Monthly streamflow forecasting with auto-regressive integrated moving average

Najah Nasir¹, Ruhaidah Samsudin¹ and Ani Shabri²
¹ Fakulti Komputeran, Universiti Teknologi Malaysia, 81310, Johor, Malaysia
² Fakulti Sains, Universiti Teknologi Malaysia, 81310, Johor, Malaysia

E-mail: najah6@live.utm.my

Abstract. Forecasting of streamflow is one of the many ways that can contribute to better decision making for water resource management. The auto-regressive integrated moving average (ARIMA) model was selected in this research for monthly streamflow forecasting with enhancement made by pre-processing the data using singular spectrum analysis (SSA). This study also proposed an extension of the SSA technique to include a step where clustering was performed on the eigenvector pairs before reconstruction of the time series. The monthly streamflow data of Sungai Muda at Jeniang, Sungai Muda at Jambatan Syed Omar and Sungai Ketil at Kuala Pegang was gathered from the Department of Irrigation and Drainage Malaysia. A ratio of 9:1 was used to divide the data into training and testing sets. The ARIMA, SSA-ARIMA and Clustered SSA-ARIMA models were all developed in R software. Results from the proposed model are then compared to a conventional auto-regressive integrated moving average model using the root-mean-square error and mean absolute error values. It was found that the proposed model can outperform the conventional model.

1. Introduction
Streamflow, a key component of the hydrologic cycle is the movement of water in canals that comes from surface runoff, groundwater and pipe discharges. Streamflow forecasts can be generated when chronological hydrographs, i.e. the data for rate of flow over time of a stream are present. Hydrologic models are basic, theoretical depictions of the hydrologic cycle components which are mainly applied for comprehension of the hydrologic processes and in hydrologic estimations. There are two notable types of hydrologic models which are process-based and stochastic models. Attempts to characterise the physical activities examined in nature are made through process-based models. Stochastic models are black box systems built on data and uses statistical and mathematical concepts to relate the input to output. Frequent techniques practised by researchers and scientists include transfer functions, regression systems, and neural networks. Recent efforts in hydrologic modelling research are focused on having a more universal approach in interpreting the behaviour of hydrologic systems so that reliable predictions can be made and important issues in managing water resources may be overcome.

Time series analysis is the systematic investigation of time series data to obtain meaningful statistical and mathematical interpretations of the series [1]. Approximation of general autocovariance structures and identification of possible parametric models during time series modelling can be achieved through the use of linear models such as auto-regressive or moving average models [2]. Theoretically, the most common models for time series forecasting are auto-regressive integrated moving average (ARIMA) models. Differencing the data or subtraction of the estimated trend and seasonal components from the
data can make a time series be stationary [3]. An ARIMA model may be regarded as a noise filter for a signal so it can then be employed for forecasting by extrapolation. Table 1 presents the research in recent years implementing assorted models for streamflow forecasting.

Table 1. Streamflow forecasting models used by past researchers.

| Year | Author            | Model                                                                 |
|------|-------------------|----------------------------------------------------------------------|
| 2011 | Samsudin et al.   | Artificial neural network (ANN), auto-regressive integrated moving     |
|      |                   | average (ARIMA), group method of data handling-least squares          |
|      |                   | support vector machine (GLSSVM),                                     |
| 2012 | Ismail et al.     | ANN, ARIMA, LSSVM, Self organising map-LSSVM                         |
| 2012 | Shabri and Suhartono | ANN, ARIMA, LSSVM, support vector machine (SVM)               |
| 2014 | Radzi et al.      | ANN, adaptive neuro-fuzzy interface system (ANFIS), SVM              |
| 2014 | Valizadeh et al.  | ANFIS                                                                |
| 2015 | Adenan and Noorani | Chaos approach                                                        |
| 2016 | Adnan et al.      | Soil and water assessment tool                                        |
| 2016 | Salaudeen et al.  | Multivariate equations                                                |
| 2016 | Yaseen et al.     | Feed-forward back-propagation neural network, radial basis function   |
|      |                   | neural network                                                        |

