J. Agr. Sci. Tech. (2020) Vol. 22(2): 563-578

Comparative Evaluation of Hybrid SARIMA and Machine Learning Techniques Based on Time Varying and Decomposition of Precipitation Time Series

L. Parviz[1*]

ABSTRACT

Accurate precipitation forecasts are much attractive due to their complexity. This study aimed to use the hybrid Seasonal Autoregressive Integrated Moving Average (SARIMA) model and machine learning techniques such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM) to improve precipitation forecasts. Time variation analysis and time series decomposition were the two concepts applied to construct the hybrid models. The performance of the two concepts was evaluated with monthly precipitation time series of two stations in northern Iran. Time variation analysis of time series was conducted with the clustering analysis, which increased the accuracy of forecasting with 20.99% decrease in the geometric mean error ratio for the two stations. SVM model decreased the forecasted error compared to ANN in the internal process of time variation analysis. Average of Mean Relative Error (MRE) were MRE_{SVM}= 0.72, MRE_{ANN}= 0.89, and Mean Absolute Error (MAE) in the two stations were MAE_{SVM}= 18.02 and MAE_{ANN}= 23.88. Therefore, SVM outperformed the ANN model. Comparison of the two hybrid models indicated that more accurate results belonged to the concept of time series decomposition (the decrease in root mean square error from time variation to time series decomposition concepts was 13.35%). Extracting the pattern of data with SARIMA-based hybrid model with time series decomposition improved the precipitation forecasting. Configurations related to nonlinear components of time series with time steps of residual had good performance (the average of agreement index was 0.9). The results suggest that the hybrid model can be a valuable and effective tool for decision processes, and time series decomposition to linear and nonlinear components has a better performance.

Keywords: Support Vector Machines, Cluster analysis, Nonlinear component, Configuration

INTRODUCTION

The complexity of precipitation leads to the need to a comprehensive model in order to forecast time series with high accuracy (Wang *et al.*, 2007; Wang *et al.*, 2014). The stochastic models based on Box and Jenkins (1976) method with widespread usage for time series forecasting were considered in many studies (Liang, 2009; Narasimha Murthy *et al.*, 2018). Among the stochastic models, seasonal autoregressive integrate moving average (SARIMA) can capture the seasonality of time series (Cryer and Chan, 2008; Box *et al.*, 2015). Despite the successes of the SARIMA model, the improvement of model to reach the optimal forecasts always has been considered (Wang *et al.*, 2014; Mo *et al.*, 2018). Two groups of univariate time series forecasting methods, stochastic models such as BATS, ARIMA_f and machine learning methods such as support vector machine (SVM) were utilized to forecast geophysical process. The average of median of absolute errors (mean) for all

[1] Faculty of Agriculture, Azarbaijan Shahid Madani University, Tabriz, Islamic Republic of Iran.
[*] Corresponding author; e-mail: Laleh_parviz@yahoo.com

[DOR: 20.1001.1.16807073.2020.22.2.19.7]

stochastic models and machine learning algorithms within the simulation experiment were 0.88 and 0.82, respectively (Papacharalampou *et al.,* 2018a). To investigate the seasonality of time series, temperature and precipitation time series were employed. Some of the used models were Autoregressive Fractionally Integrated Moving Average, and random walk. Seasonal decomposition with efficient method is most important in the modeling process such that the classical compared to automatic seasonal decomposition used by the BATS and Prophet decreased the error (Papacharalampou *et al.,* 2018b). Eleven stochastic models such as ARMA, ARFIMA and nine machine learning methods such as SVM were used for the multi-step ahead forecasting of river discharge. Median analysis of the dimensionless metrics showed that the average MAPE for all machine-learning algorithms was 24.32 and for all stochastic models was 25.97 (Papacharalampous *et al.,* 2019). The hybrid models with integration of the different models can be a good candidate in order to find the reliable results (Lee *et al.*, 2018; Mo *et al.,* 2018; Zhang *et al.*, 2018). A review of researches indicated that the combination of SARIMA model with machine learning algorithms can be done with two concepts: time variation analysis and time series decomposition (Chen and Wang, 2007; Wang *et al.,* 2014; Zeynoddin *et al.,* 2018). The interannual variations of monthly data were considered in the SARIMA model process, but the influence of intermonthly variations within each year was not considered. Therefore, the comprehensive analysis with emphasis on the time variation can help to improve the time series forecasts. The improvement of SARIMA model, ISARIMA, with time variation analysis in China decreased the mean absolute error from 11.49 to 9.41 mm (ISARIMA) (Wang *et al.,* 2014). The second concept of SARIMA-based hybrid model is related to time series decomposition to linear and nonlinear component of time series. The assumption of linear models such as SARIMA model is the linearity of time series while the hydrological process has a nonlinear structure (Tealab *et al.*, 2017). The linear nature of the SARIMA model leads to the problems in accurate modeling of complex nonlinear time series (Khandelwal *et al.,* 2015; Rathod and Mishra, 2018). Therefore, two components of time series should be taken into account, which leads to proposing several hybrid approaches (Chen and Wang, 2007; Yolcu *et al.,* 2013). Monthly rainfall forecasting with time series decomposition in tropical climate showed the improvement of rainfall predictions where the mean determination coefficient of the scenario for hybrid model was reported as 0.98 (Zaynoddin *et al.,* 2018). A combination of SARIMA and SVM (as hybrid) was used to forecast the production values. The minimum values of normalized mean square error among the hybrid, SARIMA and SVM models was related to the hybrid model (Chen and Wang, 2007). The combination of SARIMA and SVM models in the study of Ruiz-Aguilar *et al.* (2014) and Lee *et al.* (2018) rather than the single methods improved the inspection volume and atmospheric pollution forecasting, respectively. Two large datasets of short times have been used to evaluate the efficiency of random forests performance related to the variable selection in one step forecasting. Random forest can be candidate of machine learning algorithm. The first database was related to the simulated time series from a number of ARFIMA models and the second was related to time series of annual temperature. Lagged predictor variable with low number could increase the performance of random forests (Tyralis and Papacharalampous, 2017). Also, SARIMA and artificial neural network (ANN) were utilized as a hybrid model to forecast production value of the mechanical industry and the volume of passenger flows, and annual energy cost budget, respectively. The results showed the better performance of hybrid models compared to the conventional model (Jeong *et al.,* 2014; Glisˇovic´ *et al.,* 2016). Two machine learning algorithms,

[DOR: 20.1001.1.16807073.2020.22.2.19.7]

Neural Networks (NN) and SVM, were applied to forecast temperature and precipitation time series in 50 single case studies in Greece. The comparison of the study was conducted between machine learning algorithms and four classical algorithms. The median index of agreement for precipitation forecast was increased by 5.97% from NN to SVM. Also, the minimum values of RMSE (median) between stochastic and machine learning methods were related to SVM (Papacharalampous *et al.*, 2018c).

