Improvement of SVR-Based Drought Forecasting Models using Wavelet Pre-Processing Technique

Kit Fai Fung[1], *Yuk Feng* Huang[1*] and *Chai Hoon* Koo[1]

[1]Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Jalan Sungai Long, Bandar Sungai Long, Cheras, 43000 Kajang, Selangor, Malaysia

Abstract. Drought is a damaging natural hazard due to the lack of precipitation from the expected amount for a period of time. Mitigations are required to reduced its impact. Due to the difficulty in determining the onset and offset of droughts, accurate drought forecasting approaches are required for drought risk management. Given the growing use of machine learning in the field, Wavelet-Boosting Support Vector Regression (W-BS-SVR) was proposed for drought forecasting at Langat River Basin, Malaysia. Monthly rainfall, mean temperature and evapotranspiration for years 1976 – 2015 were used to compute Standardized Precipitation Evapotranspiration Index (SPEI) in this study, producing SPEI-1, SPEI-3 and SPEI-6. The 1-month lead time SPEIs forecasting capability of W-BS-SVR model was compared with the Support Vector Regression (SVR) and Boosting-Support Vector Regression (BS-SVR) models using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (R^2) and Adjusted R^2. The results demonstrated that W-BS-SVR provides higher accuracy for drought prediction in Langat River Basin.

1 Introduction

Drought is a damaging natural hazard due to the lack of precipitation from the "normal" or expected amount for a period of time, lead to the insufficient water availability to meet the human and environmental needs. As stated in The Intergovernmental Panel on Climate Change Report on extreme events [1], drought was recognised as extreme climatic events and mitigations are necessary to reduce its impacts. Forecasting and warning facilitation are widely known as important aids in managing natural hazards. Similarly, drought forecasting is crucial for its risk management and mitigation [2-4]. Compared to raw meteorological data, numerically expressed drought indices (DIs) are better in detecting the onsets and offsets of droughts [2].

Along the years, different DIs were introduced to assess the water supply deficit corresponding to the length of the precipitation shortage, including the Percent of Normal, the Standardized Precipitation Index (SPI) [5], Palmer Drought Severity Index (PDSI) [6], Crop Moisture Index (CMI) [7], Streamflow Drought Index (SDI) [8], Standardized

* Corresponding author: huangyf@utar.edu.my

Precipitation Evapotranspiration Index (SPEI) [9], etc. Among them, SPEI is a relatively new simple multi-scalar drought index developed by [9]. Apart from its simplicity, SPEI also able to represent different types of drought due to its multi-scalar characteristic and its consideration of both hydrological (precipitation) and ecological (potential evapotranspiration) variables. [9] tested the SPEI based on 11 observations from different types of climate, including tropical, monsoon, Mediterranean, semi-arid, continental, cold, and oceanic. Hence, tropical country Malaysia is considered suitable for applying the SPEI as the drought index to describe the severity of the events. Other than that, studies also showed that SPEI is a robust index to monitor and analyse droughts [10-13].

Since the forecasting of DIs serves the purpose of recovery planning, mitigation and decision-making, fast, accurate and reliable models to forecast lead-time information on the future drought occurrence and intensity are required. Drought forecasting using machine learning (ML) algorithms is an evolving research area. With the advantages of high flexibility and adaptability, they showed remarkable results in many studies over the years [14-18]. Support Vector Regression (SVR) is a popular ML approach in hydrologic forecasting. It has been used in many studies for drought prediction and papers have shown that SVR is a promising tool in drought prediction [19-24]. In addition to the use of ML models, researchers also started to produce hybrid models by combining with other techniques, such as ensemble and pre-processing technique. According to [25], the boosting ensemble technique can improve the performance of a given learning algorithm. A recent study by [4] also showed that the boosting technique is suitable for improving the performance of SVR models for the prediction of the SPI. Other than that, the use of wavelet transformation to reduce noise and produce wavelet coupled ML models have also been researched over the years [19, 26-29]. Hence, the improvement of using wavelet in SVR-based models were discussed in this paper.

