PREDICTION BASED ON COPULA ENTROPY AND GENERAL REGRESSION NEURAL NETWORK

HUANG, C. Y.[1] – ZHANG, Y. P.[2*]

[1]*School of Mathematics and Statistics, North China University of Water Resources and Electric Power, Zhengzhou, Henan 450011, China*

[2]*North China University of Water Resources and Electric Power, Zhengzhou, Henan 450011, China*

**Corresponding author*
e-mail: zhangyunpeng@ncwu.edu.cn; phone: +86-139-3826-4486

(Received 7th Jun 2019; accepted 10th Oct 2019)

Abstract. Drought prediction is the premise of effective response to the hazards of drought, a serious water-deficit phenomenon resulting from the complex interaction between climatic factors. Considering the high non-linearity of drought, this paper establishes a drought prediction model based on copula entropy and general regression neural network (GRNN). The climate indices that directly affect the formation of drought were selected as the inputs of the GRNN. The established model was applied to predict droughts for Lanzhou site of the middle reaches of the Yellow River. The results show that our model can effectively predict drought occurrence in the study area. With a strong predictive ability for drought, the proposed model is an ideal tool to forecast meteorological disasters based on climate factors.
Keywords: *Hampel, intelligence algorithms, SPI, neural network, prediction*

Introduction

Drought as one of the major natural disasters the human race faces directly threatening human life. It is now a serious factor that can restrict the sustainable development of community economy. Intensifying drought monitoring and prediction is significant for the relative authorities, in order to prevent drought and reduce losses caused by drought. However, the route that leads to drought is extremely complex, but it often develops and occurs as the result of a relatively slow process. Impacted by many factors, there has not been a relatively mature, accurate and universal model for drought prediction. The study is still in its infancy (Liu et al., 2019; Mossad and Alazba, 2018).

To predict drought, the regression model in time series is mostly used to construct future drought prediction model (Mokhtarzad et al., 2017) since it has a certain practicability. However, the meteorological forecast features significant nonlinearity. There is a highly complex nonlinear relationship between the predicted object and the predictor, but the regression model in the time series ignores this complex relationship.

By far, the Artificial Neural Network (ANN) model have been widely applied in the field of hydrology (Bello and Mamman, 2018; Garai et al., 2018; Isah et al., 2017; Mostefa et al., 2018; Qi et al., 2019; Sánchez-Escalona and Góngora-Leyva, 2018). When it is used to establish a drought prediction model, there is problem on how to select the predictors and the network models. To select predictive input factors, the practice is to consider the correlation between relevant drought predictors and drought indices, based on which to effectively predict the drought. As there is a highly complex nonlinear relationship between drought indices and prediction factors, it is considered the copula entropy can be used to determine the prediction factors; for the specific

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1441514424

network model, the Generalized Regression Neural Network (GRNN) with strong nonlinear mapping is used (Ladlani et al., 2012; McKee et al., 1993).

Based on all above, GRNN coupled with copula entropy (CE) is used to construct a drought prediction model based on CE-GRNN. The drought prediction correlation factor was selected by CE, and the GRNN was used for prediction to achieve drought prediction results.

This paper includes the following parts: Part 1 is the Introduction; Part 2 discusses the basic method and theory, including: the selection of predictors, the establishment of neural networks and the flow chart of the hybrid model. Part 3 is the Analysis of the results of the relevant sites; Part 4 gives the discussion, and the last part gives the Conclusion.

Materials and methods

Standard precipitation index

Standard Precipitation Index (SPI) (McKee et al., 1995; Shiau and Chiu, 2019) as one of drought indices was proposed by American scholar McKeed in 1993. It represents the standard deviation of the precipitation from the mean, similar to that in mathematical statistics.

