarta, Indonesia.

Received: March 12, 2023 Revised: July 27, 2023 Accepted: November 28, 2023

Keywords: Singular Spectrum Analysis; Recurrent Singular Spectrum Analysis; Vector Singular Spectrum Analysis; Internet of Things; Rainfall Patterns; Yogyakarta

1. Introduction

Information on spatial and temporal rainfall patterns is essential for hydrologists to recommend actions for effective and sustainable water resource management [1]. This planning becomes crucial in countries where food production by farmers largely depends on the monsoon season. Prolonged drought can cause crop failure [2], which may increase food prices and then inflation. For example, unusual rainfall patterns caused by El Nino affect food production and lead to inflation in some Indonesian provinces [3,4]. Internet of Things (IoT) based data collection methods can further facilitate the effectiveness of the prediction model.

One of the approaches to obtaining information on spatial and temporal rainfall patterns is identifying a local time scale to determine when rainfall occurs at a particular location based on its trend. Identification of local time scales can be

defined as detecting the time or length of a process in a particular region or area [5]. It can be detected by observing the trend of time series data characterized by the shape of the data.

Many studies have been conducted to identify rainfall trends to detect time scales [6-8]. In general, large datasets of recorded rainfall data are bound to contain noise in the rainfall measurements. This noise may be an intrinsic error structure caused by technical or human recording errors. These studies do not account for noise, potentially jeopardizing the method used.

One method for identifying the range of local time scales based on the trend is called the singular spectrum analysis (SSA). SSA essentially analyzes time series using the singular spectrum involving the eigenvalue decomposition [9]. It is a powerful non-parametric method in time series analysis [10]. A study showed that the SSA could separate the various components of time series and forecast the daily rainfall time series for longer periods in a single run [11]. In SSA, the observed time-series data are decomposed into additive components of trend and noise [12]. It is done in two stages: Stage 1 transforms a univariate time series into a trajectory matrix and obtains its eigenvalues and eigenvectors and Stage 2 separates the eigenvectors into trend and noise and reconstructs the time series components [13]. In addition to these stages, an important advantage of SSA is that it allows, after the reconstruction of the time series understudy, to produce forecasts for the reconstructed components, which is known as the SSA forecasting algorithm [14].

In SSA, selecting a window length is critical to separating the trend and noise components. The chosen window length is expected to be large enough but not greater than half of the observed time-series data [15]. However, the observed behavior of a dataset may influence the window length selection. Another challenge is that the monthly amount of rainfall in torrential rainfall time series data is roughly similar over time [16]. These two situations may cause coinciding singular values when using SSA. In this case, disjoint sets of singular values and different series components could be mixed, resulting in poor separability between the trend and noise components. Moreover, the leading components from the eigentime series are often selected subjectively by looking at the eigenvector plots. Consequently, the separability strength between the trend and noise components in SSA could be affected. As a result of the aforementioned issues, the extracted trend (ET) from SSA tends to flatten out and lacks any discernible pattern [16]. Therefore, forecasting when a torrential rainfall event will occur is difficult because determining the local time range is only sometimes obvious.

To effectively overcome the above shortcomings, we propose to use two forecasting variations of SSA, which are referred to as Recurrent SSA (SSA-R) and Vector SSA (SSA-V). The previous studies show that these methods provide better forecasts than the original SSA, especially when the data are non-stationary [17,18]. Since most of the rainfall data contain variations and seasonality, it is predicted that these methods would yield good forecasts. As a case study, we apply the two forecasting variations of SSA for predicting the monthly rainfall trends in the special region of Yogyakarta, one of the provinces in Indonesia.

2. SSA

SSA is a model-free method that can be extended to all-time series data forms, Gaussian or non-Gaussian, linear or non-linear, and stationary or non-stationary. Monthly rainfall data can be decomposed into multiple additive components through SSA, which can be described as pattern and noise components. The SSA attempts to decompose the initial sequence into a small number of independent, slowly interpretable components, various patterns, oscillating elements, and structureless noise.

2.1. Steps of SSA

The main technique of SSA consists of two complementary stages: decomposition and reconstruction, including two separate steps. In the first step, we decompose the series and reconstruct the original series in the second phase and use the reconstructed series (which is noiseless) to forecast new data points [15].

2.1.1. Stage 1: Decomposition

The decomposition stage consists of embedding and singular value decomposition (SVD) [9]. This stage aims to decompose the series to obtain the Eigen time series data.

Step 1.1: Embedding. The first step in the basic SSA algorithm is the embedding step which refers to constructing a one-dimensional series $\mathbb{X} = \mathbb{X}_n = (x_1, \ldots, x_N)$ of lengths N into a sequence of lagged vectors of size L by forming

$K=N-L+1$ lagged vectors $X_i=\{x_i,\ldots,x_{i+L-1}\}^T$ where $1\leq i \leq K$ to a multidimensional series contained in a matrix, $\mathbf{X=(X_1,\ldots,X_k)}$ called the trajectory matrix as shown in Equation 1 [9]. The rows and columns of **X** are sub-series of the original one-dimensional time series data where the lagged vectors are in the columns of this trajectory matrix. The dimension of the trajectory matrix is called the window length, L which ranges from $1\leq L \leq N$ where the upper bound is usually N/2.

$$\mathbf{X}=\left[\mathbf{X}_1:\ldots:\mathbf{X}_K\right]=\begin{pmatrix} x_1 & x_2 & x_3 & \cdots x_K \\ x_2 & x_3 & x_4 & \cdots x_{K+1} \\ x_3 & x_4 & x_5 & \ldots x_{K+2} \\ \vdots & \vdots & \vdots & \ddots \quad \vdots \\ x_L & x_{L+1} & x_{L+2} & \cdots x_N \end{pmatrix} \quad (1)$$

Step 1.2: SVD. In the second step, the trajectory matrix in Step I is decomposed to obtain its eigentime series based on their singular values using a SVD. The SVD of the trajectory matrix, **X,** is represented as

$$\mathbf{X=U^T\Sigma V} \quad (2)$$

Where $\boldsymbol{U}=(U_1,\ldots,U_L)$ is an $L \times L$ orthogonal matrix, $\mathbf{V}=(V_1,\ldots,V_k)$ is a $K \times K$ orthogonal matrix and $\mathbf{\Sigma}$ is an $L \times K$ diagonal matrix with non-negative real diagonal entries $\mathbf{\Sigma_{ii}}=\sigma_i$ for i=1,…,L. The vectors U_i are called left singular vectors, the V_i are the right singular vectors, and the σ_i are the singular values. Let $\mathbf{S=XX^T}$ where the eigenvalues of **S** are arranged in descending order such that $(\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_L\geq 0)$.

Let $d=rank\ \mathbf{X}=\max\{i,$ such that $\lambda_i>0\}$ and $V_i = X^T U_i / \sqrt{\lambda_i}\,(i=1,\ldots,d)$, then, the SVD of the trajectory matrix **X** can be written as

$$X=X_1+\cdots+X_d \quad (3)$$

Where $\mathbf{X_i} = \sqrt{\lambda_i}_i U_i V_i^T$. Note that the matrices of $\mathbf{X_i}$ are called elementary matrices if $\mathbf{X_i}$ has rank one. The collection $(\sqrt{\lambda_i}_i, U_i, V_i)$ is called the *i*th *E*igentriple of the SVD [17].

2.1.2. Stage 2: Reconstruction

There are two steps in the reconstruction stage, namely, grouping and diagonal averaging. In general, this stage aims to reconstruct the original series and use the reconstructed series for further analysis, such as forecasting.