This study predicted SPEI-1, SPEI-3 and SPEI-6 by combining boosting technique with SVR models. Then, wavelet transformation was used to de-noise the original SPEI series. The SPEI was selected because of its ability to represents different time scales of drought. The SPEIs were predicted to the lead time of 1-month because 1-month is a usual short-term lead time in drought forecasting. Prediction results of this study are useful for the water resources management in the basin.

2 Materials and methods

2.1 Study area: The Langat River Basin, Malaysia

Although Malaysia is a tropical country and receives mean annual precipitation of 2,800 mm, the rainfall amount and rain day occurrence have large variability. The extremity of rainfall intensity and occurrences cause difficulties in water resources management for both urban and agricultural use, which generally relies on direct rainwater and rainwater stored in dams. Some of the remarkable historical drought events in Malaysia includes 1991 Malacca water crisis, 1998 Klang Valley water crisis (El-Nino), and the 2014 Selangor water crisis [30].

Given the vulnerability of Malaysia in droughts, Langat River Basin was chosen as the study area. It located between Selangor and Negeri Sembilan, within latitudes 2^{o} 40’ 152” to 3^{o} 16’ 15” and longitudes 101^{o} 19’ 20” to 102^{o} 1’ 10”. It has a total area of approximately 2,400 km^2 and consists of two main dams Langat Dam and Semenyih Dam. These two dams are supplying water to the household and the industrial areas, which are Putrajaya, Hulu Langat, Kuala Langat, Sepang, Petaling and Cheras. Other than that, there are also active agricultural activities at the downstream of the basin, with oil palm plantations coveing an area of approximately 847 km^2 area. Rainfall station at Pejabat JPS Sg. Manggis (ID: s2815001) located at the central of basin downstream, and temperature station at Petaling

Jaya (ID: 48648), both with 40 years (1976 - 2015) of data were used to retrieve data as input for this study. The rainfall data was collected from the Department of Irrigation and Drainage (DID) Malaysia while the temperature data was from the Malaysian Meteorology Department (MMD). The locations of the meteorological stations are shown in Fig. 1.