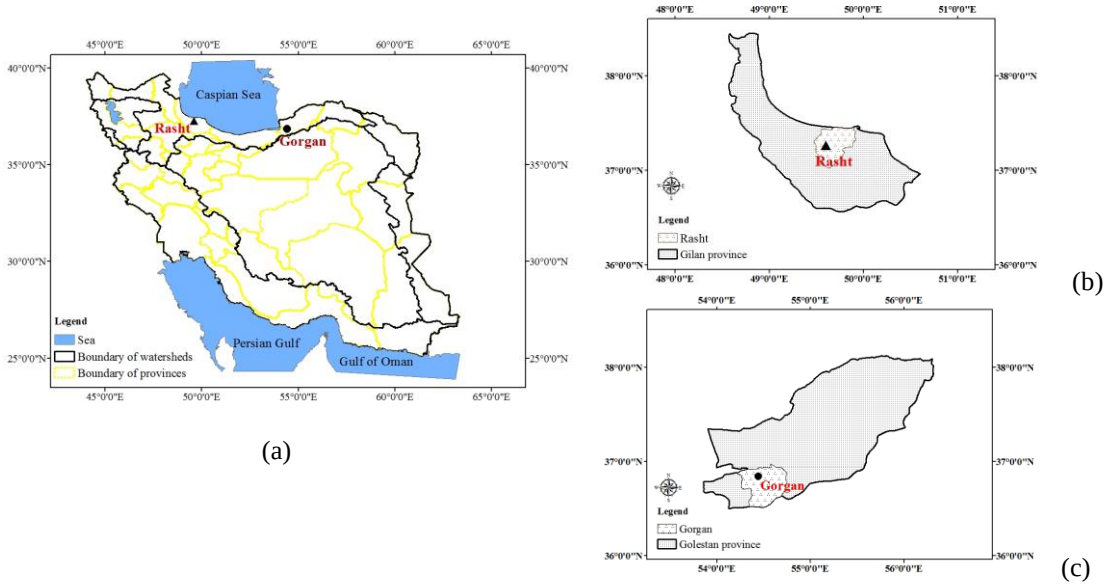
The main objective of the present study was to evaluate the performance of hybrid SARIMA and machine learning techniques based on the time variation analysis and time series decomposition concepts. The comparison of two concepts was conducted with the comprehensive evaluation criterion. Interannual and intermonthly variations of monthly time series were investigated with clustering analysis and then, ANN and SVM models were applied to model the internal process of ISARIMA model. Also, the combination of SARIMA and SVM models was considered to model the linear and nonlinear components of time series (with different configurations of residual). Eventually, the hybrid model performance was evaluated for the seasonal time series.

MATERIALS AND METHODS

The used monthly precipitation time series for evaluating the performance of hybrid models are related to Rasht and Gorgan stations. The provinces of stations are Gilan and Golestan, where the station's location is shown in Figure 1. Two monthly time series were applied to build model during 1976-2013 and to evaluate model from 2014 to 2016. The climate of stations is very wet and Mediterranean, based on the De Martonne climate classification method.

SARIMA Model

Assimilation of autoregressive and moving average term with integration term led to ARIMA model construction (Ruiz-Aguilar *et al.*, 2014). ARIMA model was extended to SARIMA due to the drawbacks of the ARIMA to model the seasonal time series (Lee *et al.*, 2018). ARIMA modification to SARIMA is conducted to consider the seasonality pattern



(a) (b) (c)

Figure 1. The position of Rash and Gorgan stations in Iran (a) and in the related provinces; Gilan (b) and Golestan (c)

[DOR: 20.1001.1.16807073.2020.22.2.19.7]

of time series; therefore, the basis of SARIMA model is related to ARIMA model (Lee *et al.,* 2018; Mo *et al.,* 2018). The SARIMA model consists of several steps with the stationary check, identification and estimation, diagnostics and prediction (Jeong *et al.,* 2014; Jadhav *et al.,* 2017). The SARIMA model with the denotation of SARIMA $(p,d,q)(P,D,Q)_s$ is described as Eq. 1.

$$\phi_p(B)\Phi_P(B^s)\,(1-B)^d(1-B^s)^D z_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \quad (1)$$

Where, z_t is expression of time series, B is the lag operator, s is the seasonal period length, p and q are the orders of autoregressive and moving average section, respectively, P and Q are the orders of seasonal autoregressive and moving average, respectively, d is the number of differencing operation, D is the number of seasonal differencing, ε_t is a white noise, $\phi_p(B)=1-\phi_1B-\ldots-\phi_pB^p$, is the regular autoregressive operator of order p, $\theta_q(B)=1-\theta_1B-\ldots-\theta qB^q$, is the regular moving average operator of order q, $\Phi_P(B^s)=1-\Phi_1B^s-\ldots-\Phi_pB^{sP}$, is the seasonal autoregressive operator of order P, $\Theta_Q(B^s)=1-\Theta_1B^s-\ldots-\Theta_QB^{sQ}$, is the seasonal moving average operator of order Q (Bas *et al.,* 2017).

SARIM Based Hybrid Model with Time Variation Analysis

The interannual variations of monthly time series are considered with SARIMA model, but the shortcoming of the model is related to the intermonthly variations. The basis of the improvement is the clustering analysis to identify the structures within the data. The relationship between clustering process and monthly time series is defined with modeling the main statistics of each cluster with the associated time series with linear regression. Then, the main statistics of each cluster in the validation period are achieved with ARIMA model and the forecasted values are substituted in the regression model to achieve monthly time series (Wang *et al.,* 2014). Yolcu *et al.* (2013) stated that modeling of time series with nonlinear models had more accurate and effective forecasts when the nonlinearity component of time series is superior to the linearity part. Wang *et al.* (2014) used the linear regression in the internal process of ISARIMA, therefore, in order to overcome the limitation of the proposed model, machine learning algorithms were used in this study (novel part of the study), and the methods used are explained in the following sections.

ANN Structure

ANN is an abstract computational model of human brain. The characteristics of a node and the node's connectivity in the network can build ANN architecture. (Weng *et al.,* 2016). The structure of model is a network with three layers of simple processing units connected by acyclic links (Zhang, 2003). It should be stated that the network may contain several intermediary layers between the input and output layers. Such intermediary layers are called hidden layers and the nodes embedded in these layers are called hidden nodes. The most common type of neural network consists of three layers (Weng *et al.,* 2016). Equation 2 can be defined to explain the relationship between the input and output variables.

$$y_t = \alpha_0 + \sum_{j=1}^{q}\alpha_j\, g\,(\beta_{oj} + \sum_{i=1}^{p}\beta_{ij}\, y_{t-i}) + \varepsilon_t \quad (2)$$

Where, y_t is output, y_{t-1}, $y_{t-2,\ldots,}$ y_{t-i} are inputs, α_j (j=0,1,2,…,q) and β_{ij} (i=0,1,2,…,p) are model parameters, p is the number of nodes and q is the number of hidden nodes(Zhang, 2003).

SVM Structure

In order to map the input data x into a higher-dimensional feature space F by nonlinear mapping, SVM can be applied. The regression approximation can be

[DOR: 20.1001.1.16807073.2020.22.2.19.7]

explained as the function estimation based on a given data set $G = \{(x_i, d_i)\}_i^n$, x_i represents the input vector, d_i is related to the desired value, n is the total number of data patterns. Therefore, the function of Eq. 3 can be an approximation of the regression function (Chen and Wang, 2007).

$$f(x) = \omega\phi(x) + b,$$
$$\phi: R^n \to F, \omega \in F, \quad (3)$$

Where, b is a scalar threshold; $\phi(x)$ is the high dimensional feature space, ω is the coefficient.