SPI is defined as follows (Sharma, 2000):

Assume that the precipitation in an area is x, the probability density function is:

$$g(x) = \frac{1}{\beta^{\gamma}} x^{\gamma-1} e^{-x/\beta} \quad \text{(Eq.1)}$$

where: $\beta > 0$ is the scale, but $\gamma > 0$ is the shape parameter; β and γ can be obtained by the maximum likelihood estimation method:

$$\gamma = \frac{1+\sqrt{1+4A/3}}{4A} \quad \text{(Eq.2)}$$

$$\beta = \frac{\bar{x}}{\gamma} \quad \text{(Eq.3)}$$

where:

$$A = \lg \bar{x} - \frac{1}{n}\sum_{i=1}^{n} \lg x_i \quad \text{(Eq.4)}$$

where: x_i represents the precipitation; $\bar{x}$ represents the mean of precipitation; n represents the length of the sequence.

After determining the parameters for the probability density function of the distribution, the cumulative probability function is expressed as

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1441514424

$$G(x) = \int_0^x g(x)dx = \int_0^x \frac{1}{\beta^\gamma} x^{\gamma-1} e^{-x/\beta} dx \quad \text{(Eq.5)}$$

Since the *Gamma* function is undefined when $x = 0$, the precipitation value 0 can be transformed by the following formula:

$$H(x) = q + (1-q)G(x) \quad \text{(Eq.6)}$$

where: q is the probability of occurrence of value 0 in the precipitation sequence, and then $H(X)$ is normalized by a Gaussian function. The normal normalization is performed on *Gamma* distribution probability, so that an approximated solution can be obtained:

$$SPI = S(t - \frac{2.52 + 0.8t + 0.01t^2}{1 + 1.43t + 0.19t^2 + 0.001t^3}) \quad \text{(Eq.7)}$$

where:, $t = \sqrt{ln\frac{1}{H(x)^2}}$, $H(x)$ is the probability obtained by the *Equation 6*, when $H(x) > 0.5$, $S = 1$; when $H(X) \leq 0.5$, $S = -1$.

Option of prediction factors based on copula entropy

Copula Entropy, as a new entropy concept defined by Ma and Sun (Chen et al., 2015; Chen and Guo, 2019; Fernando et al., 2009; Hao and Singh, 2012; Ma and Sun, 2011) in 2008, can be used to measure the full order correlation between random variables. Here, Copula entropy can describe the correlation between the target and the predictive factors. How to use the copula entropy to select predictive factors requires an effective and reliable standard. That is to say, at the time of what the copula entropy value is, relevant prediction factors can be used as the input set of the prediction model. Here, the Hampel test recommended by Fernando and May, et al. is used as the stopping criterion of the algorithm (May et al., 2008).

The Hampel test algorithm is expressed as:

$$H_j = \frac{d_j}{1.4826 d_j^{(50)}} \quad d_j = \left| C_{PMI} - C_{PMI}^{(50)} \right| \quad \text{(Eq.8)}$$

where: H_j- Hampel distance; 1.4826 - normalization factor such that H_j is equal to the standard deviation σ of the data sequence; $d_j^{(50)}$- median of d_j; $C_{PMI}^{(50)}$- median of the values C_{PMI} in a set of data; $C_{PMI}^{(50)}$- copula entropy.

According to the standard deviation 3σ criterion, when the Hampel distance is greater than 3, the input variable will be included into the set of input variables, that is, determined as the prediction factors of the model.

According to the selection criteria of copula entropy, the prediction factors of the neural network model are determined by the following procedure:

1. Filter out all possible input variables;.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1441514424

2. Construct a joint distribution of individual factors and the predicted target, and select the optimal Copula function.
3. Determine the Copula density function with the selected Copula function, calculate the Copula entropy value of each factor and the predicted target.
4. Calculate the Hampel distance between the predicted target and the factor with the Hampel algorithm.
5. According to the test criteria, the determined factors are included in the input variable set as the last input factor for the neural network.