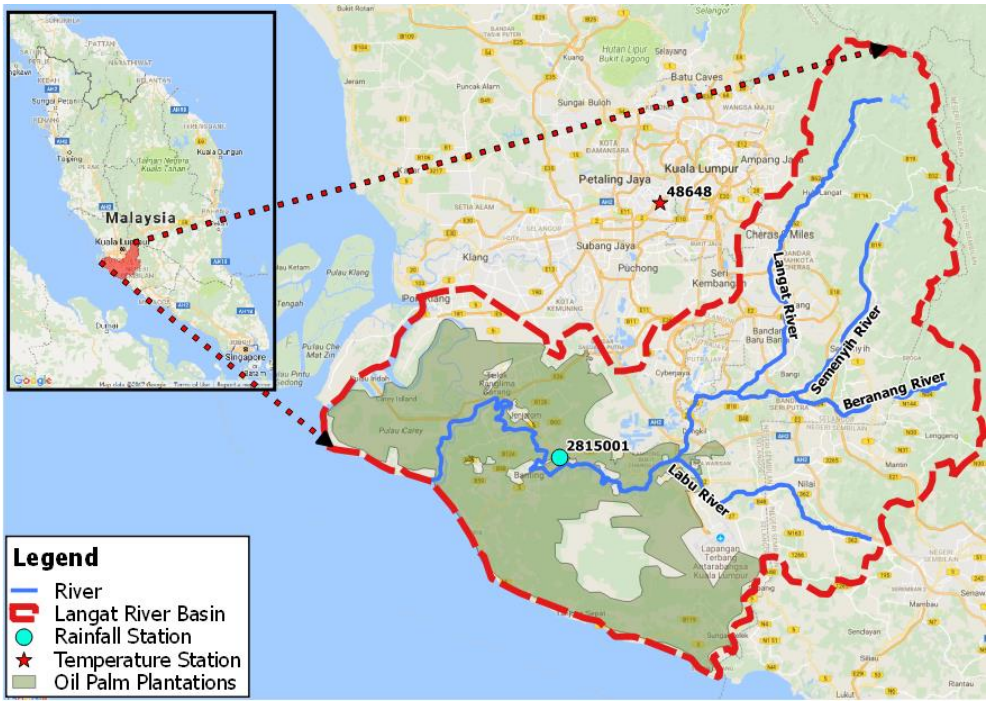

Fig. 1. Map of Langat River Basin with the locations of meteorological stations and land use

2.2 Standardized Precipitation Evapotranspiration Index (SPEI)

The Standardized Precipitation Evapotranspiration Index (SPEI) was introduced by [9]. SPEI The SPEI consists of both multi-scale nature of SPI and evapotranspiration sensitivity of PDSI using simple calculation. It has the same advantage with SPI that allows to describe droughts on multiple time scales in addition of considering the effects from evapotranspiration. In addition, since its concept solely relying on precipitation and temperature, not on soil moisture content as is the PDSI, the SPEI is also not badly affected by landscape. Hence, the SPEI is widely accepted in drought forecasting as it has a broader range of applications than other DIs.

SPEI values can be classified into seven categories (Table 1). Normal conditions are founded from the combinations of two classes: $-0.99 \leq \text{SPEI} \leq 0$ (mild drought) and $0 \leq \text{SPEI} \leq 0.99$ (slightly wet). SPEI values can be presented in positive or negative, depending on its value for greater or less than the mean value, respectively. The magnitude of the SPEI value describes the severity of the events. In this study, SPEI-1, SPEI-3 and SPEI-6 were constructed and the details on SPEI computation are shown in the work of [9].

Table 1. Categories of SPEI

Moisture Category	SPEI
Extremely Wet	2.00 and above
Very Wet	1.50 to 1.99
Moderately Wet	1.00 to 1.49
Near Normal	-0.99 to 0.99
Moderately Dry	-1.00 to -1.49
Severely Dry	*-1.50 to -1.99*
Extremely Dry	-2.00 and below

2.3 Support Vector Regression (SVR)

Support Vector Machine is a category of ML models that can detect the nonlinear characteristics of the data. Unlike other Empirical Risk Minimization based learning algorithms (eg. ANN) that minimize the error over the training data set (training error), the SVM minimizes the model's generalisation error in high dimensional space, so called Structural Risk Minimization [31]. SVMs can be categorized into two types: support vector classification (SVC) and support vector regression (SVR). Since the aim of this study is to predict SPEI, SVR which describes regression was chosen.

All SVR models were developed using 'fitrsvm' function in MATLAB, which is specialized in building Support Vector Machine in regression. The data was divided into two sets: a training set and a validation set. 80% of the data was used as the training set while the final 20% was partitioned as validation set. In the case of the nonlinear regression, an SVM uses Radial Basis Function, rbf kernels [32]. Thus, the estimation of parameter "C" and epsilon value for rbf kernel were required. The parameter C is responsible for the offsets between the model complexity and the degree of deviations (from Epsilon), whereas Epsilon determines the width for the fitting of training data [33]. These parameters were selected based on the methods suggested in [34].

2.4 Boosting-Support Vector Regression (BS-SVR)

Boosting ensemble technique was adopted in this study to enhance the prediction accuracy of SVR models. It is a method which produce sequence of models to improve the performance of a given learning algorithm, where each succeeding model focusses on the poorly trained cases in the preceding one to generate more accurate results [25]. For the BS-SVR model, the 'fitensemble' function in MATLAB was used to boost the observed SPEI. The new learners from 'fitensemble' function can be represented by:

$$y_n - \eta f(x_n) \tag{1}$$

where y_n is the observed response, $f(x_n)$ is the combined prediction from all weak learners created so far for observation x_n and η is the learning rate.

