Improved SVR machine learning models for agricultural drought prediction at downstream of Langat River Basin, Malaysia

Kit Fai Fung, Yuk Feng Huang, Chai Hoon Koo and Majid Mirzaei

ABSTRACT

Drought is a harmful and little understood natural hazard. Effective drought prediction is vital for sustainable agricultural activities and water resources management. The support vector regression (SVR) model and two of its enhanced variants, namely, fuzzy-support vector regression (F-SVR) and boosted-support vector regression (BS-SVR) models, for predicting the standardized precipitation evapotranspiration indices (SPEI) (in this case, SPEI-1, SPEI-3 and SPEI-6, at various timescales) with a lead time of one month, were developed to minimize potential drought impact on oil palm plantations at the downstream end of the Langat River Basin, which has a tropical climate pattern. Observed SPEIs from periods 1976 to 2011 and 2012 to 2015 were used for model training and validation, respectively. By applying the MAE, RMSE, MBE and R^2 as model assessments, it was found that the F-SVR model was best with the trend of improving accuracy when the timescale of the SPEIs increased. It was also found that differences in model performance deteriorates with increased timescale of the SPEIs. The outlier reducing effect from the fuzzy concept has better improvement for the SVR-based models compared to the boosting technique in predicting SPEI-1, SPEI-3 and SPEI-6 for a one-month lead time at the downstream of Langat River Basin.

Key words | agricultural drought prediction, boosting ensemble, fuzzy logic, Langat River Basin, standardized precipitation evapotranspiration index, support vector regression

Kit Fai Fung
Yuk Feng Huang (corresponding author)
Chai Hoon Koo
Department of Civil Engineering, Lee Kong Chian Faculty of Engineering and Science,
Universiti Tunku Abdul Rahman,
Jalan Bandar Sg. Long, Bandar Sg. Long, 43000 Kajang, Selangor,
Malaysia
E-mail: *huangyf@utar.edu.my*

Majid Mirzaei
Department of Civil Engineering, Faculty of Engineering,
University of Malaya,
50603 Kuala Lumpur,
Malaysia

INTRODUCTION

Drought is a damaging and little understood natural calamity (Pulwarty & Sivakumar 2014). Drought events usually develop slowly over time with their effects normally lasting for a long period of time (Wilhite *et al.* 2014). These features allow for making drought mitigation possible, albeit difficult, as the starting and ending of droughts are difficult to determine precisely. In particular, some of the rare and extreme drought events vary considerably in time and extent (Burke *et al.* 2010). These characteristics of droughts, no doubt, will further cause significant difficulties in drought mitigation. Hence, effective drought prediction and its subsequent management is important for sustainable agricultural activity and water resource management.

Despite the fact that Malaysia is located in a tropical region and receives an average of 2,800 mm of precipitation annually, the rainfall amount and rain day occurrence however, exhibit large variability. Due to these reasons, the wet and dry conditions can be extreme at times, causing difficulties in sustaining dam water storage and supply management. Some of the drastic droughts that have occurred in the basin and its surrounding areas include the 1991 Malacca water crisis, 1998 Klang Valley water crisis (El Niño) and the 2014 Selangor water crisis (Abdulah *et al.* 2014). This evidence also showed that the study area, which is located in the Langat River Basin of Peninsular Malaysia is vulnerable to droughts and improvement of its

doi: 10.2166/wcc.2019.295

drought prediction capability is required for better drought preparedness. It was reported that 60% of the Langat River Basin is used for agricultural activity (DOA 1995; JICA 2002) with oil palm being the major crop. Oil palm plantations with an approximate 847 km^2 area, are located downstream, as shown in Figure 1. Since oil palm production plays a leading and important role in Malaysia for the agricultural and industrial sectors, a study on developing an agricultural drought prediction model is important for mitigating the negative impacts from drought events. This is especially important as virtually all the oil palm estates in Peninsular Malaysia rely solely on direct precipitation for rain-fed irrigation purposes.

Agricultural drought refers to the circumstances when soil moisture is insufficient, resulting in the lack of water availability for crop growth and production (Wilhite & Glantz 1985). However, the estimation of soil moisture is always a challenging task, more so when in drought management. In order to overcome this problem, the multi-scalar drought index, namely, standardized precipitation index (SPI) has been used to describe different types of drought, including agricultural droughts (Hu *et al.* 2015; Stagge *et al.* 2015; Liu *et al.* 2016; Venkataraman *et al.* 2016). The WMO (2012) stated that the SPI with any timescale from one month to six months, could be used to define agricultural drought as soil moisture conditions respond to precipitation anomalies on a relatively short timescale. However, it is undeniable that there may be a delay between the estimated and actual condition due to the indirect estimation of soil moisture. Hence, the standardized precipitation evapotranspiration index (SPEI) has been developed to include the potential evapotranspiration

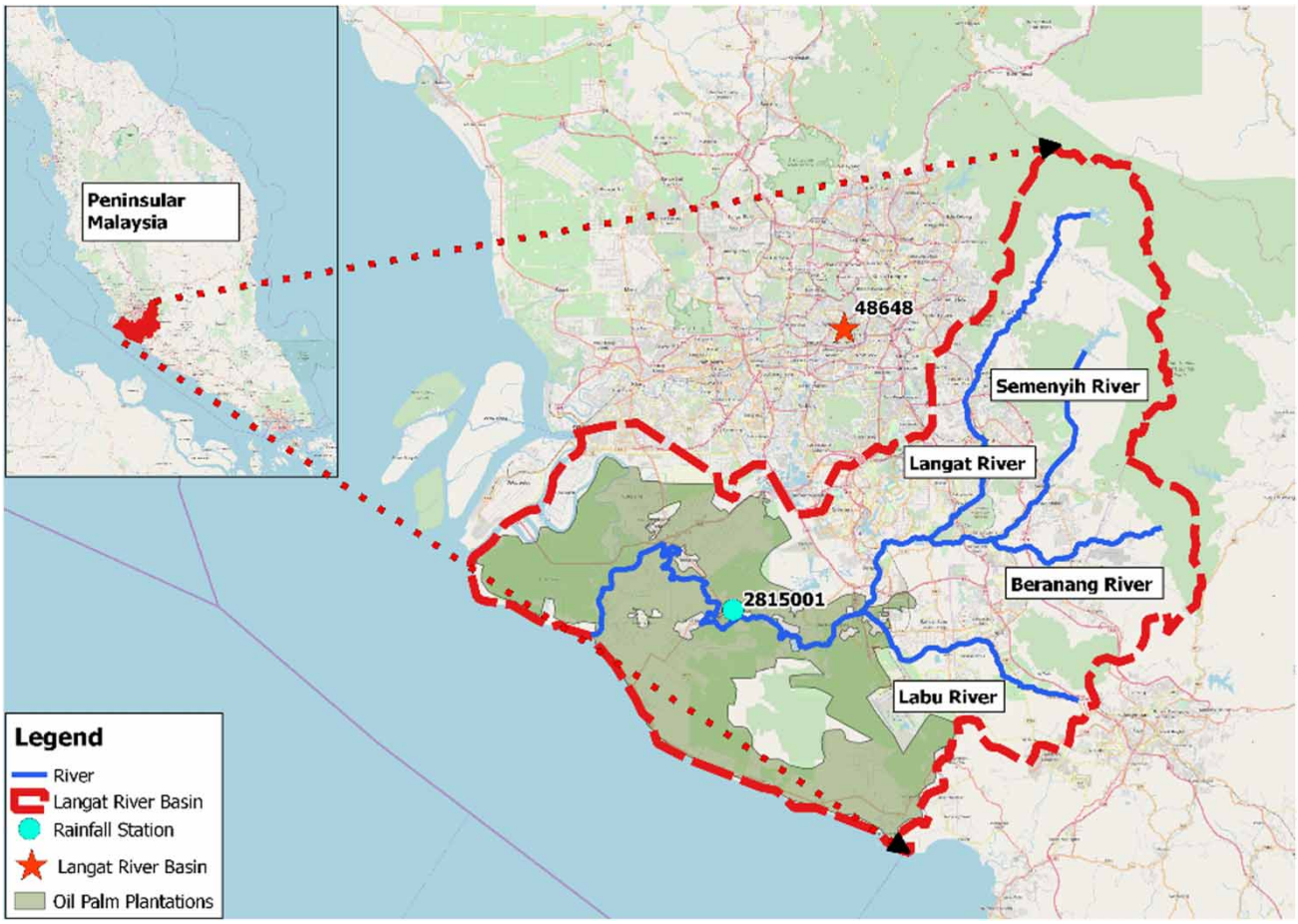

Figure 1 | Map of the Langat River Basin with the locations of meteorological stations and land use.