Establishment of NN model

The neural network is a new type of intelligent algorithm that can imitate the animal's nervous system. It depends on the neurons to link each other. there are three layers: the input layer, the hidden layer, and the output layer, each of which has its own special functions: the input layer performs factor input and processing; since input data must not be linear in real world, and sometimes is multidimensional. It is required to allow data to be trained via the hidden layer to make data visible, in order to obtain data as required. The hidden layer is therefore the core of the neural network; after the training via the hidden layer, data basically is what we want to be. Then it is required to output data via the output layer. These three layers are independent of each other, and the state of neurons in each layer can only affects those in the next layer. In this paper, the input variables are various factors related to drought prediction, and the output variables is drought prediction result. The framework of neural network is shown in *Figure 1*.

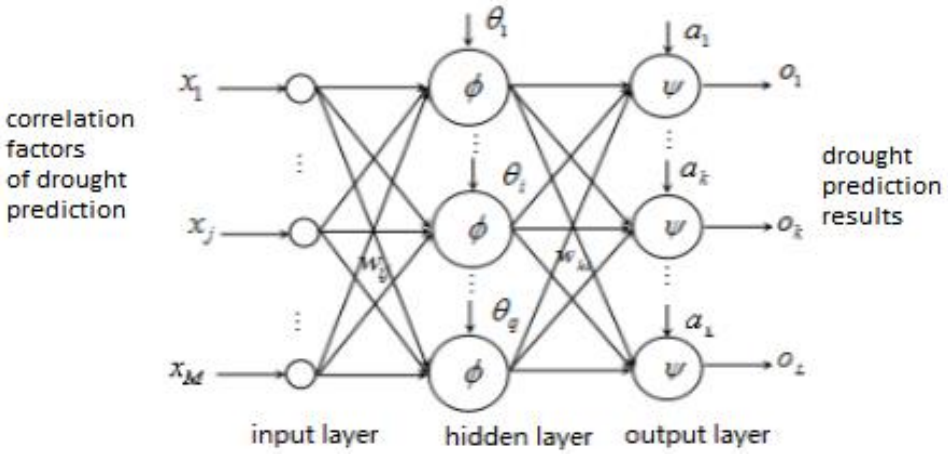

Figure 1. *Framework of neural network*

Hybrid prediction model based on Copula entropy and GRNN

When the GRNN is used to carry out the correlation prediction algorithm, the correlation between individual factors and drought object is first calculated based on the Copula entropy, and then the last input factor selected is chosen as the input layer of the neural network according to the Hampel distance. The drought prediction is also explored when appropriate. The relevant prediction process is shown in *Figure 2*.

In the process of establishing the GRNN model, more attention should be paid to the analysis of the physical meaning between prediction objects and the selection of

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1441514424

predictors. Undoubtedly, precipitation and drought are the most closely related. The SPI drought index is calculated from precipitation. Therefore, when the predictor is considered, the correlation between the nine related factors other than precipitation is studied. Among these related factors, CE method is used to select four factors which are most closely related to drought as prediction factors according to HAMPLE criterion. This coincides with the prediction content of the climate dynamics model proposed by Zeng Qingcun, an academician at the Institute of Atmospheric Physics of Chinese Academy of Sciences (Zeng, 1998).

