Pol. J. Environ. Stud. Vol. 29, No. 1 (2020), 1293-1303
DOI: *10.15244/pjoes/105972* ONLINE PUBLICATION DATE: 2019-09-18

Original Research

Application of Artificial Neural Network and Climate Indices to Drought Forecasting in South-Central Vietnam

Luong Bang Nguyen[1]*, Manh-Hung Le[2]

[1]Faculty of Water Resource Engineering, Thuyloi University, Hanoi, Vietnam
[2]Department of Engineering Systems and Environment, University of Virginia, Charlottesville, VA USA

Received: 16 January 2019
Accepted: 26 March 2019

Abstract

Widespread negative consequences of droughts related to climate indices in Vietnam have motivated many studies integrating those indices to predict the onset of drought in the region. This study aims to examine the capacity of eight climate Pacific Ocean indices as input variables for forecasting the drought index at 30 stations of south-central Vietnam during the period 1977 to 2014. The standardized precipitation evapotranspiration index (SPEI) was selected as a predicted target drought index at four multiple time scales (3, 6, 9, and 12 months). Input variable selection filters (partial correlation input selection and partial mutual information selection) were used to select the suitable climate indices as input parameters, and an artificial neural network was applied for the drought model. The results showed that partial correlation input selection selected a better optimal input set for the drought model. The west tropical Pacific index (NINOW), east central tropical Pacific index (NINO34), and south oscillation index (SOI) were climate indices that could improve the drought forecasting performances at the given study.

Keywords: south-central region of Vietnam, SPEI, ENSO, input variable selection

Introduction

Prolonged drought is one of the possible causes of famine, an extreme condition shortage of food. Also, droughts have been blamed for environmental degradation and desertification. Given the threats of drought, it is critical to forecast drought accurately with sufficient lead time to mitigate some consequences – especially in developing countries where farmer activities primarily depend on rain-fed agriculture [1]. It often requires a considerable amount of time (e.g., a few months) to recognize drought impact on socioeconomic systems. Taking this advantageous feature, a prediction of the onset of drought conditions in advance can mitigate the most adverse consequences of drought rather than other extreme events such as flooding and hurricanes [2]. The drought forecasting model can be a foundation for an effective monitoring

*e-mail: nguyenluongbang77@tlu.edu.vn

system that can support water managers, characterize droughts, and determine risk scenarios [3]. Common meteorological drought indices selected for forecasting are the standard precipitation index (SPI) [4-6] and the standard precipitation evapotranspiration index (SPEI) [7-10]. In this study, SPEI was selected since this index considers both important factors affecting droughts (i.e., rainfall and temperature) [11], and the SPEI is proved to be better than SPI in the south-central Vietnam [12].

Traditional multi-linear regression (MLR) or autoregressive integrated moving average (ARIMA) has been widely applied in predicting the drought index [13]. In 2005, Vietnam's Thuyloi University (formerly Water Resources University) developed the MeDF2005 (meteorological drought forecasting) software package to apply to Vietnam's Central Highlands and south-central region [14]. This software used the MLR technique with four NINO regions (NINOW, NINO3, NINO4, and NINO12) as predictors to build forecasting models of each month and season (three months) at 24 stations and 7 zones. The shortcoming of the linear model lies in it may be not a suitable model for long-lead-time forecasting due to its assumption of linearity between predictor and predict and [15]. Recently, data-driven modeling has been paying more attention because it is built based on finding connections between the system state variables (input, internal and output variables) without explicit knowledge of the physical behavior of the system [16]. The usefulness of data-driven modeling in drought forecasting is clear, since variables that trigger a drought may not be well understood. One of the long standing, well-known data-driven models is the artificial neural network (ANN), which has been used extensively for various hydrological application purposes [17-20]. Therefore, the capability of ANN for a complex problem such as a prediction of the onset of drought is not new but still attracts numerous studies [15,21].

Input variable selection (IVS) is a crucial step in the succession of the application of data-driven modeling. The challenge of IVS is to select the fewest input variables that best characterize the relationship of input-output while minimizing variable redundancy [22]. Predictors for drought forecasting can be divided into two groups: local variables and climate indices. Local variables are often lagged observations of quantifying drought indices or rainfall. The climate indices that measure the large scale of the ocean and atmospheric circulation are associated with the rainfall variation at distant continents across the world, inducing periods of moderate to extreme drought. Therefore, those low-frequency variation indices could become potential predictors of drought onset [17]. Selecting the best input set for drought forecasting is a challenge, since many input combinations could be created [23], and the success of implication often depends on subjective judgment or expert knowledge [24]. To overcome this shortcoming, May et al. (2005) [25] adopted iteration input variable selection based on partial correlation coefficient and partial mutual information algorithm. Although it is shown practically in several environmental issues [26,27], there is still little research on drought forecasting.

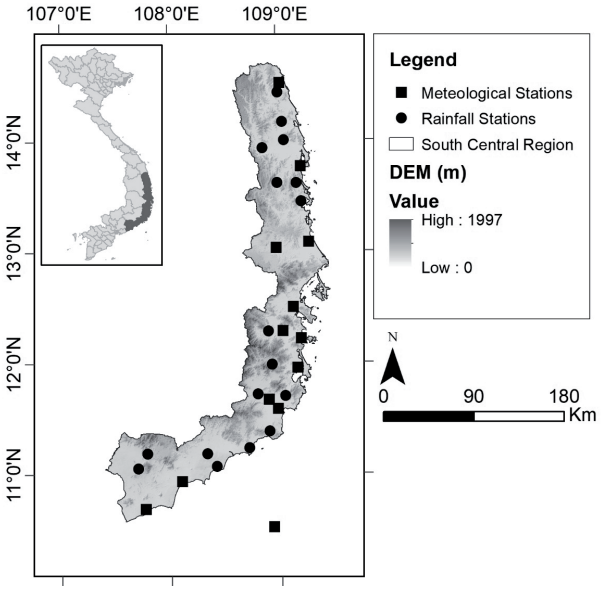

Fig. 1. Distribution of meteorological stations in south-central Vietnam.

This study selected the south-central region of Vietnam as a case study since this region is one of the most severe drought-prone areas in Vietnam. The objectives of the study are: (1) To examine the capacity of two filter algorithms (partial correlation input selection and partial mutual information selection) for selecting suitable input parameters to predict drought index using the ANN model and (2) To investigate which climate indices are important inputs for improving drought forecasting in the case study.