The algorithm of 'fitensemble' function aims to lower the mean absolute error. In every learning cycles, it increases the weightage of the weak learners from preceding models to improve the performance. Two parameters were selected in this section: the appropriate ensemble function and the number of learning cycles. Since the Least Squares Boosting (LS-Boost) fits for regression purposes [35], the "LSBoost" in MATLAB were used to carry out the tasks [36]. As for the number of learning cycles, it was selected based on iteration

procedure. Thereafter, original observed SPEIs were replaced by boosted observed SPEIs to build BS-SVR models.

2.5 Wavelet-Boosting-Support Vector Regression (W-BS-SVR)

Wavelet transformations allow time-scale representation of a given time series and its relationship for the analysis of non-stationaries. Wavelet transformations have the capability to de-noise a signal or a particular set of data in addition of revealing properties of data, such as trends, breakdown points, and discontinuities that other signal processing techniques may not able to achieve [26]. There are two categories of wavelet transforms, are known as the Continuous Wavelet Transform (CWT) and the Discrete Wavelet Transform (DWT). With the reason of CWT processed signal is often described by the information redundancy of the wavelet coefficients, DWT is usually preferred for practical applications such as forecasting [37].

In this study, the DWT is adopted as the data pre-processing technique to reduce the noise in data. The DWT is a simplified approach of the wavelet transform using an independent set of the wavelet scales, and can be represented using Equation (2) [38], as shown below:

$$\psi_{j,m}(m) = \frac{1}{\sqrt{\left|s_o^j\right|}} \sum_k \psi\left(\frac{k - m\tau_0 s_0^j}{s_0^j}\right) x(k) \tag{2}$$

where j and m are integers that control the scale and translation respectively, while $s_o^j > 1$ is a fixed dilation step and τ_0 is a translation factor that depends on the dilation step.

W-BS-SVR denotes the combination of wavelet transformation with BS-SVR. The 'a trous' wavelet is adopted as a pre-processing technique to improve the models' performance by reducing the noises in the time series. Daubechies was selected as the mother wavelet (with vanishing moment, v = 2 or db2) and the processes were performed in the MATLAB platform with the original observed SPEI used as the input for 1-D Stationary Wavelet Transform. The choice of db2 was decided based on the results of [37]. They showed that db2 yielded better prediction efficiencies for time series with long term features (e.g. monthly). During the decomposition process, the original observed SPEIs were decomposed into approximation and detail components. Then, the denoised series were used as inputs to BS-SVR models for prediction using the procedures outline in the previous section.

2.6 Performance measures

The following measures were adopted to assess the accuracy of the models in this study:

$$\text{The Mean Absolute Error (MAE)} = \sum_{i=1}^{N} \frac{|\widehat{y_i} - y_i|}{N} \tag{3}$$

$$\text{The Root Mean Square Error (RMSE)} = \sqrt{\frac{\sum_{i=1}^{N} (\widehat{y_i} - y_i)^2}{N}} \tag{4}$$

$$\text{The Coeficient of Determination } (\mathrm{R}^2) = \frac{\sum_{i=1}^{N} (\widehat{y_i} - y_i)^2}{\sum_{i=1}^{N} (\overline{y_i} - y_i)^2} \tag{5}$$

$$\text{The Adjusted } \mathrm{R}^2 = 1 - \frac{(1 - R^2)(N-1)}{N-p-1} \tag{6}$$

where $\bar{y}_i$ is the mean value taken over N, y_i is the observed value, $\hat{y}_i$ is the predicted value, N is the number of data points and p is the number of predictors. The MAE and RMSE evaluate the similarity between observed and predicted values, while the R^2 and Adjusted R^2 measures the degree of correlation among the observed and predicted values.