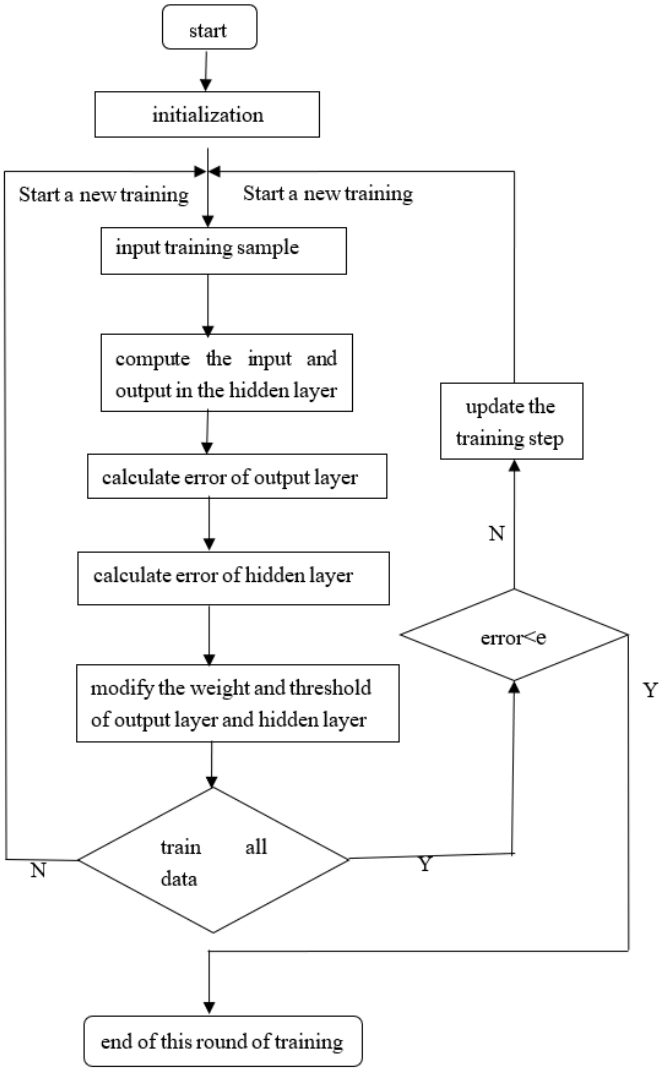

Figure 2. *Flow chart of prediction of neural network*

Results

Selection of input set

In order to verify the effectiveness of the algorithm of drought prediction in practice, the Lanzhou site in the middle reaches of the Yellow River basin of China is taken as study sample. The Lanzhou site is located at 34.29° north latitude and 110.05° east longitude. The distribution map of Lanzhou site and nearby sites is shown in *Figure 3*.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1441514424

For some reasons, only some relevant data are obtained. Therefore, meteorological data from 1957 to 2010 are used for analysis. In order to eliminate the inconsistency of data, normalization processing for homogenize data is adopted, and at the same time, interpolation method is used to process the abnormal value.

