지질학회지 제 52권 제 3호, p. 187-199, (2016년 6월)
J. Geol. Soc. Korea, v. 52, no. 3, p. 187-199, (June 2016)
DOI http://dx.doi.org/10.14770/jgsk.2016.52.3.187

ISSN 0435-4036 (Print)
ISSN 2288-7377 (Online)

기계학습 기법 기반 지하수위 변동 시계열 예측 모델의 국가지하수관측망 지하수위 자료 적용 연구

윤희성[1] · 윤필선[1] · 이은희[1,‡] · 김규범[2] · 문상호[1]
[1]한국지질자원연구원
[2]대전대학교 건설안전방재공학과

요 약

본 연구에서는 기계학습 기법인 인공신경망(ANN)과 지지벡터기계(SVM) 기반 지하수위 시계열 예측모델을 개발하고 이를 국가지하수관측망 216개소 지하수위 자료에 적용하였다. 시계열 예측 모델 적용의 목적은 결측 및 이상 자료를 보정하고 지하수위 자료에 포함된 양수 및 하천수위 변동 영향을 제거하여 강우에 의한 지하수위 변동 모의에의 적용성을 평가하는 것이다. 먼저 각 관측소에 대해 1일 직접예측 모델을 구성한 뒤 이를 이용한 반복예측 모델을 구성하고 각 모델의 예측 및 모의 능력을 검토하였다. 시계열 예측 모델 적용 결과 결측, 이상 자료, 양수 및 하천수위 영향을 효과적으로 보정하고 강우에 의한 지하수의 변동 패턴을 모의하는 것으로 나타났다. 오차지표 분석 결과 ANN 모델이 SVM 모델보다 다소 높은 직접예측능력을 보여주지만 반복예측 모델에 대한 안정성은 SVM이 높은 것을 알 수 있다. 또한 본 연구에서는 이용된 연구자료에 대한 ANN, SVM 기반 시계열 모델 적용 시 적절한 모델 파라미터 값의 범위를 제시하였다. 본 연구에서 적용한 방법론과 결과는 지하수관측망 수위 자료 이상값을 탐지하는데 이용되어 지하수관측망을 효과적으로 관리하는데 활용될 수 있다. 또한 강우에 의한 지하수위 변동을 모의함으로써 지하수위 변동법을 적용하여 지하수 함양률을 산정하는 등 지하수 자원을 관리하는데 활용될 수 있을 것으로 기대한다.

주요어: 기계학습, 지하수위, 시계열 예측 모델, 국가지하수관측망

Heesung Yoon, Pilsun Yoon, Eunhee Lee, Gyoo-Bum Kim and Sang-Ho Moon, 2016, Application of machine learning technique-based time series models for prediction of groundwater level fluctuation to national groundwater monitoring network data. Journal of the Geological Society of Korea. v. 52, no. 3, p. 187-199

ABSTRACT: In the present study, we developed artificial neural network (ANN) and support vector machine (SVM) based time series models and applied them to groundwater level time series data of 216 observatories in National Groundwater Monitoring Network. The purpose of the development and application of the time series model is to evaluate the model applicability to simulation of groundwater level fluctuation due to the rainfall by forecasting missing and abnormal data and filtering out the effect of groundwater pumping and stream water stage fluctuation. First, 1 day lead time direct prediction model for each station was built and utilized for establishing recursive prediction model. Results of time series modeling of groundwater level show that they can fill the missing data and filter out the effect of pumping and stream water fluctuation on groundwater level effectively. Results of error index analysis show that ANN models are slightly superior to SVM in direct prediction, however, SVM models are more stable for conducting the recursive prediction. Based on the result of model parameter selection process using the trial and error method, the present study suggests appropriate range of model parameter values for the given time series data of National Groundwater Monitoring Network. We expect that the applied method and results of this study can be useful for managing groundwater monitoring network by detecting abnormal groundwater level data and groundwater resources effectively by appling it to groundwater recharge estimation.

Key words: machine learning, groundwater level, time series model, national groundwater monitoring network

‡ Corresponding author: +82-42-880-3187, E-mail: eunheelee@kigam.re.kr

(Heesung Yoon, Pilsun Yoon, Eunhee Lee and Sang-Ho Moon, Korea Institute of Geoscience and Mineral Resources Daejeon 34132, Republic of Korea; Gyoo-Bum Kim, Dajeon University, Daejeon 34520, Republic of Korea)

1. 서 론

최근 기후변화로 인한 가뭄·홍수 등의 이상 기후 발생 및 물공급의 지역적 불균형에 의한 물부족 현상 발생 빈도가 증가하면서 수자원의 안정적인 공급 및 체계적인 관리에 대한 중요성이 대두되고 있다. 지하수 자원은 이러한 문제에 대한 대안이 될 수 있는 청정 수자원으로 이를 관리하기 위해 전국규모 관측망이 각 부처별 목적에 따라 운영되고 있으며 그 규모가 계속 증가하고 있다.

지하수 관측망을 통한 지하수자원의 효과적인 관리 및 운영을 위해 지하수위 변화를 분석하고 예측하는 것이 필요하다. 지하수위 예측 방법으로 크게 물리 모델링 기법 및 시계열 모델링 기법을 들 수 있다(Yoon *et al*., 2013). 물리 모델링 기법은 지하수 유동에 대한 물리적 개념을 바탕으로 수학적인 지배방정식을 세우고 이를 해석적 혹은 수치적으로 풀이하는 방법으로(Rai and Singh, 1995; Knotters and Bierkens, 2000) 대상 지역의 지하수위에 대한 장기적인 시공간 분포 모의가 가능한 반면 정확한 예측을 위해 매질의 물성값에 대한 정확한 측정 자료가 요구된다(Yoon *et al*., 2011). 시계열 모델링 기법은 지하수위 시계열 관측 자료를 출력변수로, 지하수위 변화에 영향을 주는 강우, 하천수위 등의 시계열 관측 자료를 입력변수로 하여 입·출력 변수 간의 반응 관계를 바탕으로 지하수위의 변화를 예측한다. 지하수위에 대한 시계열 모델링을 위해서는 입·출력 변수에 대한 연속적인 시계열 관측 자료가 필요하지만 대상 영역의 물성값에 대한 정보가 필요하지 않으며 특히 단기 예측에 대해 뛰어난 성능을 보이는 것으로 알려져 있다(Yoon *et al*., 2014).