(PET), which is normally used to quantify the loss of water to the atmosphere by combining the processes of evaporation from the soil and plant surfaces and transpiration from plants, to the description of drought. In view of its multi-scalar representation and with the inclusion of the consideration of the effect of PET, it has been a popular index in various drought studies (Begueria *et al.* 2014; Li *et al.* 2015; Hernandez & Uddameri 2016; Liu *et al.* 2016; Maca & Pech 2016; Xiao *et al.* 2016; Alam *et al.* 2017; Manatsa *et al.* 2017; Chen *et al.* 2018; Soh *et al.* 2018).

Given the advantages of SPEI, it was chosen in this study to describe agricultural drought conditions for the study area, with the timescales of one month, three months and six months. According to the USDA (2016), drought stress lasting more than 8 weeks in Malaysia will usually result in reduced flowering and fruit production during the subsequent 12-month period. This statement is supported by actual historical events, as the moisture deficits that occurred between January and March 2016 were reported to possibly have caused the 8% reduction in crop yield then, although it was quickly recovered with near-normal rainfall between May and July. A newsworthy item was that the newest oil palm hybrid was reported to not only acquire improved drought tolerance, but could tolerate (survive) the maximum of 57 no-rain days Silva *et al.* (2017). Hence, SPEI-1, SPEI-3 and SPEI-6 were proposed based on the reported moisture sensitivity of oil palm in Malaysia.

Drought monitoring and the early warning instrument are important phases to manage droughts (Bachmair *et al.* 2016). An approach for drought prediction concerns the application of machine learning models. Since the characteristics of droughts are difficult to determine, machine learning models, well known for their high flexibility and adaptability, have been used to predict droughts that have different durations, frequencies and intensities. Moreover, the use of machine learning has shown outstanding performance and accuracy (Ozger *et al.* 2011; Belayneh *et al.* 2014; Masinde 2014; Deo *et al.* 2016; Prasad *et al.* 2017; Liang *et al.* 2018). One of the foremost machine learning models used for drought prediction is support vector regression (SVR), which is a family of data-driven type supervised machine learning models. It has been used in several studies for drought prediction and a host of papers have validated that the SVR approach is a promising tool for drought prediction (Chiang & Tsai 2012, 2013; Belayneh & Adamowski 2013; Jalili *et al.* 2014; Jalalkamali *et al.* 2015; Borji *et al.* 2016). In view of the data-driven characteristics inherent in SVR models, it was used to develop the agricultural drought prediction model in this study area, in tandem with using the drought index SPEI.

According to Freund & Schapire (1996), the boosting ensemble technique can improve the performance of a given learning algorithm. The boosting ensemble technique is a method that combines multiple weak learners to produce predictions with higher accuracy, after measuring the pseudo residuals between the predicted and observed values. However, due to the rapid and multi-directional growth of machine learning models in the hydrological field, the application of boosting-ensemble machine learning model is very limited even though it has shown promising results. For example, a recent study by Belayneh *et al.* (2016) showed that the boosting technique is suitable for improving the performance of SVR models for the prediction of the SPI. Hence, this study aimed to explore further the application of the boosting ensemble technique in drought prediction.

According to the standard practice, the option available for the modeller towards the problem of outliers is to discard them from the data sets through careful reasoning and selection. Although these data points are treated as the redundant outliers that may cause undesired errors in the modelling processes, it is inevitable that they are part of the observed values to describe the event. In order to cater for both situations, the concept of fuzzy logic is normally applied to define the grey zone. For this reason, a method called the fuzzy SVR has been developed so that different input data points can provide different contributions to the learning of decision surface based on respective fuzzy membership values, which indicates their importance among the data sets. They have shown outstanding performances in predicting runoff (Wiriyarattanakul *et al.* 2009) and in other applications (Chaudhuri & Kajal 2011; Allaoua & Laoufi 2013; Hung 2016; Edwin & Somasundaram 2016). Given the effectiveness of both techniques in improving the prediction accuracy of machine learning models, the motivation for this study is to improve the agricultural drought predictions with the SVR model by hybridizing it with the

boosting ensemble technique and fuzzy membership values, namely, F-SVR and BS-SVR models.

To the best knowledge of the authors, the SVR-based drought prediction models coupled with fuzzy or boosting technique, using SPEI as predictor, have not been previously carried out for the Langat River Basin. Since the study area is the downstream of Langat River Basin that has a similar humid and warm tropical climate as the basin, it is fascinating to develop the aforementioned agricultural drought prediction models, to predict the wet and dry conditions by considering both the simultaneous changes in precipitation and PET. In order to evaluate the improvements of the fuzzy and boosting technique to the SVR models, the models were all developed by the method of producing SPEI-1, SPEI-3 and SPEI-6 of one-month ahead (lead time). Hence, the expected targeted results of this study are the improved one-month lead time predictions of SPEI with various timescales from the SVR, BS-SVR and F-SVR models.

STUDY AREA AND DATA SET

Study area and data acquisition

The Langat River Basin with an approximate total area of 2,400 km^2 is located over two Peninsular Malaysia states, Selangor and Negeri Sembilan, within latitudes 2° 40′ 15″ N to 3° 16′ 15″ N and longitudes 101° 19′ 20″ E to 102° 1′ 10″ E (Juahir *et al.* 2011). The precipitation data were retrieved from the Department of Irrigation and Drainage (DID) Malaysia, while the temperature data were from the Malaysian Meteorology Department (MMD). It was observed that the main agricultural activity in Langat River Basin is oil palm plantations, located at the downstream of the basin with an approximate area of 847 km^2, as shown in Figure 1. Hence, the rainfall station at Pejabat JPS Sg. Manggis (ID: s2815001) located at the centre of the basin downstream, and temperature station at Petaling Jaya (ID: 48648), both with 40 years (1976–2015) of data, were used to generate the SPEIs to represent the agricultural drought conditions at the downstream agricultural area of the basin (Figure 1), which fulfils the minimum required density of one station per 575–900 km^2 for non-mountainous areas (WMO 2008).

Standardized precipitation evapotranspiration index (SPEI)

The SPEI (for each specific period and for a specific lead time) is a new simple multi-scalar drought index developed by Vicente-Serrano *et al.* (2010). The study developed and tested the SPEI based on 11 observations from different parts of the world, which include tropical, monsoon, Mediterranean, semi-arid, continental, cold, and oceanic climates. Hence, a tropical country like Malaysia is considered a suitable candidate for applying the SPEI as the drought index to represent the severity of the events. The SPEI has a simple calculation algorithm based on the original SPI calculation but combines precipitation and temperature data. Compared to the SPI, that only uses precipitation as inputs, the SPEI is calculated based on the difference between precipitation and potential evapotranspiration (PET), as shown below in Equation (1):

$$D_i = P_i - PET_i \tag{1}$$

PET represents the amount of moisture loss through evaporation and transpiration when there is sufficient or ample availability of water. Mavromatis (2007) proved that the Thornthwaite method is sufficient to calculate the PET. This simple approach with the advantage of only requiring temperature data was adopted for evaluating PET, as suggested by Vicente-Serrano *et al.* (2010). The PET (mm) is obtained via Equation (2), as shown below:

$$PET = 16K\left(\frac{10T}{I}\right)^m \tag{2}$$

where T = monthly mean temperature (°C); I = the heat index of the sum of 12-month i index, where i index is described in Equation (3); $m = 6.75\times10^{-5}\times I^3 + 7.75\times10^{-7}\times I^2 + 1.79\times10^{-2}\times I + 0.492$; K = correction factor as a function of latitude and month as described in Equation (4):

$$i = \left(\frac{T}{5}\right)^{1.514} \tag{3}$$

$$K = \frac{N}{12}\left(\frac{NDM}{30}\right) \tag{4}$$

where NDM = total days in the month and N = maximum number of sun hours.