Study Area and Materials

Study Area

The South Central Region of Vietnam (SCR) has a total of 27500 km^2, stretching from 10°50'N - 14°50'N and 107°50'E - 109°20' E (Fig. 1). The topography of the SCR has a west-east gradient with high elevation in the west and narrow flat plains along the coastal east. The average annual temperature ranges 26-27.3ºC, with maximum temperature of up to 40-42ºC [28]. The temperature resource in the SCR is equal to the Central Highlands of Vietnam, less than the southern region, but much higher than the northern regions of Vietnam. The annual rainfall is typical in a range from 1200-1600 mm. However, some micro-climate areas at the narrow plains to the south of the SCR have less than 800 mm annual rainfall [28].

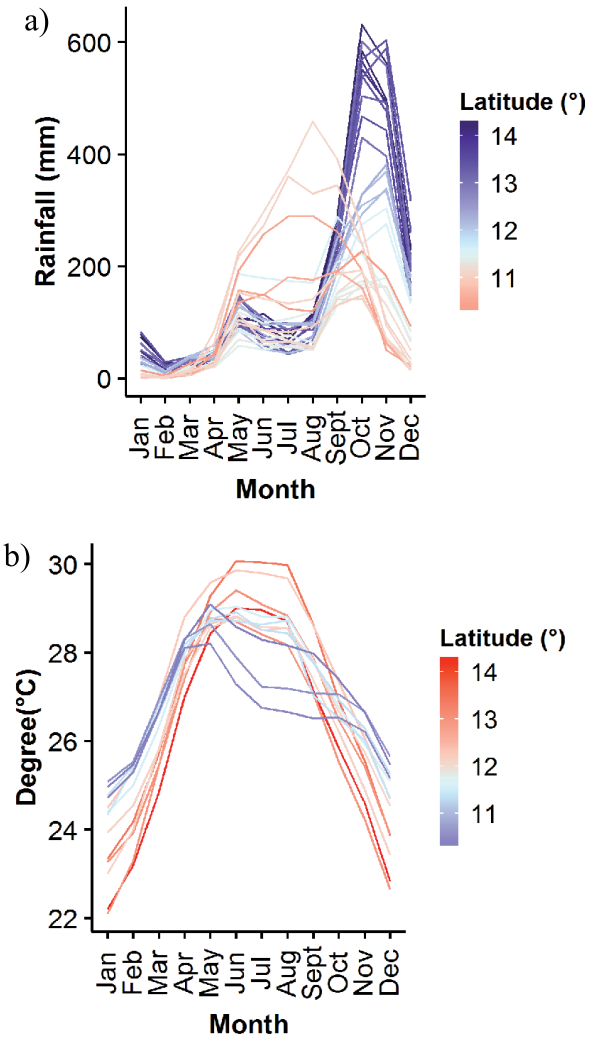

Fig. 2. Average monthly a) rainfall at 30 rainfall stations and b) temperature at 13 temperature stations over the SCR.

Data

Meteorological Data

In this study, 30 monthly rainfall and 13 monthly temperature datasets were collected from the Vietnam Meteorological and Hydrological Administration from 1977 to 2014. For a consistent data length for all datasets, missing values were filled using inverse distance weighting [28]. The SCR has two distinct seasons: dry and wet. The wet season varies along the latitude (Fig. 2). At 12° northward, a high rainfall amount is observed from September to December while below that degree, the rainfall season is from May to October. Similarly, monthly temperature value reaches their maximumsin June at stations above 12°N and in May at stations below 12°N.

Climate Indices (CIs)

Table 1 lists eight climate indices (CIs) used as potential candidates for input variables of the drought forecasting model. In a similar region, the use of ENSO indices as input for drought forecasting are common in previous studies [14, 29, 30]. The use of PDO as one of CIs' candidate input variables for drought forecasting is new for the study area. The PDO index represents a spatial pattern of sea surface temperature anomalies (SSTA) of the North Pacific Ocean (poleward 20°N). Although the PDO index is not an ENSO index, it reflects the impact of El Niño and La Niña on spatial patternsin the North Pacific. The main difference between ENSO and PDO indices lies in their time scales: while ENSO is a typical interannual event, the time scale of the PDO is decadal.

Methods

Drought Index: Standardized Precipitation Evapotranspiration Index

The multiple-month scales SPEI drought index (3, 6, 9, and 12 months) were calculated and considered as observational data for the drought forecasting model. The SPEI index was calculated as the desired time series (e.g., time series) fitted to a probability distribution, which is then transformed into a normal distribution, leading to a zero mean of SPEI value. The more negative the SPEI value, the drier the condition. -1.0, -1.5, and -2.0 are thresholds that represent starts of dry, severely dry, and extremely dry conditions, respectively.

Three steps of SPEI calculation [11] are described as below:

Table 1. CIs used as potential input variables for drought forecasting.

No.	Index	Abbr.	No.	Index	Abbr.
1	El Niño Modoki Index	EMI	5	Southern Oscillation Index	SOI
2	West Tropical Pacific Index	NINOW	6	Bivariate ENSO time series	BEST
3	Central Tropical Pacific Index	NINO4	7	Multivariate ENSO Index	MEI
4	East Central Tropical Pacific Index	NINO34	8	Pacific Decadal Oscillation	PDO

Note: NINOW was obtained from NOAA Asia-Pacific Data Research Center website (http://apdrc.soest.hawaii.edu/las/v6/constrain?var=287), the remaining indices were obtained from the NOAA Earth System Research Laboratory Physical Sciences Division website (http://www.esrl.noaa.gov/psd/data/climateindices/list/).

Step 1: Fitting distribution for time series G

$$G_{y,m}^{k} = \sum_{t=13-k+m}^{12} D_{y-1,t} + \sum_{t=1}^{m} D_{y,t} \quad (1)$$

With:

$$D_{y,m} = P_{y,m} - PET_{y,m} \quad (2)$$

…where k is time scale (i.e.,3, 6, 9, and 12 months), m and y represent month and year respectively, and P and PET represent precipitation and potential evapotranspiration respectively. In this study, the Thornthwaite method was used to calculate PET since this method requires only temperature. The fitting function of time series G is usually a three-parameter log-logistic distribution. The cumulative probability distribution function is calculated as below:

$$F(g) = \left[1 + \left(\frac{\alpha}{g-\gamma}\right)^{\beta}\right]^{-1} \quad (3)$$

…where α, β, and γ are shape, scale, and original parameter.