The coefficients ω and b can be estimated with minimizing (Chen and Wang, 2007).

$$R_{SVM}(C) = R_{emp} + \frac{1}{2}\|\omega\|^2 = C \times \frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(d_i, y_i) + \frac{1}{2}\|\omega\|^2 \quad (4)$$

$$L_\varepsilon(d, y) = \begin{cases} |d - y| - \varepsilon & |d - y| \geq \varepsilon \\ o & otherwise \end{cases} \quad (5)$$

The term of $C \times \frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(d_i, y_i)$ in Eq. 4 can be representative of empirical error (risk) which its estimation is related to the e-insensitive loss function in Eq. (5). In order to get samples of the decision function in Eq. (3) with fewer data points, the loss function has been applied (Chen and Wang, 2007). Faced with error, the regularized constant C calculates the penalty with determination of trade-off between the empirical risk and the regularization term and that can help to improve the prediction of regression (Chen and Wang, 2007).

SARIMA Based Hybrid Model with Time Series Decomposition

SARIMA and machine learning algorithms have capabilities to model the linear and nonlinear problems. Modeling the complex nonlinear time series with SARIMA model may not be sufficient (Zhang, 2003). The procedure of hybrid model can be described by the time series, which can be comprised of the linear autocorrelation structure and a nonlinear component with the general form that is expressed as Eq. 6.

$$y_t = L_t + N_t \quad (6)$$

Where, L_t is representative of linear characteristic from SARIMA model, N_t denotes the non- linear characteristic (Lee *et al.*, 2018).

Two factors of equation 6 can be estimated from the data. The first step of hybrid model is related to apply SARIMA to model the linear component and then the residual of the linear model has the nonlinear relationship. The residual derived from SARIMA model can be explained as Eq. 7.

$$\varepsilon_t = y_t - \hat{L}_t \quad (7)$$

Where, $\hat{L}_t$ is representative of the forecasted value of SARIMA model at time t.

The residual analysis is important to find the nonlinear pattern of data (as the second step). The different configuration of residuals modeling can be expressed as equations 8-11.

$$\varepsilon_t = f(\varepsilon_{t-1}, \varepsilon_{t-2}, ..., \varepsilon_{t-n}) + e_t \quad (8)$$
$$\varepsilon_t = f(\varepsilon_{t-1}, \varepsilon_{t-12}) + e_t \quad (9)$$
$$y_t = f(y_{t-1}, y_{t-12}, \hat{L}_t) + e_t \quad (10)$$
$$y_t = f(y_{t-1}, y_{t-12}) + e_t \quad (11)$$

Where, f is the nonlinear function which can be achieved by machine learning techniques, e_t is the random error.

Equations 8 and 9 can be identified as $\hat{N}_t$, therefore, the forecasted values can be achieved by summation of linear and nonlinear components (Lee *et al.*, 2018; Zhang, 2003). Figure 2 shows the functional flowchart of two hybrid models.

Evaluation Criterion

In order to evaluate the efficiency of hybrid models, several statistical tests were used in the validation period (Adamowski and Karapataki, 2010; Niedbala and

[DOR: 20.1001.1.16807073.2020.22.2.19.7]

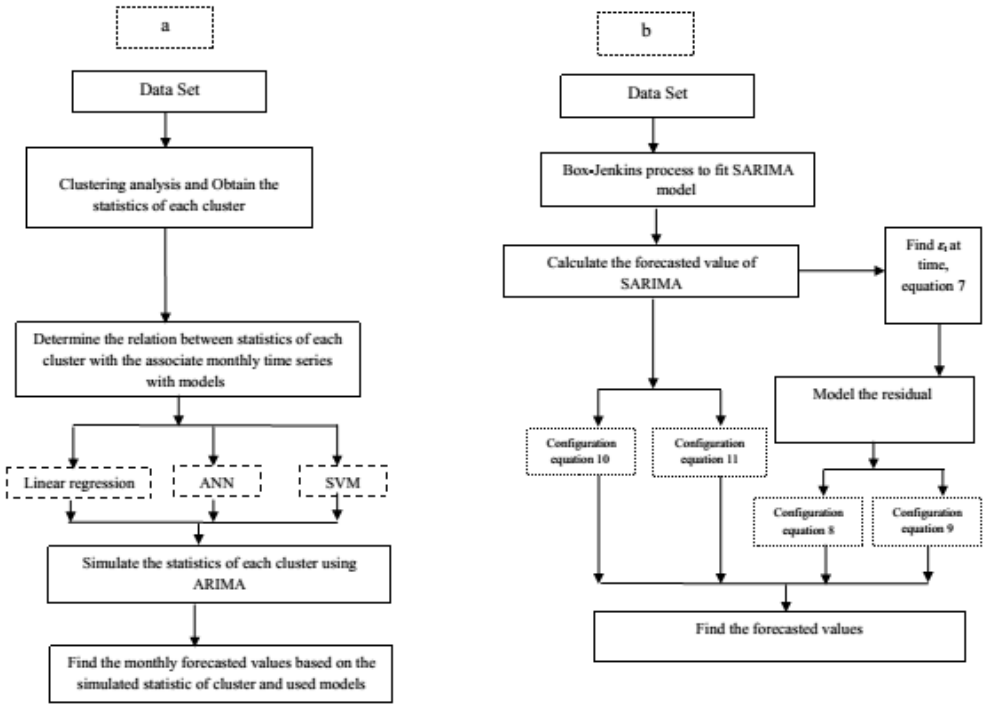


Figure 2. Schematic structure of two hybrid models; time variation analysis (a) and time series decomposition (b)

Kozlowski, 2019; Taghadomi-Saberi and Razavi, 2019).

$$RMSE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(O_i - P_i)^2} \quad (12)$$

Root Mean Square Error (13)

$$RRMSE = \frac{RMSE}{\overline{O}}$$

Relative Root Mean Square Error

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - O_i|$$

Mean Absolute Error (14)

(15)

$$MRE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{P_i - O_i}{O_i}\right|$$

Mean Relative Error

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{P_i - O_i}{O_i}\right| \times 100$$

Mean Absolute percentage Error (17)

$$d = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)}{\sum_{i=1}^{N}(|P_i - \overline{O}| + |O_i - \overline{O}|)^2}$$

Index of Agreement

$$dm = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)}{\sum_{i=1}^{N}(|P_i - \overline{O}| + |O_i - \overline{O}|)} \quad (18)$$

Modified Index of Agreement

$$UI = \frac{\left[\sum_{i=1}^{N}(P_i - O_i)^2\right]^{0.5}}{\left[\sum_{i=1}^{N}(O_i)^2\right]^{0.5} + \left[\sum_{i=1}^{N}(P_i)^2\right]^{0.5}} \quad (19)$$

$$UII = \frac{\left[\sum_{i=1}^{N}(P_i - O_i)^2\right]^{0.5}}{\left[\sum_{i=1}^{N}(O_i)^2\right]^{0.5}} \quad (20)$$

Geometric Mean Error Ratio

$$GMER = \exp\left[\frac{1}{N}\sum_{i=1}^{N}\ln\left(\frac{P_i}{O_i}\right)\right] \quad (21)$$

Where, P_i is the predicted value, O_i is the observed value, $\overline{O}$ is the mean of observed value, i represents the time, N is sample size of validation period.