전통적인 지하수위 시계열 예측 모델링 연구로는 Box and Jenkins (1976)에 의해 고안된 선형 모델인 자기회귀누적이동평균 모형 및 전이함수 잡음 모형을 적용하는 연구들이 국내외에서 수행되어 왔다(Tankersley *et al*., 1993; Yi and Lee, 2004; Yi *et al*., 2004). 최근에는 입·출력 간의 비선형성을 고려할 수 있는 기계학습 기법을 활용한 지하수위 시계열 예측 모델 개발 및 적용 연구가 진행되고 있다(Coulibaly *et al*., 2001; Coppola *et al*., 2005; Gill *et al*., 2007; Behzad *et al*., 2010; Yoon *et al*., 2011). 시계열 모델을 이용한 예측 방법은 크게 직접예측(direct prediction)과 반복예측(recursive prediction)의 두 가지로 나눌 수 있다(Ji *et al*., 2005; Herrera *et al*., 2007). 직접예측은 자기회귀성분, 즉 지하수위를 포함한 입력자료에 대해 과거 관측값만을 이용하여 예측을 수행하는 방법으로 지하수위 예측 분야의 적용에 있어 비교적 긴 관측주기 자료나 실시간 관측 시스템 자료에 적용시 유용하다. 그러나 각 미래 예측 시간(lead time)에 모델을 구성해야하기 때문에 장기 예측에는 적합하지 않다. 반복예측은 일반적으로 최소 단위 미래 예측 시간에 대한 직접예측 모델을 구성하고 입력자료 중 자기회귀성분에 대해 예측된 값을 반복적으로 예측하는 방법으로 직접예측에 비해 단기 예측능력은 비교적 떨어지지만 장기 예측 및 임의 입력 자료에 대한 지하수위 변화 모의가 가능하다(Yoon *et al*., 2016). 현재까지 진행된 시계열 모델 기반 지하수위 변화 예측 연구들은 직접예측 방법을 이용한 사례가 대부분이었다.

일반적으로 자연적인 지하수위 변동은 강우에 의해 발생한다. 지하수 관측망 자료를 이용한 지하수자원의 체계적인 관리를 위해서는 강우에 의한 지하수 변화를 파악할 필요가 있다. 따라서 지하수 시계열 자료에 포함된 결측, 측정 오류 등 이상자료 및 양수 등에 의한 인위적인 변화를 보정할 필요가 있다. 또한 지하수위 관측소가 하천에 인접한 경우 지하수위는 강우의 직접적인 함양 외에 하천수위 변화에 의한 영향을 동시에 받을 수 있다. 이러한 경우 지하수자원 관리를 위한 지하수 함양 평가에 있어 오차 발생의 원인이 될 수 있으므로 하천수위 변화에 의한 영향을 제거할 필요가 있다(Koo *et al*., 2013; Yoon *et al*., 2015). 이와 같이 지하수위 자료에 대한 이상자료 보정 및 양수, 하천수위 변화 영향 제거를 위해서는 직접예측 뿐 아니라 반복예측 기법까지 고려한 시계열 모델 구축이 필요하다.

위와 같은 필요성과 관련하여 본 연구에서는 대표적인 기계학습 구조인 인공신경망과 지지벡터기계 기반 지하수위 예측 시계열 모델을 작성하고 이

를 국가지하수관측망 216개소 관측소의 지하수위 자료에 대한 적용성을 평가하였다. 각 관측소에 대한 직접예측 및 반복예측 모형을 작성하여 예측 결과를 비교하고 강우에 대한 지하수위 변화 모의 가능성을 검토하였다.

2. 연구방법

2.1 인공신경망

인공신경망(Artificial Neural Network: ANN)은 수많은 시냅스 결합을 이용한 반응 전달을 통해 복잡한 문제들을 병렬처리로 효과적으로 처리하는 인간의 뇌 구조를 모방하여 고안된 연산체계이다. 일반적으로 ANN의 구조는 입력층, 은닉층, 출력층의 세 개의 층으로 구성되어 있다. 각 층은 다수의 노드들로 이루어져 있으며 층간 노드들은 특정 연결강도로 연결되어 있다(그림 1(a)). 입력층의 노드에 입력 자료가 주어졌을 때 은닉층, 출력층을 거쳐 출력값이 연산되는 과정을 피드포워드 과정(feedforward process)라 하는데 이에 대한 수학적 표현은 다음의 식과 같다.

$$y_j = F\left(\sum_{i=1}^{l} w_{ji}\, x_i + b_j\right) \tag{1}$$

여기서 x_i는 이전 층 i번째 노드값, y_j는 현재 층 j번째 노드값, b_j는 현재 층 j번째 노드의 편중값, w_{ji}는 x_i와 y_j의 연결강도, l은 이전 층의 노드 개수이다. F는 각 층에 할당되어 해당 층 노드에서 연산된 값을 다음 층 노드에 전달해주기 위한 전이함수이다. 본 연구에서는 일반적으로 가장 빈번하게 이용되는 로그-시그모이드(log-sigmoid) 함수 및 선형함수를 각각 은닉층 및 출력층 전이함수로 설정하였다. 이와 같은 피드포워드 과정을 통해 연산된 출력값은 관측값과의 오차가 발생한다. ANN 모델 구성의 궁극적인 목적은 주어진 관측 자료에 대해 출력값과 관측값의 오차가 최소화하도록 연결강도 및 편중값을 결정하는 것이다. 이 과정을 학습이라 하는데 대표적인 학습 방법은 Rumelhart *et al.* (1986)이 제안한 역전파 알고리즘(Backpropagation algorithm)으로 ANN의 출력값과 관측값 사이의 오차를 이용하여 피드포워드 과정의 역방향, 즉 출력층에서 입력층 방향으로 연결강도 및 편중값을 수정한다. 역전파 알고리즘의 수학적 표현은 다음과 같다.

$$E^k = \sum_{i}^{N} \left(\mathrm{Obs}_i^k - \mathrm{Cal}_i^k\right)^2 \tag{2}$$

$$w^{k+1} - w^k = \eta\left(w^k - w^{k-1}\right) + (1-\eta)\,\gamma\left(-\frac{\partial E^k}{\partial w^k}\right) \tag{3}$$

$$b^{k+1} - b^k = \eta\left(b^k - b^{k-1}\right) + (1-\eta)\,\gamma\left(-\frac{\partial E^k}{\partial b^k}\right) \tag{4}$$

여기서 E^k는 k번째 반복 연산에서 발생하는 관측값(Obs_i^k)와 예측값(Cal_i^k) 사이의 오차, γ는 학습속도를 나타낸다. η는 ANN의 지역해 탐색 문제를 보완