According to Vicente-Serrano *et al.* (2010), the SPEI can easily be obtained as the standardized values of $F(x)$ as expressed in Equation (5):

$$SPEI = W - \frac{C_0 + C_1 W + C_2 W^2}{1 + d_1 W + d_2 W^2 + d_3 W^3} \quad (5)$$

where $W = [-2 \ \ln(P)]^{0.5}$ for $P \leq 0.5$; $P = 1 - F(x)$; $C_0 = 2.515517$, $C_1 = 0.802853$, $C_2 = 0.010328$; $d_1 = 1.432788$, $d_2 = 0.189269$ and $d_3 = 0.001308$. If the value of P is greater than 0.5, then P will be substituted by $1 - P$ and the sign of the final SPEI is reversed. The probability distribution $F(x)$ is calculated with Equation (6):

$$F(x) = \left[1 + \left(\frac{\alpha}{x - \gamma}\right)^{\beta}\right]^{-1} \quad (6)$$

where α represents the scale, β represents the shape, γ represents the origin parameters, for D values in the range ($\gamma > D < \alpha$). They can be determined using the L-moment method (Ahmad *et al.* 1988) with Equations (7)–(9):

$$\beta = \frac{2w_1 - w_0}{6w_1 - w_0 - 6w_2} \quad (7)$$

$$\alpha = \frac{(w_0 - 2w_1)\beta}{\Gamma\left(1 + \frac{1}{\beta}\right)\Gamma\left(1 - \frac{1}{\beta}\right)} \quad (8)$$

$$\gamma = w_0 - \alpha\Gamma\left(1 + \frac{1}{\beta}\right)\Gamma\left(1 - \frac{1}{\beta}\right) \quad (9)$$

where Γ is the gamma function of β and w_ι ($\mathrm{l} = 0, 1, 2\ldots$) can be computed by probability weighted moments (PWMs) through the L-moment method (Hosking & Wallis 1997):

$$w_\iota = \frac{1}{n}\sum_{i=1}^{n} x_i \left(1 - \frac{i - 0.35}{n}\right)^{\iota} \quad (10)$$

where x_i is the ordered random sample ($x_1 < x_2 \ldots < x_\mathrm{n}$) of D and n represents the sample size.

Since the SPEI has a similar computation process to the SPI, the drought severity defined by SPEI shares the same categories as the SPI, as shown in Table 1 (Li *et al.* 2015). Similar to the SPI, the SPEI can also be computed over different timescales, allowing for the evaluation of the type of droughts affecting shortage of water resources. The SPEI-1 (over a one month period), SPEI-3 (three months) and SPEI-6 (six months) were calculated for the study of agricultural droughts at the downstream of the Langat River Basin. With different timescales, the sensitivity of the SPEIs inevitably varies. Thus, the average moving range (AMR) was adopted to estimate the variations in the series (Montgomery & Runger 2014), as shown in Equation (11):

$$AMR = \frac{1}{m-1}\sum_{i=2}^{m} |X_i - X_{i-1}| \quad (11)$$

where X represents the values of SPEI and m represents the data size.

METHODOLOGY

Support vector regression (SVR)

Support vector machines (SVM) were introduced by Vapnik (1995) in an effort to characterize the properties of learning machines so that they can generalize well to unseen data (Kisi & Cimen 2011). The learning task is insensitive to the relative number of training examples in positive and negative classes. Compared to the artificial neural networks (ANN), the SVM is less prone to overfitting as it seeks to minimize the generalization error, while the ANN seeks to

Table 1 | Categories of SPEI (Li *et al.* 2015)

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

minimize the training error (Chiang & Tsai 2013; Belayneh *et al.* 2016; Borji *et al.* 2016). Thus, the SVM was chosen over the ANN. The SVM can be separated into two types: support vector classification (SVC) and support vector regression (SVR). This study is primarily concerned with the prediction of the SPEI and hence, the SVR was chosen.

In regression estimation with the SVR, the purpose is to estimate a functional dependency $f(\vec{x})$ between a set of sampled points $X = \{\overrightarrow{x_1}, \overrightarrow{x_2}, \overrightarrow{x_3}, \ldots, \overrightarrow{x_i}\}$ taken from R^n and target values $Y = \{y_1, y_2, y_3, \ldots, y_i\}$ with $y_i \in \mathrm{R}$ (the input and target vectors (x_i and y_i) refer to the monthly records of the SPEI index). Assuming that these samples have been generated independently from an unknown probability distribution function $P(\vec{x}, y)$ and a class of functions (Vapnik 1995):

$$F = \{f | f(\vec{x}) = (\overrightarrow{W}, \vec{x}) + B_s : \overrightarrow{W} \in R^n, R^n \to R\} \tag{12}$$

where $\overrightarrow{W}$ and B_s are coefficients that have to be estimated from the input data. The main objective is to find a function $f(\vec{x}) \in F$ that minimizes a risk function. According to Cimen (2008), a regularized risk function with the smallest steepness among the functions is used for the SVR and can be expressed as:

$$R_{reg}[f(\vec{x})] = C_C \sum_{x_i \in X} l_{\in}(y_i - f(\overrightarrow{x_i})) + \frac{1}{2}\|\vec{W}\|^2 \tag{13}$$

where C_C is a positive constant that influences a trade-off between an approximation error and the regression vector $\|\overrightarrow{W}\|$ is a design parameter. The loss function in this expression, which is called an $\in$-insensitive loss function ($l_{\in}$), has the advantage that it does not need all the input data for describing the regression vector $\|\overrightarrow{W}\|$; a more detailed description is given in Cimen (2008).

Herein, the SPEIs of years 1976–2011 (90%) were used for subsequent mandatory training and of years 2012–2015 (10%) for validation thereafter. In order to produce predictions with one-month lead time, the input data were targeted with the SPEIs of one month ahead during the training process. In the case of the nonlinear regression, a SVM uses radial basis function (rbf) kernels (Kecman 2001). Thus, the parameter 'C' and epsilon value for rbf kernel were estimated based on Equations (14) and (15), suggested by Cherkassky & Ma (2004):

$$C = \max(|\bar{y} + 3\sigma_y|, |\bar{y} - 3\sigma_y|) \tag{14}$$

where $\bar{y}$ and σ_y are the mean and standard deviation of the y values of training data (targeted data):

$$\varepsilon = 3\sigma\sqrt{\frac{\ln n}{n}} \tag{15}$$

where σ is the standard deviation of noise and n is the data size.

Boosting-support vector regression (BS-SVR)

As aforementioned, the idea of using a boosting technique is to improve the performance of the SVR model. Boosting is an ensemble method, which attempts to boost the performance of a given learning algorithm (Schapire 1990; Freund & Schapire 1996). As suggested by Friedman (1999), the least squares boosting (LS-boost), which is a variant of gradient boosting, fits for regression of the time series. The main idea behind the algorithm is to produce a sequence of models so that each subsequent model concentrates on the training cases that were not well predicted by the previous one (Freund & Schapire 1996). The process always gets initialized using the mean value for the whole data set to be boosted as the first prediction. Thereafter, pseudo residuals between the predicted and observed values will be estimated and used as the indication to decide the number of weak learners to be combined in order to improve the predictions in the next step. In other words, the ensemble creates a new learner every step by observing the difference between the observed response and the accumulated prediction of all learners created previously. Hence, every new learner was fitted as:

$$y_n - \eta f(x_n) \tag{16}$$

where y_n is the observed response, $f(x_n)$ is the aggregated prediction from all weak learners created so far for observation x_n and η is the learning rate. The algorithm of boosting fits to minimize the differences between input and targeted data. Since the learning rate is constant for the whole

aggregated prediction from all weak learners, the number of weak learners will be increased in order to reduce the pseudo residuals. In other words, initial estimation with high pseudo residuals will require a higher number of weak learners to be combined in order to improve the prediction accuracy. However, this process may result in overfitting when the number of weak learners is too high and fits the data perfectly. Hence, the appropriate number of learning cycles had to be selected carefully without compromising on the generalization of the training process.

In this study, the 'LSBoost' function from Ensemble Learning Toolbox in MATLAB was utilized to combine weak learners and generate a more accurate ensemble. The process to generate accurate and generalized boosted values was iterative and, hence, it was carried out with the aid from 'resume' function in MATLAB. After that, boosted SPEIs were produced and imported to the 'fitrsvm' function in MATLAB together with targeted observed SPEIs with lead time of one month, as inputs for training and validation of SVR.