The smaller values of RMSE, RRMSE, MAE, MRE, MAPE, UI, UII and proximity of d to one is indicative of better performance of the model. Sensitivity of RMSE to outliers rather than MAE is high (Papacharalampous *et al.*, 2019). 0%<MAPE<10% is indicative of very accurate prediction and MAPE greater than 50% is indicative of inaccurate forecast (Lee *et al.*, 2018). MAPE can be defined as

[DOR: 20.1001.1.16807073.2020.22.2.19.7]

scale-independent criteria, which can be an advantage for model comparison across different dataset (Papacharalampous *et al.*, 2019). GMER greater than 1 indicates overestimation and GMER <1 shows underestimation. To evaluate the accuracy of forecasting, the used criterion is denoted as UI and the quality of forecasting is denoted as UII. UI= 0 and UII=0 are indicative of prefect forecast (Zeynoddin *et al.* 2018).The values of agreement index near 1 shows the better agreement of simulated and observed values. For comparison of the performance of hybrid and single models, the accuracy improvement criteria was proposed with Eq. 22 (Chen and Zhu, 2013).

$$AI = \frac{S - S_h}{S} \times 100 \quad (22)$$

Where, S and S_h are the MAE of single and hybrid model. AI greater than 0 is indicative of the best performance of hybrid model, AI less or equal to 0 shows that hybrid model dose not outperform the single model (Chen and Zhu, 2013).

RESULTS

Two monthly time series from Rash and Gorgan stations were used to evaluate the hybrid models performance. The statistical characteristics of monthly time series in each station are presented in Table 1. The application of SARIMA and SARIMA-based hybrid model with the time variation analysis implemented to ANN and SVM models, and SARIMA-based hybrid model with time series decomposition implemented to SVM model were utilized to forecast monthly precipitation time series.

The Box and Jenkins methodology for model construction consist of some steps: Model identification, parameter estimation, and diagnostic checking (Zhang *et al.*, 2018). The first modeling step of SARIMA was examined with the autocorrelation function and seasonal Mann-Kendal test in order to find the randomness and seasonality of time series, respectively. The best-fitted SARIMA model are $SARIMA(3,0,2)\times(1,1,2)_{12}$ and $SARIMA(0,1,1)\times(1,1,2)_{12}$ for the Rasht and Gorgon stations, respectively.

SARIMA-Based Hybrid Model with Time Variation Analysis

Hierarchical clustering algorithm was applied for cluster analysis. From the popular agglomerative methods of hierarchical algorithm, Ward's method was used to divide the precipitation time series to different groups with similar hydrological patterns in a group. The results of clustering analysis are represented in a dendrogram, which is displayed in Figure 3. The y-axis of dendrogram shows the distance between the clusters (Liu and Ge, 2018), and the Euclidean distance was used. The x-axis can represent the objective of hierarchical clustering, which is the monthly precipitation.

The results of clustering are: Rasht station; month 1, 10 (cluster 1), month 5,6,11

Table 1. Some statistical characteristics of time series (period of training: 1976-2013, period of testing: 2014 2016).

statistical characteristics	Gorgan		Rasht	
	Training	Testing	Training	Testing
Mean (mm)	45.061	35.41	112.054	101.6
Standard deviation (mm)	31.7	22.036	94.05	79.67
Minimum	0	2.5	0	3.2
Maximum	166.8	98.6	601.4	370.2
Skewness	0.748	0.69	1.46	1.34
Kortosis	0.24	0.488	2.93	2.56
S.E. mean	1.55	3.83	4.4	13.27
Coefficient of variation	0.7	0.62	0.83	0.78

[DOR: 20.1001.1.16807073.2020.22.2.19.7]

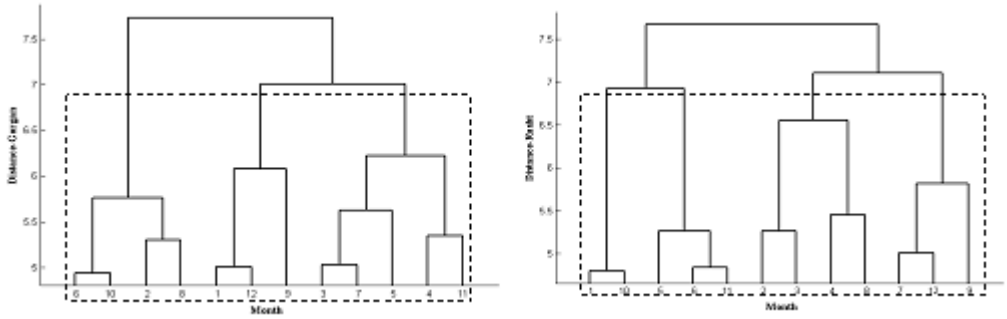


Figure 3. Dendrograms of hierarchical clustering for monthly precipitation time series in Gorgan and Rasht stations.

(cluster 2), month 7, 9, 12 (cluster 3) and month 2, 3, 4, 8 (cluster 4)-Gorgan station; month 2, 6, 8, 10 (cluster 1), month 1, 9, 12 (cluster 2) and month 3,4, 5, 7, 11 (cluster 3). Linear regression was used in the internal process of ISARIMA to model the statistics of each cluster (maximum, minimum, truncated mean) with the associated time series of each cluster in the Wang *et al.* (2014) study. In order to consider the nonlinearity relationship between the statistics of each cluster and the monthly precipitation time series, ANN and SVM models were applied in this study. The sensitivity analysis of SVM was related to kernel functions and penalty parameter. The investigated kernel functions were linear, polynomial, Gaussian radial basis and sigmoid functions. The used network of this study is back propagation neural network algorithm. ANN sensitivity analysis was related to the activation functions of hidden-output layers and number of neurons. The used activation functions were logistic sigmoid, tangent sigmoid, and pure linear. Decrease in RMSE from logistic sigmoid-pure linear to tangent sigmoid-logistic sigmoid is 20.23% in Gorgan station. ANN and SVM performance comparison is shown in Figure 4.

The clustering analysis can improve the forecasted values; for example, decrease in RMSE and MRE from SARIMA model to ISARIMA is 21.29% and 70.27% in Rasht and 7.83% and 13.77% in Gorgan station, respectively. The performance of SARIMA model implemented to clustering analysis improved the evaluation criterion rather than the SARIMA model. The clustering analysis significantly improved precipitation simulation rather than the SARIMA model with the increase of forecasting accuracy to 21% (Wang *et al.*, 2014). The evaluation criterion comparison indicated that the optimum case of

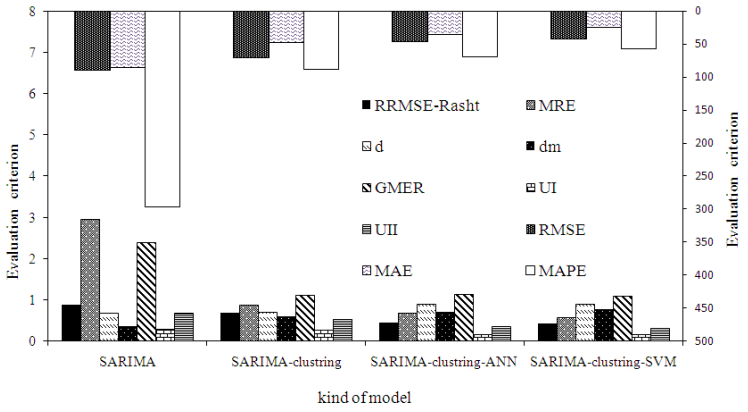


Figure 4. The evaluation criterion related to the comparison of ANN and SVM models performance in Rasht station with hybrid model based on time variation analysis

[DOR: 20.1001.1.16807073.2020.22.2.19.7]

criteria is related to the SARIMA model based on clustering analysis with implementation to SVM model. For example, the RMSE, RRMSE, MAE and GMER decrease from SARIMA to ISARIMA-SVM model is 52.4%, 52.29%, 71.05% and 54.16% in Rasht station. It should be noted that in each station and for all models, GMER is greater than one, which is indicative of overestimation. SVM model based on the evaluation criterion has better results than the ANN model. The evaluation criterion such as RMSE and efficiency coefficient indicated the outperformance of SVM compared to ANN model for precipitation prediction in Hamadan station (Hamidi *et al.*, 2014). The evaluation criterion such as RMSE, MAE and efficiency coefficient indicated that SVM outperformed ANN model. Therefore, clustering analysis and the application of machine learning techniques for internal process of SARIMA improved the accuracy of precipitation simulation. The performance of accuracy improvement for comparing the hybrid model and single model performance is shown in Figure 5.

