Journal of Hydrology (NZ) 56 (2): 69-77 2017
© New Zealand Hydrological Society (2017)

An integrated hidden Markov model with an adaptive exponential weighting scheme for forecasting a meteorological drought index

Si Chen,[1] Dong-Hyeok Park,[1] Dongkyun Kim[2] and Tae-Woong Kim[3]

[1] *Department of Civil and Environmental Engineering, Hanyang University, Seoul 04763, Korea*

[2] *Department of Civil Engineering, Hongik University, Seoul 04066, Korea*

[3] *Department of Civil and Environmental Engineering, Hanyang University, Ansan 15588, Korea. Corresponding author: twkim72@hanyang.ac.kr*

Abstract

Drought is a naturally occurring climate phenomenon that significantly affects human and environmental activity, and can be considered one of the most widespread and destructive natural disasters. This makes drought forecasting vital for drought risk reduction. In this study, a hidden Markov model (HMM) with an adaptive exponential weighting (AEW) scheme (HMM-AEW) was proposed to develop a new framework to forecast the standardized precipitation index (SPI) as a meteorological drought measurement considering historically-similar patterns of the SPI. The model was applied to a monthly SPI series for a sub-basin of Han River in South Korea that covered more than 30 years. The performance of the HMM-AEW was measured using two commonly-used statistical criteria. The results indicated that the HMM-AEW is able to forecast most of the key points (i.e., fluctuation and extreme points) of the SPI series, with root mean squared error values of 0.612 and 0.368 and adjusted R-square values of 0.671 and 0.670, respectively. Furthermore, in terms of drought category, the proposed model exhibited a satisfying forecasting performance with hit rates of 80% and 50% for fluctuation and extreme points, respectively.

Introduction

Drought, which is known as a period of persistent abnormally dry weather, can have devastating effects on ecosystems and human life. Drought has been reported as ranking first worldwide among natural disasters in terms of number of people directly affected (Hewitt, 1997) and is the second most geographically-extensive global hazard after flood (Thomas *et al.*, 2016). Due to the complex nature of drought, one of the difficulties in mitigating its effects is accurately predicting drought conditions. Hidden Markov models (HMMs), which are a class of statistical models built on a probabilistic framework with wide practical applications, have been widely applied in various areas, including speech processing, bioinformatics, econometrics, and finance. Due to their strong statistical foundation, ability to deal with data robustly, computational efficiency, and ability to predict similar patterns (Hassan *et al.*, 2007), HMMs have become a widely-used tool to analyse and forecast hydro-meteorological time series. An HMM is a doubly-embedded stochastic process with an unobservable (hidden) underlying stochastic process, but can be observed through another set of stochastic processes that produce the sequence of observations. This feature fits well in the drought forecasting process since the underlying mechanisms of drought are relatively uncertain. Moreover, the historical data record that is very similar to the current data pattern will provide valuable information for forecasting drought conditions in the near future. Thus, considering the strengths of HMM, the objective of this study is to propose an improved HMM with an adaptive exponential weighting (AEW) scheme (HMM-AEW) for calculating the Standardized Precipitation Index (SPI), a commonly-used measurement of meteorological drought, based on selected historical similar patterns.

Study area and data

The Han River is located in the middle of the Korean Peninsula and is Korea's largest river. A sub-basin (code #1001) of the Han River was selected as the study area (Figure 1). The selected sub-basin is composed of mountains and valleys, and its drainage area is 2447.85 km^2 with an average slope of 51.09%, an effective basin width of 13.56 km, and an average elevation of 748.29 m (WAMIS, 2016). The in-situ monthly precipitation time series from 1979 to 2011 was obtained from Korean Water Management Information System (WAMIS).