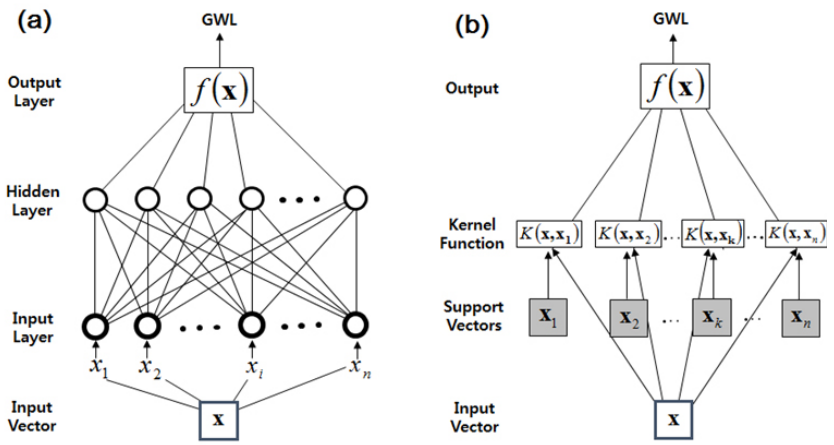

Fig. 1. Structures of (a) ANN and (b) SVM (modified from Yoon *et al.*, 2011).

하기 위해 제안된 모멘텀을 나타낸다(Rumelhart *et al*., 1986). ANN 모델 구축을 위해 결정해야할 모델 파라미터는 은닉층 노드 개수(NUMHID), 학습속도(γ), 모멘텀(η)의 세 가지이다. 본 연구에서는 ANN 모델 파라미터 설정을 위해 위의 각 모델 파라미터에 대해 [5, 10], [0.0001, 0.01], [0.0, 0.9]의 범위에서 총 125개 조합에 대한 모델을 생성하고 이 중 모델 구성 단계에서의 오차를 최소화하는 파라미터를 선정하였다. 이를 위하여 모델 구성에 필요한 자료를 학습(Training) 및 보정(Calibration)의 두 단계로 나누었다. 역전파 알고리즘에 의한 연결강도 및 편중값 업데이트는 학습 단계 자료를 통해 수행하고 모델 파라미터는 보정 단계 자료 적용 결과의 오차가 최소화되는 그룹으로 선정되도록 하였다. 이러한 과정은 모델이 훈련 자료에 과도학습(over training)되는 현상을 방지하고 입출력 자료가 발생하는 해당 시스템의 반응 관계를 적절하게 학습할 수 있도록 도움을 줄 수 있다(Yoon *et al*., 2011).

2.2 지지벡터기계

지지벡터기계(Support Vector Machine: SVM)는 1995년 Vapnik에 의해 제안된 방법으로 자료를 분류(classification)하는데 있어 분류 경계와 자료와의 여백을 최대화한다는 개념을 바탕으로 한다. ANN은 입력 자료의 각 성분들이 입력층의 입력노드에 할당되지만 SVM의 경우 모델을 구성하기에 적합한 것으로 판정된 자료가 벡터의 형태로 모델 입력층을 구성하는데 이러한 모델 구성 벡터를 지지벡터(support vector)라 한다. SVM에서는 새로운 입력 벡터에 대해 지지벡터들과의 연산을 통해 출력값을 예측하도록 구성되어 있다(그림 1 (b)). 따라서 SVM 모델 구성의 최종 목적은 주어진 자료에 대해 가장 적합한 지지벡터를 선정하고 각 지지벡터들의 연결강도 및 편중값을 결정하는 것이다. 일반적으로 지지벡터들과 임의의 입력벡터와의 관계 연산을 위해 다음과 같은 가우시안(Gaussian) 형태의 커널함수가 사용된다.

$$K(\mathbf{x}_s, \mathbf{x}) = \exp\left(-\frac{\| \mathbf{x}_s - \mathbf{x} \|^2}{2\sigma^2}\right) \tag{5}$$

여기서 $\mathbf{x}_s$는 지지벡터, $\mathbf{x}$는 새로운 입력 벡터, σ는 가우시안 함수 파라미터이다.

SVM의 연산함수(f) 및 SVM 구조 선정을 위한 목적함수는 각각 식(6), 식(7)와 같이 주어진다(Vapnik, 1995).

$$f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b \tag{6}$$

$$\text{maximize} \quad \frac{1}{2}\| \mathbf{w} \|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{7}$$

$$\text{subject to} \quad \begin{cases} y_i - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i \\ \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i) + b - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

여기서 $\mathbf{w}$는 연결강도 벡터, b는 편중값, ϕ는 비선형 전이함수를 나타낸다. C는 경험오차 어느 정도 고려할 것인가를 나타내는 길항 파라미터, ξ는 모델 연산 오차, ϵ은 오차 허용율을 의미한다. 본 연구에서는 SVM의 학습방법으로 순차적 최소규모 최적화 알고리즘(sequential minimal optimization algorithm: SMO)을 이용하였다. SMO는 위의 목적함수를 라그랑주 승수(Lagrangian multiplier)를 이용하여 다음의 식과 같이 변환한다.

$$\text{maximize} \quad \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\mathbf{x}_i, \mathbf{x}_j) \\ -\epsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*) \end{cases} \tag{8}$$

$$\text{subject to} \begin{cases} \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0 \\ 0 \leq \alpha_i, \alpha_i^* \leq C \end{cases}$$

여기서 α는 라그랑주 승수를 나타낸다. SMO에서는 $\alpha_i - \alpha_i^*$를 β_i로 치환하고 임의로 선정된 두 개의 β에 대해 식 (8)를 2차 함수 최대값 산정 문제로 변환하여 그 해석해를 순차적으로 풀이한다. 이러한 과정을 통해 0값이 아닌 연결강도, 즉 β를 갖는 벡터를 지지벡터로 선정하게 된다(Platt, 1999; Scholkopf and Smola, 2002). SVM 모델 파라미터는 오차 길항 파라미터(C), 오차허용률(ϵ), 가우시안함수 파라미터(σ)이고 ANN 모델 구성 방법에서와 같이 각 파라미터에 대해 [2.0, 10.0], [0.06, 0.12], [0.5, 2.5]의 범위에서 총 125개 조합에 대한 모델을 생성하고 이 중 모델 구성 단계에서의 오차를 최소화하는 파라미

터를 선정하였다.