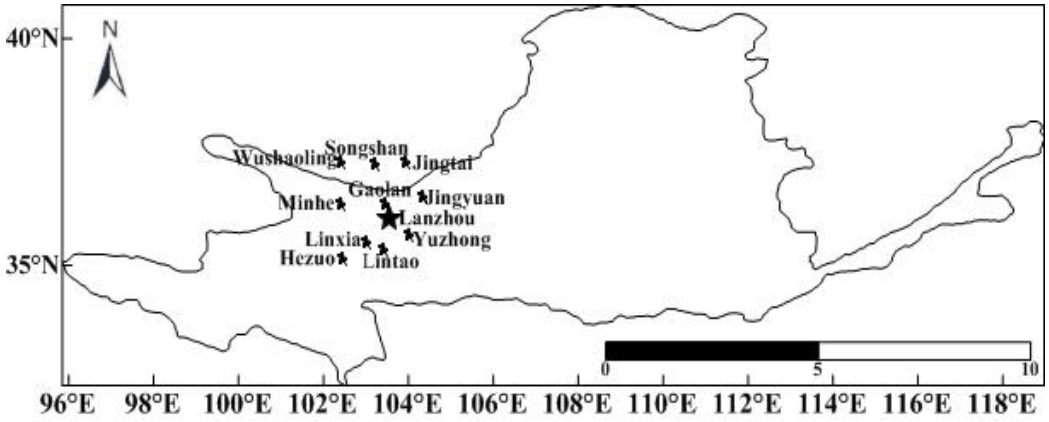

Figure 3. *Distribution map of Lanzhou site*

For selecting drought prediction factors, the following nine meteorological drought indices: wind speed, wind direction, air temperature, precipitation anomaly, temperature departure, vapor pressure, relative humidity, percentage of sunshine, sunshine hours, are used as initial factors and labeled as independent variables to calculate the Copula entropy, then n factors closely related to drought index of SPI were screened out based on the Hampel test which value is greater than 3 as the last set C. Finally, the variables of last set *C* are used as an input variable into the neural network for training. *Table 1* lists the copula value and the associated Hampel distance calculated by the above method from meteorological data of Lanzhou site.

Table 1. *Entropy value and Hampel list*

Factor (SPI)	Mutual entropy	Copula entropy	Hampel distance
Precipitation nomaly	0.29	-0.29	2.82
Wind speed	0.16	-0.16	3.05
Air temperature	0.39	-0.39	3.47
Temperature nomaly	0.09	-0.09	2.89
Vapor pressure	0.45	-0.45	3.31
Sunshine hours	0.07	-0.07	2.82
Relative humidity	0.29	-0.29	3.41
Percentage of sunshine	0.16	-0.16	2.77
Wind direction	0.01	-0.01	1.65

According to the Hampel test criteria, when the Hampel distance is greater than 3, the selected factors are used as the last input factors. As shown in *Table 1*, there are meteorological factors with Hampel distance greater than 3, including four factors: wind speed, air temperature, vapor pressure, and relative humidity. Therefore, these four factors are used as the last input set of variables of GRNN for drought prediction.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1441514424

Predicted results

The predictive analysis of drought is made with the Lanzhou site SPI as the prediction target, and the selected factors, i.e. wind speed, air temperature, vapor pressure and relative humidity, are input into the input layer of the neural network for correlation prediction. And in his paper, the error of the actual value and the predicted value is used to test the performance of the algorithms. The image of predicted results is shown in *Figure 4*. The error map for the actual and predicted values of the drought indices is shown in *Figure 5*, and the violin plot of the predicted results is shown in *Figure 6*.

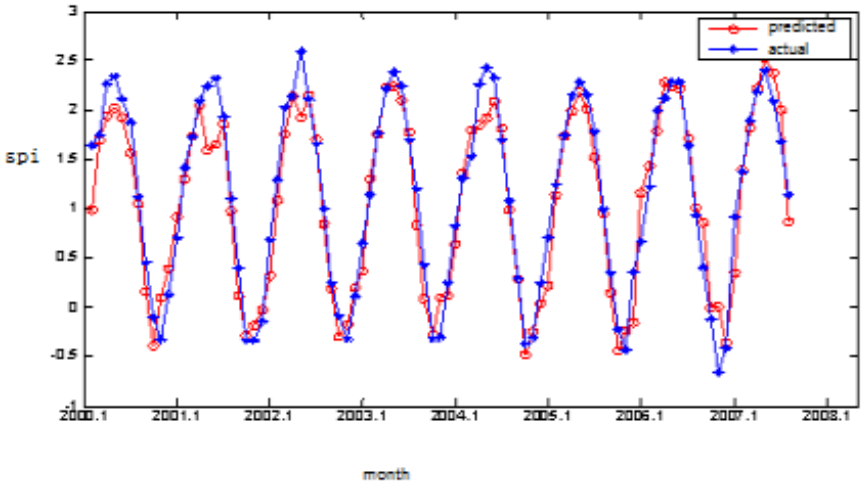

Figure 4. *Actual and predicted SPI value*

The red line in *Figure 4* represents the predicted value output, and the blue line represents the actual value. As shown in *Figure 4*, the fitting effect is better at most points, but the prediction effect at the critical values is relatively poor. In *Figure 5*, the value of the horizontal coordinate corresponds to the time coordinate in *Figure 4*, and it describes the error value of the corresponding time. As seen from the error results in *Figure 5*, the prediction effect is better.