3 Results and discussion

The performances of the SVR, BS-SVR and W-BS-SVR models were evaluated using the commonly used performance measures, namely the MAE, the RMSE, the R^2 and the Adjusted R^2 for all three SPEIs with lead times of 1-month. The results showed that the predictions generated by the SVR models varied for SPEIs of different time scales, as shown in Table 2. The performances of the models increase when the time scales increase, especially for the SPEI-3 and the SPEI-6. This showed the increasing generalization ability of the models when the time scale of the SPEIs increases. Based on the estimated Average Moving Range (indicate variations in a series) value for each of the SPEIs, the SPEI-1 has the highest value of 1.0942, followed by the SPEI-3 and then the SPEI-6 with the values of 0.6472 and 0.5622, respectively. Based on the drastic improvement in the performance measures from prediction of the SPEI-1 to the SPEI-3 and gentler improvement from prediction of the SPEI-3 to the SPEI-6, what is certain is that the SVR models have better efficiency for the longer term SPEIs.

For the BS-SVR models, the adoption of the boosting technique resulted in improved performances compared to the standalone SVR model. Based on the results, it was observed that the optimal number of learning cycles to create the lowest generalization error were 313, 206, 195 respectively for the SPEI-1, the SPEI-3 and the SPEI-6. Since MATLAB trains one weak learner for every template object at every learning cycle, the results of decreasing optimum number of learning cycles showed that the learning process is getting easier when the AMR of the series decrease. Similar to SVR models, the performance of BS-SVR models also increases when the time scales of the SPEIs increase. With the evidences of decreasing optimum number of learning cycles and improving accuracy when the lead time increase, it was hypothesized that wavelet transformation which reduce noise can produce even better accuracy. As expected, W-BS-SVR models produced the most accurate results compared to the other two models (Table 2).

The improvements caused by wavelet pre-processing were drastic for SPEI-1, with all performance measures showing W-BS-SVR outperformed the other two models. However, the effects of de-noising seemed to become less significant when the time scales of the SPEI increases. This was observed when the W-BS-SVR model only had significant improvement on correlation (R^2 and Adjusted R^2) in predicting SPEI-3 and no significant improvement in predicting SPEI-6. These results showed that the denoising effects from wavelet become less effective when SPEIs has higher variations for smaller time scales. Other than that, by applying wavelet pre-processing technique, the optimum number of learning cycles in boosting stage have also reduced, to 290, 182, 179 respectively for SPEI-1, SPEI-3 and SPEI-6. These results showed that wavelet pre-processing technique is also effective in improving the learning accuracy.

Table 2. SVR, BS-SVR and W-BS-SVR results

Time Scales	Models	Training				Validation			
		MAE	RMSE	R^2	Adj R^2	MAE	RMSE	R^2	Adj R^2
SPEI-1	SVR	0.421	0.510	0.909	0.909	0.459	0.557	0.933	0.932
	BS-SVR	0.284	0.340	0.983	0.983	0.308	0.368	0.977	0.976
	W-BS-SVR	0.188	0.232	0.984	0.984	0.205	0.249	0.987	0.987
SPEI-3	SVR	0.130	0.192	0.976	0.976	0.152	0.246	0.950	0.949
	BS-SVR	0.117	0.158	0.986	0.986	0.105	0.182	0.973	0.973
	W-BS-SVR	0.118	0.148	0.991	0.991	0.090	0.158	0.982	0.982
SPEI-6	SVR	0.091	0.139	0.986	0.986	0.103	0.176	0.973	0.973
	BS-SVR	0.092	0.130	0.989	0.989	0.099	0.148	0.987	0.987
	W-BS-SVR	0.092	0.129	0.990	0.990	0.100	0.147	0.987	0.987