Data pre-processing is needed because data in the real world is incomplete, noisy and inconsistent. Data is considered as incomplete when they lack certain attributes or values, is noisy when containing outliers or errors and is said to be inconsistent if they contain discrepancies [4]. Quality data are the basis for quality decisions but if quality data are nonexistent, there may be no quality mining results [4] and no reliable data warehouses or databases. In time series analysis, singular spectrum analysis (SSA) is an established, powerful and novel technique. SSA has a combination of elements from classical time series, dynamical systems, multivariate geometry, multivariate statistics and signal processing [5], hence is applicable to diverse areas for numerous practical problems such as economics, mathematics and physics. SSA could be successfully applied to apparently complex time series with a notable structure [6]. Model performance can be improved by employing SSA for pre-processing time series data [7] as it is able to efficiently reconstruct the time series after identifying, extracting and eliminating its noise components [8].

2. Auto-Regressive Integrated Moving Average

An auto-regressive integrated moving average (ARIMA) model, generalised from an auto-regressive moving average (ARMA) model, are commonly benefitted by those from the field of statistics, econometrics and time series analysis, where both models are fitted to time series data to improve knowledge of the data or calculate new points in the series. Although the concept behind ARIMA models was established much earlier, the systematic method for utilising the technique was detailed in [9]. ARIMA forecasting and Box-Jenkins forecasting have ever since been referred to the same procedure. ARIMA models are used in cases where signs of non-stationarity appear in the data, in which preliminary differencing can be employed to reduce it [10].

The AR part specifies that the relevant variable is regressed on its lagged values while the MA part of ARIMA denotes that the regression error is a linear combination of error terms whose values appeared contemporaneously and at different times in the past. The I in ARIMA, standing for “integrated”, implies that the data have been substituted with the difference between current and preceding values. The purpose of these features is to tailor the model to the data appropriately. Non-seasonal ARIMA models are usually expressed as $ARIMA(p, d, q)$ wherein parameters p, d, and q are non-negative integers, p is the number of time lags, d is the differencing count and q is the order of the MA model. Seasonal ARIMA models are typically designated as $ARIMA(p, d, q)(P, D, Q)_m$, in which the number of periods per season is referred as m whereas the seasonal part of AR, differencing, and MA expressions are the uppercase $P, D, Q$ [10].
A real number time series $Y_t$ where $t$ is an integer, can be presented as an $ARMA(p,q)$ model by

$$
(1 - \sum_{i=1}^{p} \alpha_i L^i) Y_t = (1 + \sum_{i=1}^{q} \theta_i L^i) \varepsilon_t \tag{1}
$$

whereby $L$ is the lag operator, $\alpha_i$ is the parameter of the AR part, $\theta_i$ is the parameter of the MA part and $\varepsilon_t$ is the error term. Suppose that the polynomial $\left(1 - \sum_{i=1}^{p} \alpha_i L^i\right)$ has a unit root, i.e. a factor $(1 - L)$ of multiplicity $d$, it can then be transcribed as follows.

$$
(1 - \sum_{i=1}^{p} \alpha_i L^i) = (1 - \sum_{i=1}^{p-d} \phi_i L^i) (1 - L)^d
$$

The polynomial factorisation property of an $ARIMA(p,d,q)$ is specified as $p = p' - d$, given by

$$
(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - L)^d Y_t = (1 + \sum_{i=1}^{q} \theta_i L^i) \varepsilon_t \tag{3}
$$

Therefore, it becomes a distinct case of an $ARMA(p + d, q)$ having the AR polynomial with $d$ unit roots. Figure 1 describes the method for forecasting by applying ARIMA.

**Figure 1.** Auto-regressive integrated moving average common procedure.

### 3. Singular Spectrum Analysis

Decomposing the initial time series into several series is the gist of singular spectrum analysis (SSA), in which each element can be classified as either a periodic component, trend or noise [11]. This is followed by a reconstruction of the initial series and the reconstructed series can then be used for prediction of future data. The SSA technique consists of four steps namely, embedding, singular value decomposition (SVD), grouping and diagonal averaging, whereby the first two steps make up the decomposition stage while the others are under the reconstruction stage. A brief SSA procedure is provided below, mainly following the discussion by [6]. The association between every component of SSA is shown in figure 2.