2.3 시계열 예측 모델

앞서 서론에서 언급한 바와 같이 시계열 모델을 이용한 예측 방법은 크게 직접예측과 반복예측으로 나눌 수 있다(Ji *et al.*, 2005; Herrera *et al.*, 2007). 지하수위 변동에 대한 주요 외부 입력 변수로 강우를 고려한 직접예측 방법은 다음과 같이 표현할 수 있다.

$$\hat{g}_{t+h} = F_h(\mathbf{x}), \quad \mathbf{x} = \begin{Bmatrix} p_{t-a+1}, p_{t-a+2}, \cdots, \\ p_t, \ g_{t-b+1}, g_{t-b+2}, \cdots, g_t \end{Bmatrix} \tag{9}$$

여기서 $\hat{g}_t$, g_t, p_t는 각각 시간 t에서의 지하수위 예측값, 지하수위 관측값 및 강우 관측값, F는 ANN 혹은 SVM을 이용한 시계열 모델, $\mathbf{x}$는 모델 입력 성분, h는 미래 예측 시간(lead time), a 및 b는 각각 강우 및 지하수위에 대한 과거 관측 자료 사용 정도, 즉 지연시간(lag time)을 나타낸다. 본 연구에서는 입출력 시계열의 교차상관분석 결과를 참고하여 모델 구성의 편의를 위해 각 지연시간을 3일로 고정하였고 1일 직접예측 시계열 모델($h=1$)을 고려하였다. 1일 직접예측 시계열 모델을 이용한 반복예측 방법은 다음과 같이 표현된다.

$$\hat{g}_{t+1} = F_1(\mathbf{x}), \quad \mathbf{x} = \begin{Bmatrix} p_{t-a+1}, p_{t-a+2}, \cdots, \\ p_t, \ \hat{g}_{t-b+1}, \hat{g}_{t-b+2}, \cdots, \hat{g}_t \end{Bmatrix} \tag{10}$$

지하수위 변동이 강우 외 인접한 하천수위의 영향을 받는 경우 입력 성분에 하천수위 과거 관측 자료를 추가하여 모델을 구성할 수 있다. 본 연구의 목적 중 하나는 국가지하수관측망 지하수위 자료에 대해 강우에 대한 지하수위 변동을 모의하는 것이다. Yoon *et al.* (2015)은 시계열 모델을 활용하여 반복예측시 하천수위 성분을 고정값으로 설정함으로써 지하수위에 대한 하천수위 영향을 제거하는 방법을 제안한 바 있다. 따라서 지하수위가 하천에 인접해 있고 하천수위 시계열 자료 확보가 가능한 경우 하천수위 제거 방법을 활용하여 지하수위를 모의하였다(그림 2).

3. 연구자료

본 연구에서는 국가지하수관측망 중 216개 관측소에서 측정된 2000년부터 2011년 사이의 지하수위 관측 자료를 이용하였다. 또한 ANN 및 SVM 기반 시계열 예측 모델의 입력 자료로 지하수관측소 인근 기상청 강우 자료를 이용하였고 하천에 인접한 지하

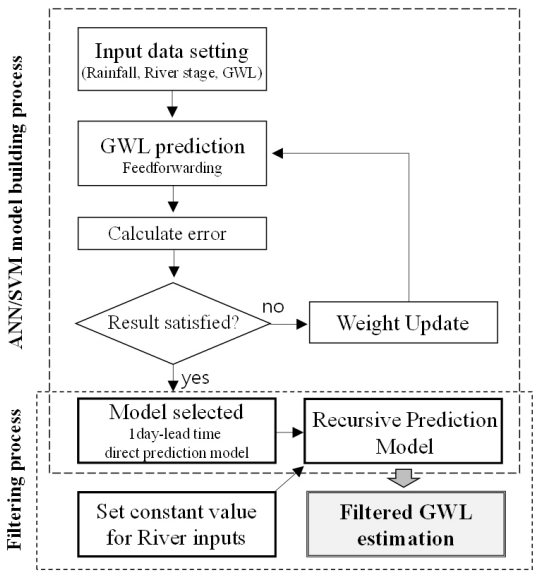

Fig. 2. Flow chart of the process for filtering out the effect of stream water fluctuation using time series models (modified from Yoon *et al.*, 2015).

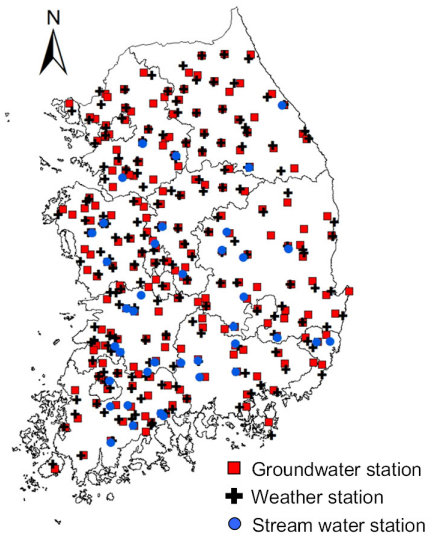

Fig. 3. Location of observatories of groundwater level, weather and stream water level (source of groundwater level: www.gims.go.kr, rainfall: www.kma.go.kr, stream water level: www.wamis.go.kr).

수관측소 중 하천수위 자료가 지하수위 관측자료 구간과 동일하게 존재하는 경우에는 하천수위 시계열 자료를 입력 자료에 추가하였다. 본 연구에서 강우 외에 하천수위를 추가 고려한 관측소는 34개소이다.

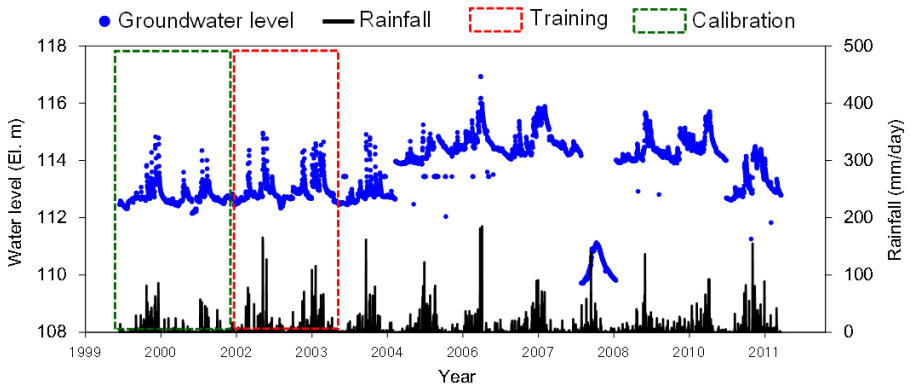

Fig. 4. An example of observed time series data of groundwater level and rainfall and allocation of data for model training and calibration (GeosanGeosan observatory).