Step 2: The cumulative probability of time series is computed relative to the fitting distribution function.

$$H(g) = F(g) \quad (4)$$

…where $H(g)$ is the cumulative probability of time series G.

Step 3: The cumulative probability is transformed into the standard normal variable, and the SPEI is estimated (Z value).

$$Z = \begin{cases} -\left(W - \dfrac{C_0 + C_1 W + C_2 W}{1 + d_2 W + d_2 W^2 + d_3 W^3}\right) & 0 < H(g) \le 0.5 \\ +\left(W - \dfrac{C_0 + C_1 W + C_2 W^2}{1 + d_2 W + d_2 W^2 + d_3 W^3}\right) & 0.5 < H(g) \le 1 \end{cases} \quad (5)$$

…where:

$$Z = \begin{cases} \sqrt{-2\,ln(H(g))} & 0 < H(g) \le 0.5 \\ \sqrt{-2\,ln(1 - H(g))} & 0.5 < H(g) \le 1 \end{cases} \quad (6)$$

…and:

$$C_0 = 2.515517, C_1 = 0.802853, C_2 = 0.010328$$
$$d_1 = 1.432788, d_2 = 0.189269, d_3 = 0.001308$$

Drought Forecasting Flowchart

The flowchart of drought forecasting is presented in Fig. 3a). We divided input data into two groups. Group 1 was a set of models in which the input variable was local variables (lagged observation of drought index and rainfall). Group 2 was a set of models in which the input variable included local variables and CIs. After creating an initial candidates (IC_i) set, two filter algorithms – partial correlation input selection (PCIS) and partial mutual information selection (PMIS) procedures – were chosen to reduce input variables to form reduction candidates (RC_i). Both of those follow a forward selection strategy in which one variable is selected at each iteration. Details for each step of the flowchart can be described as below.

The *SPEI* drought index can be predicted from initial candidates IC_i as follows:

$$SPEI_t = F(IC_i) \quad (7)$$

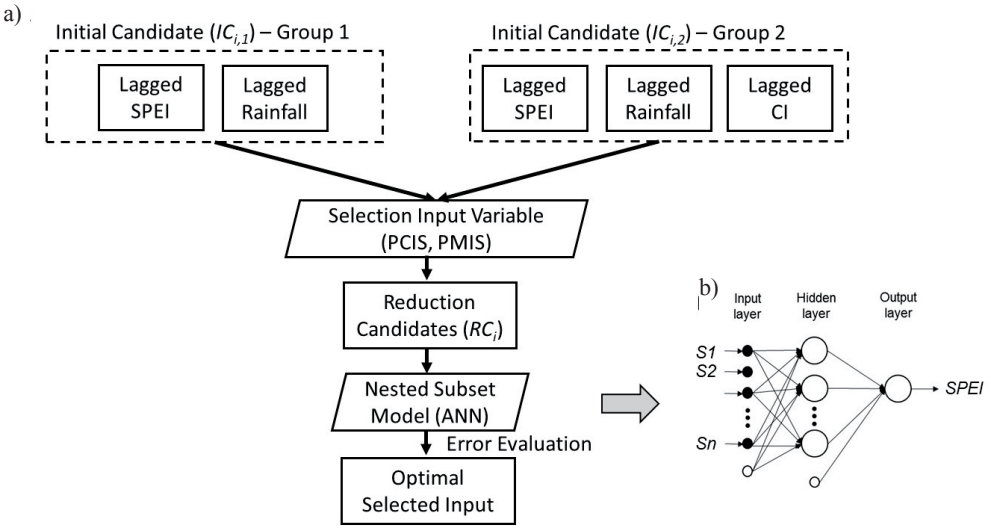

Fig. 3. a) Flowchart of drought forecasting and b) scheme of multilayer perceptron neural network; S1, S2, …Sn are rank selected input variables using PCIS or PMIS filters.

…where $F(.)$ is a data-driven model such as ANN; IC_i is formed by a number of input candidate pools $X_n(t-1, t-2,..., t-i)$ where X_n generally indicates previous observation of SPEI, rainfall or CIs, and $t-i$ indicates lagged time of the above-mentioned variables. We selected $i = 3$ as maximum lagged time for local candidates [31] and $i = 12$ as maximum lagged time for CIs [32].

The Pearson linear correlation coefficient measures the strength and direction of the linear relationship between each input variable and the output. While this coefficient is useful to illustrate linear dependence between the independent variable IC_i and the dependent variable $SPEI_t$, this relationship cannot account for redundancy in the candidate pool IC_i. To consider this redundancy, the partial correlation coefficient is used instead. This algorithm is a coefficient to describe the relationship, for example between $X_{1,t-i}$ and $SPEI_t$ while eliminating the effect of a set of reduction candidates $RC_k\{S_{1,t-i}, S_{2,t-i}\}$. The computation of the partial correlation coefficient is defined as a Pearson correlation coefficient between two residuals – linear regression of $X_{1,t-i}$ with $S_{1,t-i}$, $S_{2,t-i}$, and linear regression of $SPEI_t$ with $S_{1,t-i}$, $S_{2,t-i}$, as formulas (8) – (11):

$$X_{1,t-i} = a_{01} + a_{11}S_{1,t-i} + a_{21}S_{2,t-i} + \varepsilon \quad (8)$$

$$r_{X_{1,t-i}} = X_{1,t-i} - \left[a_{01} + a_{11}S_{1,t-i} + a_{21}S_{2,t-i}\right] \quad (9)$$

$$SPEI_t = a_{02} + a_{12}S_{1,t-i} + a_{22}S_{2,t-i} + \varepsilon \quad (10)$$

$$r_{SPEI_t} = SPEI_t - \left[a_{02} + a_{12}S_{1,t-i} + a_{22}S_{2,t-i}\right] \quad (11)$$

…where a_{xy} are regression coefficients, ε is error term, $r_{X_{1,t-i}}$ and r_{SPEI_t} are residuals in multi-linear regression models for $X_{1,t-i}$ and $SPEI_t$ respectively on other variables.