Step 2.1: Grouping. In the grouping step, the trajectory matrix is split into two groups based on the trend and noise components. The indices set $\{1,\ldots,d\}$ is partitioned into m disjoint subsets $I_1,\ldots,I_m$, corresponding to splitting the elementary matrices into m groups. Set $I=\{i_1,\ldots,i_p\}$, then the resultant matrix $\mathbf{X}_I$ is defined as

$$\mathbf{X_I=X_{i1}}+\cdots+X_{ip} \quad (4)$$

The resultant matrices are computed for $I=I_1,\ldots,I_m$ and substituted in Equation (4). The expansion is defined as

$$\mathbf{X=X_{I1}}+\cdots+X_{lm}$$

where the trajectory matrix is represented as a sum m of resultant matrices. The choice of the sets $\mathbf{I=I_1,\ldots,I_m}$ is known as eigentriple grouping [17].

Step 2.2: Diagonal averaging. The final step in the SSA transforms each matrix of the grouped decomposition (5) into a new series of length T.

Let **Y** be an $L \times K$ matrix with elements y_{ij}, $1\leq i \leq L$, $1\leq j\leq \mathbf{K}$. Set $L^*=\min(L,K)$, $K^*=\max(L,K)$ and $N=L+K-1$. Let $y_{ij}^*=y_{ij}$ if $L < K$ and $y_{ij}^* = yij$ otherwise. By making the diagonal averaging, the **Y** matrix is transferred into the $y_1,\ldots,y_N$ using the formula.

$$y_k = \begin{cases} \frac{1}{k}\sum_{m=1}^{k} y^*_{m,k-m+1} & 1 \le k < L^* \\ \frac{1}{L^*}\sum_{m=1}^{L^*} y^*_{m,k-m+1} & L^* \le k \le K^* \\ \frac{1}{N-k+1}\sum_{m=k-k^*+1}^{N-K^*+1} y^*_{m,\,k-m+1} & K^* < k \le N \end{cases} \quad (6)$$

Diagonal averaging in Equation (6) applied to a resultant matrix $\mathbf{X}_{Ik}$ produces the reconstructed Series $\tilde{\mathbb{X}}_T^{(k)} = (\tilde{x}_1^{(k)},\ldots,\tilde{x}_T^{(k)})$. Hence, the initial series $\mathbb{X}_N = \{y_1, y_2,\ldots, y_N\}$ is decomposed into a sum of m reconstructed series, $x_n = \sum_{k=1}^{m} \tilde{x}_n^{(k)}$. The reconstructed series produced by the elementary grouping will be called the elementary reconstructed series [17].

After the two main stages of decomposition and reconstruction, other features are considered while applying SSA. Figure 1 shows the steps involved in the SSA algorithm.

2.1.3. Stage 3: Forecasting

In SSA forecasting, the time series must follow a linear recurrent formula (LRF). A time series $Y_N = (y1,\ldots y_N)$ satisfies LRF of order d if

$$y_n = a_1 y_{n-1} + a_1 y_{n-2} + \cdots + a_d y_{n-d}, n = d+1,\ldots,N \quad (7)$$

The repeating equations' dimension (or, more precisely, order) is often unknown. The class of LRFs-governed series is rather large and has significant practical implications [17]. For example, an infinite series is regulated by an LRF if and only if it can be expressed as a linear combination of exponential, polynomial, and harmonic series products. Natural recurrent continuation is possible because each term in a series ruled by LRFs is equivalent to a linear combination of numerous previous terms. Naturally, the coefficients of this linear combination may also be utilized to determine the continuation. It is critical to note that seeking an LRF with a small size is not required. Indeed, any other LRF that governs the series will generate the same continuation.

2.2. SSA-R and SSA-V

In this study, SSA-R and SSA-V are used for forecasting purposes since it is a popular approach when predicting time series data. These algorithms are described as follows [18,19].

Step 1: SSA-R. Let $v^2 = \pi_1^2 + \ldots + \pi_j^2$ where π_j is the last component of the eigenvector $U_j(j=1,\ldots,r)$. Denoting $v^2 = \sum_{j-1}^{r} \pi_j^2$, we define the coefficient vector $\mathfrak{R}$ as:

$$\mathfrak{R} = \frac{1}{1-v^2}\sum_{j=1}^{r} \pi_j \boldsymbol{U}_j^{\nabla} \quad (8)$$

Moreover, suppose for any vector $U \in R^L$ denoted by $U^{\nabla} \in R^{L-1}$, the vector consists of the first L–1 components of the vector $\boldsymbol{U}$. Let $y_{N+1},\ldots,y_{N+M}$ show the M terms of the SSA recurrent forecast. Then, the M-step ahead forecasting procedure can be obtained by the following formula:

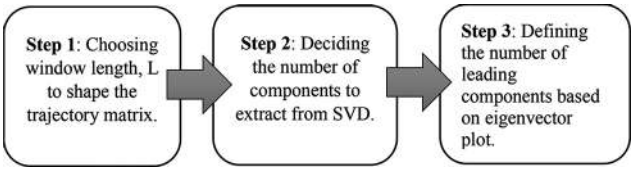

Figure 1: The steps of the SSA algorithm.

$$\hat{y}_i = \begin{cases} \tilde{y}_i, & i = 1,\ldots,N \\ \Re^N Z_i, & i = N+1,\ldots,N+M \end{cases} \tag{9}$$

step $\mathbf{Z}_i = \left[\hat{y}_{i-L+1},\ldots,\hat{y}_{i-1}\right]^N$ and $\tilde{y}_1,\ldots,\tilde{y}_N$, are the reconstructed time series values and can be attained from the 4th step above.

Step 2: SSA-V. Consider the following matrix

$$\Pi = V^{\nabla}(V^{\nabla})^{T} + (1-v^2)AA^T \tag{10}$$

where, $\mathbf{V}^{\nabla} = \left[\boldsymbol{U}_1^{\nabla},\ldots,\boldsymbol{U}_r^{\nabla}\right]$ Now, consider the linear operator:

$$\theta^{(v)}: L_r \longmapsto R^L \tag{11}$$

and,

$$\theta^{(v)}\boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}^{\nabla} \\ \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{U}^{\nabla} \end{pmatrix} \tag{12}$$

Define vector $\mathbf{Z}_i$ as follows:

$$\boldsymbol{Z}_i = \begin{cases} \tilde{\boldsymbol{X}}_i & \text{for } i = 1,\ldots,K \\ \theta^{(v)}\boldsymbol{Z}_{i-1} & \text{for } i = K+1,\ldots,K+h+L-1 \end{cases} \tag{13}$$

Where $\tilde{\boldsymbol{X}}_i$'s are reconstructed columns of the trajectory matrix after grouping and eliminating noise components. Now, by constructing matrix $Z=[Z_1,\ldots,Z_{K+h+L-1}]$ and performing diagonal averaging, we obtain a new series $y_1,\ldots,y_{N+h+L+1}$, where $y_{N+1},\ldots,y_{h+L}$ form the terms of the SSA vector forecast.

3. Materials and Methods

The IoT advanced technology provides an integrated platform for rainfall data collection. This study utilizes the monthly rainfall data in the special region of Yogyakarta, a province located at 7°47'S 110°22'E. The monthly rainfall data for 50 years (1970–2019) from 25 stations were obtained from the Indonesian Meteorology, Climatology, and Geophysics Agency and used for the observed time-series data. The rainfall dataset contains 600 months, which is considered sufficient for identifying rainfall patterns. Figure 2 shows the 25 rainfall stations in the Special Region of Yogyakarta, Indonesia.

Indonesia's rainfall patterns occur in three main areas: monsoon region (type A), equatorial region (type B), and local climate region (type C) [14]. The monsoon region (type A) is typically dominant in Indonesia, covering nearly the entire country. The wet northwest monsoon influences the rainfall in this area, which peaks from November to March, while the dry southeast monsoon affects a trough from May to September, allowing a clear distinction between dry and rainy seasons. The equatorial region (type B) has two peaks, from October to November and from March to May. A shift to the north and the south from the ITCZ or equinox point (culmination) of the sun influences this pattern. Another type of rainfall pattern that occurs in Indonesia is the local climate region (type C), with a peak in June to July and a trough from November to February.