Fuzzy-support vector regression (F-SVR)

Fuzzy logic was also adopted in this study for its mathematical modelling ability to incorporate imprecision and tolerance towards uncertainty. Classically (Boolean or crisp set theory), membership of an element x in a set A, is defined by the value of either 1 (true) or 0 (false) to each individual in the universal set X, which also means 'every proposition is either true or false'. However, fuzzy logic violates both 'excluded middle' and 'contradiction' laws (Klir & Yuan 2008). According to Zadeh (1965), the true values of variables may be any real number between 0 and 1, which can be done by using fuzzy membership function (FMF) to assign membership value (or degree/grade of membership) between 0 and 1 to every point in the input space (universe of discourse). However, the fuzzy membership function varies for different types of data and the importance to be evaluated on.

The main idea of F-SVR model in this study is to carry out predictions with the adoption of the 'tolerance towards uncertainty and imprecision' on the outliers. Hence, the SPEIs were fuzzified using the fuzzy membership function suggested by Lin & Wang (2002) to reduce the effects of outliers. Since the equations require the data sets to be separated into positive and negative classes, the SPEIs were separated into wet (SPEI >0) and dry (SPEI < or = 0) periods to carry out the tasks. Equation (17) shows the radius of class and Equation (18) shows the fuzzy membership functions:

$$\begin{aligned} r_+ &= max|x_+ - x_i| \quad \textit{for positive class} \\ r_- &= max|x_- - x_i| \quad \textit{for negative class} \end{aligned} \tag{17}$$

where r_+ and r_- are the radius of positive and negative classes, x_+ and x_- are the mean of positive and negative classes.

$$\begin{aligned} S_i &= 1 - |x_+ - x_i|/(r_+ + \sigma) \quad \textit{for positive class} \\ S_i &= 1 - |x_- - x_i|/(r_- + \sigma) \quad \textit{for negative class} \end{aligned} \tag{18}$$

where S_i is the fuzzy membership values.

With these, the fuzzy membership values, S_i that correspond to each observed SPEI data point were produced. Thereafter, fuzzy membership values, observed SPEI and targeted observed SPEI with lead time of one month were imported to the 'fitrsvm' function in MATLAB for training and validation of the SVR. The generated fuzzy membership values were used as additional inputs together with the SPEIs (two input variables) to transform training points from $\{(x_1, y_1), \ldots, (x_i, y_i)\}$ to $\{(x_1, S_1, y_1), \ldots, (x_i, S_i, y_i)\}$. Figure 2 shows the flowchart for the development of models.

Models' performance evaluation

The performances of the models were evaluated using the four standard performance measures including the mean absolute error (MAE), root mean square error (RMSE), mean bias error (MBE) and the coefficient of determination (R^2), as in Equations (19)–(22), respectively. According to the research on comparison of hydrological performance measures done by Dawson *et al.* (2007), they suggested that one must not rely on a single performance measure but should select as per the requirement of the modelling. The MAE, RMSE and MBE were used to measure different types of information about the forecasting capabilities of the model in overall data sets. Basically, the MAE measures the goodness-of-fit of the predicted values to the observed values in an equal manner and regardless of sign; while the RMSE measures the deviations from the observed values relevant to high magnitude, where high(er) weightage is given to the high(er) magnitude events or vice versa. Meanwhile,

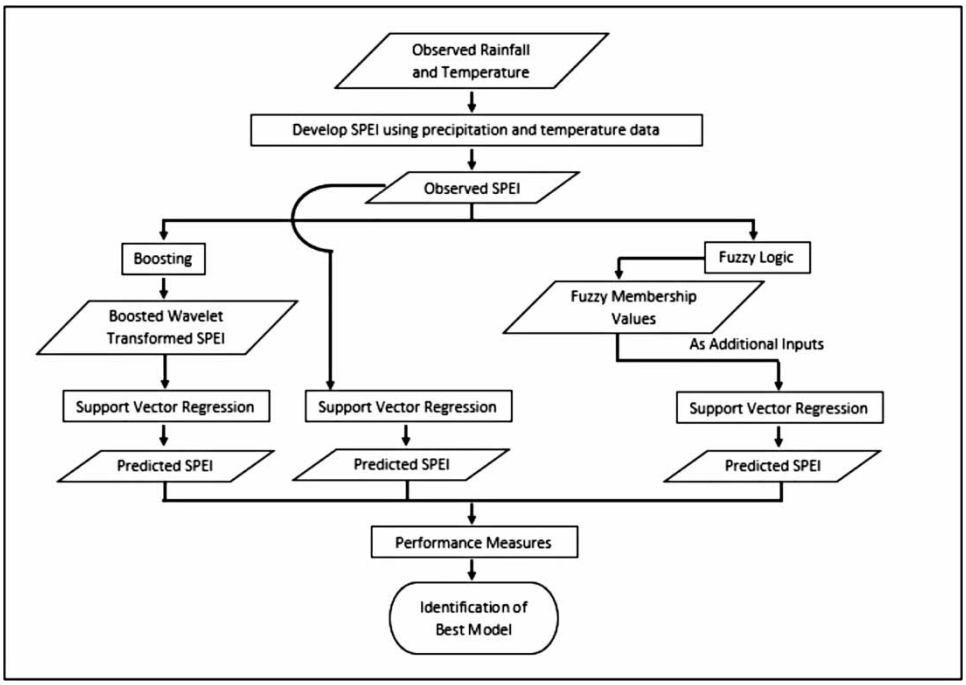

Figure 2 | Development of models.

the MBE is another goodness-of-fit measure that is similar to the MAE but with the consideration of positive and negative sign. This allows the MBE to tell if the model's overall prediction is either overfitted (positive) or underfitted (negative) to the observed values. As for R^2, it describes the proportion of the total statistical variance in observed values that can be explained by the drought model. It ranges from 0.0 (poor model) to 1.0 (perfect model) and records the degree of association between observed and predicted values:

$$MAE = \sum_{i=1}^{N} \frac{|\widehat{y_i} - y_i|}{N} \tag{19}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\widehat{y_i} - y_i)^2}{N}} \tag{20}$$

$$MBE = \sum_{i=1}^{N} \frac{\widehat{y_i} - y_i}{N} \tag{21}$$

$$R^2 = \frac{\sum_{i=1}^{N} (\widehat{y_i} - \overline{y_i})}{\sum_{i=1}^{N} (y_i - \overline{y_i})^2}, \text{ where } \overline{y_i} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{22}$$

where $\overline{y_i}$ is the mean value taken over N, $\widehat{y_i}$ is the predicted value, y_i is the observed value, N is the number of samples.

RESULTS AND DISCUSSION

Development of models

For this paper, all the three proposed models, namely, the SVR, the F-SVR and the BS-SVR, results were generated for the station s2815001 (Pejabat JPS Sg. Manggis) located at the Langat Basin. These models were used to predict the SPEI-1, the SPEI-3 and the SPEI-6, respectively, each on a one-month lead time frame. Before the development of the models, the precipitation and temperature data were used to generate the SPEI-1, SPEI-3 and SPEI-6. For the development of SPEIs, the developed series is shown in Figure 3. Compared to the SPEI-1, it can be observed that both the SPEI-3 and SPEI-6 are less sensitive to the changes in monthly precipitation and/or temperature within the long-term record. Since the SPEI-3 and SPEI-6 are longer cumulative indices than SPEI-1, the onset of drought only becomes obvious at over a longer time frame. The AMR

was used to describe the variations caused by the sensitivity of each of the SPEIs. As expected, the SPEI-1, which is most sensitive to changes acquired the highest average moving range values of 1.0942. For the SPEI-3 and the SPEI-6, that are less sensitive, they were characterized respectively, with lower average moving range values of 0.6472 and 0.5622.