Mutual information measures a reduction of uncertainty in knowing from gaining knowledge of each input variable. However, MI raises the same problem as the Pearson linear correlation coefficient in a matter of redundancy. To address this issue, partial mutual information is developed to reveal a true relationship between $SPEI_t$ and $X_{n,t-i}$ while taking away the effect of other interactive factors such as formulas (12) – (14) [33]:

$$PMI = \iint f_{X'SPEI'}(X', SPEI') log\left[\frac{f_{X'SPEI'}(X', SPEI')}{f_{X'}(X') f_{SPEI'}(SPEI')}\right] dxdy \quad (12)$$

$$X' = X - E[X|S] \quad (13)$$

$$SPEI' = SPEI - E[SPEI|S] \quad (14)$$

…where $E[.]$ denotes an expectation operation and, $f_{X'}$, $f_{SPEI'}$ and $f_{X'SPEI'}$ represent marginal and joint distribution functions. As suggested from Galelli et. al. (2014) [34], the Gaussian density function was applied to estimate density function. The PMI criterion is analogous as a partial correlation coefficient since X' and $SPEI'$ generally represent the residual information of $X_{n,t-i}$ and $SPEI_t$ on the conditional S, which is a set of reduction candidates RC_i.

In short, PCIS or PMIS algorithms can follow as below [25].

– At the first iteration, the partial correlation (PC) or partial mutual information (PMI) value were calculated for each candidate input variable from IC_i using Eq. (8) - (14) with an empty reduction candidate set RC_i. The input variable with the highest PC or PMI was added to RC_i. It is noticed that there are two sets, one for PCIS algorithm and another for PMIS algorithm.
– At the next iteration, the PC value was calculated by formulas (8) - (11), and PMI value was calculated by formulas (12) - (14) on condition of selected input variables in the corresponding RC_i. The variable with the highest value PC or PMI was added for each RC_i.

The stopping criteria are determined based on the coefficient of determination R^2 when the new adding input variable could not be improved or even reduced, or the maximum iteration is reached. PCIS and PMIS algorithms use a program running on R software provided by Galelli et. al. (2014) [34]. This program is available online at http://ivs4em.deib.polimi.it/?page_id=7.

The reduction candidates set $RC_k\{S_{1,t-i}, S_{2,t-i}\}$ then comprised overlapping sets (i.e., nested subset) by incrementally adding variables. Those were then fed into the artificial neural network (ANN) model to select the optimal subset using statistical evaluation.

ANN is one of the most widely used artificial intelligence techniques because such a model could be built based on a highly nonlinear relationship without any prior knowledge. Among a pool of ANN approaches, a multilayer perceptron feed-forward neural network was used in this study due to its popularity. The architecture of the model includes three layers: input, hidden and output layer (Fig. 3b). Weights and bias connect each layer, but no weight is assigned between nodes within layers. The multilayer perceptron is worked in such a wayto minimize the error between output values of model and target values by updating the weights between each node.

An activation function of a node is used to determine the output of that node on giving a set of input. The $\varphi(x)$ is a common activation function used in ANN [35], with an output between -1 and 1. This function is described as follows:

$$\varphi(x) = \frac{2}{1 + e^{-2x}} - 1 \quad (15)$$

Before running the ANN model, all data series was rescaled to the range of $[a, b]$ by the following formula:

$$x_{rescaled} = \frac{(b-a)(x-x_{min})}{x_{max}-x_{min}} + a \quad (16)$$

…where $x_{rescaled}$ is the rescaled value of x, and x_{max} and x_{min} are maximum and minimum values of original time series respectively. A range [-0.9, +0.9] was selected since we can extrapolate unseen data that may have larger values than the available data.

To validate the ANN model, we split data into two subsets: training and verification. The training data was from 1977 to 2000 while the verification data was from 2001 to 2014. Additionally, the number of hidden nodes affects model performance [17]. Therefore, the optimal hidden node is found by trial and error procedure, which varies the hidden node number in a certain range.

Error Evaluation

To compare observed SPEI and predicted SPEI we used coefficient of determination () and root mean square error (RMSE) [32, 36]. The formulas for those statistical metrics are:

$$R^2 = \left(\frac{\sum_{j=1}^{N}(SPEI_{o,j} - \overline{SPEI_o})(SPEI_{p,j} - \overline{SPEI_p})}{\sqrt{\sum_{j=1}^{N}(SPEI_{o,j} - \overline{SPEI_o})^2 \sum_{j=1}^{N}(SPEI_{p,j} - \overline{SPEI_p})^2}} \right)^2 \quad (17)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(SPEI_{p,j} - SPEI_{o,j})^2} \quad (18)$$

…where N is the total number of samples; $SPEI_{o,j}$ and $SPEI_{p,j}$ represent monthly SPEI values for the observed data and the predicted data, respectively; and $SPEI_o$ and $SPEI_o$ represent the mean of the corresponding variables. R^2 score ranges from 0 to 1, R^2 is closer to 1, more fit between observation and prediction. *RMSE* score reflects deviations of the predictions from observation with perfect value as 0.

Results and Discussion

Performance of PCIS and PMIS Filter Algorithms

There is a total of 8640 ANN models (1440 models for Group 1 input set and 7200 models for Group 2 input set)run to find the optimal input sets (Table 2). Group 1 initial input set has six variables on which we ran six nested models for each station. The Group 2 initial input set has 102 variables (six local variables + 8x12 CIs). However, to reduce the running workload, we only selected the first 30 input variables after filtering to run the nested models.

Performance of two filter algorithms (PCIS, PMIS) on the verification set is presented in Table 3. To examine the efficiency of two filter algorithms, Table 3 also included performances of models using maximum input parameters for each group. The advantage of PCIS and PMIS filter algorithms is clear. Both algorithms not only provided a reduction in the input dimension solution but also increased the performance of the ANN model. While there was a slight increase in performance before and after using filter algorithms for Group 1, a significant enhancement performance was observed before and after using filter algorithms for Group 2. The reason seems to be that full input set of Group 2 had a high degree of redundancy or irrelevant variables, serving as adding noise and complexity during the training period, leading to poor performance for the verification set. On the other hand, a large portion of potential candidates had a chance to find more relevant input variables to better explain the variation of output target. In inter-comparison between two filters, PCIS was slightly better to find optimal input sets than PMIS. The average scores of PCIS filter with Group 2 input set regarding SPEI3, SPEI6, SPEI9, and SPEI12 prediction were 0.55, 0.75, 0.82, and 0.89 respectively; while those figures of PMIS filter were 0.54, 0.74, 0.82, and 0.88 respectively. Similarity, the RMSE score was slightly lower with PCIS filter as compared to the PMIS filter. The average RMSE scores of PCIS filter with Group 2 input set regarding SPEI3, SPEI6, SPEI9, and SPEI12

Table 2. Description of ANN models to predict four different SPEI time scales.