#### 3.1. Step 1: Embedding

A one-dimensional time series $Y_t = (y_1, ..., y_T)$ is transferred into the multi-dimensional series $X_1$, ..., $X_K$ with $X_i = (y_{i}, ..., y_{i+L-1})'$ in $R^L$, where $K = T - L + 1$ and $2 \leq L \leq T$. This step produces the trajectory matrix $X = [X_1, ..., X_K] = (x_{i,j})_{i,j=1}^{L,K}$ which is a Hankel matrix, meaning that all diagonal elements are equal.

#### 3.2. Step 2: Singular value decomposition

For matrix $XX'$, designate its eigenvalues in decreasing order by $\lambda_1, ..., \lambda_L$ and by $U_1, ..., U_L$ its orthonormal system of eigenvectors corresponding to the eigenvalues, that is, $(U_i, U_j) = 0$ for $i \neq j$ and $\|U_i\| = 1$. $(U_i, U_j)$ is the inner product of vectors $U_i$ and $U_j$ whereas $\|U_i\|$ is the norm of $U_i$. Write $d = \max(i, \text{such that} \lambda_i > 0) = \text{rank} X$, $V_i = X'U_i(\sqrt{\lambda_i})^{-1}$ and the SVD of the trajectory matrix $X = X_1 + \cdots + X_d$, such that $X_i = \sqrt{\lambda_i} U_i V_i'$ ($i = 1, ..., d$). $X_i$ are elementary matrices while $U_i$ and $V_i$ represents the left and right eigenvectors of the trajectory matrix. For matrix $X$, its $i$-th eigentriple is $(\lambda_i, U_i, V_i)$ and the singular values are $\sqrt{\lambda_i}$ ($i = 1, ..., d$) whereas its matrix spectrum is the set $\{\sqrt{\lambda_i}\}$.
3.3. Step 3: Grouping
Divide the elementary matrices $X_i$ into groups and sum them within the group. Let $I = \{i_1, \ldots, i_p\}$ be indices for matrix $X_{i_1}$, then $X_I$ is defined as $X_{i_1} + \cdots + X_{i_p}$. The representation $X = X_{i_1} + \cdots + X_{i_m}$ relates to separation of the indices $K = 1, \ldots, m$ into disjoint subsets $I_1, \ldots, I_m$. The selection process of sets $I_1, \ldots, I_m$ is called the eigentriple grouping.

3.4. Step 4: Diagonal averaging
Every matrix $I$ is transferred into an additive component of the original series $Y_T$. Get the $k$-th term of the resulting series by averaging all $z_{ij}$ elements of matrix $Z$, wherein $i + j = k + 2$. This method is called diagonal averaging or Hankelisation of matrix $Z$, resulting in $HX = \bar{X}_{i_1} + \cdots + \bar{X}_{i_m}$ is obtained by employing diagonal averaging on all $X = X_{i_1} + \cdots + X_{i_m}$, in which $\bar{X}_{i_k} = HX$.

4. Data
The data used in this research is the Hydrological Data published by the Water Resources Management and Hydrology Management from the Department of Irrigation and Drainage Malaysia. Of all available hydrological archives, only the monthly Streamflow Records for Sungai Muda (Jeniang) from 1960 to 2009, the 1974-2009 records of Sungai Muda (Syed Omar) and Sungai Ketil (Kuala Pegang) records in 1980 until 2009, are appropriate for this study. The dataset obtained have distinct total observations such that there are correspondingly 600, 432 and 360 data for Sungai Muda (Jeniang), Sungai Muda (Syed Omar) and Sungai Ketil (Kuala Pegang). It was decided that a training to testing ratio of 9:1 would be used for this research. Figure 3 displays the plots of monthly streamflow for Sungai Muda (Jeniang), Sungai Muda (Syed Omar) and Sungai Ketil (Kuala Pegang).