As mentioned in the section 'Study area and data set', the fuzzy membership values (S_i) were used in this study to represent the degree of importance of all data (Figure 4)

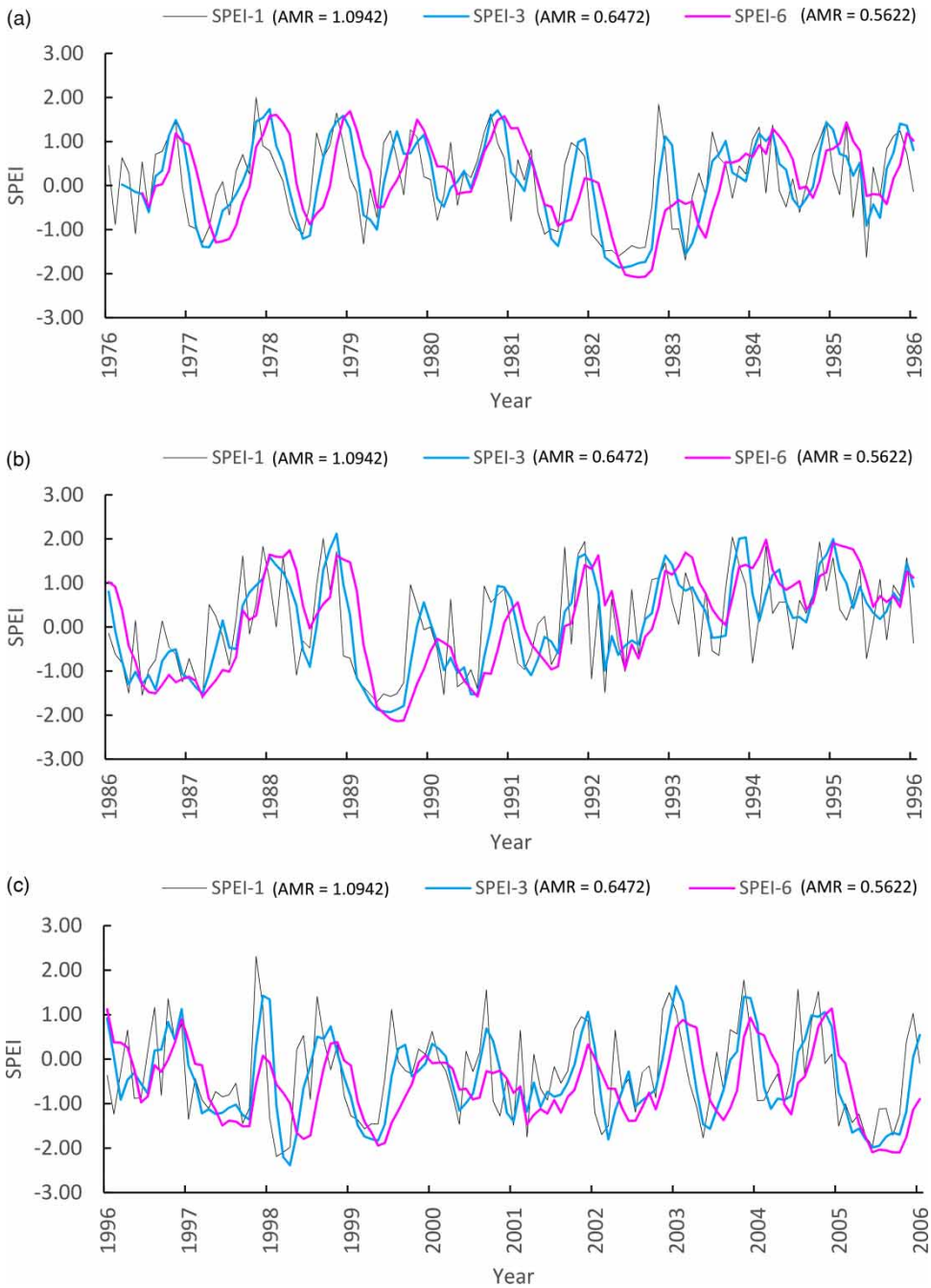

Figure 3 | Developed SPEI-1, SPEI-3 and SPEI-6: (a) 1976–1985, (b) 1986–1995, (c) 1996–2005 and (d) 2006–2015. *(Continued.)*

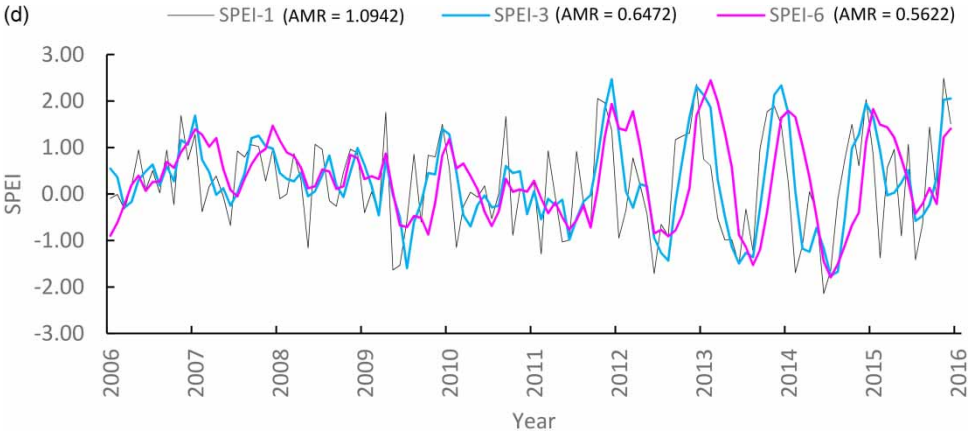

Figure 3 | Continued.

and adopted as additional input in the SVR to reduce the effects of outliers. From Figure 4, it was observed that when the SPEIs is closed to the mean value, it will acquire a high S_i, and vice versa. For example, the SPEI-6 with a value of −0.82 at time step 64, attained an S_i of 0.99 (largest is 1.00) because it is close to the negative mean of −0.83. Otherwise, the SPEI-6 with a value of −2.08 at time step 75, only attained an S_i of 0.46 because of its large difference from negative mean. The same goes for the S_i of the SPEI-1 and the SPEI-3, as shown in Figure 4. With these results, it was shown that the adopted fuzzy membership functions have the ability to estimate the degree of importance for each point.

For the boosting ensemble, since the algorithm is to improve the performance of the models by improving the learning effects from the weak learner, the problem of overfitting may occur when the number of learning cycles becomes too high. Thus, the selection of the appropriate number of learning cycles is important for a generalized model. For this study, the optimal number of learning cycles to create the lowest generalization error were 313, 206 and 195, respectively, for the SPEI-1, the SPEI-3 and the SPEI-6. It was observed that the number of learning cycles decreased when the timescales of the SPEI series increased. At every learning cycle, MATLAB trains one weak learner for every template object in learners. Thus, the increasing number of learning cycles to reach the optimum stage also indicates that the number of weak learners were increasing for the decreasing timescales. Hence, the results of the boosting ensembles are also indicating that the training difficulties are getting lower when the timescales of SPEIs increases.

Performance of the models

With the optimal parameters of each model being determined, the prediction results from each model were generated and their performances were evaluated based on the measures of MAE, RMSE, MBE and R^2 between the observed and predicted SPEIs, as tabulated in Table 2.

Based on the results, it can be observed that the overall performance of the models increased with timescale, from SPEI-1 to SPEI-6. For example, the validation results in terms of MAE, RMSE, MBE, R^2 increased from the range of 0.325–0.559, 0.372–0.644, −0.017–0.133, 0.796–0.828 in SPEI-1 to 0.126–0.170, 0.159–0.202, −0.026–0.040, 0.824–0.854 in SPEI-3, and further to 0.105–0.144, 0.137–0.187, −0.023–0.011, 0.866–0.903 in SPEI-6. Hence, the prediction accuracy of the models increases when the timescales increase, especially for the SPEI-3 and the SPEI-6. On the other hand, it was found that the estimated AMR (indicate variations in a series) value of 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. In addition, 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, it is further deemed that the prediction capability of the models is affected by the variation in the series.