Further evaluation of the models was done with a time-series plot of data in the validation period (Fig. 2a to Fig. 2b). As clearly illustrated in all three figures, the predicted SPEIs generated by each model closely mirrored the pattern of the observed SPEIs. There was also no noticeable delay between the observed and predicted SPEIs. This shows that the SVR-based models have no time-shift error in this study and are ideal for the prediction of droughts for the Langat River Basin. However, the results of SVR models that underpredicted the values of SPEIs also showed that improvements to generate better predictions are essential (Fig. 2a). Fig. 2a also showed that the BS-SVR models always tend to over predict the extremes, compared to the other two models. This may due to its algorithm of assigning higher weightage to weak learners when new ensembles were produced. For this case, the extremes may have being treated as weak learners in the process and caused the problem of overprediction. As for the W-BS-SVR models, the problem of over-prediction by BS-SVR was lessen due to the reduced of noises in the series by wavelet denoising process. As for Fig. 2b and Fig. 2c, with the closely mirrored patterns and smaller difference between the predicted and observed values, it was concluded that the prediction accuracy of all three models have improved due to the increased time scales or reduced variations of the SPEIs.

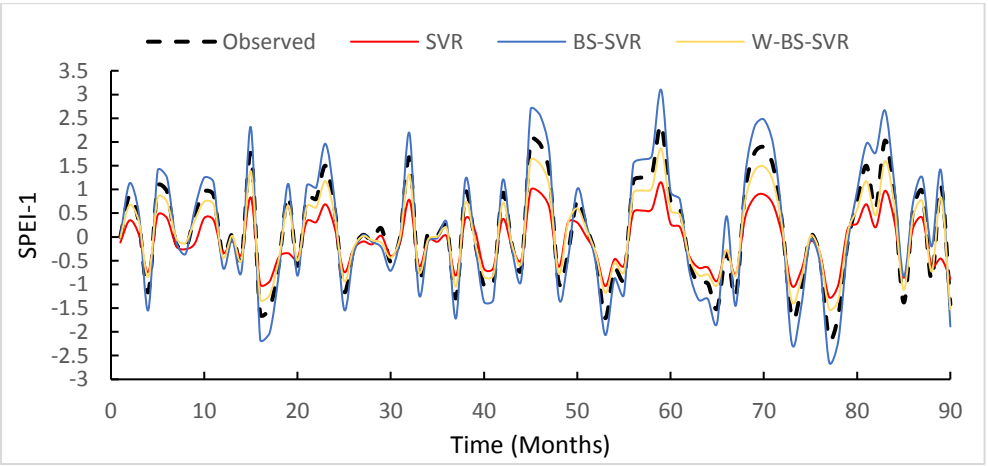

Fig. 2a. Prediction results for SVR, BS-SVR and W-BS-SVR (SPEI-1)

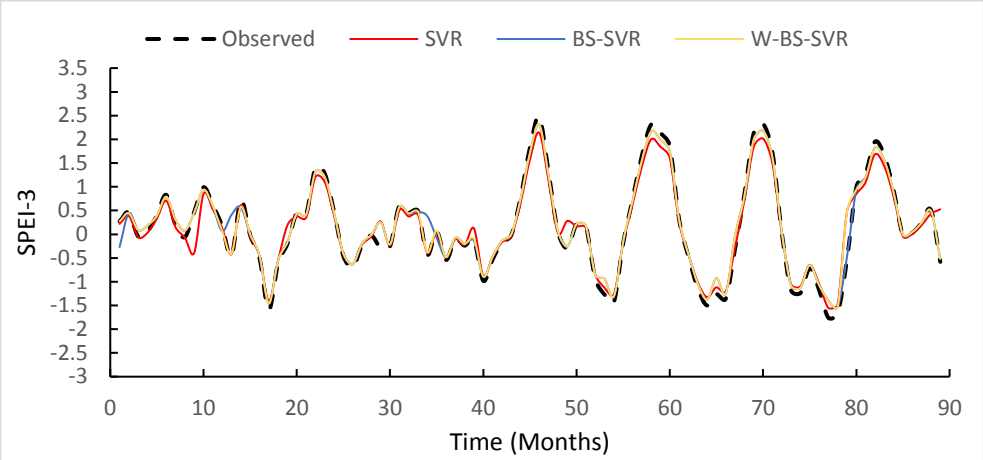

Fig. 2b. Prediction results for SVR, BS-SVR and W-BS-SVR (SPEI-3)