Methodology

Standardized Precipitation Index

The SPI developed by McKee *et al.* (1993) is a tool that allows determination of the dryness at a given time scale (temporal resolution) for any rainfall station with historic data. The SPI is a z-score representing departure from the mean in standard deviation units. One advantage of using the SPI is that it can be used more effectively than rainfall itself in spatial and temporal analysis of drought. The SPI allows drought analyses at different time scales (e.g., 3, 6, 12, and 24 months). In this study, the SPI-3, which is tied to short-term drought conditions with a 3-month time scale, was adopted to quantify the severity of meteorological droughts. A drought category was assigned according to the SPI value as follows: non-drought (D0) when SPI ≥ 0; slight drought (D1) when -1.0 ≤ SPI < 0; moderate drought (D2) when -1.5 ≤ SPI < -1.0; severe drought (D3) when -2.0 ≤ SPI < -1.5; and extreme drought (D4) when SPI ≤ -2.0.

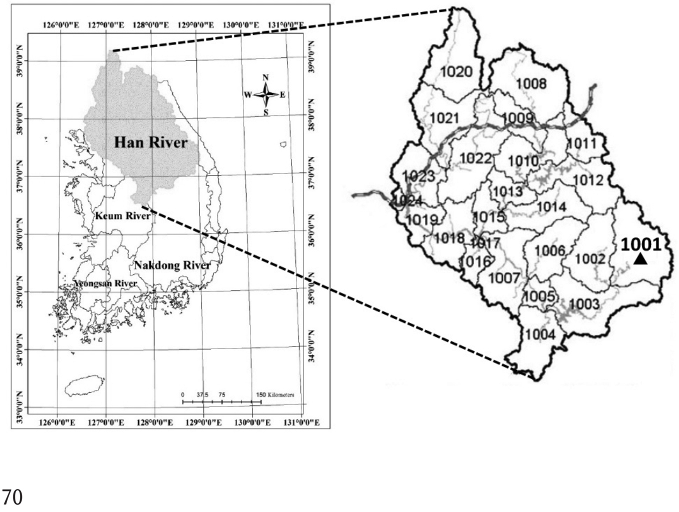

Figure 1 – Location of the study area (sub-basin 1001 marked with triangle). The dotted line indicates the Military Demarcation Line (MDL).

Forecast target

The forecast target in this study contained two kinds of data, i.e., the fluctuation points and extreme points of SPI series, during the validation periods to assess model performance. These points play a key role in the drought evolution process. An example time series plot is shown in Figure 2. If $y_{t-1} \leq y_t \geq y_{t+1}$ or $y_{t-1} \geq y_t \leq y_{t+1}$, t is defined as an extreme point in the time series (solid circles dots in Figure 2). The first point following the extreme point is defined as the fluctuation point (triangles in Figure 2). All other points are general points.

HMM-AEW based forecasting

Hidden Markov model (HMM)

An HMM mainly consists of a discrete-time bivariate process $\{(S_t, O_t)\}$ with $\{S_t\}$ being a latent or hidden Markov chain and $\{O_t\}$ being the observations at time *t*. $\{S_t\}$ and $\{O_t\}$ have the following two properties:

$$P\left(S_t \middle| S_{1:t-1}\right) = P\left(S_t \middle| S_{t-1}\right), t = 2,3,\ldots \quad (1)$$

$$P\left(O_t \middle| O_{1:t-1}, S_{1:t}\right) = P\left(O_t \middle| S_t\right), t = 1,2,\ldots \quad (2)$$

Equation 1 can be interpreted as an unobserved 'state process', $\{S_t\}$ satisfying the first-order Markov property, and Equation 2 shows that the 'state-dependent process' $\{O_t\}$ is directly governed by $\{S_t\}$. Taken together, the properties of HMM indicate that the joint likelihood function of the observations O_t and the unobserved sequence of states S_t is given by:

$$L\left(S_{1:T}, O_{1:T}\right) = P(S_1)P(O_1 | S_1)\prod_{t=2}^{T} P(S_t = j | S_{t-1} = i)P(O_t | S_t) \quad (3)$$

where $P(O_t|S_t)$ is defined as the emission probability. We assumed a Gaussian emission distribution $O_t|S_t \sim N(\mu_z, \sigma_z^2)$ where μ_z and σ_z^2 are the state-specific mean and variance of the Gaussian process, respectively. $P(S_t=j|S_{t-1}=i)$ is the transition probability from state *i* to state *j*, and there are *m* hidden states. A graphical interpretation of the HMM is shown in Figure 3. In this study, the SPI dataset served as the observations for the HMM, while the underlying drought states were treated as the hidden states in the model.