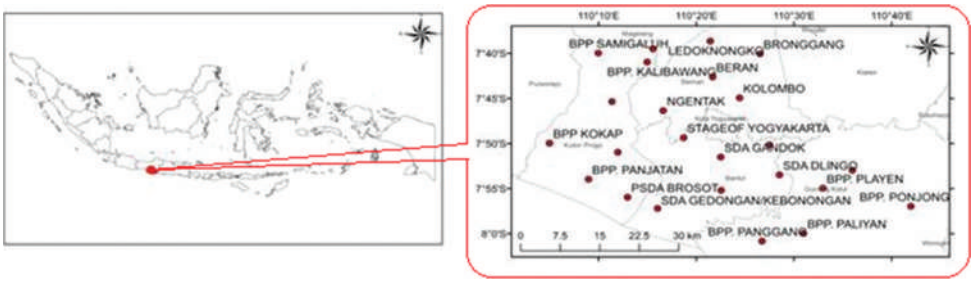

Figure 2: Location of rainfall stations in Yogyakarta.

This study focused entirely on the SSA-V and SSA-R, so it is important to briefly comment on the computational complexity associated with the two SSA forecasting approaches. Furthermore, this discussion could be useful for cases when both approaches are very similar in terms of computation as they both rely on the SSA *L* and *r* for decomposition and reconstruction and LRF for generating the forecast, as shown in Figure 3. As such, in terms of computational complexity, there is no major distinguishable factor, and both approaches will take a similar computation time to generate forecasts.

In summary, recurrent forecasting directly conducts recurrent continuation (using LRF), while vector forecasting works with *L*-continuation. When it comes to approximation continuation, the two forecasting algorithms often provide results that differ. A flow chart of the developed forecasting model based on SSA is shown in Figure 3. The monthly rainfall data in this study were analyzed using an Rssa package in the R program [20].

4. Results and Discussion

In the initial stage of this study, the rainfall data were decomposed into components using the SSA model, which required the identification of the window length (*L*) and ET as parameters pair. Here, *L* denotes the compromise between statistical confidence and information. The suitable *L* value should resolve the oscillations embedded in the original signal.

The performance of the SSA results was determined by assessing the w-correlation at a distinct window length, *L*. The w correlation calculated the separability among the reconstructed time series' noise, trend, and seasonal components. Here, $L = N/2$, $N/5$, $N/10$, and $N/20$, representing $L = 13$, 25, 50, and 125, respectively, for *N* based on 250 monthly rainfall data on Station 22 had been selected for Dry and Rainy seasons. Meanwhile, $L = 5$, 10, 20, and 50, respectively, for *N* based on 100 monthly rainfall data on station 22, had been selected for the inter-monsoon season, as shown in Figures 4-6. The scales were selected to fit the time series data, apart from striking a balance to achieve a proper lag vector sequence.

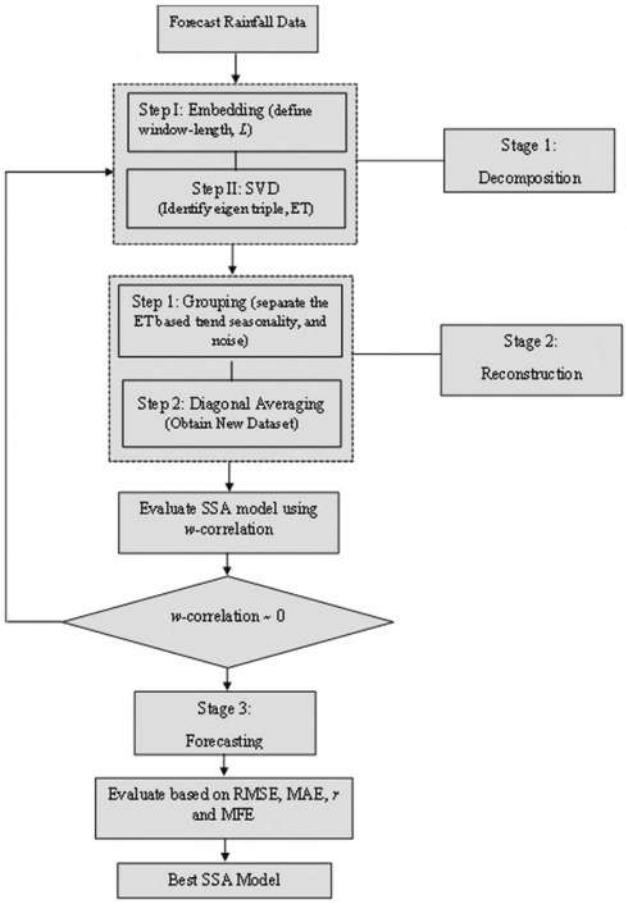

Figure 3: Framework of singular spectrum analysis.

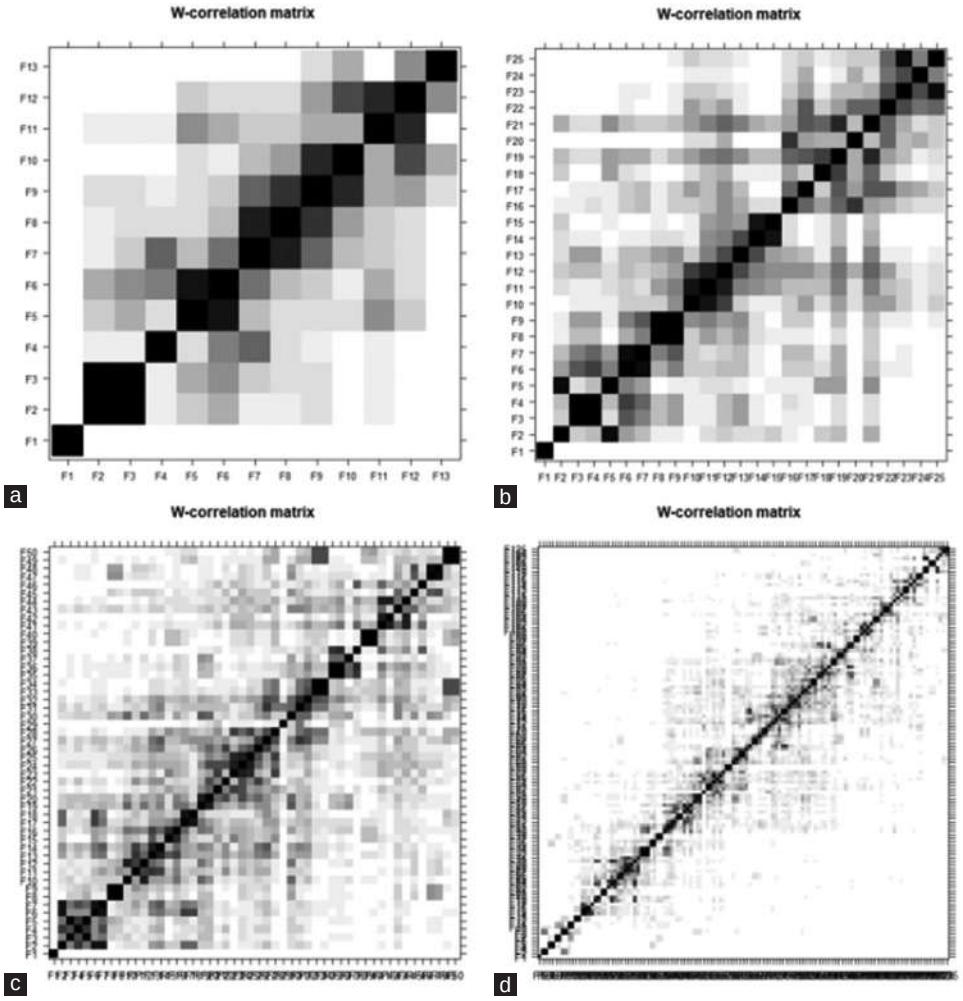

Figure 4: w-correlation plot dry season using SSA with varied windows length (a) L = 13, (b) L = 25, (c) L = 50, and (d) L = 125.