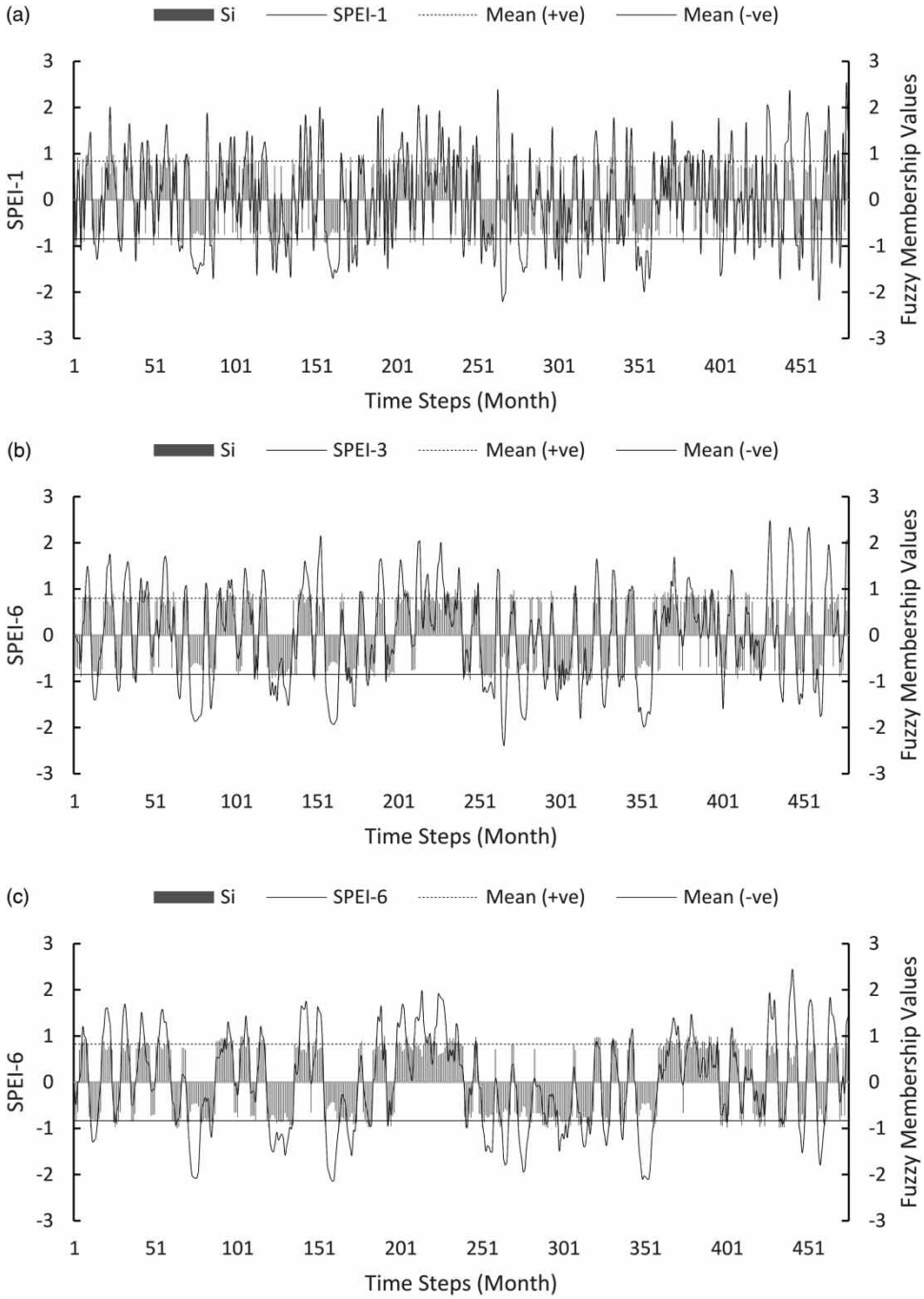

Figure 4 | Fuzzy membership values of SPEI: (a) SPEI-1, (b) SPEI-3 and (c) SPEI-6.

For the comparison between models, the results show that the overall performance of the BS-SVR and F-SVR models have been improved as compared to the standalone SVR model. For example, the MAE of the models in predicting SPEI-1 reduced from 0.559 in SVR to 0.543 in BS-SVR and 0.325 in F-SVR during the validation period. These results of BS-SVR and F-SVR models outperforming the standalone SVR model were also shown in other performance measures, such as RMSE and R^2 (Table 2). These indicate that both the boosting ensemble technique and

fuzzy membership values have successfully improved the prediction capability of the SVR model, as the prediction errors have shown to be reduced together with the increase in correlation between predicted and observed values.

It was also observed that the differences in performance between the BS-SVR model and the F-SVR model is getting less significant over the increasing timescales. Taking into account the fact that the variation in series increases over the increasing timescales, it is deemed that the advantage of fuzzy membership values over boosting ensemble technique is getting smaller when the variation in series decreases. By reviewing the algorithms of each approach, it is reasonable to have this conclusion as the fuzzy membership values in this study were estimated in the effort to reduce the effect of outliers by assigning lower fuzzy membership values to the points further from class mean, while the boosting ensemble technique improved the predictions by combining weak learners. Hence, when the variation in the series decreases over increasing timescales, the advantage of fuzzy membership values over boosting ensemble technique also decreases.

Further evaluation of the models was also carried out using a time series plot of data in the validation period (Figure 5). As clearly illustrated in Figure 5, 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 agricultural droughts for the downstream of Langat River Basin. However, the evidence of the SVR models that underpredicted the values of SPEIs also show that improvements to generate better predictions were needed. Figure 5 also shows that the BS-SVR model always tends to overpredict the extremes, e.g., 10-month time step in Figure 5(a), 11-month time step in Figure 5(b) and 18-month time step in Figure 5(c). This phenomenon can be explained by reviewing the algorithm in the models.

As mentioned, the boosting process was initialized using the overall mean as the first prediction. Thereafter, pseudo residuals between the predicted and observed values were estimated and used as the indication to decide the number of weak learners to be combined in order to improve the predictions in the next step. However, this initial estimation may produce high pseudo residuals at extreme values due to the zero or near to zero mean generated from the cancel-off effect between the positive and negative values during the calculation of mean. Thereafter, higher weightage will be assigned to the extreme values during the training process and cause overprediction at those points, which is in agreement with the graphical illustration shown in Figure 5. However, this problem was avoided in the F-SVR model as the fuzzy membership values used were generated with reference to respective mean from positive and negative classes. In other words, the effects from each point were altered using the fuzzy membership values with reference to its class mean. As compared to the initial mean generated in the BS-SVR model, the class mean in the F-SVR model has better representation on the original

Table 2 | Performance measures of SVR, F-SVR and BS-SVR models

		Training				Validation			
Time scales	**Models**	**MAE**	**RMSE**	**MBE**	**R^2**	**MAE**	**RMSE**	**MBE**	**R^2**
SPEI-1	SVR	0.439	0.520	−0.082	0.838	0.559	0.644	−0.130	0.796
	F-SVR	0.256	0.304	−0.013	0.872	0.325	0.372	−0.017	0.828
	BS-SVR	0.427	0.506	−0.087	0.864	0.543	0.626	0.133	0.821
SPEI-3	SVR	0.129	0.155	−0.009	0.864	0.170	0.202	−0.036	0.824
	F-SVR	0.107	0.113	−0.025	0.899	0.126	0.159	−0.026	0.854
	BS-SVR	0.112	0.121	0.014	0.891	0.133	0.172	0.040	0.846
SPEI-6	SVR	0.118	0.132	0.012	0.912	0.144	0.187	−0.023	0.866
	F-SVR	0.101	0.112	−0.012	0.950	0.105	0.137	−0.036	0.903
	BS-SVR	0.106	0.124	0.013	0.941	0.114	0.146	0.011	0.894