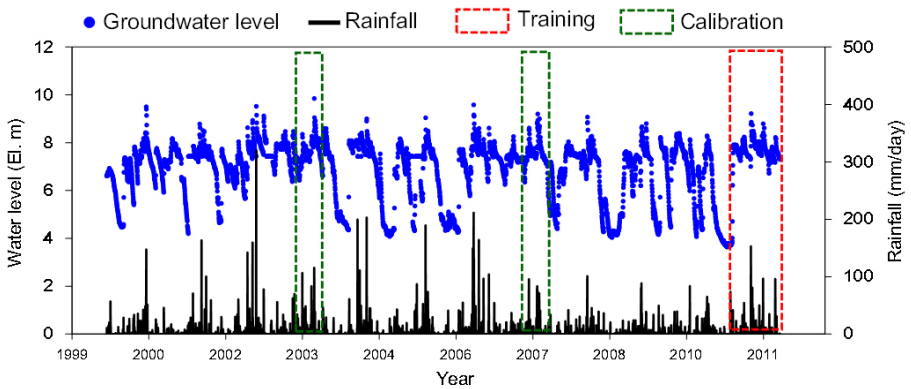

Fig. 5. An example of observed time series data of groundwater level and rainfall and allocation of data for model training and calibration (DonghaeGwiun observatory).

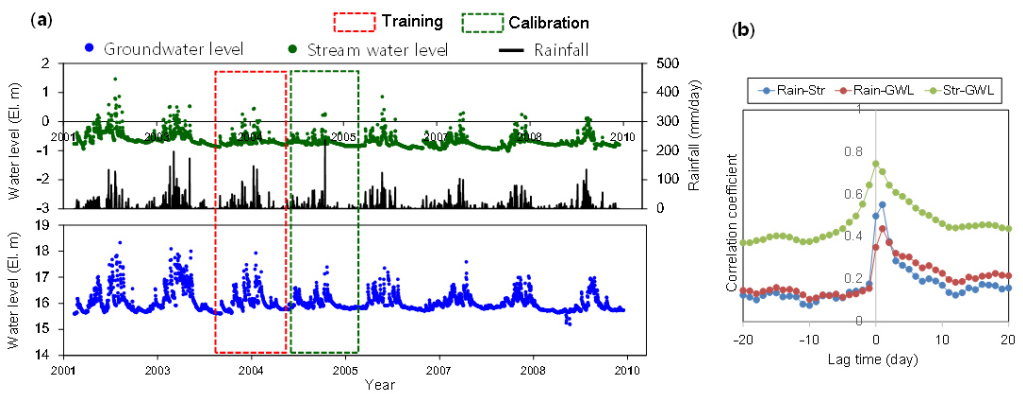

Fig. 6. An example of (a) observed time series data of groundwater level, stream water level and rainfall and allocation of data for model training and calibration (UlsanBeomseo observatory) (b) with the result of cross correlation analysis.

그림 3은 적용한 지하수관측소, 강우 및 하천수위 관측소의 위치를 보여준다.

그림 4는 충북괴산괴산 관측소의 강우 및 지하수위 변동 시계열 자료를 보여준다. 전체 구간에 대해 강우에 대한 지하수위 변화가 뚜렷하게 관찰된다. 그러나 관측 중반 이후 다수의 결측 및 이상 자료가 지속적으로 관찰되는 것을 볼 수 있다. 강원동해귀운 관측소의 자료의 경우 매년 반복적이고 주기적인 양수의 영향이 나타나고 있고 다수의 결측자료가 존재한다(그림 5). 경남울산범서 관측소의 경우 대곡천 및 태화강과 약 40 m, 300 m 거리에 인접해 있으며 약 800 m 거리에 태화강 조동관측소에서 하천수위를 측정하고 있다. 경남울산범서 관측소의 강우, 하천, 지하수위 시계열 자료 및 교차상관분석 결과를 보면 강우-지하수위의 반응보다 하천수위-지하수위의 반응이 보다 빠르고 크게 나타남을 알 수 있다(그림 6). 위의 예시들의 경우 강우에 의한 지하수 변동 양상을 파악하기 어려워 추세를 분석하거나 지하수위 변동법에 기반한 지하수 함양률 산정 시 오류를 범할 수 있다. 이를 고려할 때 지하수위 시계열 예측 모델을 이용하여 강우에 대한 지하수위 변동 모의가 필요한 경우는 다음의 세 가지로 정리할 수 있다.

1) 지하수위 자료에 결측 및 센서 오류 등의 이상 자료가 빈번한 경우
2) 양수 등 반복적이고 인위적인 요인이 포함된 경우
3) 하천에 인접하여 강우 외 하천수위 영향을 크게 받는 경우

본 연구에서는 시계열 관측 자료 중 강우에 의한 지하수위 변동 양상이 비교적 뚜렷하게 나타나는 구간 자료을 이용하여 지하수위 시계열 1일 직접 예측 모델 구성하고 이를 활용한 반복예측을 통해 강우에

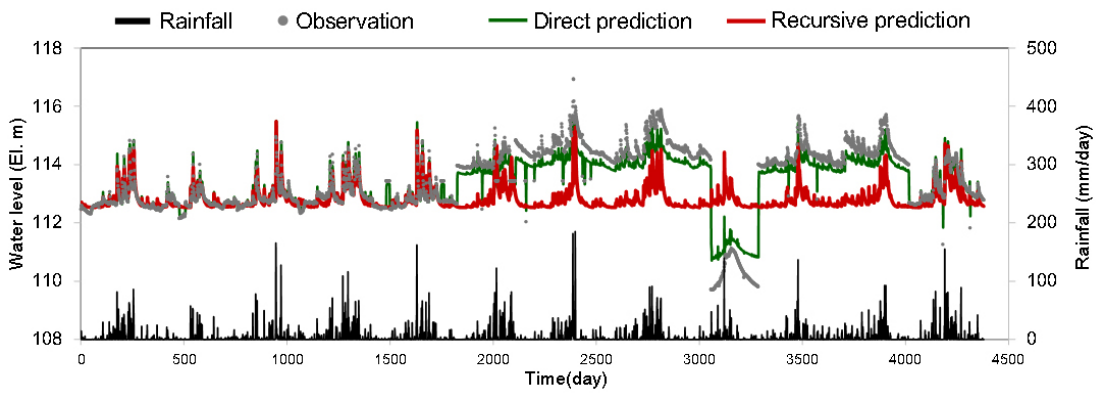

Fig. 7. The result of groundwater level prediction using time series models for GeosanGeosan observatory data.