The graphs in Figures 4a-d, 5a-d, and 6a-d illustrate the heat-plot of different window lengths, *L*, based on w-correlations using the SSA approach. The heat plot of w-correlation for the reconstructed components based on a white-black scale ranges between 0 and 1. Huge correlation values among the reconstructed components exhibited the possibility of the components forming a group while corresponding to the same component. As illustrated in Figure 4, the shade of each square represents the w-correlation strength between the two components. Meanwhile, Figures 4, 5, and 6b-d portray the tendency of the components to form a correlation with other components despite signifying a weak correlation. As a result, the components of trends are still mixed with noise and seasonal components in SSA. Then, they were corrected by the small window length of $L = 13$ for dry and rainy seasons and $L = 5$ for inter-monsoon seasons, as shown in Figures 4a, 5a, and 6a for better separability for every season.

Table 1 presents the w-correlation and root mean square error (RMSE) for forecasting results using vector forecasting (SSA-V) and recurrent forecasting (SSA-R) from monthly rainfall data at different window lengths. The results show that the w-correlation illustrates a decreasing pattern as the window length decreases for all rainfall seasons. This result implies that different window lengths affect the separability of components in the dataset. It also shows that SSA pointed out the lowest average w-correlation at a small window length, L = N/20, which shows the strongest separability between the reconstructed components as it is closest to zero.

The lowest RMSE was observed from $L = T/20$ for every season for SSA-R, which had the smallest value among other *L*, indicating its suitability based on the short-time series of the outbreak data. According to these results, the SSA-R model is commendably efficient for forecasting monthly rainfall data in Yogyakarta, Indonesia. Meanwhile, the high RMSE values were reported in this study due to the high model variance for the small sample set. For SSA-V, it showed that the values of RMSE were large compared to the SSA-R. From the RMSE values, the SSA-V was unsuitable for

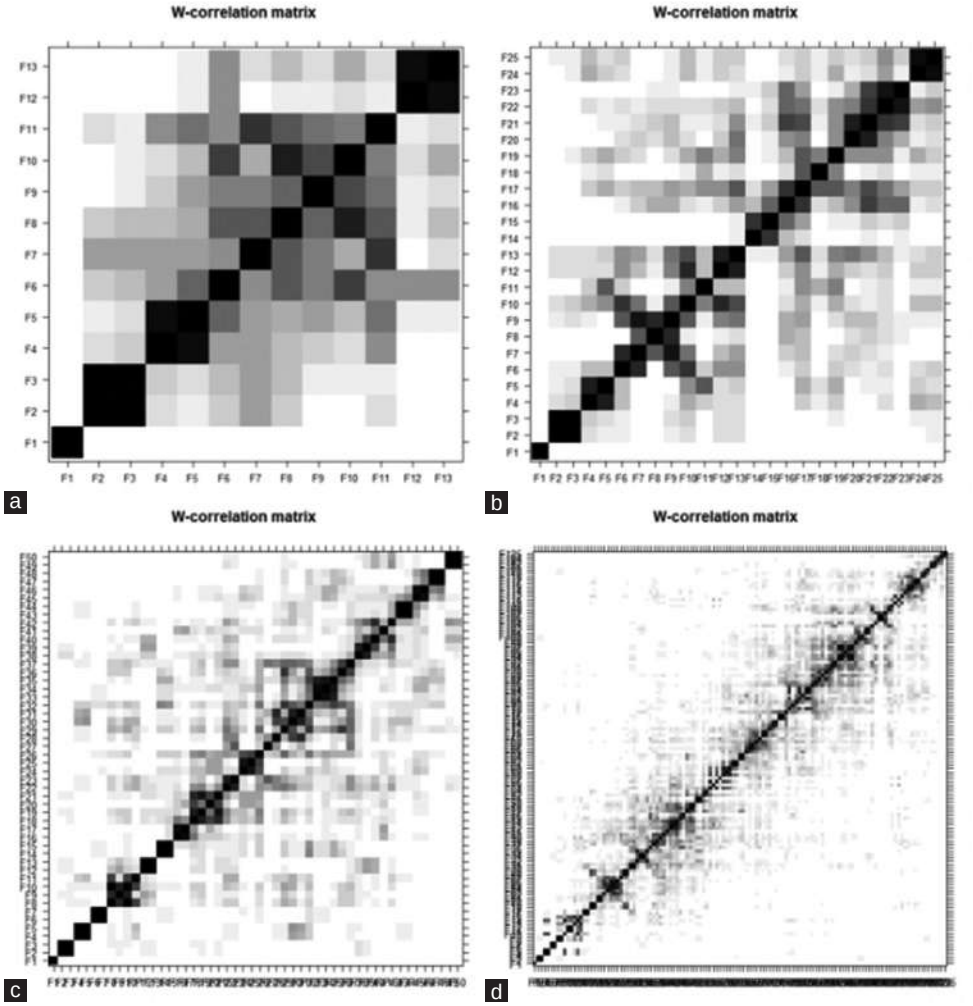

Figure 5: w-correlation plot rainy season using SSA with varied windows length (a) L = 13, (b) L = 25, (c) L = 50, and (d) L = 125.

forecasting Yogyakarta's rainfall data compared to the SSA-R. This result is contrary to the simulation study which was conducted by.

The plot of 12 and four main eigenvectors is displayed in Figure 7. Such a plot is beneficial to choosing an appropriate group for the components of time series data, especially to separate the noise, trend, and seasonal components. The retrieved information may be further analyzed in the step of grouping in SSA-R. The trend and seasonal components were identified from the eigenvector plot, in which they had sine waves indicated by the cycles found in the graph (high frequency). Meanwhile, the noise component was represented by the saw-tooth found in the graph (low frequency). The leading eigenvector has nearly continual coordinates, thus corresponding to a pure smoothing by the Bartlett filter.

The reconstruction result by each of the thirteen and five *ET* is presented in Figure 8. In Figure 8a, the two figures verified the compatibility of the first and second *ET* with the trend. They also verified the compatibility of the 2nd, 3rd, and 13th *ET* with seasonality, whereas the remaining *ET* had the noise component and were thus irrelevant to trend and seasonality for the dry season. For Figure 8b, the two figures verified the compatibility of the first and second *ET* with the trend, and the second and third *ET* were seasonal. For Figure 8c, two figures verified the compatibility of the first and second *ET* with the trend, and the second and the third *ET* were seasonal. In contrast, the remaining *ET* had the noise component and were thus irrelevant to trend and seasonality.

Figure 9a demonstrates the components of the reconstructed time series plot from the trend extracted through SSA-R for monthly rainfall data in the special region of Yogyakarta for the dry season. The reconstructed series is the new dataset derived from the original data, which is clear from noise. It is crucial for SSA to ensure that the forecasting results are precise and accurate. The component of trend in the time series data was used to observe the occurrence of the cases' trend and pattern, as it was randomly tabulated as per monthly rainfall data [Figure 9]. In Figure 9a, the