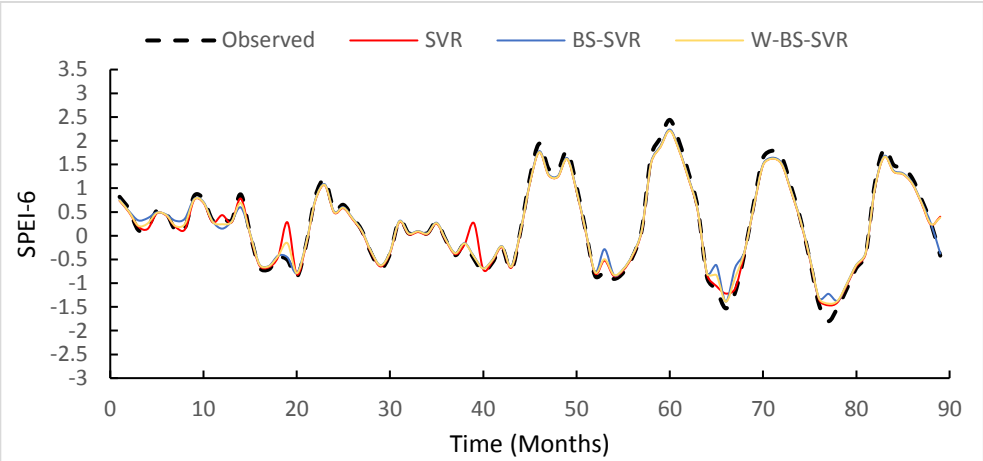

Fig. 2c. Prediction results for SVR, BS-SVR and W-BS-SVR (SPEI-6)

4 Conclusion

This study adopted SPEI as the drought index for drought analysis at Langat River Basin. This paper proposes the application of the W-BS-SVR approach for the modelling of SPEI-1, SPEI-3 and SPEI-6. The W-BS-SVR model was obtained by combining three methods: discrete wavelet transforms, boosting ensemble technique and SVR model. The performance of the proposed W-BS-SVR model was compared to prediction using standalone SVR and boosted SVR models. Comparison of the results indicated that the W-BS-SVR model performed more effectively than the SVR and BS-SVR models. The wavelet denoising effect has reduced the redundancies in the data, causing the results of lower number of learning cycles and better accuracies in W-BS-SVR models. Thus, this study concluded that capability of SVR and BS-SVR model in predicting SPEI-1, SPEI-3 and SPEI-6 are found to be improved when wavelet transformation is adopted for data de-noising purpose.