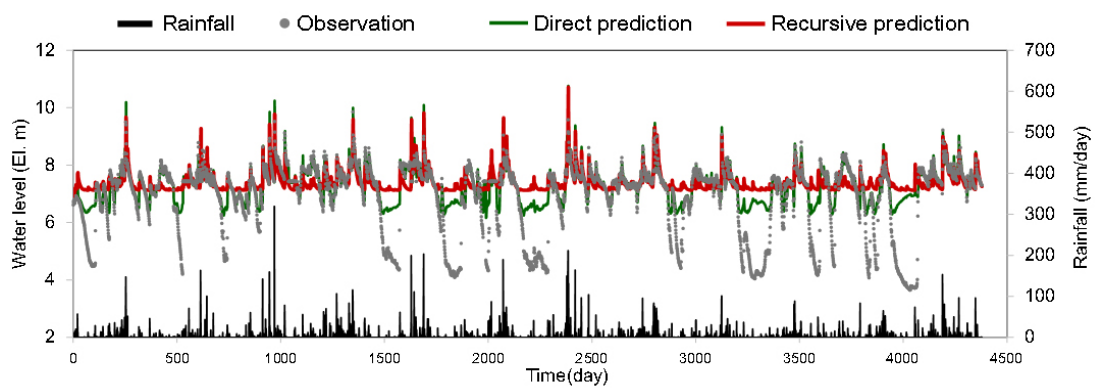

Fig. 8. The result of groundwater level prediction using time series models for DonghaeGwiun observatory data.

의한 지하수위 변동을 모의하고자 하였다.

4. 연구결과 및 토의

4.1 지하수위 시계열 예측 모델 적용 예시

그림 7, 8, 9는 각각 충북괴산괴산, 강원동해귀운, 경남울산범서 관측소에 대한 지하수위 시계열 예측 모델의 직접 및 반복예측 결과를 보여준다. 충북괴산괴산 관측소 자료에 대한 ANN 모델 적용 결과 직접예측은 센서 위치 변동의 원인으로 판단되는 이상 자료를 따라가며 예측하는 현상을 보여준다. 이는 직접예측 방법이 과거 관측 지하수위를 반복적으로 입력 성분으로 이용하고 있기 때문에 나타는 현상이다. 그러나 반복예측의 경우 지하수위 이상 변동을 따라가지 않고 강우에 대한 지하수위 변화를 효과적으로 모의하고 있는 것을 볼 수 있다(그림 7). 강원동해귀운 관측소 자료에 대한 SVM 모델 적용 결과 직접예측은 양수의 원인으로 보이는 지하수위 하강 패턴을 어느 정도 따라가며 예측하는 양상을 보여주지만 반복예측의 경우 강우에 의한 지하수위 변동 양상을 효과적으로 모의하고 있다(그림 8). 경남울산범서 관측소에 대해서는 ANN 모델의 반복예측 기법을 이용하여 하천수위 영향 제거 방법을 적용하였다. 직접예측 결과 지하수위 변동과 거의 일치하는 것으로 모의되어 적합한 직접예측 모델이 구성되었음을 알 수 있다. 이를 이용한 반복예측 결과 강우-지하수위의 상관성은 유지하면서 변동폭이 크게 줄어들었음을 볼 수 있다(그림 9).

4.2 국가지하수관측망 적용 결과

본 연구에서는 국가지하수관측망 216개소에 대해 ANN 및 SVM 기반 지하수위 시계열 1일 직접예측 및 반복예측 모델을 구축하고 그 적용성을 비교 평가하였다. 예측 모델의 오차분석을 위해 평균절대백분율오차(Mean Absolute Percentage Error: MAPE), 평균절대상대오차(Mean Absolute Relative Error: MARE), 평균제곱근편차(Root Mean Square Error: RMSE), 상관계수(Correlation Coefficient: CORR)을 오차지표로 설정하였다. 각 오차지표의 수학적 표현은 다음의 식과 같다.

Table 1. Error statistics of direct prediction results.

Model		ANN			SVM	
Statistics	MED	AVR	STD	MED	AVR	STD
MAPE (%)	0.090	0.400	1.018	0.319	1.167	1.778
MARE (%)	1.601	1.843	1.111	1.408	1.590	1.730
RMSE (m)	0.098	0.138	0.122	0.109	0.174	0.211
CORR	0.971	0.953	0.056	0.964	0.930	0.116

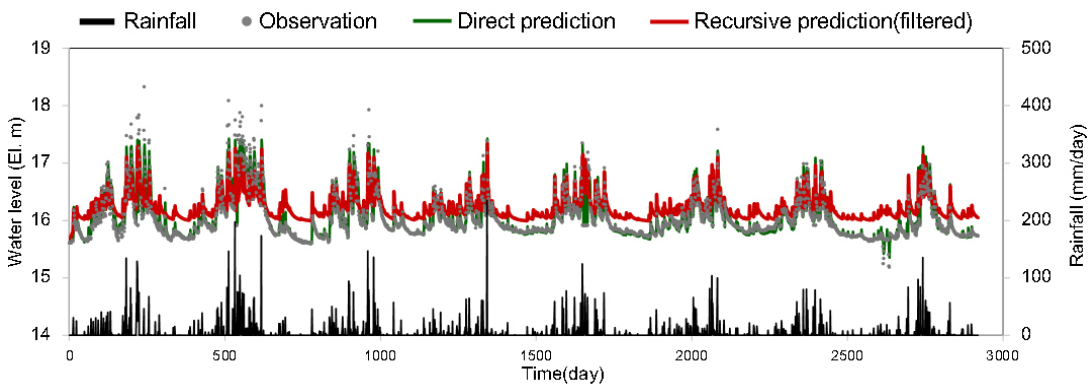

Fig. 9. The result of groundwater level prediction using time series models for UlsanBeomseo observatory data.

$$\mathrm{MAPE}(\%)=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{|\mathrm{Obs}_i-\mathrm{Cal}_i|}{|\mathrm{Obs}_i|}\right)\times 100 \quad (11)$$

$$\mathrm{MARE}(\%)=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{|\mathrm{Obs}_i-\mathrm{Cal}_i|}{|\mathrm{Obs}^{max}-\mathrm{Obs}^{min}|}\right)\times 100 \quad (12)$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{Obs}_i-\mathrm{Cal}_i)^2} \quad (13)$$

$$\mathrm{CORR}=\frac{\frac{1}{N}\sum_{i=1}^{n}(\mathrm{Obs}_i-\overline{\mathrm{Obs}_i})(\mathrm{Cal}_i-\overline{\mathrm{Cal}_i})}{\sqrt{\frac{1}{N}(\mathrm{Obs}_i-\overline{\mathrm{Obs}_i})^2}\sqrt{\frac{1}{N}(\mathrm{Cal}_i-\overline{\mathrm{Cal}_i})^2}} \quad (14)$$

여기서 N은 자료 수, Obs^{max}, Obs^{min}는 각각 관측값의 최대값과 최소값, $\overline{\mathrm{Obs}}$, $\overline{\mathrm{Cal}}$는 각각 관측값과 예측값의 평균값을 의미한다.