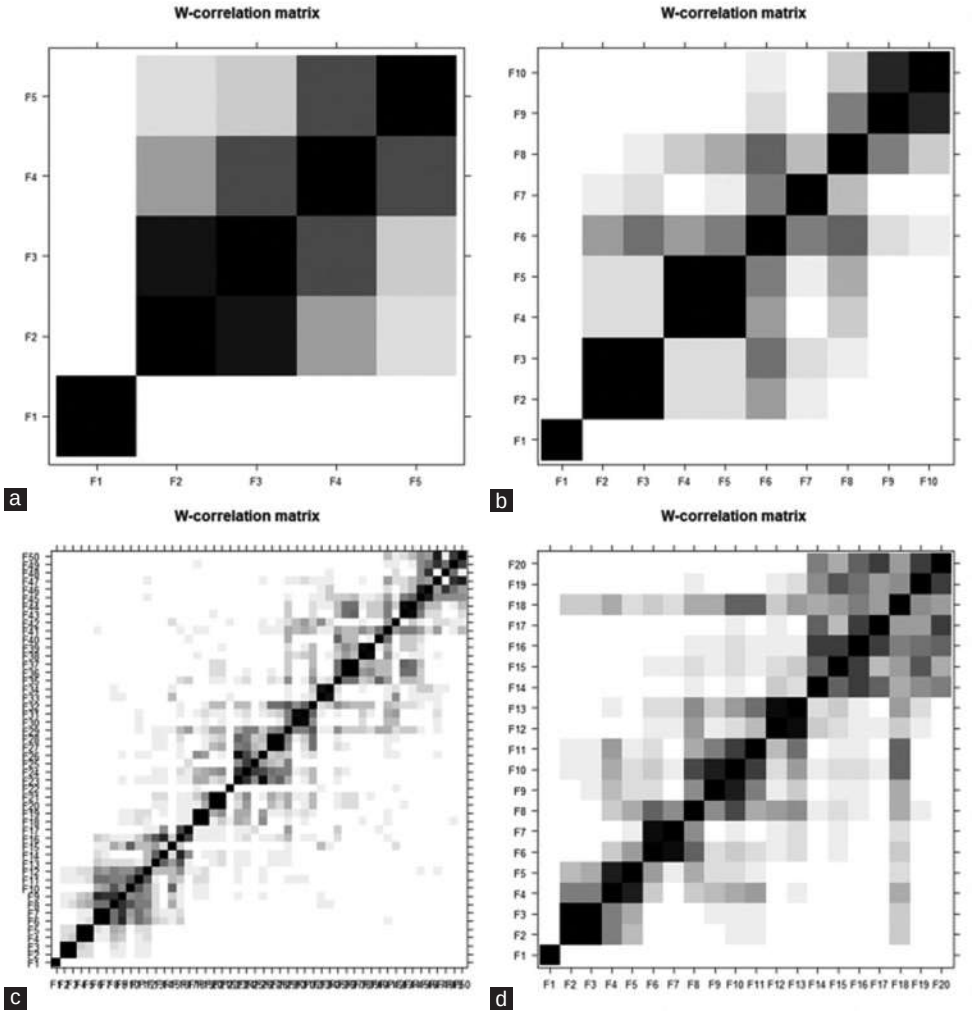

Figure 6: w-correlation plot inter-monsoon season using SSA with varied windows length (a) L = 5, (b) L = 10, (c) L = 20, and (d) L = 50.

Table 1: Comparison of Singular Spectrum Analysis Prediction Performance (*w*-correlation and root mean square error/RMSE) for Several Window Lengths (L)

Window Length, *L*	Dry season			Rainy season			Inter-monsoon season		
	W	RMSE r-f	RMSE v-f	W	RMSE r-f	RMSE v-f	W	RMSE r-f	RMSE v-f
N/20 = 13	0.35	6.115775*	6.849672	0.31	6.112175*	6.561225	0.21	5.181001*	8.271247
N/10 = 25	0.43	6.203654	6.789889	0.56	6.199486	6.274657	0.49	5.329542	8.368689
N/5 = 50	0.55	6.306773	6.437030	0.68	6.224686	6.268396	0.63	5.425216	6.632526
N/2 = 125	0.86	6.350186	6.529682	0.89	6.247369	6.253536	0.98	5.640391	5.698108

trend was precisely generated by a leading *ET*, which coincided with the initial reconstructed component exhibited in Figure 8a. The dashed and straight lines on the plot denote the reconstructed series based on the ET component from SSA and the rainfall data original time series data, respectively.

For proper identification of seasonal series components, the graph of eigenvalues and scatterplots of eigenvectors were applied. To determine the seasonal series components using an eigenvalues plot, several steps were produced by approximately equal eigenvalues. Figure 9b portrays the plot of the logarithms of the 13 singular values for the rainfall data in the Special Region of Yogyakarta. It clearly showed that the step produced by approximately equal eigenvalues corresponded to a sine wave. The scatterplot of eigenvectors displays the regular polygons yielded by a pair of eigenvectors to demonstrate that the series components have produced seasonality components. Based on

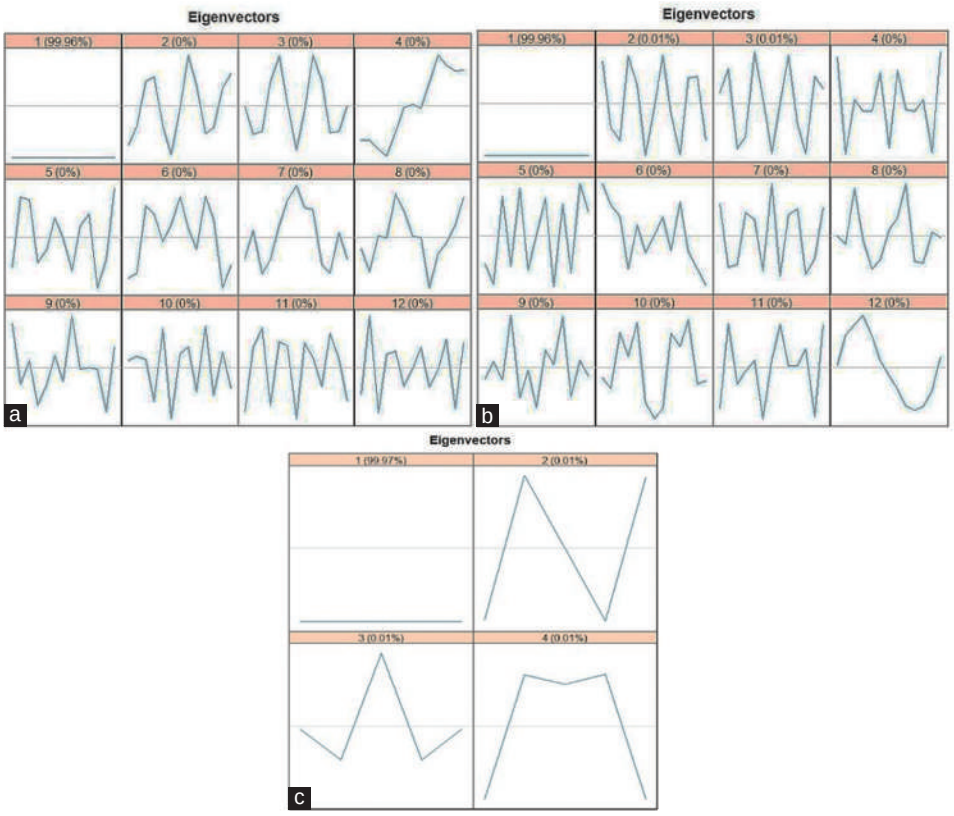

Figure 7: Eigenvectors plot using singular spectrum analysis (a) dry season, (b) rainy season, and (c) inter-monsoon season.

Figure 9c, there were two pairs of eigenvectors produced by regular polygons by components two and three and components five and six. This issue confirmed that the seasonality influenced the rainfall data for the dry season since both figures had a sine wave.

Figure 10a demonstrates the components of the reconstructed time series plot from the trend extracted through SSA-R for monthly rainfall data in the special region of Yogyakarta for the rainy season. The reconstructed series is the new dataset derived from the original data, which is clear from noise. The component of trend in the time series data was used to observe the occurrence of the cases' trend and pattern, as it was randomly tabulated as per monthly rainfall data [Figure 10]. Figure 10b portrays the plot of the logarithms of the 13 singular values for the rainfall data in the special region of Yogyakarta. It clearly showed that the step produced by approximately equal eigenvalues corresponded to a sine wave. The scatterplot of eigenvectors displays the regular polygons yielded by a pair of eigenvectors to demonstrate that the series components have produced seasonality components. Only one pair of eigenvectors produced regular polygons by components two and three [Figure 10c]. This confirmed that the seasonality influenced the rainfall data for the rainy season since both figures have sine waves.