Parameter learning, including the number of hidden states, transition probability matrix, and emission distribution parameters, was performed using a Bayesian approach with a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm (Green, 1995). This

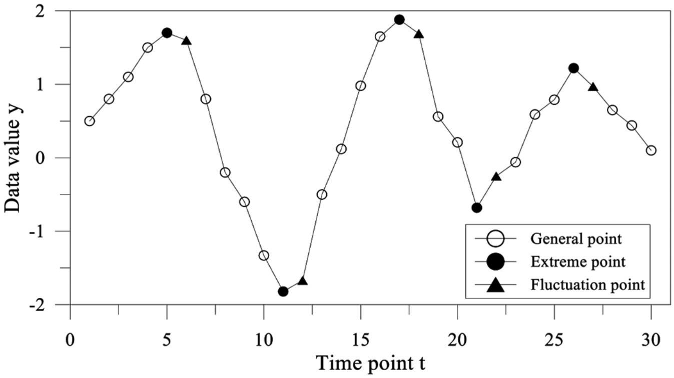

Figure 2 – An example SPI time series demonstrating fluctuation and extreme points.

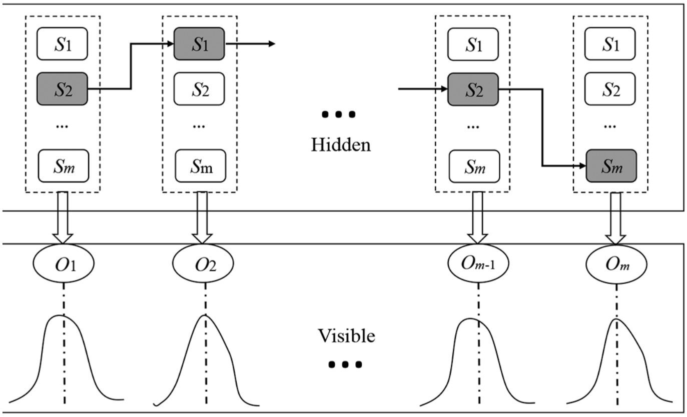

Figure 3 – Graphical representation of the HMM. *S* represents the hidden state variable and *O* represents the observation that is assumed to be governed by the corresponding state.

allows for the HMM with a varying number of hidden states to produce the posterior probability that represents the confidence in each model for selection. The optimal number of hidden states within the model can be selected with the largest model posterior probability, instead of using the usual model selection criterion of Akaike or Bayesian information. Details about the RJMCMC sampling and the prior distribution setting of each parameter can be found in Chen *et al.* (2017). Note that for a given model, the variances of emission distributions for all hidden states were constrained to be equal to provide a more robust model, which favours the hidden state with the mean value closer to the observed data at times when the data is far from either states' mean (McFarland *et al.*, 2011). For the selected best fitted model, according to Olivier *et al.* (2005), we obtained the log-likelihood estimator of each observed data point, as shown in Equation 3, via the forward algorithm and then decoded the posteriori most likely hidden state sequence via the Viterbi algorithm.

Similar patterns

It is generally known that a historical condition reflected in a data pattern will exhibit repetition in time in most cases (Jiang *et al.*, 2016). Therefore, we adopted two assumptions for the SPI data pattern. The first assumption is that the SPI data pattern in a massive historical dataset satisfies the characteristics of repetition. The second assumption is that the next time point (T+1) of the current point (T) during a forecasting period exhibits a similar data pattern to the next time point (t+1) of the referential point (t) during the historical period.