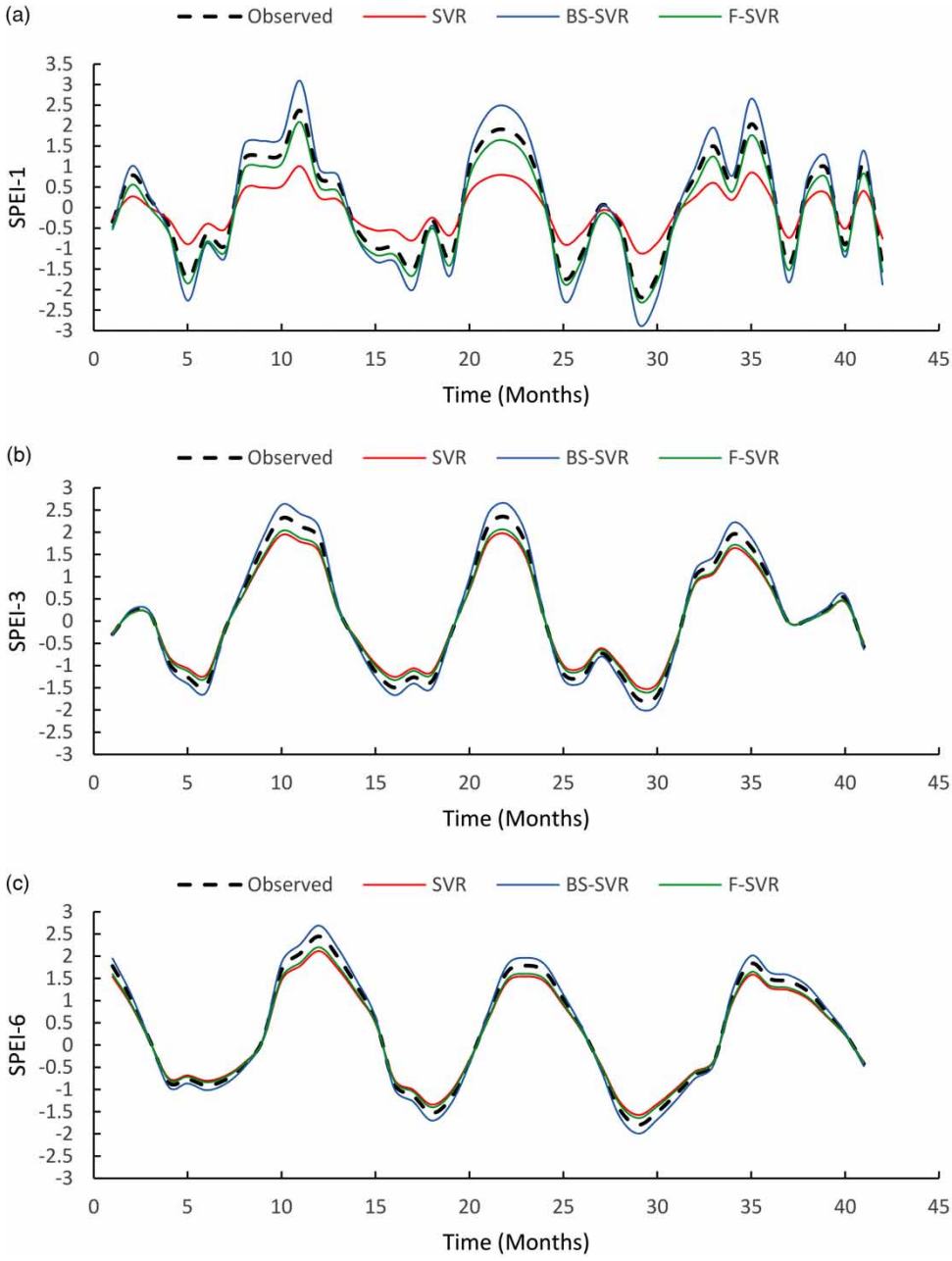

Figure 5 | Prediction results for SVR, F-SVR and BS-SVR: (a) SPEI-1, (b) SPEI-3 and (c) SPEI-6.

characteristics of the data and will not cause higher assignation of weightage to the extreme values. Hence, overprediction did not happen in the F-SVR model, whereas there were some minor underpredictions, as shown in Figure 5, which may be due to the reduced effects from data points caused by assignation of fuzzy membership values.

CONCLUSIONS

This study was done to assess the efficiency of the various SVR-based models in predicting one-month lead time agricultural drought at the downstream end of the Langat River Basin, which has a tropical climate pattern. Drought indices, namely, the SPEI-1, the SPEI-3 and the SPEI-6

were adopted to describe drought severity. Apart from the standalone SVR model, this study also combined separately, the concepts of fuzzy logic and the boosting ensemble technique with the SVR model in order to improve the accuracy in predicting the SPEIs. Based on the performance measures determined, it was observed that the performance of all three models has the trend of improving accuracy with increasing timescale of the SPEIs: most accurately for the SPEI-6, followed by the SPEI-3 and then lastly the SPEI-1. In view of previous findings and the decreasing variation in series across increasing timescale of SPEIs, the authors maintain that the difficulties of the models to fit the training data are affected by the variations in the series. Both the F-SVR and the BS-SVR models were found to consistently give better predictions than the standalone SVR model; suggesting that both the F-SVR model and the BS-SVR model can improve the prediction capability of the SVR model by reducing the outlier effects and creating ensembles from weak learners, respectively. Nevertheless, between these two better models, the F-SVR model being a notch better, showed better performance measures with the lower prediction error shown in MAE, RMSE and higher correlation shown in R^2. In view of the algorithms in the F-SVR model and the finding of improving accuracy over decreasing variations in series, the authors concluded that the advantage of the F-SVR model over the BS-SVR model is due to its outliers' reducing effect, which reduces the training difficulties due to the variations by assigning lower fuzzy membership values to the points further from the class mean. Future work should be carried out using these methods in other study areas to ensure the models' robustness, especially when longer lead time is required due to the differences in climatic conditions. Attempts to include wavelet transformation as the data smoothing technique to improve the performance of these models could be considered given the finding in this study showing that variations in training series are affecting the prediction capability of the models.

ACKNOWLEDGEMENTS

The authors would like to express their sincere appreciation to the Universiti Tunku Abdul Rahman, Bandar Sungai Long, Cheras, 43000 Kajang, Selangor, Malaysia for funds allocated to this project.

REFERENCES

Abdulah, N., Juhaimi, J. & Abdul Rahman, K. 2014 *Capacity Development to Support National Drought Management Policy*. The UN-Water Decade Programme on Capacity Development (UNW-DPC), Hanoi, Vietnam.

Ahmad, M. I., Sinclair, C. D. & Werritty, A. 1988 Log-logistic flood frequency analysis. *Journal of Hydrology* **98**, 205–224.

Alam, N. M., Sharma, G. C., Moreira, E., Jana, C., Mishra, P. K., Sharma, N. K. & Mandal, D. 2017 Evaluation of drought using SPEI drought class transitions and log-linear models for different agro-ecological regions of India. *Physics and Chemistry of the Earth, Parts A/B/C* **100**, 31–43.

Allaoua, B. & Laoufi, A. 2013 A novel sliding mode fuzzy control based on SVM for electric vehicles propulsion system. *Energy Procedia* **36**, 120–129.

Bachmair, S., Stahl, K., Collins, K., Hannaford, J., Acreman, M., Svoboda, M., Knutson, C., Smith, K. H., Wall, N., Fuchs, B., Crossman, N. D. & Overton, I. C. 2016 Drought indicators revisited: the need for a wider consideration of environment and society. *Wiley Interdisciplinary Reviews: Water* **3**, 516–536.

Begueria, S., Vicente-Serrano, S. M., Reig, F. & Latorre, B. 2014 The standardized precipitation evapotranspiration index (SPEI) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. *International Journal of Climatology* **34** (10), 3001–3023.

Belayneh, A. & Adamowski, J. 2013 Drought forecasting using new machine learning methods. *Journal of Water and Land Development* **18**, 3–12.

Belayneh, A., Adamowski, J., Khalil, B. & Ozga-Zielinski, B. 2014 Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural networks and wavelet support vector regression models. *Journal of Hydrology* **508**, 418–429.

Belayneh, A., Adamowski, J., Khalil, B. & Quilty, J. 2016 Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction. *Atmospheric Research* **172–173**, 37–47.

Borji, M., Malekian, A., Salajegheh, A. & Ghadimi, M. 2016 Multi-time-scale analysis of hydrological drought forecasting using support vector regression (SVR) and artificial neural networks (ANN). *Arabian Journal of Geosciences* **9**, 725.

Burke, E., Perry, R. & Brown, S. 2010 An extreme value analysis of UK drought and projections of change in the future. *Journal of Hydrology* **388** (1–2), 131–143.

Chaudhuri, A. & Kajal, D. 2011 Fuzzy Support Vector Machine for bankruptcy prediction. *Applied Soft Computing* **11** (2), 2472–2486.

Chen, S., Zhang, L., Liu, X., Gao, M. & She, D. 2018 The use of the SPEI and TVDI to access temporal-spatial variations in drought conditions in the middle and lower reaches of the Yangtze River Basin, China. *Advances in Meteorology* **2018**, 1–11.

Cherkassky, V. & Ma, Y. 2004 Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks* **17** (2004), 113–126.

Chiang, J. L. & Tsai, Y. S. 2012 Reservoir drought prediction using support vector machines. *Applied Mechanics and Materials* **145**, 455–459.

Chiang, J. L. & Tsai, Y. S. 2013 Reservoir drought prediction using two-stage SVM. *Applied Mechanics and Materials* **284–287**, 1473–1477.

Cimen, M. 2008 Estimation of daily suspended sediments using support vector machines. *Hydrological Sciences Journal* **53**, 656–666.

Dawson, C. W., Abrahart, R. J. & See, L. M. 2007 Hydrotest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. *Environmental Modelling & Software* **22**, 1034–1052.

Deo, R. C., Tiwari, M. K., Adamowski, J. F. & Quilty, J. M. 2016 Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. *Stochastic Environmental Research and Risk Assessment* **31**, 1211–1240.