Figure 11 demonstrates the components of the reconstructed time series plot from the trend extracted through SSA-R for monthly rainfall data in the special region of Yogyakarta for the inter-monsoon season. The reconstructed series is the new dataset derived from the original data, which is clear from noise. The trend component in the time series data was used to observe the occurrence of the cases' trends and patterns, as it was randomly tabulated as per monthly rainfall data [Figure 11].

As mentioned in the previous section, the monthly rainfall data were first decomposed and reconstructed using the SSA model. The next step in this study was to predict the future rainfall data in the special region of Yogyakarta. In this stage, an SSA forecasting algorithm, recurrent forecasting, was used accordingly. From hereafter, the model is known as SSA-RF. Table 2 presents the summary statistics from the experiment analysis of SSA-RF at several window lengths.

Table 2 shows that the best performances can be obtained from $L = 13$ and $L = 5$ as they have the lowest mean absolute error (MAE) and mean forecast error (MFE) for dry season, rainy season, and inter-monsoon season. For the Dry

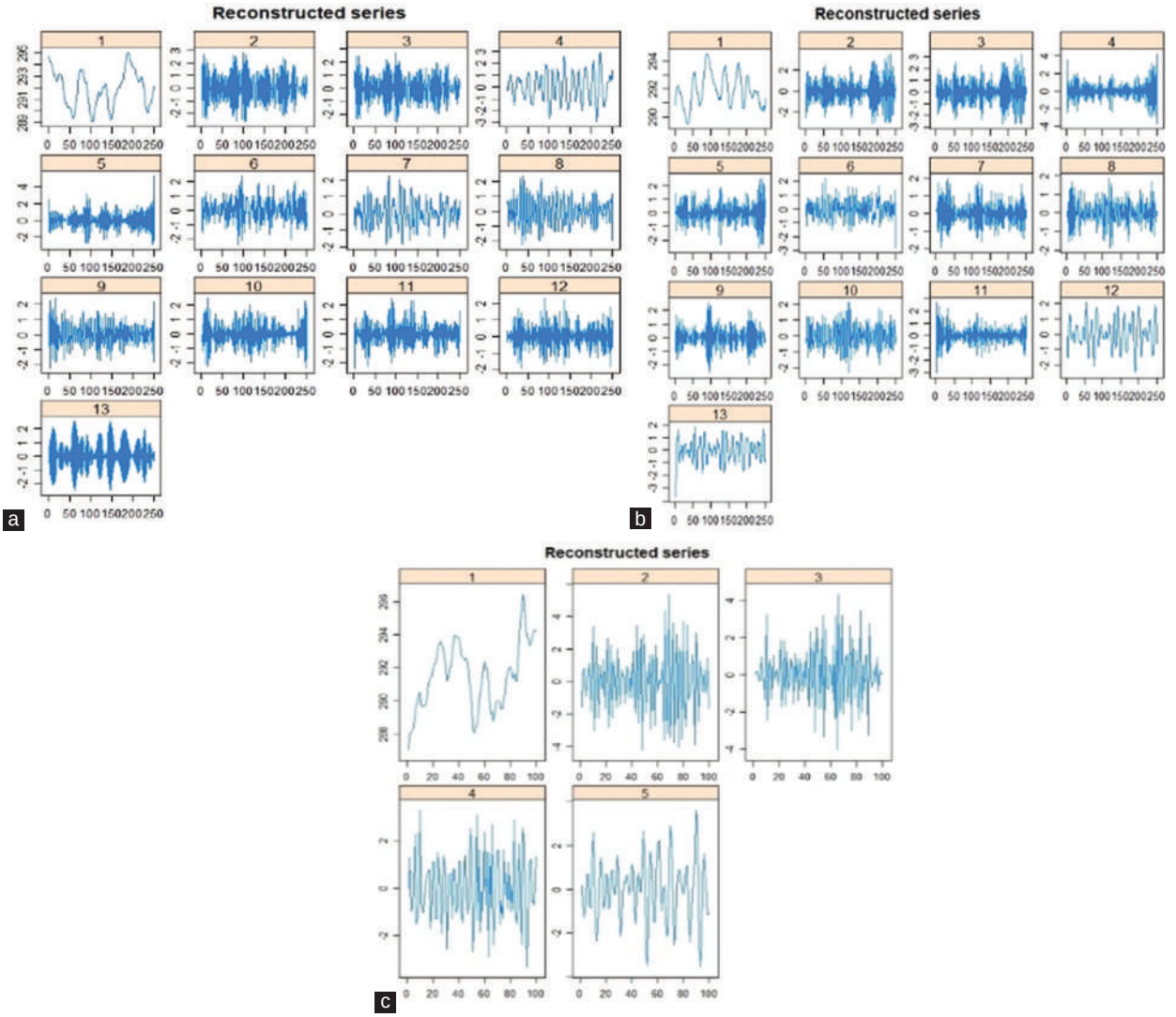

Figure 8: First stage: Elementary reconstructed series (a) dry season (l=13), (b) rainy season (l=13), and (c) inter-monsoon season (L=5).

Season, L = 13 has the lowest MAE of 4.958932. Moreover, the MFE shows that the SSA-RF algorithm with L = 13 tends to over-forecast monthly rainfall data by −0.06748%. Meanwhile, for the rainy season, L = 13 has the lowest MAE of 4.82617, and the MFE shows that the SSA-RF algorithm with L = 13 tends to under-forecast monthly rainfall data by 0.034885%. For the inter-monsoon, L = 5 has the lowest MAE of 4.169522, and the MFE shows that the SSA-RF algorithm with L = 5 tends to under-forecast monthly rainfall data by 0.039892%.

Next, the SSA-RF models were used to predict future cases starting from 2020 to 2025. At the time of this study, the historical rainfall data from 1970 to 2019 were used, and the future 60 months ahead of rainfall data had been predicted accordingly. Following the leading component time series, the SSA-R was applied, respectively, to the time series of those components. As a result of the experiments with various values, it was decided to use L = 13 for the dry and rainy seasons and L = 5 for the inter-monsoon season to proceed with the forecast step. Figure 12a-c illustrates the confirmed rainfall data from 1970 to 2019 for Station 22 for every season and the forecasted monthly rainfall data until 2025. It is worth noting that the figures display a noticeable but faint constant pattern for the rainy and dry seasons from 2020 onward. For the inter-monsoon season, the pattern increased from 2020 onward. Although the SSA-R models produced encouraging statistics based on historical data and a lower under-forecast value, they failed to capture the extreme values in the special region of Yogyakarta's monthly rainfall data, which leads to inconsistent forecasting patterns.

Some limitations of this study should be emphasized when using the SSA-R model in assessing the rainfall data in the special region of Yogyakarta. The SSA-R model works best when the data exhibit a stable or consistent pattern over time