Figure 2. Data transformations at each stage of singular spectrum analysis.

Figure 3. Graphs of streamflow per month for the respective rivers.
5. Results
This research was conducted using the ARIMA model in combination with the SSA data pre-processing technique and also experimented with the clustering of eigenvector pairs during execution of SSA. Results obtained from the models SSA-ARIMA and Clustered SSA-ARIMA were compared to a basic ARIMA model. All models were constructed in the R software environment by utilising suitable packages downloadable from the CRAN repository. The performance of each model was computed through the root-mean-square error (RMSE) and mean absolute error (MAE) formulas, which are typically expressed by equations (4) and (5) respectively. The term \( n \) is a representation for the number of data, \( a \) is the actual yield while \( p \) signifies the predicted yield in the following equations.

\[
\text{RMSE} = \left( \frac{1}{n} \sum_{i=1}^{n} (a_i - p_i)^2 \right)^{1/2}
\]

\[
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |a_i - p_i|
\]

According to the ratio of 9:1, each of the one-dimensional time series data was divided into training and testing sets such that for Sungai Muda (Jeniang), there were 540 observations for training and 60 for testing, while the training to testing ratio for Sungai Muda (Syed Omar) was 389:43 observations, whereas the Sungai Keti (Kuala Pegang) dataset was separated into groups of 324 and 36 observations respectively. For ease of conduct, all ARIMA models in this research were designed by applying the auto.arima() function from the ‘forecast’ package on each training set. The automatically generated models were then supplied to Arima() for prediction purposes, which in turn also produces the accuracy calculated from the results of the forecasts.

Methodologically, models SSA-ARIMA and Clustered SSA-ARIMA are very much the same as developing an ARIMA model, just that they require the implementation of SSA technique before building the ARIMA model. SSA was employed with the ‘Rssa’ package from available libraries in the R environment software. Fundamentally, SSA was performed on every set of data using appropriate functions built into the package, that is, decomposition was through ssa() whereas visual inspection of plots and reconstruct() were for reconstruction. The window length \( L \) chosen for every single dataset as a parameter for ssa() was \( L = (n/2) - 12 \), such that \( L \) should be smaller than \( n/2 \) but also divisible by the period, and \( n \) is the number of data. Eigenvector pairs were grouped by examining either the w-correlation matrix plot or the pairs of eigenvector plot graphically, in which eigenvectors were paired together for reconstruction if they are regularly shaped in the eigenvector plot or have a high correlation between them in the correlation plot. The set of data series resulting from SSA was then individually modelled with ARIMA following the steps aforementioned and each of the error terms calculated were then summed to obtain only one value of RMSE and MAE.

The pairs of eigenvector were also tried for clustering during the implementation of SSA in such a way that apart from choosing suitable pairs, groups of pairs that seemed to be quite highly correlated are put together. This was attainable by looking at the w-correlation matrix plots of the series after running the ssa() function to determine a suitable number of clusters to be made, which is then sent as a parameter for the grouping.auto() function available from the ‘Rssa’ package. Subsequent steps after this phase of grouping are similar to implementing the SSA-ARIMA model. There was also a trend extraction step performed before conducting the SSA technique on the initial time series for Clustered SSA-ARIMA.

The performance of ARIMA, SSA-ARIMA and Clustered SSA-ARIMA models is measured for comparison to each other. The RMSE for ARIMA is 18.479, SSA-ARIMA is 13.315 and Clustered SSA-ARIMA is 12.020 with usage of the Sungai Muda (Jeniang) data. MAE computed for models forecasting Sungai Muda (Jeniang) data are 14.399, 10.815 and 9.348 for ARIMA, SSA-ARIMA and Clustered SSA-ARIMA respectively. Applying Sungai Muda (Syed Omar) data, ARIMA gives an RMSE value of 40.678 and SSA-ARIMA returned 40.640, while Clustered SSA-ARIMA yielded 38.857. Values of 32.995, 31.946 and 29.845 were found for the computation of MAE on ARIMA, SSA-ARIMA and Clustered SSA-ARIMA models after forecasting for Sungai Muda (Syed Omar). The RMSE for ARIMA is 7.260, SSA-ARIMA is 6.897 and Clustered SSA-ARIMA is 6.497 with usage of
the Sungai Ketil (Kuala Pegang) data. MAE computed for models forecasting Sungai Ketil (Kuala Pegang) data are 5.467, 5.352 and 5.274 for ARIMA, SSA-ARIMA and Clustered SSA-ARIMA respectively. The values calculated for each model can be viewed in table 2.