The accuracy improvement of models is positive in Figure 5, and they are in the range of better performance of the model and the maximum value of accuracy improvement is related to the ISARIMA model with implementation to SVM model.

SARIMA-Based Hybrid Model with Time Series Decomposition

The function of hybrid model in this section was approximated with SARIMA model and the machine learning techniques with consideration of the linearity and nonlinearity of precipitation time series. The SARIMA-based hybrid model with time series analysis implemented to SVM model had the minimum error and better performance in the previous section and that is the reason for SVM selection in hybrid model based on time series decomposition. Therefore, the hybrid model is a combination of SARIMA and SVM models. Different configurations were utilized with the composition of residuals, linear component, and predicted values. The results of comparison related to the different configurations are listed in Table 2 (C in Gorgan is 0.01 and in Rasht is 2).

The configuration composed of residuals has the minimum error with high d in Table 2. For example, RMSE and MRE decrease from configuration with time series and residual to configuration with residuals are, respectively, 44.17% and 51.19% in Gorgan station. Lee *et al.* (2018) used SARIMA-SVM for atmospheric pollution forecasting, and the nonlinear model was the composition of residuals. The sensitivity analysis in each model is one of the most important steps. For example, in Gorgan station for the forth configuration, the RMSE with linear kernel function decreased 10.62% with C=1 relative to the radial basis function with C=2. The sensitivity analysis has a more important role in the model performance; for example, RMSE decrease from linear kernel function (C= 0.01) to

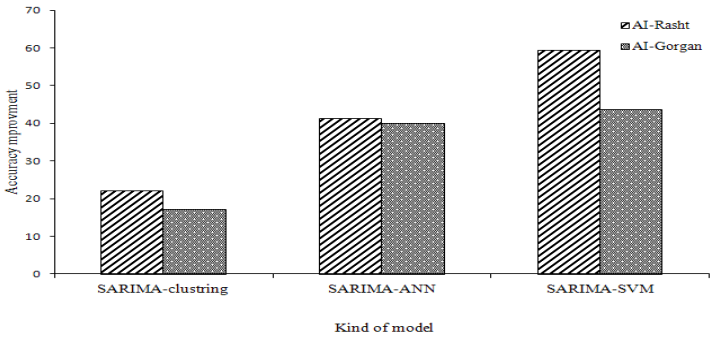


Figure 5. The accuracy improvement of SARIMA model and hybrid model based on time series analysis implement to ANN and SVM models.

[DOR: 20.1001.1.16807073.2020.22.2.19.7]

Table 2. Comparison of the used configurations of SARIMA-based hybrid model with time series decomposition.

Station	Gorgan							Rasht						
Evaluation criteria	RMSE	RRMSE	MRE	MAE	d	ui	uii	RMSE	RRMSE	MRE	MAE	d	ui	uii
$\varepsilon_t = f(\varepsilon_{t-3}, \varepsilon_{t-2}, \varepsilon_{t-1}) + e_t$	18	0.5	0.94	12.94	0.74	0.22	0.43	45.62	0.44	0.79	31.1	0.89	0.17	0.35
$\varepsilon_t = f(\varepsilon_{t-1}, \varepsilon_{t-12}) + e_t$	14.18	0.39	0.82	9.55	0.84	0.16	0.34	37.48	0.36	0.46	24.25	0.93	0.15	0.29
$y_t = f(y_{t-12}, y_{t-1}, \varepsilon_t) + e_t$	25.4	0.7	1.68	20.07	0.41	0.27	0.6	48.72	0.47	0.78	31.08	0.85	0.21	0.37
$y_t = f(L_t, y_{t-12}, y_{t-1}) + e_t$	23.88	0.665	1.57	18.93	0.43	0.26	0.57	49.66	0.48	0.82	32.8	0.84	0.21	0.38
$\varepsilon_t = f(\varepsilon_{t-1}) + e_t$	18.25	0.5	0.92	13.14	0.73	0.22	0.43							

sigmoid kernel function (C= 2) was 45.41% in Rasht station. According to the best configuration of Table 2, two hybrid models were compared with some evaluation criterion, as shown in Figure 6.

The hybrid model with time series decomposition has the minimum error and better performance. The RMSE, MRE, and UI decreases of hybrid-model-based time variation to hybrid-model-based time series decomposition are 11.14%, 19.29%, and 13.29% for Rasht time series. The related values of Gorgan station are 18.73%, 5.74%, and 15.78%. UI and UII of SARIMA-based hybrid model with time series decomposition reached 0, which is indicative of forecasting accuracy and quality improvement. Time series decomposition leads to decrease in MAPE, which is indicative of accurate prediction. Also, the index of agreement and modified index of agreement for hybrid-model-based time series decomposition are increased close to 1. Time series decomposition to linear and nonlinear components led to the improvement of time series simulation in many studies such as Ruiz-Aguilar *et al.* (2014), Chen and Wang (2007), and Lee *et al.* (2018) for forecasting inspection volume, production values of machinery industry, and atmospheric pollution, respectively. GMER in all models and all stations are greater than one, which is indicative of overestimations of forecasted values. The higher and positive values of AI are related to the SARIMA-based hybrid model with time series decomposition, which indicates the outperformance of hybrid model rather than the single model. The monthly comparison indicated that minimum

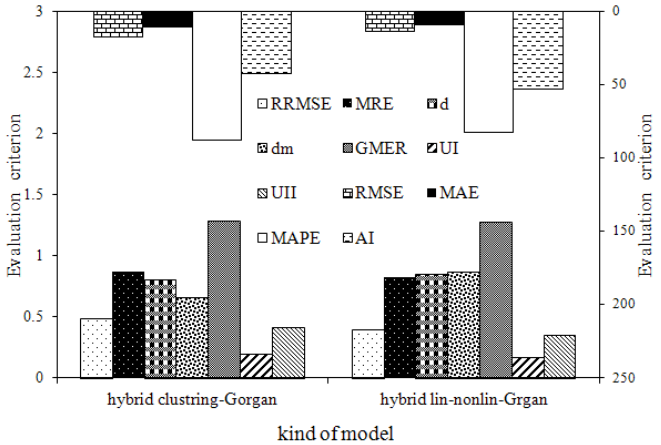


Figure 6. Comparison the performance of hybrid models with evaluation criterion in Gorgan station.

[DOR: 20.1001.1.16807073.2020.22.2.19.7]

RRMSE of months is related to November in Rasht (RRMSE= 0.08) and October (RRMSE= 0.11) in Gorgan. It can be stated that months with high precipitation have low error ($RRMSE_{June\text{-}Rasht}$=0.62; $RRMSE_{June\text{-}Gorgan}$=0.46). The monthly observed and simulated precipitation time series with two hybrid models are shown in Figure 7.