The SPI time series during the training period used in this study was chosen to validate the above assumptions. Four representative similar patterns are plotted in Figure 4, in which the corresponding similar data pattern can be clearly shown for each current point. The data patterns at time point T+1 and t+1 are indeed similar, which indicates that the assumptions are reasonable. Consequently, the similar patterns in historical data were assumed to provide useful information for the SPI forecasting in this study.

Forecasting based on similar patterns weighted by Adaptive Exponential Weighting (AEW)

Once the HMM is trained, the obtained log-likelihood estimator, which reflects the degree of similarity of data patterns, can be utilised to discover the time point with a similar pattern in the historical data to the current data:

$$\hat{t} = \arg\min_t \{|LL_T - LL_t|\},\ 1 \le t \le T\text{-}1 \qquad (4)$$

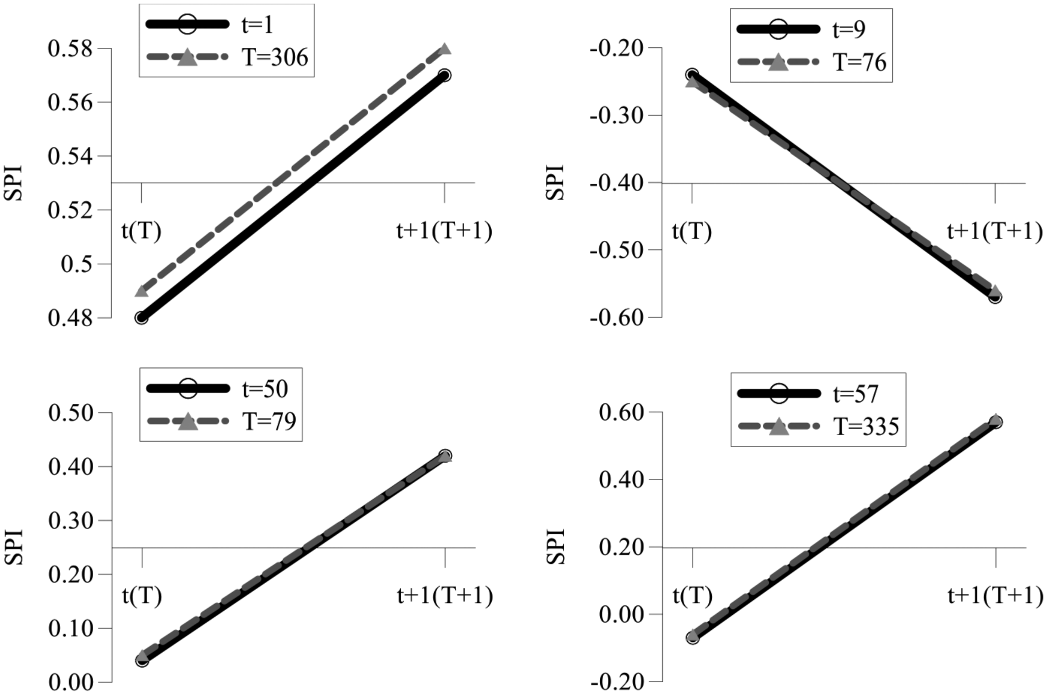

Figure 4 – Representative similar patterns of SPI value between the current and historical data pattern.

where LL_T and LL_t are log-likelihood values at the current time point (T) and historical time point (t), respectively.

Depending on only one similar pattern may lead to local optimal problems; therefore, several similar patterns from the historical data were considered for the forecasting in order to make the forecast robust. The time between historical and current data patterns also plays an important role in forecasts. Similar patterns closer in time were assumed to have more impact on one another. Therefore, in order to consider both the similarity and timeliness of similar patterns in the forecasting, the proposed AEW scheme is expressed as follows. First, m previous neighbour patterns before the current time point were selected to test the model and find the optimal weights in the AEW scheme, and n similar patterns were chosen to perform the forecast. In this study, three previous neighbour patterns and four similar patterns were used in the forecasting. $D_{m\times n}$ is the matrix of log-likelihood values reflecting the degree of similarity of similar patterns, and each row of D was ranked in a decreasing order. $T_{m\times n}$ is the matrix of time point values with each element corresponding to the elements in matrix D. $F_{m\times n}$ is the matrix of forecasts (i.e., data at the next time point) by each element in matrix T. An exponential weighting W^D and W^T for similarity and timeliness, respectively, was employed:

$$W^D = (\omega_{ij}^D)_{m\times n} = (\exp(\lambda^D D_{ij}) / \sum_j \exp(\lambda^D D_{ij}))_{m\times n} \quad (5)$$

$$W^T = (\omega_{ij}^T)_{m\times n} = (\exp(\lambda^T T_{ij}) / \sum_j \exp(\lambda^T T_{ij}))_{m\times n} \quad (6)$$

where λ^D and λ^T are the exponential parameters to be estimated. The mean absolute percentage error (*MAPE*) was selected as a criterion to evaluate the testing performance and find the optimal exponential parameters:

$$MAPE=\sum_{i=1}^{m}\left|O_{T-m+i}-[r\sum_{j=1}^{n}W^{D}(i,j)\cdot P(i.j)+(1-r)\sum_{j=1}^{n}W^{T}(i,j)\cdot P(i.j)]\right|/O_{T-m+i}\times\frac{100}{m} \tag{7}$$

where $O_{T\text{-}m+i}$ is the observed value at time $T\text{-}m\text{-}i$ and r is the ratio of similarity/timeliness. Parameters including λ^D, λ^T, and r that minimize the *MAPE* value were selected as the optimal values and were then used for forecasting at the current time *T*. The particle swarm optimization (Kennedy, 2010), which is a population-based stochastic evolutionary algorithm for global optimization, was used to adaptively search for the most suitable weighting parameters.

Performance evaluation

The root mean square error (*RMSE*) and the adjusted coefficient of determination (R_{adj}^2) were utilized to evaluate the performance of the forecast value:

$$RMSE=\sqrt{\frac{1}{n}\sum_{t=1}^{n}(O_t-F_t)^2} \tag{8}$$

$$R_{adj}^2=1-\frac{n-1}{n-k-1}\times\frac{\sum_{t=1}^{n}(O_t-F_t)^2}{\sum_{t=1}^{n}(O_t-\bar{O})^2} \tag{9}$$

where O_t and F_t are the observed and forecasted value at time *t*, respectively, *n* is the testing data size, and *k* is the number of variables in the model. In order to evaluate the performance of the categorical drought forecast expressed in terms of the five drought classes, the hit rate (*H*), defined as the number of hits divided by the total number of drought events observed, was adopted as a measure of discrimination of the proposed model.

Results and discussion

The SPI data with a 3-month time scale has a total of 393 monthly data points from 1979 to 2011. We selected the first 356 data points as the training dataset and used the last 37 data points as forecast targets, which included ten fluctuation points and six extreme (dry) points, for validating the model performance. For each forecast target, the forecast was made based on the current pattern O_T with the observations $\{O_1, \ldots, O_{T\text{-}4}\}$ forming the training SPI dataset for parameter learning in the HMM. Three previous neighbour patterns $\{O_{T\text{-}3}, O_{T\text{-}2}, O_{T\text{-}1}\}$ were used for the exponential weighting parameter optimization, and four similar historical patterns $\{O_{t1}, \ldots, O_{t4}\}$ were selected to perform the forecasting. Figure 5 presents a representative forecast for a fluctuation point and an extreme point, respectively, based on the four similar patterns. The forecast appeared similar to a weighted mean of all the similar patterns. The overall estimated parameters for forecasting these two points are listed in Table 1.