Input variables	Group 1	Group 2
Stations	30	30
SPEI time scales (*SPEI3, SPEI6, SPEI9, SPEI12*)	4	4
Filter Algorithms (*PCIS, PMIS*)	2	2
Maximum input variables	6	102
Nested models	6	30
Total running models	1440	7200

Table 3. Performance of ANN models using two filters (PCIS and PMIS)on verification data set at four different SPEI time scales.

	IC.G1		IC.G2		G1.PCIS		G1.PMIS		G2.PCIS		G2.PMIS	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
SPEI3												
Max	0.68	0.85	0.58	1.03	0.68	0.81	0.68	0.81	0.67	0.80	0.67	0.80
Min	0.27	0.58	0.14	0.68	0.33	0.57	0.33	0.56	0.40	0.57	0.39	0.58
Mean	0.50	0.71	0.37	0.83	0.53	0.68	0.53	0.68	0.55	0.68	0.54	0.68
SPEI6												
Max	0.82	0.61	0.74	1.16	0.85	0.63	0.85	0.65	0.85	0.73	0.85	0.65
Min	0.60	0.43	0.12	0.59	0.61	0.42	0.61	0.42	0.61	0.42	0.61	0.42
Mean	0.72	0.53	0.42	0.82	0.74	0.52	0.74	0.52	0.75	0.52	0.74	0.51
SPEI9												
Max	0.90	0.60	0.85	1.33	0.90	0.53	0.90	0.53	0.90	0.56	0.90	0.59
Min	0.63	0.33	0.04	0.39	0.66	0.33	0.66	0.33	0.62	0.31	0.66	0.32
Mean	0.80	0.45	0.51	0.77	0.82	0.43	0.82	0.43	0.82	0.43	0.82	0.44
SPEI12												
Max	0.93	0.76	0.93	1.02	0.93	0.49	0.93	0.48	0.93	0.47	0.93	0.50
Min	0.75	0.28	0.43	0.29	0.76	0.28	0.76	0.28	0.74	0.27	0.76	0.28
Mean	0.88	0.39	0.81	0.54	0.88	0.36	0.88	0.35	0.89	0.35	0.88	0.35

Note IC Initial input candidates; G1 Group 1; G2 Group 2.

prediction were 0.66, 0.52, 0.43, and 0.35 respectively while those figures of PMIS filter were 0.68, 0.51, 0.44, and 0.35 respectively. Inter-comparison between different SPEI time scales prediction showed that long-term time scales (i.e., SPEI9, SPEI12) prediction achieved better accuracy than short-term time scales (i.e., SPEI3, SPEI6). This was likely because of the high variability of short-term time scales, which caused difficulty to predict from ANN models.

We averaged the SPEI time series for all stations at multiple time scales and did the ANN forecast for those time series using two filter algorithms. The time series plots and residual density (i.e., predicted SPEI minus observed SPEI) on verification period is presented in Fig. 4. All four predicted model SPEI values followed the observation SPEI values at four time scales: SPEI3, SPEI6, SPEI9, and SPEI12. The predicted SPEI using PMIS filter with the Group 1 input set seems to underestimate SPEI values, especially for SPEI6, and SPEI12 time series. The predicted SPEI using Group 2 input dataset was best for following the observations trend. Regarding residual density, both predicted SPEI values as input parameters were Group 2 dataset, and the maximum density values were closest to zero values, with a slightly better density of error shape from the PCIS filter.

Analysis of Input Parameters

Figs 5-6 present the frequency appearances of 10 input parameters in optimal input sets. Both filters always selected lagged SPEI as important input parameters regardless of time scales. It was likely that lagged SPEI always had the highest correlation with current SPEI compared to other inputs. Lagged rainfall seems to be important for long-rather than short-term time scales. For example, 20% and 5% were the frequency appearance of lagged rainfall as in optimal input set to predict SPEI3 and SPEI6 using the PCIS filter respectively. Similarity, those figures for PMIS filter were only 10% and 45%. Regarding SPEI9 and SPEI12, the frequency appearance of lagged rainfall in the optimal input set was higher with values larger than 50% for both filters. Regarding CI variables, NINOW, NINO34, and SOI were important input parameters using PCIS filters as each had higher than 50% optimal input set. Similarly, using the PMIS filter, NINOW and NINO34 were also important input parameters, but for medium- to long-term time scales (SPEI6, SPEI9, and SPEI12). It is noticed that SOI was the most important CI among others as the frequency of this index in the optimal input set was very high with both filters, with frequency appearances larger than 60%. Typically, for SPEI6, SPEI9, and SPEI12 prediction