Figure 5. *Error of actual drought and predicted values*

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1441514424

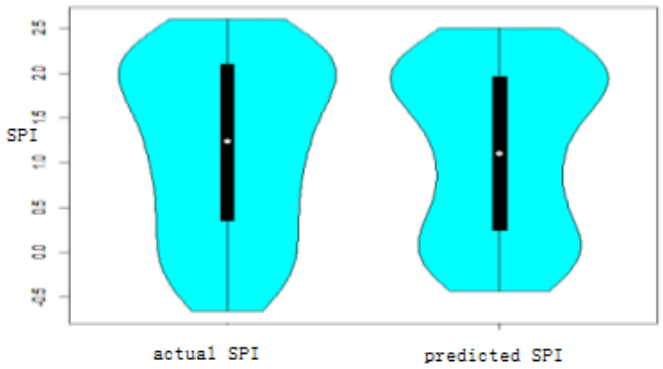

Figure 6. *Violin plot of actual and predicted results*

As shown in *Figure 6,* the simulated and actual values for the drought indices of Lanzhou site can well fit the nuclear density and box plot. The median of the simulated values is slightly lower than but substantially consistent with the actual value. It lies in half-way down the box, basically to be consistent with the actual situation. The maximum of the simulated values is substantially equal to that of the actual values, and the minimum value is relatively low.

As shown in *Figures 4*, *5* and *6,* the results simulated for the drought prediction in Lanzhou site further show that the hybrid algorithm based on Copula entropy and GRNN has better prediction effect.

Discussion

Traditional prediction method is designed for linear sequences, while predicted results from methods for nonlinear drought sequences are often unsatisfactory. It is extremely important to improve the prediction accuracy of nonlinear time series. Copula entropy can well describe how well the correlation between nonlinear drought sequences and drought predictors, and the neural network is also proven to be effective in predicting nonlinear drought time series.

Conclusion

Since the drought presents a complex nonlinearity, the traditional drought prediction algorithm has a relatively wide error. In contrast, the GRNN has a good nonlinear approximation performance, so that it features high robustness and fault tolerance. With copula entropy, the nonlinear dependence between factors and objects can be depicted. Therefore, this paper predicts the meteorological drought indices by establishing a hybrid model based on Copula entropy and GRNN. The results reveal that the well-established drought prediction model can directly predict the drought time sequence at a high accuracy rate. This model has a favorable generalization performance.

Although GRNN improves the prediction accuracy to some extent, but it is easy to fall into the extreme value, and the mapping ability of the non-stationary sequence is insufficient, which affects the accuracy of prediction to a certain extent. Therefore, some relevant parameter optimization algorithm (for example, the fruit fly optimization algorithm) can be considered to optimize the parameters of the neural network to further improve the prediction accuracy.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1441514424

Acknowledgements. This research was supported by Key Scientific Research Projects Plan of Henan Higher Education Institutions (19A120008). Sincere gratitude is extended to the editor and anonymous reviewers for their professional comments and corrections, which greatly improved the presentation of the paper.