국가지하수관측망 216개소 지하수위 자료에 대한 ANN 및 SVM 모델의 직접예측 적용 결과 평균값을 기준으로 MAPE 1.2%이하, MARE 1.9%이하, RMSE 0.18 m 이하, CORR 0.93 이상의 높은 예측결과를 보여주었다(표 1). 두 모델의 비교에서는 MARE를 제외하고 ANN 모델이 다소 높은 예측 결과를 보여주었으며 각 모델 간 오차지표의 분포도 유사한 것으로 나타났다(그림 10). 반복예측의 경우 전체 모델에 대해 오차지표가 평균값 기준 MAPE

Table 2. Error statistics of recursive prediction results.

Model	ANN			SVM		
Statistics	MED	AVR	STD	MED	AVR	STD
MAPE (%)	0.462	2.981	8.074	0.557	2.522	6.384
MARE (%)	7.890	14.877	35.613	9.133	10.679	6.903
RMSE (m)	0.317	0.738	1.814	0.352	0.533	0.470
CORR	0.700	0.681	0.180	0.660	0.620	0.199

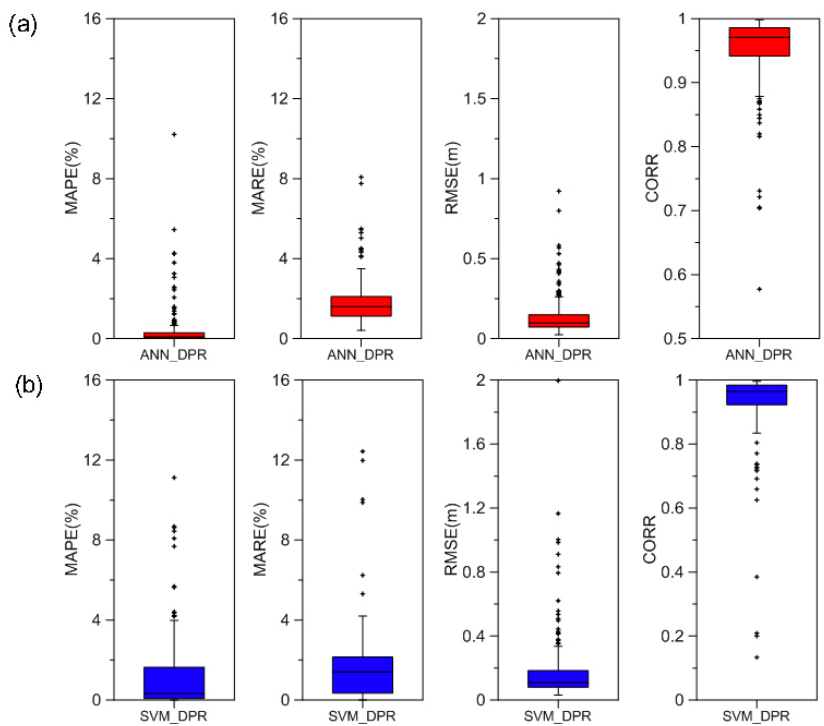

Fig. 10. The distribution of error criteria for direct prediction models: (a) ANN, (b) SVM.

3.0%이하, MARE 14.9% 이하, RMSE 0.74 m 이하, CORR 0.62 이상의 값을 보여주었다(표 2). ANN 및 SVM 모델의 비교에서 역시 MARE를 제외한 오차지표에서 ANN 모델의 오차지표 평균값이 SVM에 비해 다소 우위에 있는 것으로 나타난다. 그러나 반복예측에 대한 오차지표의 분포를 살펴보면 MARE와 RMSE에 대해 ANN의 분포가 SVM에 비해 매우 큰 것을 알 수 있다(그림 11).

반복예측의 경우 예측된 지하수위 값을 입력 성분으로 반복적으로 이용하기 때문에 해당 시스템의 입출력에 대한 반응관계를 적절하게 반영하지 못하는 직접예측 모델이 결정될 경우 예측 기간이 길어짐에 따라 오차 축적에 의한 모의 오류를 발생시킬 가능성이 있다(Herrera *et al.*, 2007; Yoon *et al.*, 2016). 본 연구의 목적은 반복예측 기법을 통해 이상자료 및 강우 외 요인을 제거하는 것이므로 반복예측 오차가 크다고 해서 반드시 모델의 효용성이 떨어진다고 할 수는 없다. 그러나 직접예측 모델의 오차에 대한 반복예측 모델의 오차의 분포가 크다는 것은 반복예측에 적합하지 않은 직접예측 모델이 선택되는 확률이 높다는 것, 즉 반복예측 모델 안정성이 낮다는 것을 의미한다(Yoon *et al.*, 2011). Yoon *et al.* (2016)은 이러한 모델 안정성을 판단하는 지표로 직접예측 모델 오차에 대한 반복예측 모델 오차의 비율을 제안한 바 있다. 본 연구에서는 이를 참고하여 다음과 같이 RPR-DPR Ratio를 정의하고 각 모델 및 관측소에 대한 값을 산정하였다.

$$\text{RPR}-\text{DPR Ratio} = \frac{\text{RMSE of Recursive Prediction}}{\text{RMSE of Direct Prediction}} \quad (15)$$

RPR-DPR Ratio가 낮고 분포가 좁을수록 안정적

Table 3. Statistics of RPR-DPR Ratio.