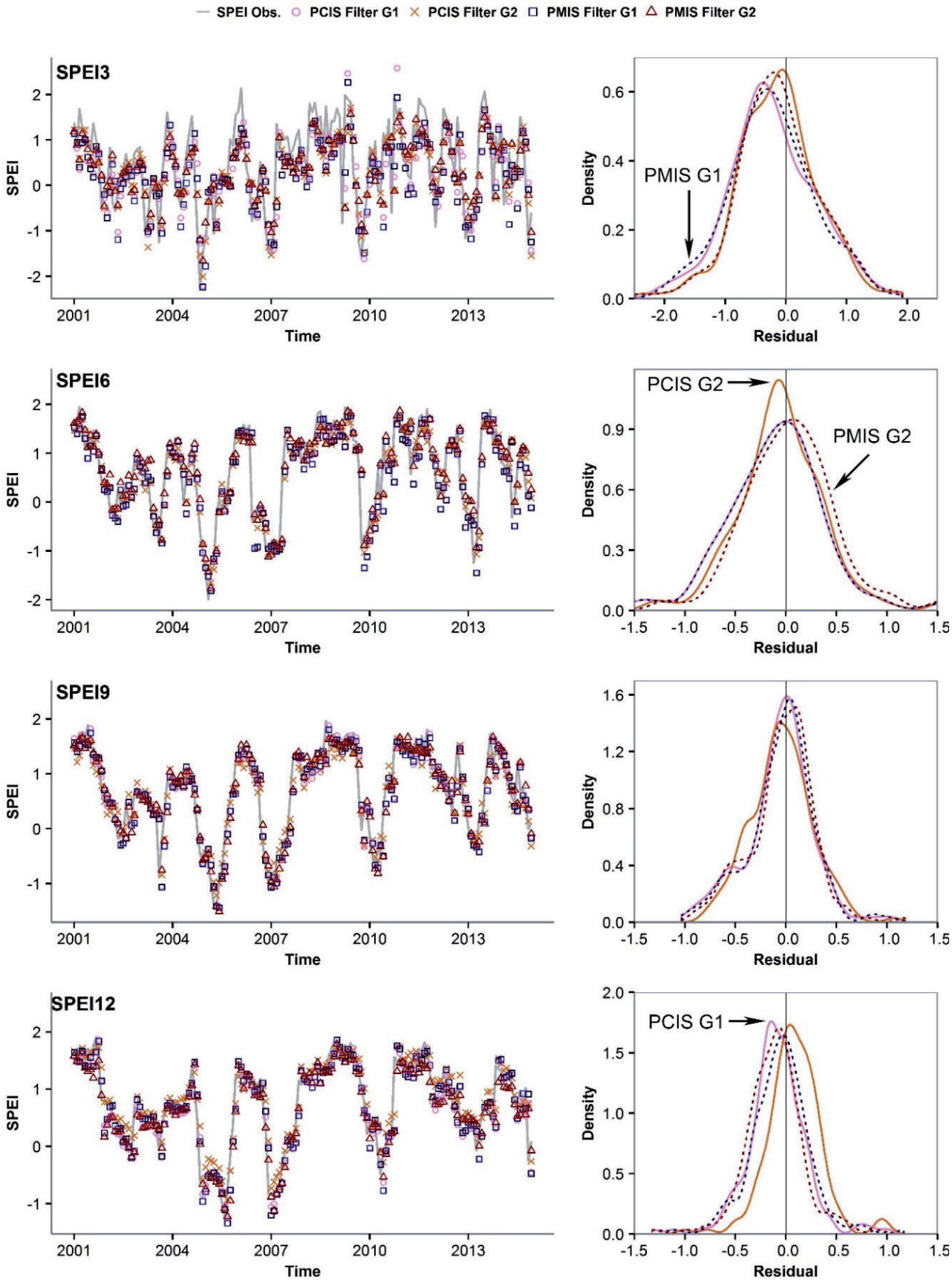

Fig. 4. Comparison average multiple-SPEI time series over the SCR between observations and predicted SPEI using PCIS and PMIS with different group input sets.

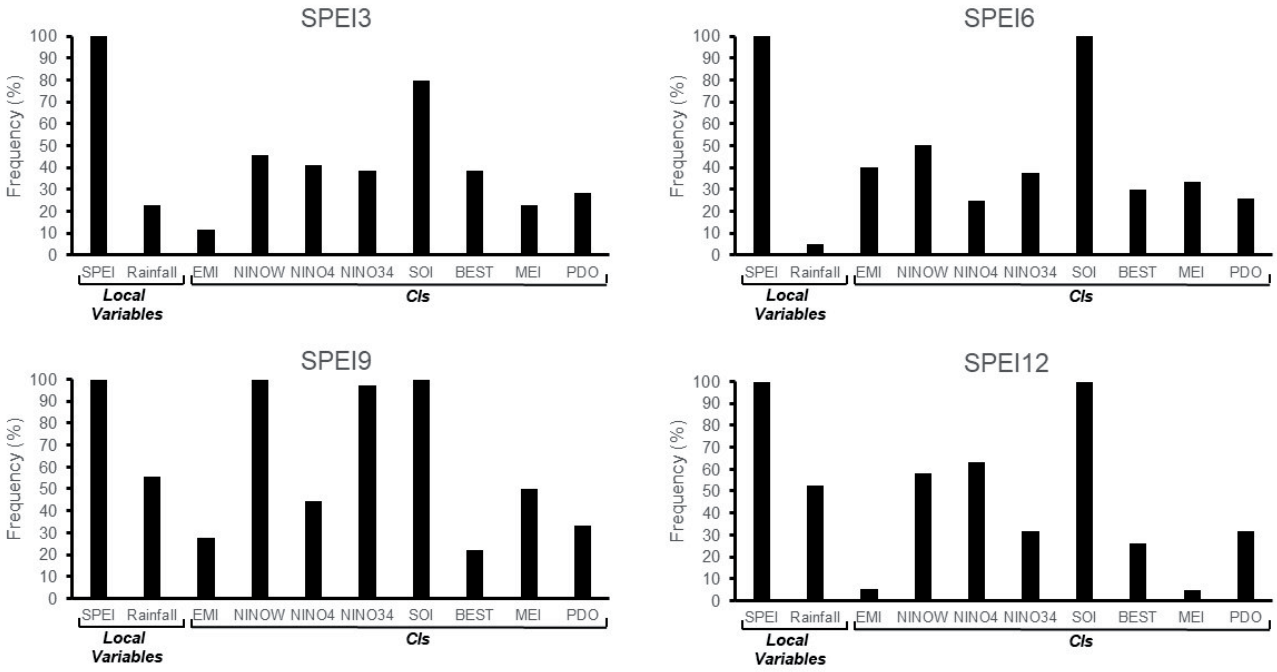

Fig. 5. Frequency appearance of different input variables in optimal input set using PCIS filter.