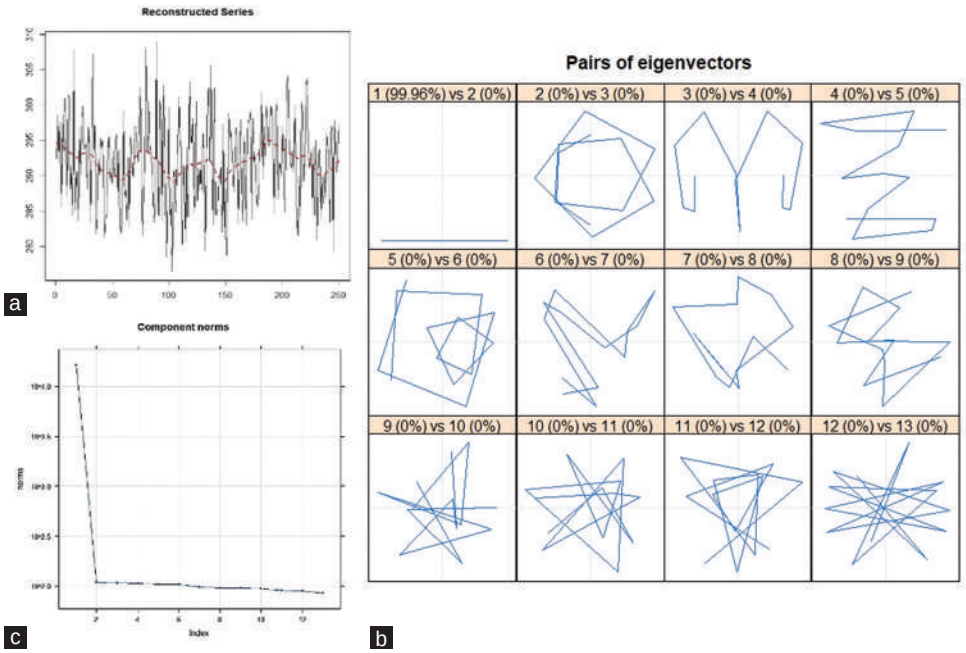

Figure 9: (a) Data of rainfall reconstructed components from extracted trends using SSA at L = 13, (b) logarithms of thirteen eigenvalues, (c) plots of eigenvectors (EV) pairs: 1-EV and 2-EV, 2-EV and 3-EV, 3-EV and 4-EV, 4-EV and 5-EV, 5-EV and 6-EV, 6-EV and 7-EV,7-EV and 8-EV,8-EV and 9-EV,9-EV and 10-EV,10-EV and 11-EV,11-EV and 12-EV, as well as 12-EV and 13-EV for dry season rainfall data.

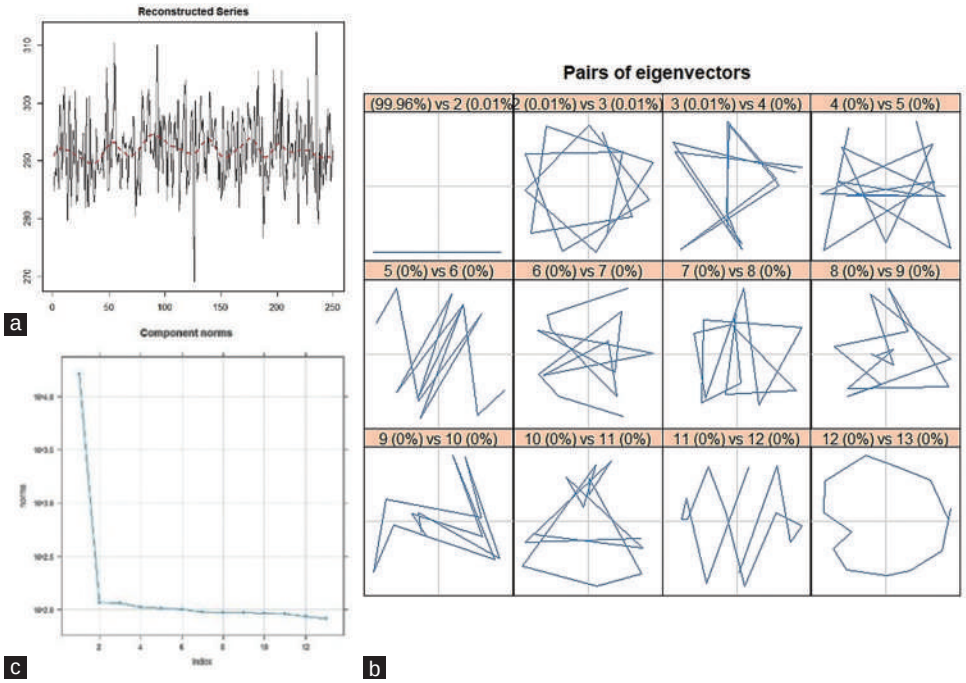

Figure 10: (a) Data of rainfall reconstructed components from extracted trends using SSA at L = 13, (b) logarithms of thirteen eigenvalues, (c) plots of eigenvectors (EV) pairs: 1-EV and 2-EV, 2-EV and 3-EV, 3-EV and 4-EV, 4-EV and 5-EV, 5-EV and 6-EV, 6-EV and 7-EV,7-EV and 8-EV,8-EV and 9-EV,9-EV and 10-EV,10-EV and 11-EV,11-EV and 12-EV, as well as 12-EV and 13-EV for rainy season rainfall data.

with a minimum number of outliers. Therefore, it can help to obtain accurate and precise results for future predictive rainfall data. The sudden spike in data leads to the low performance of forecasting results using this predictive SSA-R model. The SSA-R model mainly projects future values using historical time series data for short-term forecasts. The

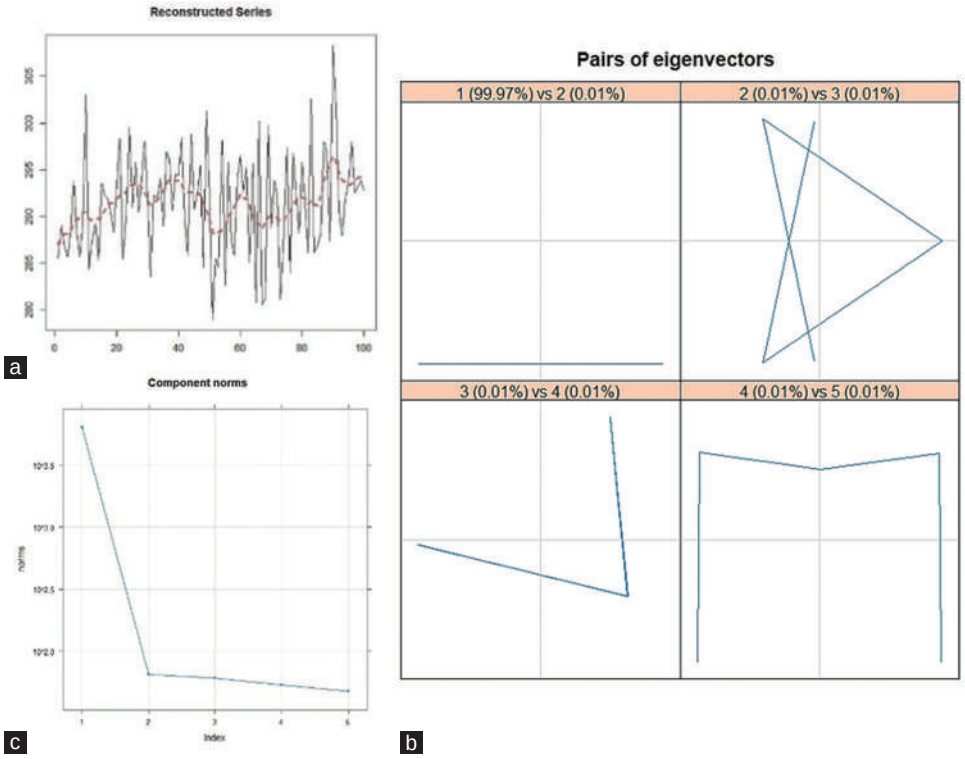

Figure 11: (a) Data of rainfall reconstructed components from extracted trends using SSA at L = 5, (b) logarithms of five eigenvalues, (c) plots of eigenvectors (EV) pairs: 1-EV and 2-EV, 2-EV and 3-EV, 3-EV and 4-EV, as well as 4-EV and 5-EV for inter-monsoon rainfall data.

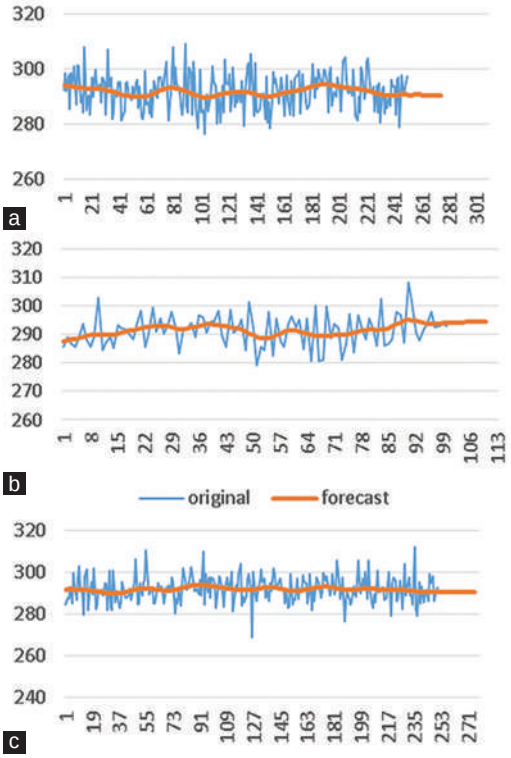

Figure 12: Predicted SSA-RF and observed rainfall data in the special region of Yogyakarta for the three different seasons. (a) Dry season, (b) rainy season, and (c) inter-monsoon season.