The forecast results of the HMM-AEW for all selected forecast targets, including the evaluation criteria *RMSE*, R_{adj}^2 and hit rate (*H*) results, are listed in Table 2. The proposed approach exhibited a better performance for the forecast of extreme points than fluctuation points in terms of the RMSE value. However, the R_{adj}^2 values for the two forecast targets were almost the same. Although several points with a positive (negative) value were forecasted as a value with an opposite sign, most of them can be accurately forecasted in terms of drought category according to the

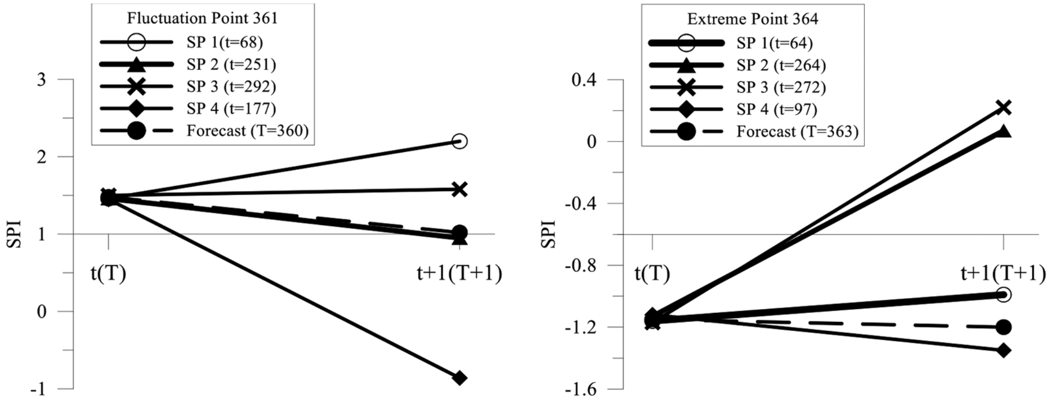

Figure 5 – Forecasts for fluctuation point 361 and extreme point 364 based on their historical similar patterns.

Table 1 – Results of parameter estimation in forecasting representative points.

Data point (#)	**HMM parameters**						**Weighting parameters**		
	Number of hidden states	**Transition probability matrix**	**Emission distribution**				λ^D	λ^T	r
			Mean			**Variance**			
			State1	**State2**	**State3**				
361	3	$\begin{bmatrix} 0.808 & 0.177 & 0.015 \\ 0.180 & 0.706 & 0.114 \\ 0.164 & 0.272 & 0.564 \end{bmatrix}$	-0.758	0.348	1.552	0.339	0.01	0.1	0.74
364	3	$\begin{bmatrix} 0.802 & 0.183 & 0.015 \\ 0.191 & 0.691 & 0.118 \\ 0.149 & 0.303 & 0.548 \end{bmatrix}$	-0.764	0.344	1.520	0.341	0.05	0.01	0.10

SPI classification. The hit rate for forecasted fluctuation and extreme points was 80% and 50%, respectively. Forecasts with large bias, such as fluctuation point 369, may be due to limitations of the historical patterns. All similar historical patterns for current point 368 had a following month value that was positive, which led to a positive forecast value, while the observation was -1.10. This main drawback of the proposed model occurs when the forecast pattern is very different from all historical similar patterns, which is a limitation of analogue forecasting, as it is virtually impossible to find a perfect analogue. However, this problem could be alleviated with a longer observation period and therefore more training data for the model. A long training dataset for the HMM is more likely to obtain the optimal model parameters that best describe the observation process and estimate the log-likelihood-based similarity. Moreover, as more observation data are archived, the chances of finding a 'good match' analogue for the current data pattern should improve, and the forecasting performance is expected to improve.

Table 2 – Forecast performance of HMM-AEW evaluated by statistical criteria

	Fluctuation Points (#)										*RMSE*	R_{adj}^2	*H*
	361	**365**	**369**	**376**	**380**	**382**	**384**	**386**	**388**	**391**			
O (DC)	1.02 (D0)	-1.79 (D3)	-1.10 (D3)	2.52 (D0)	0.41 (D0)	0.69 (D0)	-0.39 (D1)	0.75 (D0)	-0.17 (D1)	0.33 (D0)	0.615	0.666	80%
F (DC)	1.06 (D0)	-1.52 (D3)	0.47 (D0)	2.02 (D0)	-0.39 (D1)	0.74 (D0)	-0.76 (D1)	0.85 (D0)	-0.63 (D1)	0.36 (D0)			

	Extreme (dry) Points (#)						*RMSE*	R_{adj}^2	*H*
	364	**370**	**372**	**379**	**383**	**387**			
O (DC)	-1.89 (D3)	-1.72 (D3)	-0.19 (D1)	-0.15 (D1)	-0.80 (D1)	-0.30 (D1)	0.368	0.669	50%
F (DC)	-1.20 (D2)	-1.75 (D3)	0.22 (D0)	0.24 (D0)	-0.69 (D1)	-0.26 (D1)			

* O is observed SPI data; F is forecasted SPI data; DC is drought category.