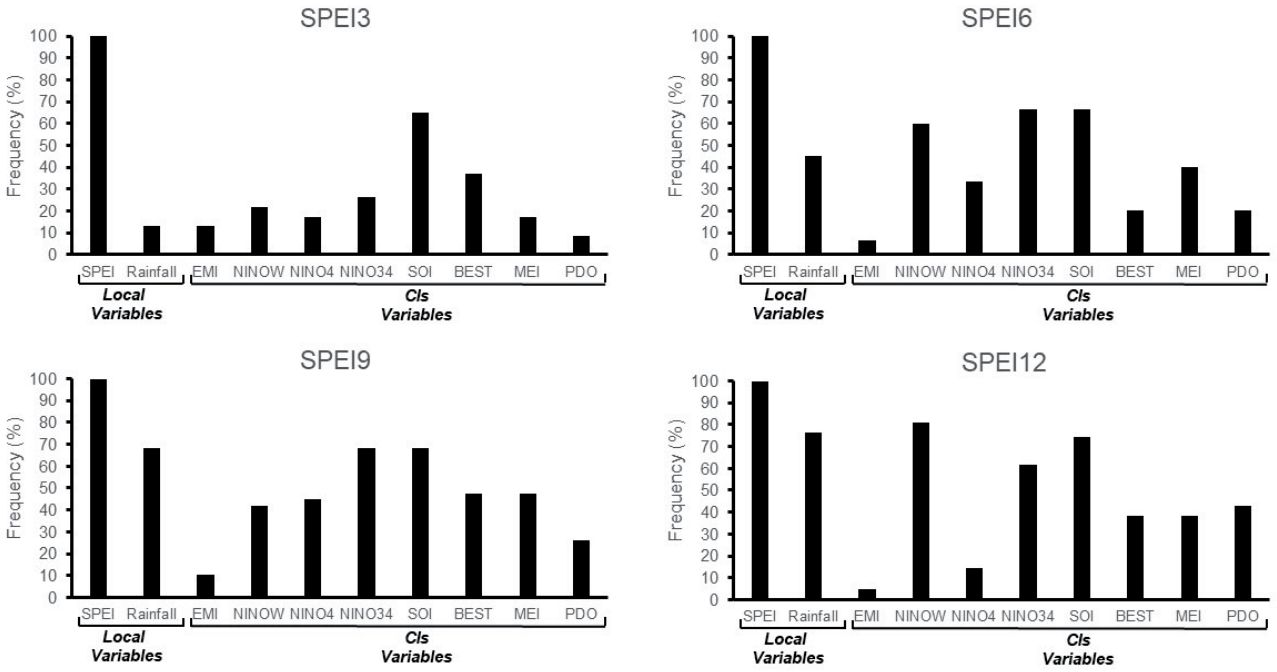

Fig. 6. Frequency appearance of different input variables in optimal input set using PMIS filter.

using PCIS filter, SOI appeared in all optimal input sets. The importance of NINO34 as the input of drought forecasting for the case study was confirmed in previous literature [14,30]. However, in this study the SOI index was a possible new input variable for drought prediction in Vietnam. In fact, previous studies showed that its role in drought conditions in Australia was claimed [37]. The distance from the region's calculated climate indices to the case study might affect their appearances in the drought prediction models. For example, NINOW had more frequent appearance in the predicted models than PDO. NINOWis calculated as an average sea surface temperature in the Western Pacific region (0°N-15°N, 130°E-150°E), which is geographically closer to the case study than the region calculated Pacific decadal oscillation (20°N-65°N, 120°E-105°W). On the other hand, some previous research confirmed the importance of PDO in predicting

drought for its neighbor areas such as the USA [38] or South Korea [17].

Conclusions

Climate indices such as ENSO indices are known as an important factor in triggering drought in many regions around the globe. This study aims to integrate eight climate indices (i.e., EMI, NINOW, NINO4, NINO34, SOI, BEST, MEI, and PDO) in the drought forecasting model for south-central Vietnam (SCR). The SPEI was selected as a predicted target drought index at four multiple time scales (3-month, 6-month, 9-month, and 12-month) at 30 stations over the SCR during the period 1977-2014. Since the potential input variables were large (up to 102 inputs and 12 lagged times), input variable selection filters (partial correlation input selection – PCIS and partial mutual information selection – PMIS) were used to select the suitable climate indices as input variables, and artificial neural network was applied for the drought model.

By evaluating the performance of PCIS and PMIS filter algorithms on selecting the optimal input set, it seems to us that models using PCIS filter input data set achieve moderately better prediction than those using the PMIS filter. Moreover, the faster computation time of PCIS, compared with that of PMIS, suggested the recommendation about the use of PCIS rather PMIS. This finding is also like the same suggestion from the work of Tran et al. (2015) [39].

For analyzing the frequency appearance of input parameters in the optimal input set, NINOW, NINO34, and SOI were the climate indices that appeared most often. This means that including those indices could enhance prediction accuracy and assist in designing guidelines for drought mitigation plans.

Acknowledgments

The first author gratefully appreciates Thuyloi University for financial support. The authors thank the Vietnam Meteorological and Hydrological Administration for their meteorological data provided in this study. The authors also thank the anonymous reviewers for their valuable comments to improve our paper's quality.

Conflict of Interest

The authors declare no conflict of interest.

References

1. BELAYNEH A., ADAMOWSKI J. Drought Forecasting: Artificial Intelligence Methods. In Exploring Natural Hazards, Chapman and Hall/CRC; 207, **2018**.
2. SPINONI J., NAUMANN G., CARRAO H., BARBOSA P., VOGT J. World drought frequency, duration, and severity for 1951-2010. International Journal of Climatology, **34** (8), 2792, **2014**.
3. HARO-MONTEAGUDO D., SOLERA A., ANDREU J. Drought early warning based on optimal risk forecasts in regulated river systems: Application to the Jucar River Basin (Spain). Journal of Hydrology, **544**, 36, **2017**.
4. ALIZADEH M.R., NIKOO M.R. A fusion-based methodology for meteorological drought estimation using remote sensing data. Remote sensing of environment, **211**, 229, **2018**.
5. AHMADEBRAHIMPOUR E., AMINNEJAD B., KHALILI K. Application of global precipitation dataset for drought monitoring and forecasting over the Lake Urmia basin with the GA-SVR model. International Journal of Water, **12** (3), 262, **2018**.
6. CHOUBIN B., MALEKIAN A., GOLSHAN M. Application of several data-driven techniques to predict a standardized precipitation index. Atmósfera, **29** (2), 121, **2016**.
7. SOH Y., KOO C., HUANG Y., FUNG K. 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, **2018**.
8. MANATSA D., MUSHORE T., LENOUO A. Improved predictability of droughts over southern Africa using the standardized precipitation evapotranspiration index and ENSO. Theoretical and applied climatology, **127** (1-2), 259, **2017**.
9. MACA P., PECH P. Forecasting SPEI and SPI drought indices using the integrated artificial neural networks. Computational intelligence and neuroscience, **2016**, 14, **2016**.
10. CHEN Y.D., ZHANG Q., XIAO M., SINGH V.P., ZHANG S. Probabilistic forecasting of seasonal droughts in the Pearl River basin, China. Stochastic environmental research and risk assessment, **30** (7), 2031, **2016**.
11. BEGUERÍA S., VICENTE-SERRANO S.M., REIG F., LATORRE B. Standardized precipitation evapotranspiration index (SPEI) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. International Journal of Climatology, **34** (10), 3001, **2014**.
12. LE H.M., CORZO G., MEDINA V., MERCADO V.D., NGUYEN B.L., SOLOMATINE D.P. A Comparison of SpatialeTemporal Scale Between Multiscalar Drought Indices in the South Central Region of Vietnam. Spatiotemporal Analysis of Extreme Hydrological Events, 143, **2018**.
13. HAN P., WANG P., TIAN M., ZHANG S., LIU J., ZHU D. In Application of the ARIMA models in drought forecasting using the standardized precipitation index, International Conference on Computer and Computing Technologies in Agriculture; Springer; 352, **2012**.
14. NGUYEN Q.K. Drought investigation and risk reduction on Vietnam South Central Region and Central Highlands (in Vietnamese); Vietnam Ministry of Science and Technology: Ho Chi Minh City, 2005.
15. MOUATADID S., RAJ N., DEO R.C., ADAMOWSKI J.F. Input selection and data-driven model performance optimization to predict the Standardized Precipitation and Evaporation Index in a drought-prone region. Atmospheric research, **212**, 130, **2018**.