References

1. C. B. Field. *Managing the risks of extreme events and disasters to advance climate change adaptation: Special report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge (2012)
2. A. K. Mishra, V. P. Singh. J. Hydrol. **391**, 202-216 (2010).
3. A. K. Mishra, V. P. Singh. J. Hydrol. **403**, 157-175 (2011).
4. A. Belayneh, J. Adamowski, B. Khalil, J. Quilty. Atmos. Res. **172-173**, 37-47 (2016).
5. T.B. McKee, N.J. Doesken, J. Kleist. (1993) The Relationship of Drought Frequency and Duration to Time Scales. In: Proc. 8th Conf. on Applied Climatology, 17–22 January, Americ Meteorol Soc, Mass, pp 179–184
6. W.C. Palmer. *Meteorological drought. Research Paper No. 45.* US Weather Bureau, Washington (1965).
7. I. Bordi, A. Sutera (2007) Drought monitoring and forecasting at large scale. In: Rossi G (ed) Methods and tools for drought analysis and management. Springer, Dordrecht, pp 3–27
8. I. Nalbantis, G. Tsakiris. Water Resour Manag. **23**, 881-897 (2009).
9. S.M. Vicente-Serrano, S. Beguería, J.I. López-Moreno. J. Clim. **23**, 1696-1718 (2010).
10. J. Lorenzo-Lacruz, S.M. Vicente-Serrano, J.I. López-Moreno. J. Hydrol. **386**, 13-26 (2010).
11. J.I. López-Moreno, S.M. Vicente-Serrano, J. Zabalza, S. Beguería, J. Lorenzo-Lacruz, C. Azorin-Molina, E. Morán-Tejeda. J. Hydrol. **477**, 175-188 (2012).
12. Z. Liu, Y. Wang, M. Shao, X. Jia, X. Li. J. Hydrol. **534**, 281-299 (2016).
13. M. Xiao, Q. Zhang, V.P. Singh, L. Liu. J. Hydrol. **534**, 297-406 (2016).
14. M. Ozger, A.K. Mishra, V.P. Singh. Int J Climatol. **31**, 2021-2032 (2011).
15. A. Belayneh, J. Adamowski, B. Khalil, B. Ozga-Zielinski. J. Hydrol. **508**, 418-429.(2014).
16. M. Masinde. Mitig Adapt Strat Gl. **19(8)**, 1139-1162 (2014).
17. R.C. Deo, M.K. Tiwari, J.F. Adamowski, J.M. Quilty. Stoch Env Res Risk A. **31**, 1211-1240 (2016).
18. R. Prasad, R.C. Deo, Y. Li, T. Maraseni. Atmos Res. **197**, 42-63 (2017).
19. A. Belayneh., J. Adamowski. J Water Land Dev. **18**, 3-12 (2013).
20. J.L. Chiang, Y.S. Tsai. Appl Mech Mater. **145**, 455-459 (2012).
21. J.L. Chiang, Y.S. Tsai. Appl Mech Mater. **284-287**, 1473-1477 (2013).
22. M. Jalili, J. Gharibshah, S.M. Ghavami, M. Beheshtifar, R. Farshi. IEEE T Comput. **63**, 90-101 (2014).
23. A. Jalalkamali, M. Moradi, N. Moradi. Int J Environ Sci Te. 12, 1201–1210 (2015).
24. M. Borji, A. Malekian, A. Salajegheh, M. Ghadimi. Arab J Geosci. **9**, 725 (2016).
25. Y. Freund, R.E. Schapire. Mach Learn. 148-156 (1996).
26. J. Adamowski, K. Sun. J Hydrol. **390**, 85-91 (2010).
27. M. Ozger, A.K. Mishra, V.P. Singh. J Hydrometeorol. **13**, 284-297 (2012).
28. A.D. Mehr, E. Kahya, M. Ozger. J Hydrol. **517**, 691-699 (2014).
29. S. Djerbouai, D. Souag-Gamane. Water Resour Manag. **30**, 2445-2464 (2016).

30. N. Abdulah, J. Juhaimi, K. Abdul Rahman. *Capacity Development to support National Drought Management Policy.* Hanoi (2014).
31. V. Vapnik. *The Nature of Statistical Learning Theory.* Springer Verlag, New York, USA (1995).
32. V. Kecman. *Learning and Soft Computing*. MIT Press, London (2001).
33. O. Kisi, M. Cimen. J Hydrol. **399**, 132-140 (2011).
34. V. Cherkassky, Y. Ma. Neural Netw. **17(2004)**, 113-126 (2004).
35. J.H. Friedman. Ann Stat. **29(5)**, 1189-1232 (1999).
36. A. Cordiner. *Adaboost Toolbox: A Matlab Toolbox for Adaptive Boosting. Advanced Multimedia Research Lab Oratory Information Communications Technology Research Institute.* University of Wollongong, Australia (2009).
37. R. Maheswaran, R. Khosa. Comput Geosci. **46**, 284-295 (2011).
38. B. Cannas, A. Fanni, G. Sias, S. Tronci, M.K. Zedda. Geophysical Research Abstracts. River flow forecasting using neural networks and wavelet analysis. **7** (2005).