| Model                  | Sungai Muda (Jeniang) | Sungai Muda (Syed Omar) | Sungai Ketil (Kuala Pegang) |
|------------------------|------------------------|-------------------------|----------------------------|
|                        | RMSE       | MAE       | RMSE   | MAE       | RMSE   | MAE       |
| ARIMA                  | 18.479     | 14.399    | 40.678 | 32.995    | 7.260  | 5.467     |
| SSA-ARIMA              | 13.315     | 10.815    | 40.640 | 31.946    | 6.897  | 5.352     |
| Clustered SSA-ARIMA   | 12.020     | 9.348     | 38.857 | 29.845    | 6.497  | 5.274     |

6. Conclusions
With the help of streamflow forecasts, a variety of water users and administrators who rely upon various water resources including streams or rivers, lakes, wetlands and the ocean, will be able to enhance their decision making and water management capabilities. In addition to other organisational and statistical tools, streamflow forecasts enable informed decisions related to economic management, flood mitigation, and environmental consideration of the water resources system [12], such as water distributions, environmental watering, cropping strategies and managing drought, to name a few.

It was revealed in this research that the application of SSA pre-processing technique onto data can improve forecasting capabilities of the ARIMA model. Results were found to be better when the pairs of eigenvectors were made into collections during implementation of SSA, specifically before reconstructing the series. Thus, forecasts from these models may aid in the management of water resource systems and enable more accurate decisions be made on irrigation water allocations and environmental flows [13]. More investigations could be performed by researchers on modelling different datasets using various models coupled with other pre-processing methods.

Acknowledgments
The Malaysia Ministry of Higher Education has conferred Universiti Teknologi Malaysia the support for this research under the vote number R.J130000.7826.4F681 and R.J130000.7828.4F875.

References
[1] Shumway R H and Stoffer D S 2010 Time Series Analysis and Its Applications with R Examples (New York: Springer Science+Business Media) section 1.1 pp 1-2
[2] Breidt J 2005 J. Am. Stat. Assoc. 100 pp 348-9
[3] Brockwell P J and Davis R A 2016 A general approach to time series modeling Introduction to Time Series and Forecasting (New York: Springer-Verlag) section 1.3.3 p 14-5
[4] Kamber M, Han J and Pei J 2012 Data preprocessing Data Mining: Concepts and Techniques (Amsterdam: Elsevier) chapter 3 pp 83-124
[5] Zhigljavsky A A 2010 Stat. Interface 3 255-8
[6] Golyandina N, Nekrutkin V and Zhigljavsky A A 2001 Analysis of Time Series Structure: SSA and Related Techniques (Chapman and Hall / CRC)
[7] Gao Y, Qu C and Zhang K 2016 Energies 9 pp 757-85
[8] Hassani H and Mahmoudvand R 2013 Int. J. Energy Stat. 1 pp 55-83
[9] Box G E P and Jenkins G M 1970 Time Series Analysis: Forecasting and Control (San Francisco: Holden-Day)
[10] Hyndman R J and Athanasopoulos G 2014 Stationarity and differencing Forecasting: Principles and Practice (Melbourne: OTexts) section 8/1 pp 213-21
[11] Hassani H 2007 J. Data Sci. 5 pp 239-57
[12] Abudu S, King J P and Pagano T C 2009 J. Hydrol. Eng. 15 612-23
[13] Chiew F H S, Zhou S L and McMahon T A 2003 J. Hydrol. 270 pp 135-44