16. MOUNT N.J., MAIER H.R., TOTH E., ELSHORBAGY A., SOLOMATINE D., CHANG F.-J., ABRAHART R. Data-driven modelling approaches for socio-hydrology: opportunities and challenges within the Panta Rhei Science Plan. Hydrological Sciences Journal, **61** (7), 1192, **2016**.
17. LEE J., KIM C.-G., LEE J., KIM N., KIM H. Application of artificial neural networks to rainfall forecasting in the Geum River basin, Korea. Water, **10** (10), 1448, **2018**.
18. YUAN F., BERNDTSSON R., UVO C.B., ZHANG L., JIANG P. Summer precipitation prediction in the source region of the Yellow River using climate indices. Hydrology Research, **47** (4), 847, **2016**.
19. NOORI N., KALIN L. Coupling SWAT and ANN models for enhanced daily streamflow prediction. Journal of Hydrology, **533**, 141, **2016**.
20. ALIZADEH M.J., KAVIANPOUR M.R., KISI O., NOURANI V. A new approach for simulating and forecasting the rainfall-runoff process within the next two months. Journal of hydrology, **548**, 588, **2017**.
21. ALI Z., HUSSAIN I., FAISAL M., NAZIR H.M., HUSSAIN T., SHAD M.Y., MOHAMD SHOUKRY A., HUSSAIN GANI S. Forecasting drought using multilayer perceptron artificial neural network model. Advances in Meteorology, **2017**, **2017**.
22. GUYON I., ELISSEEFF A. An introduction to variable and feature selection. Journal of machine learning research, **3** (Mar), 1157, **2003**.
23. MOREIRA E.E., PIRES C.L., PEREIRA L.S. SPI drought class predictions driven by the North Atlantic Oscillation index using log-linear modeling. Water, **8** (2), 43, **2016**.
24. SANTOS J.F., PORTELA M.M., PULIDO-CALVO I. Spring drought prediction based on winter NAO and global SST in Portugal. Hydrological Processes, **28** (3), 1009, **2014**.
25. MAY R.J., MAIER H.R., DANDY G.C., FERNANDO T.G. Non-linear variable selection for artificial neural networks using partial mutual information. Environmental Modelling & Software, **23** (10-11), 1312, **2008**.
26. WANG L., YAN Y., WANG X., WANG T. Input variable selection for data-driven models of Coriolis flowmeters for two-phase flow measurement. Measurement Science and Technology, **28** (3), 035305, **2017**.
27. SHAO X., YE G., HAN R. In Variable Selection Based on Hybrid Ant Colony Optimization-Simulated Annealing for Spectroscopy Quantitative Analysis of Underground Cable Tunnel, 2018 37th Chinese Control Conference (CCC); IEEE; 8091, **2018**.
28. LE M.H., PEREZ G.C., SOLOMATINE D., MEDINA V. In Studying the impact of infilling techniques on drought estimation - A case study in the South Central Region of Vietnam, 2017 Seventh International Conference on Information Science and Technology (ICIST); 292, **16-19 April, 2017**.
29. LE M.H., PEREZ G.C., SOLOMATINE D., NGUYEN L.B. Meteorological Drought Forecasting Based on Climate Signals Using Artificial Neural Network – A Case Study in Khanhhoa Province Vietnam. Procedia Engineering, **154**, 1169, **2016**.
30. NGUYEN V., LI Q., NGUYEN L. Drought forecasting using ANFIS-a case study in drought prone area of Vietnam. Paddy and water environment, **15** (3), 605, **2017**.
31. NGUYEN L.B., LI Q.F., NGOC T.A., HIRAMATSU K. Adaptive Neuro-Fuzzy Inference System for Drought Forecasting in the Cai River Basin in Vietnam. J Fac Agric Kyushu Univ, **60** (2), 405-415, **2015**.
32. MORID S., SMAKHTIN V., BAGHERZADEH K. Drought forecasting using artificial neural networks and time series of drought indices. International Journal of Climatology, **27** (15), 2103, **2007**.
33. BOWDEN G.J., DANDY G.C., MAIER H.R. Input determination for neural network models in water resources applications. Part 1 – background and methodology. Journal of Hydrology, **301** (1-4), 75, **2005**.
34. GALELLI S., HUMPHREY G.B., MAIER H.R., CASTELLETTI A., DANDY G.C., GIBBS M.S. An evaluation framework for input variable selection algorithms for environmental data-driven models. Environmental Modelling & Software, **62**, 33, **2014**.
35. REZAEIANZADEH M., TABARI H., YAZDI A.A., ISIK S., KALIN L. Flood flow forecasting using ANN, ANFIS and regression models. Neural Computing and Applications, **25** (1), 25, **2014**.
36. BELAYNEH A., ADAMOWSKI J., KHALIL B. Short-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet transforms and machine learning methods. Sustainable Water Resources Management, **2** (1), 87, **2016**.
37. AZMI M., RÜDIGER C. Validating the data fusion-based drought index across Queensland, Australia, and investigating interdependencies with remote drivers. International Journal of Climatology, **38** (11), 4102, **2018**.
38. ÖZGER M., MISHRA A.K., SINGH V.P. Low frequency drought variability associated with climate indices. Journal of Hydrology, **364** (1-2), 152, **2009**.
39. TRAN H., MUTTIL N., PERERA B. Selection of significant input variables for time series forecasting. Environmental Modelling & Software, **64**, 156, **2015**.