Table 2: SSA-RF Prediction Performance of Several Window Lengths (L) with Two Forecasting Categories (Under-Forecast (UF) and Over-forecast (OF))

Window Length, *L*	Dry season			Rainy season			Inter-monsoon season		
	MAE r-f	MFE r-f		MAE r-f	MFE r-f		MAE r-f	MFE r-f	
N/20 = 13	4.958932	−0.06748	UF	4.82617	0.034885	OF	4.169522	0.039892	OF
N/10 = 25	5.031003	−0.0557	UF	4.886513	0.006157	OF	4.364632	0.13405	OF
N/5 = 50	5.081387	−0.0559	UF	4.903435	0.100276	OF	4.398098	0.074445	OF
N/2 = 125	5.128451	−0.26737	UF	4.927836	0.246166	OF	4.607249	−0.09635	UF

MFE: Mean forecast error, MAE: Mean absolute error

recurrent forecasting approach is a better contender than the vector approach for forecasting both short- and medium-time series data of the SSA. However, under such scenarios, it is advisable that users also evaluate the performance of the forecasting SSA approach on their data to arrive at a complete picture. Although the SSA can capture the pattern of the rainfall data, its ability to predict the rainfall data accurately still needs to be investigated further. The different observed behavior of a dataset might influence the window length selection.

5. Conclusion

This study proposed an enhanced SSA model for identifying large rainfall data in the Special Region of Yogyakarta using SSA-R and SSA-V. The SSA-R was the best model compared to the SSA-V, as shown with the smallest RMSEs. From the SSA-R, the best performances were found in the window lengths of $L = 13$ and $L = 5$, resulting in the lowest MAE and MFE for dry season, rainy season, and inter-monsoon season. The SSA-RF models were also used to forecast monthly rainfall data from 2020 until 2025 for station 22 in every season. The forecast results showed faint constant patterns for the rainy and dry season, while the increased pattern occurred in the inter-monsoon season.

References

[1] S. Al-Azzawi, and A. M. Hasan, "A New 4D Hidden Hyperchaotic System with Higher Largest Lyapunov Exponent and its Synchronization," *International Journal of Mathematics, Statistics, and Computer Science*, vol. 2, pp. 63-74, 2023, doi: 10.59543/ijmscs.v2i.8469.

[2] N. Obeid, "On the Product and Ratio of Pareto and Erlang Random Variables," *International Journal of Mathematics, Statistics, and Computer Science*, vol. 1, pp. 33-47, 2023, doi: 10.59543/ijmscs.v1i.7737.

[3] M. Khaleghi, H. Zeinivand, and S. Moradipour, "Rainfall and River Discharge Trend Analysis: A Case Study of Jajrood Watershed, Iran," *International Bulletin of Water Resources and Development*, vol. 2, no. 3, pp. 7-8, 2014.

[4] A. Mondal, S. Kundu, and A. Mukhopadhyay, "Rainfall Trend Analysis by Mann-Kendall Test: A Case Study of North-eastern Part of Cuttack District, Orissa," *International Journal of Geology, Earth and Environmental Sciences*, vol. 2, no. 1, pp. 70-78, 2012.

[5] J. B. Elsner, and A. A. Tsonis, "*Singular Spectrum Analysis: A New Tool in Time Series Analysis*," Springer Science+Business Media, New York, 1996.

[6] R. Mahmoudvand, and P. C. Rodrigues, "A New Parsimonious Recurrent Forecasting Model in Singular Spectrum Analysis," *Journal of Forecasting*, vol. 37, no. 2, pp. 191-200, Mar. 2018, doi: 10.1002/for.2484.

[7] P. Unnikrishnan, and V. Jothiprakash, "Daily Rainfall Forecasting for One Year in a Single Run using Singular Spectrum Analysis," *Journal of Hydrology*, vol. 561, pp. 609-621, Jun. 2018, doi: 10.1016/J.JHYDROL.2018.04.032.

[8] M. C. R. Leles, J. P. H. Sansão, L. A. Mozelli, and H. N. Guimarães, "Improving Reconstruction of Time-series Based in Singular Spectrum Analysis: A Segmentation Approach," *Digital Signal Processing*, vol. 77, pp. 63-76, Jun. 2018, doi: 10.1016/J.DSP.2017.10.025.

[9] N. Golyandina, and A. Zhigljavsky, "*Singular Spectrum Analysis for Time Series*," Springer Verlag, Berlin, 2013. doi: 10.1007/978-3-642-34913-3.

[10] H. Hassani, and D. Thomakos, "A Review on Singular Spectrum Analysis for Economic and Financial Time Series," *Statistics and its Interface*, vol. 3, pp. 377-397, 2010.

[11] S. M. Shaharudin, N. Ahmad, and F. Yusof, "Effect of Window Length with Singular Spectrum Analysis in Extracting the Trend Signal on Rainfall Data," *AIP Conference Proceedings*, vol. 1643, no. 1, pp. 321-326,

Feb. 2015, doi: 10.1063/1.4907462.
[12] S. M. Shaharudin, N. Ahmad, and N. H. Zainuddin, "Modified singular Spectrum Analysis in Identifying Rainfall Trend Over Peninsular Malaysia," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, no. 1, pp. 283-293, 2019, doi: 10.11591/ijeecs.v15.i1.pp283-293.
[13] "*Special Region of Yogyakarta*." Available from: https://en.wikipedia.org/wiki/special_region_of_yogyakarta [Last accessed on 2021 Dec 12].
[14] E. Aldrian, and R. Dwi Susanto, "Identification of Three Dominant Rainfall Regions within Indonesia and their Relationship to Sea Surface Temperature," *International Journal of Climatology*, vol. 23, no. 12, pp. 1435-1452, Oct. 2003, doi: 10.1002/JOC.950.
[15] H. Hassani, "Singular Spectrum Analysis: Methodology and Comparison," *Journal of Data Science*, vol. 5, no. 2, pp. 239-257, 2007.
[16] L. J. Rodríguez-Aragón, and A. Zhigljavsky, "Singular Spectrum Analysis for Image Processing," *Statistics and its Interface*, vol. 3, no. 3, pp. 419-426, 2010, doi: 10.4310/SII.2010.V3.N3.A14.
[17] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky, "*Analysis of Time Series Structure: SSA and Related Techniques*," CRC Press, Boca Raton, 2001.
[18] M. Ghodsi, H. Hassani, D. Rahmani, and E. S. Silva, "Vector and Recurrent Singular Spectrum Analysis: Which is Better at Forecasting?" *Journal of Applied Statistics*, vol. 45, no. 10, pp. 1872-1899, Jul. 2017, doi: 10.1080/02664763.2017.1401050.
[19] S. Milleana Shaharudin, "Predictive Modelling of Covid-19 Cases in Malaysia based on Recurrent Forecasting-singular Spectrum Analysis Approach," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1.4, pp. 175-183, 2020, doi: 10.30534/ijatcse/2020/2691.42020.
[20] N. Golyandina, and A. Korobeynikov, "Basic Singular Spectrum Analysis and Forecasting with R," *Computational Statistics and Data Analysis*, vol. 71, pp. 934-954, Mar. 2014, doi: 10.1016/J.CSDA.2013.04.009.