DOA 1995 *Land use of Selangor and Negeri Sembilan*. Department of Agriculture (DOA), Kuala Lumpur, Malaysia.

Edwin, D. J. & Somasundaram, K. 2016 Evolutionary fuzzy SVR modeling of weld residual stress. *Applied Soft Computing* **42**, 423–430.

Freund, Y. & Schapire, R. E. 1996 Experiments with a new boosting algorithm. In: *Proceedings of the Thirteenth International Conference on Machine Learning*, Bari, Italy, pp. 148–156.

Friedman, J. H. 1999 Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29** (5), 1189–1232.

Hernandez, A. E. & Uddamri, V. 2016 Standardized precipitation evaporation index (SPEI)-based drought assessment in semi-arid south Texas. *Environmental Earth Sciences* **71** (6), 2491–2501.

Hosking, J. R. M. & Wallis, J. R. 1997 *Regional Frequency Analysis – An Approach Based on L-Moments*. Cambridge University Press, Cambridge, UK.

Hu, Y. M., Liang, Z. M., Liu, Y. W., Wang, J., Yao, L. & Ning, Y. 2015 Uncertainty analysis of SPI calculation and drought assessment based on the application of Bootstrap. *International Journal of Climatology* **35** (8), 1847–1857.

Hung, J. C. 2016 Fuzzy support vector regression model for forecasting stock market volatility. *Journal of Intelligent & Fuzzy Systems* **31** (3), 1987–2000.

Jalalkamali, A., Moradi, M. & Moradi, N. 2015 Application of several artificial intelligence models and ARIMAX model for forecasting drought using the Standardized Precipitation Index. *International Journal of Environmental Science and Technology* **12**, 1201–1210.

Jalili, M., Gharibshah, J., Ghavami, S. M., Beheshtifar, M. & Farshi, R. 2014 Nationwide prediction of drought conditions in Iran based on remote sensing data. *IEEE Transactions on Computers* **63**, 90–101.

JICA 2002 *The Study on the Sustainable Groundwater Resources and Environmental Management for the Langat Basin in Malaysia*. Final Report, vol. 3: Supporting Report. Japan International Cooperation Agency (JICA), Tokyo, Japan.

Juahir, H., Zain, S. M., Yusoff, M. K., Tengku Hanidza, T. I., Mohd Armi, A. S., Toriman, M. E. & Mokhtar, M. 2011 Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. *Environmental Monitoring and Assessment* **173** (1–4), 625–641.

Kecman, V. 2001 *Learning and Soft Computing*. MIT Press, London, UK.

Kisi, O. & Cimen, M. 2011 A wavelet-support vector machine conjunction model for monthly streamflow forecasting. *Journal of Hydrology* **399**, 132–140.

Klir, G. J. & Yuan, B. 2008 *Fuzzy Sets and Fuzzy Logic, Theory and Applications*. Prentice Hall, Upper Saddle River, NJ, USA.

Li, B., Zhou, W., Zhao, Y., Ju, Q., Yu, Z., Liang, Z. & Acharya, K. 2015 Using the SPEI to assess recent climate change in the Yarlung Zangbo River Basin, South Tibet. *Water* **7**, 5474–5486.

Liang, Z., Li, Y., Hu, Y., Li, B. & Wang, J. 2018 A data-driven SVR model for long-term runoff prediction and uncertainty analysis based on the Bayesian framework. *Theoretical and Applied Climatology* **133** (1–2), 137–149.

Lin, C. F. & Wang, S. D. 2002 Fuzzy support vector machines. *IEEE Transactions on Neural Networks and Learning Systems* **13** (2), 464–471.

Liu, Z., Wang, Y., Shao, M., Jia, X. & Li, X. 2016 Spatiotemporal analysis of multiscalar drought characteristics across the Loess Plateau of China. *Journal of Hydrology* **534**, 281–299.

Maca, P. & Pech, P. 2016 Forecasting SPEI and SPI drought indices using the integrated artificial neural networks. *Computational Intelligence and Neuroscience* **2016**, 1–17.

Manatsa, D., Mushore, T. & Lenouo, A. 2017 Improved predictability of droughts over the Southern Africa using the standardized precipitation evapotranspiration index and ENSO. *Theoretical and Applied Climatology* **127** (1–2), 259–274.

Masinde, M. 2014 Artificial neural networks models for predicting effective drought index: factoring effects of rainfall variability. *Mitigation and Adaptation Strategies for Global Change* **19** (8), 1139–1162.

Mavromatis, T. 2007 Drought index evaluation for assessing future wheat production in Greece. *International Journal of Climatology* **27**, 911–924.

Montgomery, D. C. & Runger, G. C. 2014 *Applied Statistics and Probability for Engineers*. Wiley, New York, USA.

Ozger, M., Mishra, A. K. & Singh, V. P. 2011 Estimating Palmer Drought Severity Index using a wavelet fuzzy logic model

based on meteorological variables. *International Journal of Climatology* **31**, 2021–2032.

Prasad, R., Deo, R. C., Li, Y. & Maraseni, T. 2017 Input selection and performance optimization of ANN-based streamflow forecasts in the drought-prone Murray Darling Basin region using IIS and MODWT algorithm. *Atmospheric Research* **197**, 42–63.

Pulwarty, R. S. & Sivakumar, M. V. K. 2014 Information systems in a changing climate: early warnings and drought risk management. *Weather and Climate Extremes* **3**, 14–21.

Schapire, R. E. 1990 The strength of weak learnability. *Machine Learning* **5**, 197–227.

Silva, P. A., Cosme, V. S., Rodrigues, K. C. B., Detmann, K. S. C., Leao, F. M., Cunha, R. L., Festucci Buselli, R. A., DaMatta, F. M. & Pinheiro, H. A. 2017 Drought tolerance in two oil palm hybrids as related to adjustments in carbon metabolism and vegetative growth. *Acta Physiologiae Plantarum* **39**, 58.

Soh, Y. W., Koo, C. H., Huang, Y. F. & Fung, K. F. 2018 Application of artificial intelligence models for the prediction of standardized precipitation evapotranspiration index (SPEI) at Langat River Basin, Malaysia. *Computers and Electronics in Agriculture* **144**, 164–173.

Stagge, J. H., Kohn, I., Tallaksen, L. M. & Stahl, K. 2015 Modeling drought impact occurrence based on meteorological drought indices in Europe. *Journal of Hydrology* **530**, 37–50.

USDA 2016 MALAYSIA: El Nino Takes a Bite Out of 2015/16 Palm Oil Production. In: *International Production Estimates Division (IPAD)*. Office of Global Analysis (OGA), USA.

Vapnik, V. 1995 *The Nature of Statistical Learning Theory*. Springer Verlag, New York, USA.

Venkataraman, K., Tummuri, S., Medina, A. & Perry, J. 2016 21st century drought outlook for major climate divisions of Texas based on CMIP5 multimodel ensemble: implications for water resource management. *Journal of Hydrology* **534**, 300–316.

Vicente-Serrano, S. M., Beguería, S. & López-Moreno, J. I. 2010 A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *Journal of Climate* **23** (7), 1696–1718.

Wilhite, D. A. & Glantz, M. 1985 Understanding the drought phenomenon: the role of definitions. *Water International* **10** (3), 111–120.

Wilhite, D. A., Sivakumar, M. V. K. & Pulwarty, R. 2014 Managing drought risk in a changing climate: the role of national drought policy. *Weather and Climate Extremes* **3**, 4–13.

Wiriyarattanakul, S., Auephanwiriyakul, S. & Theera-Umpon, N. 2009 Runoff Forecasting Using Fuzzy Support Vector Regression. In: *2008 International Symposium on Intelligent Signal Processing and Communication Systems. Thailand*, 8–10 February 2009, Bangkok, Thailand.

WMO 2008 *Guide to Hydrological Practices*. World Meteorological Organization (WMO) and Global Water Partnership (GWP), Geneva, Switzerland.

WMO 2012 *Standardized Precipitation Index User Guide*. World Meteorological Organization (WMO), Geneva, Switzerland.

Xiao, M., Zhang, Q., Singh, V. P. & Liu, L. 2016 Transitional properties of droughts and related impacts of climate indices in the Pearl River basin, China. *Journal of Hydrology* **534**, 397–406.

Zadeh, L. A. 1965 Fuzzy sets. *Information and Control* **8**, 338–353.

First received 10 December 2018; accepted in revised form 28 May 2019. Available online 27 June 2019