REFERENCES

[1] Bello, A. A., Mamman, M. B. (2018): Monthly rainfall prediction using artificial neural network: A case study of Kano, Nigeria. – Environmental and Earth Sciences Research Journal 5(2): 37-41.
[2] Chen, L., Guo, S. (2019): Copulas and Its Application in Hydrology and Water Resources. Chap. 10. – Springer Water Series. Springer, Singapore, pp. 237-271.
[3] Chen, L., Singh, V. P., Guo, S., Zhou, J. Z., Zhang, J. H. (2015): Copula-based method for multisite monthly and daily streamflow simulation. – Journal of Hydrology 528: 369-384.
[4] Fernando, T. M. K. G., Maier, H. R., Dandy, G. C. (2009): Selection of input variables for data driven models: an average shifted histogram partial mutual information estimator approach. – Journal of Hydrology 367: 165-176.
[5] Garai, D., Agrawal, H., Mishra, A. K., Kumar, S. (2018): Influence of initiation system on blast-induced ground vibration using random forest algorithm, artificial neural network, and scaled distance analysis. – Mathematical Modelling of Engineering Problems 5(4): 418-426.
[6] Hao, Z., Singh, V. P. (2012): Entropy-copula method for single-site monthly streamflow simulation. – Water Resources Research 48(6): 6604.
[7] Isah, O. R., Usman, A. D., Tekanyi, A. M. S. (2017): A hybrid model of PSO algorithm and artificial neural network for automatic follicle classification. – International Journal Bioautomation 21(1): 43-58.
[8] Ladlani, I., Houichi, L., Djemili, L., Heddam, S. (2012): Modeling daily reference evapotranspiration (ET0) in the north of Algeria using generalized regression neural networks (GRNN) and radial basis function neural networks (RBFNN): a comparative study. – Meteorology & Atmospheric Physics 118(3-4): 163-178.
[9] Liu, Q., Zhang, G. L., Ali, S., Wang, X. P., Wang, G. D., Pan, Z. K., Zhang, J. H. (2019): SPI-based drought simulation and prediction using ARMA-GARCH model. – Applied Mathematics and Computation 355: 96-107.
[10] Ma, J., Sun, Z. (2011): Mutual information is copula entropy. – Tsinghua Science and Technology 16(1): 51-54.
[11] May, R. J., Maier, H. R., Dandy, G. C., Fernando, T. M. K. G. (2008): Non-linear variable selection for artificial neural networks using partial mutual information. – Environmental Modeling & Software 23: 1312-1326.
[12] McKee, T. B., Doesken, N. J., Kleist, J. (1993): The relationship of drought frequency and duration to time scales. – 8th Conference on Applied Climatology, Anaheim, CA, pp. 179-184.
[13] Mckee, T. B. N., Doeskin, N. J., Kleist, J. (1995): Drought Monitoring with Multiple Time Scales. – American Meteorological Society, Dallas, TX, pp. 233-236.
[14] Mokhtarzad, M., Eskandari, F., Vanjani, N. J., Arabasadi, A. (2017): Drought forecasting by ANN, ANFIS, and SVM and comparison of the models. – Environmental Earth Sciences 76(21): 729-735.
[15] Mossad, A., Alazba, A. A. (2018): Determination and prediction of standardized precipitation index (SPI) using TRMM data in arid ecosystems. – Arabian Journal of Geosciences 11(6): 132-139.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1441514424

[16] Mostefa, T., Tarak, B., Hachemi, G. (2018): An automatic diagnosis method for an open switch fault in unified power quality conditioner based on artificial neural network. – Traitement du Signal 35(1): 7-21.
[17] Qi, J. X., Jiang, G. Z., Li, G. F., Sun, Y. (2019): Surface EMG hand gesture recognition system based on PCA and GRNN. – Neural Computing and Applications (4): 1-9.
[18] Sánchez-Escalona, A. A., Góngora-Leyva, E. (2018): Artificial neural network modeling of hydrogen sulphide gas coolers ensuring extrapolation capability. – Mathematical Modelling of Engineering Problems 5(4): 348-356.
[19] Sharma, A. (2000): Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 - A strategy for system predictor identification. – Journal of Hydrology 239(1): 232-239.
[20] Shiau, J. T., Chiu, Y. F. (2019): Wavelet-based detection of time-frequency changes for monthly rainfall and SPI series in Taiwan. – Asia-Pacific Journal of Atmospheric Sciences 55(4): 657-667.
[21] Zeng, Q. C. (1998): A mathematic model of climate dynamics suitable for modern mathematical analysis. – Scientia Atmospherica Sinica 22(4): 408-417.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1441514424