The R^2 of the fitted line comparison demonstrates its increase from SARIMA-based hybrid model with time variation analysis to SARIMA-based hybrid model with time series decomposition, which is indicative of better performance of the latter model. The observed precipitation values of Rasht is greater than Gorgan station and the performance evaluation of SARIMA-based hybrid model with time series decomposition shows that RRMSE and MRE of Rasht is lower than Gorgan station. The minimum precipitation values of Rasht station in the validation period occurred in June 2014, where the minimum value of forecasted precipitation time series using SARIMA-based hybrid with time variation analysis and SARIMA-based hybrid with time series decomposition occurs on that date. The maximum precipitation value of Gorgan station in the validation period occurs in March of 2014 and the minimum value of forecasted precipitation time series with the two hybrid models occurs on that date. Furthermore, the performance of hybrid model was investigated in the seasonal time step, and the evaluation criterion comparison and R^2 of the fitted line are shown in Figure 8.

Results of the seasonal time series forecasting with SARIMA-based hybrid using time series decomposition have higher accuracy, where the index of agreement in Rash and Gorgan stations are, respectively, 0.97 and 0.91, both in the optimum range of criteria. The R^2 of fitted line increases with time series decomposition to linear and nonlinear components. For example, the R^2 of fitted line in Rasht station increased from 0.845 to 0.924 and from 0.639 to 0.792 in Gorgan station. The GMER is greater than 1 for the two hybrid models, which indicates the overestimation of the forecasted time series. MAPE has lower values, which can be indicative of accurate estimation. For air quality forecasting, the hybrid model was able to process not only month or a season, but also the whole year (Diaz-Robles *et al.*, 2008).

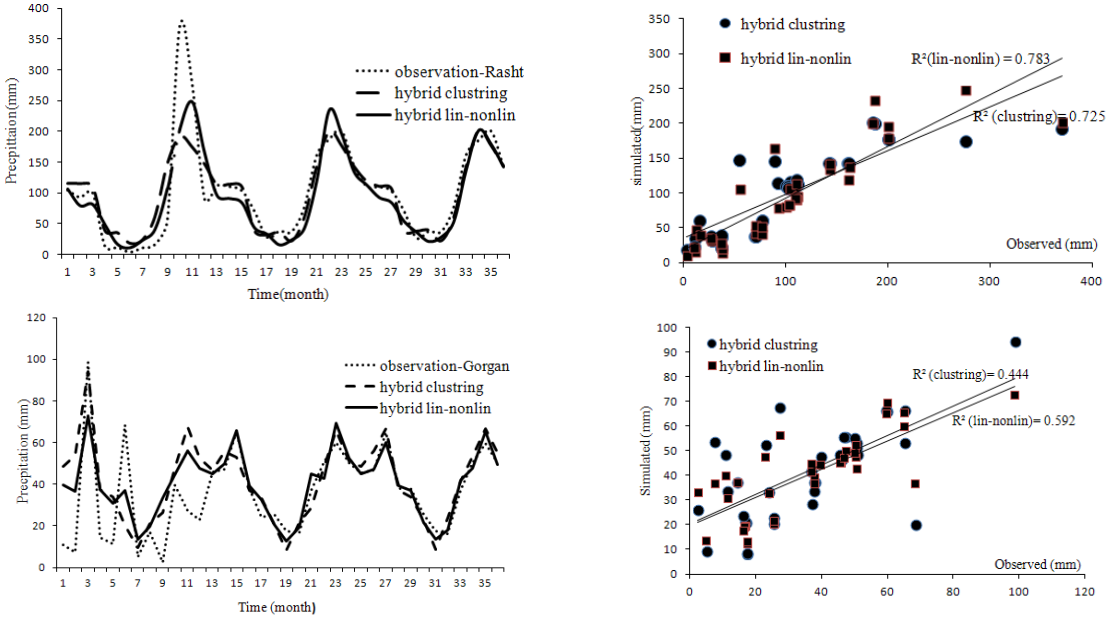


Figure 7. Monthly observed and simulated precipitation time series with different models (a), and scatter plot of simulated and observed data (b).

[DOR: 20.1001.1.16807073.2020.22.2.19.7]

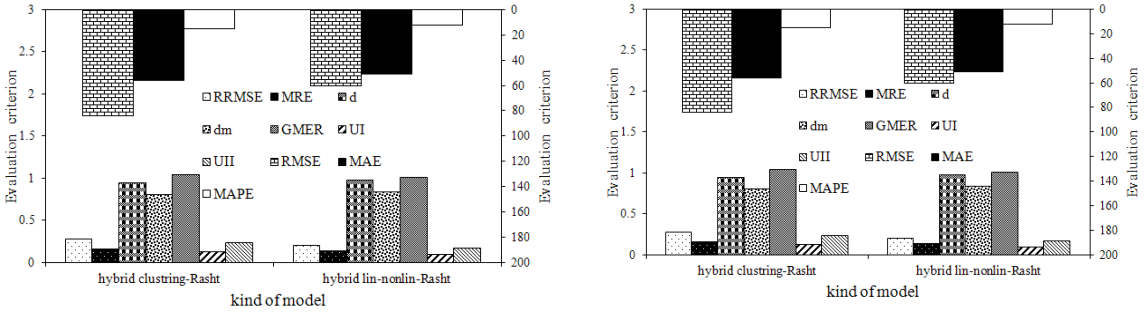


Figure 8. Comparison of the seasonal performance of models with evaluation criterion.

DISCUSSION

The results indicated that the clustering analysis could improve the SARIMA model performance because more information can be extracted from the used time series. Also, the time series with similar hydrological characteristics are grouped in a cluster that led to high performance of SARIMA based on a systematic grouping. In order to find the similarity among time series, clustering analysis is the main subject. The measurement of the distance in the space determined by observed time series can be used to find the similarity of cluster members. The ISARIMA model implement to SVM and ANN models had better results compared to ISARIMA model implement to linear regression. Linear regression can consider the relationships of a pre-specified functional form, and linear regression may not be sufficient for accurate prediction in order to model the nonlinear nature of time series (Adamowski and Karapataki, 2010). SVM model had better performance for precipitation time series simulation compared to ANN model. Comparing the performance of NN and SVM for case studies with high number was conducted and, in most cases, SVM had better results (Papacharalampous *et al.*, 2018c). The better performance of machine learning algorithms was seen in the study of Tyralis and Papacharalampous (2017) with two large datasets. The success of SVM model can be related to the principle of structural risk minimization instead of empirical risk minimization, as the other techniques such as ANN (Ruiz-Aguilar *et al.*, 2014). Also, the generalization capability for relating the input to the desired output is more noticeable and some advantages of SVM models have low sensitivity to small training sets and noisy data and the avoidance of over fitting. SVM is a powerful option in nonlinear time series, especially with an unknown distribution (Hamidi *et al.*, 2014). Adopting Kernal by SVM models led to the increase in the model efficiency in nonlinearity of time series modeling (Naguib and Darwish, 2012). Overfitting can be seen for ANN model when training takes too long time. For any pattern, this means that a model should be used to consider the noise as part of the pattern, but this problem cannot be seen for SVM model (Selvanayaki and Somasundaram, 2015). The results show that SVM sensitivity analysis is one of the most important steps, and Hamidi *et al.* (2014) stated that SVM model can be a robust model if appropriate Kernal function and related parameters are selected. Generally, the evaluation criterion comparison shows that the two hybrid models improved the accuracy of forecasts compared the SARIMA model. Chen and Wang (2007) indicated that the hybrid models performance are superior to the individual models in terms of both prediction error and directional change detestability. Ruiz-Aguilar *et al.* (2014) indicated that hybrid model based on artificial intelligent systems was an effective tool for powerful decision making. The