Conclusions

This study investigated the capability of the HMM-AEW scheme for forecasting a meteorological drought index, the SPI. We constructed a reversible jump Markov chain Monte Carlo algorithm for inference on the model parameters and a particle swarm optimization evolution algorithm to optimize the weighting scheme for forecasting. The forecast based on the proposed model was carried out for the SPI series observed at a Han River sub-basin in South Korea. Both fluctuation and extreme points of the SPI data series during the validation period were forecasted. The results indicated that the HMM-AEW is a useful tool for SPI forecasting over the study area, witnessed by evaluation via statistical evaluation criteria *RMSE*, R_{adj}^2 and hit rate (*H*) values. The proposed scheme can be generalised to other sites and variables for possible applications in other hydrological process forecasting. However, it is worth noting that the forecasts were performed based on historical similar patterns, which may generate large bias for the performance when the forecast pattern is very different from all of the historical similar patterns. Thus, future work is needed to improve the model to reduce bias by accounting for the inherent uncertainty involved in the data gathering and model parameter learning processes to achieve improved probabilistic forecasting.

Acknowledgements

This work was supported by grants from the National Research Foundation (NRF-2016R1D1A1A09918872) and the Water Management Research Program (17AWMP-B083066-04) of the Korean government.

References

Chen, S.; Shin, J.Y.; Kim, T.W. 2017. Probabilistic forecasting of drought: A hidden Markov model aggregated with the RCP 8.5 precipitation projection. *Stochastic Environmental Research and Risk Assessment 13(5)*: 1061-1076.

Green, P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82(4)*: 711-732.

Hassan, M.R.; Nath, B.; Kirley, M. 2007. A fusion model of HMM, ANN and GA for stockmarket forecasting. *Expert Systems with Applications 33(1)*: 171-180.

Hewitt, K. 1997. *Regions at Risk. A Geographical Introduction to Disasters*. Addison Wesley Longman Limited, England.

Jiang, P.; Liu, X.J.; Zhang, X.Y. 2016. A framework based on hidden Markov model with adaptive weighting for microcystins forecasting and early-warning. *Decision Support Systems 84*: 89-103.

Kennedy, J. 2010. Particle swarmoptimization. *In:* Sammut, C.; Webb, G.I. *(eds.) Encyclopedia of Machine Learning*. Springer Science & Business Media. pp. 760-766.

McFarland, J.M.; Hahn, T.T.G.; Mehta, M.R. 2011. Explicit-duration hidden Markov model inference of up-down states from continuous signals. *PLoS ONE*, 6(6), e21606.

McKee, T.B.; Doesken, N.J.; Kleist, J. 1993. The relationship of drought frequency and duration to time scales. *Proceedings of the 8th Conference on applied climatology.* American Meteorological Society, Boston.

Olivier, C.; Eric, M.; Tobias, R. 2005. *Inference in Hidden Markov Models (Springer Series in Statistics).* Springer, Lund University, Sweden.

Thomas, T.; Jaiswal, R.K.; Galkate, R.; Nayak, P.C.; Ghosh, N.C. 2016. Drought indicators-based integrated assessment of drought vulnerability: a case study of Bundelkhand droughts in central India. *Natural Hazards 81(3)*: 1627-1652.

WAMIS (Water Management Information System). 2016. http://wamis.go.kr/eng/. Accessed October 2016.

Manuscript received 29 March 2017; accepted for publication 16 October 2017