[DOR: 20.1001.1.16807073.2020.22.2.19.7]

model basis on the linearity and nonlinearity of time series by taking advantages of two models (in term of decomposition time series in linear and nonlinear components) can improve the model efficiency. Time series with monthly variation has linear and nonlinear component and this hybrid method can be effective. Compassion of the two hybrid models indicated that applying monthly time series was more accurate than the statistics of monthly clusters. Therefore, decomposition of time series with an efficient method has more importance to forecast climatological parameters, which was proved in the study of Papacharalampous *et al.* (2018b) with large number of samples. The comparison of stochastic models and machine learning algorithms in the study of Papacharalampous *et al.* (2018a), Papacharalampous *et al.* (2019) and Papacharalampous and Tyralis (2018) for large scale studies in most cases showed the high performance of machine learning algorithms. However, it should be noted that the scale is more important. According to the research of Papacharalampous *et al.* (2018a), the different stochastic models had dissimilar error criteria. Tyralis and Papacharalampous (2018) used two approaches: past information of time series and using exogenous predictor variables alongside with the use of the endogenous ones, and the usefulness of the two approaches was proved. Rasht station, with high precipitation, had minimum RRMSE and MRE compared to Gorgan station. Also, in this study, eleven evaluation criterions were utilized for model evaluation and the trend of evaluation criterion is favorable. Each criterion investigates the performance of model from different aspects. For example, RMSE can measure the goodness of fit related to high precipitation and MAE can represent the goodness of fit belonging to moderate precipitation (Hamidi *et al.*, 2014). The coordination of all criterions is indicative of high performance of time series decomposition for hybrid models.

CONCLUSION

Accurate precipitation forecasting is always a challenging problem, which is more attractive in many fields. SARIMA model is one of the most popular and well-known models for precipitation forecasting. Therefore, in this study, two SARIMA-based hybrid models were compared, which could improve the efficiency of SARIMA model. Time variation analysis and time series decomposition are the two concepts that were used to construct the hybrid models. ANN and SVM models were applied to complete the internal process of ISARIMA instead of linear regression, which had been used in the previous studies, and application of machine learning algorithm is prominent point of the research. The ISARIMA model with implementation to ANN and SVM increased the accuracy of forecasting. The simulation of SARIMA-based hybrid model with time series decomposition reaches the observed values, rather than the hybrid model with time variation analysis. The reason for that can be related to the nature of monthly time series with the inclusion of linearity and nonlinearity of time series, which the hybrid model could extract the governing pattern of data with accuracy. SARIMA model has shortcomings for modeling the nonlinear component of time series and, therefore, SVM model can solve this problem. The mentioned hybrid model takes advantages of SARIMA and SVM models in linear and nonlinear modeling, which is appropriate for modeling complex phenomena. To achieve the highly accurate forecasts in the field of SARIMA-based hybrid model with time series decomposition, other decomposition methods such as empirical mode decomposition (EMD) and wavelet transform methods should be investigated.

REFERENCES

1. Adamowski, J. and Karapataki,. C.H. 2010. Comparison of Multivariate Regression and

[DOR: 20.1001.1.16807073.2020.22.2.19.7]

Artificial Neural Networks for Peak Urban Water-Demand Forecasting: Evaluation of Different ANN Learning Algorithms. *J. Hydrol. Engin.*, **15**: 729-743.
2. Bas, M., Ortiz, J., Ballesteros, L. and Martorell, S. 2017. Evaluation of a Multiple Linear Regression Model and SARIMA Model in Forecasting 7Be Air Concentrations. *Chemosphere*. **177**: 326-333.
3. Belaid, S. and Mellit, A. 2016. Prediction of Daily and Mean Monthly Global Solar Radiation Using Support Vector Machine in an Arid Climate. *Energy Convers. Manag.*, **118**: 105-118.
4. Box, G.E.P. and Jenkins, G.M. 1976. Times Series Analysis -Forecasting and Control. Prentice-Hall, Englewood Cliffs.
5. Box, G.E.P., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. 2015. Time Series Analysis: Forecasting and Control, John Wiley & Sons.
6. Chen, K. Y. and Wang, C.H. 2007. A Hybrid SARIMA and Support Vector Machines in Forecasting the Production Values of the Machinery Industry in Taiwan. *Expert Syst. App.*, **32**: 254-264.
7. Chen, X. and Zhu, S. 2013. Improved Hybrid Model Based on Support Vector Regression Machine for Monthly Precipitation Forecasting. *J. Comput.*, **8**(1): 232-238.
8. Cryer, J.D. and Chan, K.S. 2008. Time Series Analysis with Application in R, 2nd Ed. Springer, New York, p 491.
9. Diaz-Robles, L., Ortega, J.C., Fu, J.S., Reed, G.D., Chow, J.C., Watson, J.G. and Moncada-Herrera, J.A. 2008. A Hybrid ARIMA and Artificial Neural Network Model to Forecast Particular Matter in Urban Areas: The Case of Temuco-Chile. *Atmos. Environ.*, **42**: 8331-8340.
10. Du, J., Yayun, L., Yu, Y. and Yan, W. 2017. A Prediction of Precipitation Data Based on Support Vector Machine and Particles Swarm Optimization (PSO-SVM) Algorithms. *Algorithm*, **10**(75): 1-15.
11. Glisˇovic´, N., Milenkovic´, M., Bojovic´, N., Sˇvadlenka, L. and Z. Avramovic´. 2016. A hybrid model for forecasting the volume of passenger flows on Serbian railways. *Operational Res.*, **16**: 271-285.
12. Hamidi, O., Poorolajal, J., Sadeghifar, M., Abbasi, H., Maryanaji, Z., Faridi, H. R. and L. Tapak. 2014. A comperative study of support vector machines and artificial neural network for predicting precipitation in Iran. Theor. *Appl. Climatol.*, **119**: 723-731.
13. Jadhav, V., Chinnappa Reddy, B. V. and G. M. Gaddi. 2017. Application of ARIMA model for forecasting agricultural prices. *J. Agr. Sci. Tech.*, **19**: 981-992.
14. Jeong, K., Koo, C., and T. Hong. 2014. An estimation model for determination the annual energy cost budget in educational facilities using SARIMA (seasonal autoregressive integrated moving average) and ANN (artificial neural network). *Energy*, **71**: 71-79.
15. Khandelwal, I., Adhikari, R., and G. H. Verma. 2105. Time series forecasting using Hybrid ARIMA and ANN models based on DWT decomposition. *Procedia Comput. Sci.*, **48**: 173-179.
16. Lee, N-UK, Shim, J-S., Ju, Y. W. and S-Ch. Park. 2018. Design and implementation of the SARIMA-SVM time series analysis algorithm for the improvement of atmospheric environment forecast accuracy. *Soft Comput.*, **22**: 4275–4281.
17. Liang, Y. U. 2009. Combining seasonal time series ARIMA method and neural networks with genetic algorithms for predicting the production value of the mechanical industry in Taiwan. *Neural Comput. App.*, **18**: 833-841.
18. Liu, Y, and Z. Ge. 2018. Weighted random forests for fault classification in industrial processes with hieratchical clustering model selection. *J. Process Control*, **64**: 62-70.
19. Mo, L., Xie, L., Jiang, X., Teng, G., Xu, L. and J. Xiao. 2018. GMDH-based hybrid model for container throughput forecasting: selective combination forecasting in nonlinear subseries. *App. Soft Comput.*, **62**: 478-490.
20. Naguib, I. A. and H. W. Darwish. 2012. Support vector regression and artificial neural network models for stability indicating analysis of mebeverine hydrochloride and sulpiride mixtures in pharmaceutical preparation: A comparative study. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. *Spectrochimica Acta Part A.*, **86**: 515-526
21. Narasimha Murthy, K. V., Saravana, R. and K. Vijaya Kumar. 2018. Modeling and forecasting rainfall patterns of southwest monsoons in North–East India as a

[DOR: 20.1001.1.16807073.2020.22.2.19.7]

SARIMA process. *Meteorol. Atmos. Phys.*, **130**: 99-106.
22. Niedbala, G. and R. J. Kozlowski. 2019. Application of Artificial Neural Networks for Multi-Criteria Yield Prediction of Winter Wheat. *J. Agr. Sci. Tech.*, **21**: 51-6.
23. Papacharalampous, G., Tyralis, H. and D. Koutsoyiannis. 2018a. One-step ahead forecasting of geophysical processes within a purely statistical framework. *Geosci. Lett.*, **5**(12).
24. Papacharalampous, G., Tyralis, H. and D. Koutsoyiannis. 2018b. Predictability of monthly temperature and precipitation using automatic time series forecasting methods. *Acta Geophysic.*, **66**(4): 807–831.
25. Papacharalampous, G., Tyralis, H. and D. Koutsoyiannis. 2018c. Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from Greece. *Water Resourc. Manage.*, **32**(15): 5207–5239.
26. Papacharalampous, G. and H.Tyralis. 2018. Evaluation of random forests and Prophet for daily streamflow forecasting. *Adv. Geosci.*, **45**: 201–208.
27. Papacharalampous, G., Tyralis, H. and D. Koutsoyiannis. 2019. Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. *Stochastic Environ. Res. Risk Assess.*, 1-34.
28. Rathod, S and G. C. Mishra. 2018. Statistical Models for Forecasting Mango and Banana Yield of Karnataka, *India. J. Agr. Sci. Tech.*, **20**: 803-816
29. Ruiz-Aguilar, J. J., Turias, I. J., Jimenez-Come, M. J. and M. Mar Cerban, 2014. Hybrid Approaches of support vector regression and SARIMA models to forecast the inspections volume. *Int. Conf. Hybrid Artificial Intelligence Syst.*, 502-514.
30. Selvanayaki, K. S. and R. Somasundaram. 2015. An improved approach for detection and classification of vehicles in video using support vector machines. *ARPN J. Engin. App. Sci.*, **10**(10): 4690-4700.
31. Taghadomi-Saberi, S. and S. J. Razavi.2019. Evaluating Potential of Artificial Neural Network and Neuro-Fuzzy Techniques for Global Solar Radiation Prediction in Isfahan, Iran. *J. Agr. Sci. Tech.*, **21**(2): 295-307.
32. Tealab, A., Hefny, H. and A. Badr. 2017. Forecasting of nonlinear time series using artificial neural network. *Future Comput. Informatics J.*, **1**: 9.
33. Tyralis, H. and G. Papacharalampous. 2018. Large-scale assessment of Prophet for multi-step ahead forecasting of monthly streamflow. *Adv. Geosci.*, **45**: 147–153.
34. Tyralis, H. and G. Papacharalampous. 2017. Variable selection in time series forecasting using random forests. *Algorithms*, **10**(4): 114.
35. Wang, H. R., Wang, C., Lin, X. and J. Kang, 2014. An improved ARIMA model for precipitation simulations. *Nonlin. Processes Geophys.*, **21**: 1159-1168.
36. Wang, H. R., Ye, L. T. and C. M. Liu. 2007. Problems in wavelet analysis of hydrologic series and some suggestion on improvement. *Prog. Nat. Sci.*, **17**: 80-86.
37. Weng, C., Huang, T. and R. Han. 2016. Disease prediction with different types of neural network classifiers. *Telematic. Inform.*, **33**:277–292.
38. Yolcu, U., Egrioglu, E. and C. H. Aladag. 2013. A new linear and nonlinear artificial neural network model for time series forecasting. *Decision Support Syst.*, **54**: 1340-1347.
39. Zeynoddin, M., Bonakdari, H., Azari, A., Ebtehaj, I., Gharabaghi, B. and H. Riahi Madavar. 2018. Novel hybrid linear stochastic with non-linear extreme learning machine methods for forecasting monthly rainfall a tropical climate. *J. Environ. Manage.*, **222**: 190-206.
40. Zhang, G. 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomput.*, **50**: 159-175.
41. Zhang, J., Wei, Y. M., Li, D., Tan, Z. and J. Zhou. 2018. Short term electricity load forecasting using a hybrid model. *Energy*, **158**: 774-781.

[DOR: 20.1001.1.16807073.2020.22.2.19.7]

ارزیابی مقایسه ترکیب SARIMA و یادگیری ماشین بر پایه تغییرات زمانی و تفکیک سری زمانی بارندگی

ل. پرویز

چکیده

پیش‌بینی دقیق بارندگی با توجه به پیچیدگی ماهیت آن بسیار مورد توجه است. در این تحقیق از مدل ترکیبی خودهمبسته - میانگین متحرک تلفیق شده فصلی (SARIMA) و الگوریتم یادگیری ماشین مانند شبکه عصبی مصنوعی (ANN) و ماشین بردار پشتیبان (SVM) جهت توسعه پیش‌بینی بارندگی استفاده شد. دو مفهوم تحلیل تغییرات زمانی و تفکیک سری زمانی به بخش خطی و غیرخطی جهت ساخت مدل ترکیبی استفاده شدند. مقایسه عملکرد دو مفهوم با سری زمانی ماهانه بارندگی در دو ایستگاه در شمال ایران مورد ارزیابی قرار گرفت. تحلیل تغییرات زمانی سری‌های زمانی با آنالیز خوشه‌ای انجام شد که منجر به افزایش دقت پیش‌بینی با کاهش ۲۰/۹۹٪ نسبت میانگین هندسی خطا در دو ایستگاه شد. مدل SVM در برابر ANN خطای پیش‌بینی را کاهش داد (متوسط میانگین خطای نسبی (MRE) و میانگین خطای مطلق (MAE) در دو ایستگاه برابر با MRE_{SVM}= 0.72, MRE_{ANN}= 0.89 MAE_{SVM}= 18.02 MAE_{ANN}= 23.88)، بنابراین مدل SVM دارای عملکرد بهتری نسبت به ANN است. مقایسه عملکرد دو مدل ترکیبی بیانگر دقت بیشتر مفهوم تفکیک سری زمانی است (کاهش خطای جذر میانگین مربعات از مفهوم تغییرات زمانی به تفکیک سری زمانی به ترتیب برابر با ۱۳/۳۵٪ بود.). استخراج الگوی داده‌ها با مدل ترکیبی SARIMA با تفکیک سری زمانی، پیش‌بینی سری زمانی را توسعه داد. برخی از ساختارهای مربوط به بخش غیرخطی سری زمانی مورد آزمایش قرار گرفت که ساختاری با گام-های زمانی مختلف باقی‌مانده‌ها دارای عملکرد خوبی بود (میانگین ضریب همسانی =۰/۹). همچنین عملکرد بهتر مدل ترکیبی در سری زمانی فصلی نیز مورد تایید قرار گرفت. نتایج نشان دادند که مدل هیبرید ابزار کارا و موثری در فرآیند تصمیم‌گیری است و تفکیک سری زمانی به دو بخش خطی و غیر خطی دارای عملکرد بهتری است.

[DOR: 20.1001.1.16807073.2020.22.2.19.7]