Model	ANN	SVM
MIN	1.115	0.871
MAX	122.007	20.146
MED	3.045	2.954
AVR	6.373	3.956
STD	14.422	3.238

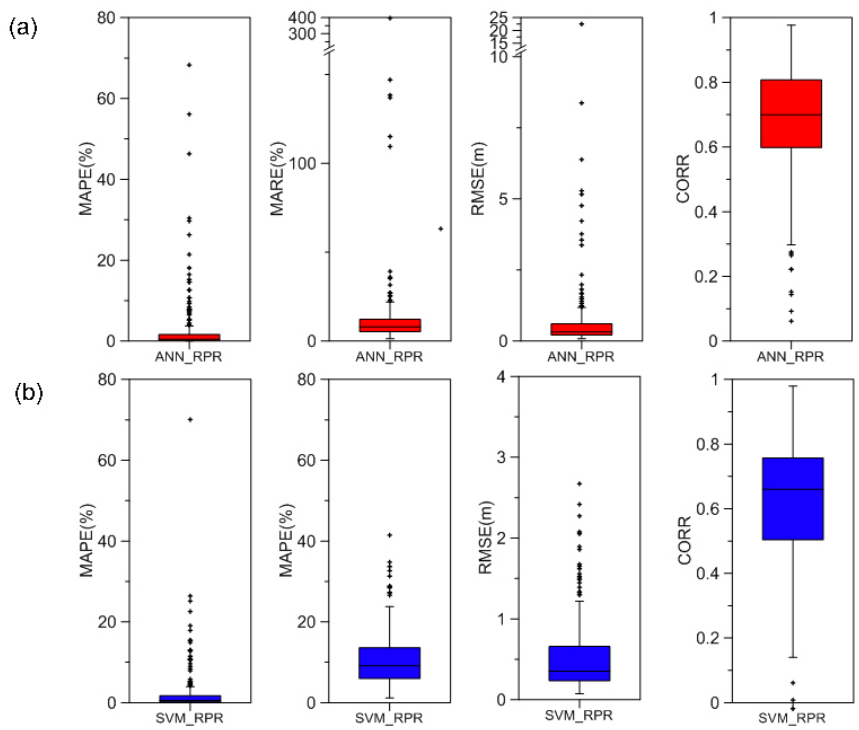

Fig. 11. The distribution of error criteria for recursive prediction models: (a) ANN, (b) SVM.

으로 반복예측 모델이 구성되었을 가능성이 높음을 의미한다. RPR-DPR Ratio 산정 결과 ANN 모델에 대해 약 1.12 ~ 122.00, SVM 모델에 대해 약 0.87 ~ 20.15의 범위를 보이는 것으로 계산되었다(표 3; 그림 12). 이는 본 연구 자료에 대해 SVM이 ANN 에 비해 반복예측 모델을 구성하는데 보다 안정적일 수 있음을 시사한다.

기계학습법 기반 시계열 예측 모델링에서는 앞서 언급한 바와 같이 매질의 물성값 등에 대한 정보가 필요하지는 않지만 적용 자료에 가장 적합한 모델 파라미터를 결정해야한다. 모델 파라미터 결정을 위해 일반적으로 시행착오법이 이용되어 왔다. 효율적인 모델 구축을 위해 시행착오법으로 모델 파라미터를 탐색하는데 있어 자료에 적절한 탐색 구간을 설정하는 것이 중요하다. 본 연구에서는 국내 지하수위 자료 적용에 있어 주로 선택되는 모델 파라미터 구간을 평가하기 위해 선택된 모델 파라미터 분포를 검토하였다. 각 모델 별 선택된 모델 파라미터 값의 1사분위에서 4사분위 사이 값을 살펴보면 ANN 모델에 대해 은닉층 노드 개수(NUMHID) [4, 10], 학습속도(γ) [0.0005, 0.001], 모멘텀(η) [0.0, 0.5]의 범위에서 주로 선택되었음을 알 수 있다. SVM 모델 파라미터에 대해서는 오차 길항 파라미터(C) [4.0, 8.0], 오차허용률(ϵ) [0.06, 0.10], 가우시안 함수 파라미터(σ) [1.5, 2.5]의 분포에서 주로 선택되었음을 알 수 있다(그림 13). 이후 본 연구 결과를 활용하여 보다 다양한 지하수위 자료 및 넓은 파라미터 탐색구간에 대한 적용이 지속적으로 이루어진다면 국내 지하수위 자료 적용을 위한 적정 모델 파라미터 설정 구간을 제시하는데 도움이 될 것이다. 또한 지하수위 변동 패턴에 따라 그룹을 나누고 그룹 별 파라미터 분포 특성에 대한 고찰이 필요할 것으로 판단된다.

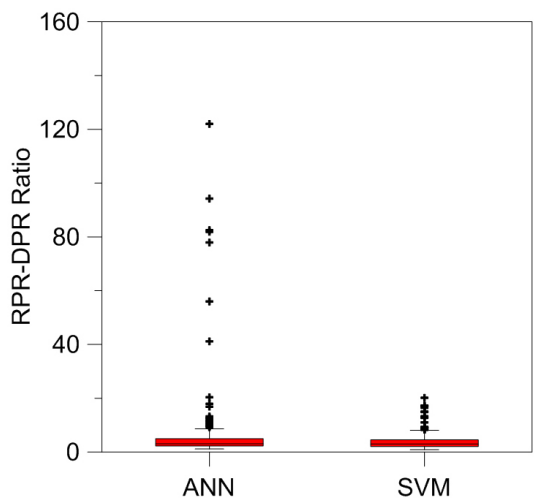

Fig. 12. The distribution of RPR-DPR Ratio.

5. 요약 및 결론

본 연구에서는 대표적인 기계학습 기법인 ANN과 SVM을 이용하여 국가지하수관측망 216개소 관측소에 대한 지하수위 시계열 예측 모델을 구축하고 직접예측 및 반복예측을 통해 강우에 대한 지하수위 모의에의 적용성을 평가하였다. 시계열 예측 모델 적용 결과 결측 및 이상 자료, 양수 등 인위적인 요인, 하천수위 영향을 받는 경우 강우에 의한 지하수위 모의를 효과적으로 수행한다고 판단된다. 각 모델의 오차지표 분석 결과 직접예측 및 반복예측에 대해 ANN 모델이 SVM 모델보다 다소 높은 예측 능력을 보여주었다. 본 연구에서 적용한 반복예측 방법의 경우 지하수위 자료에 포함된 이상 자료 및 강우 외 요인을 제거하여 보정하는 것이 목적이므로

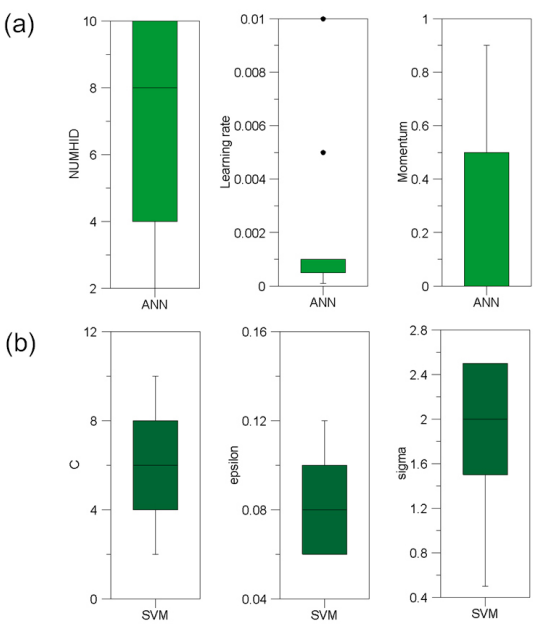

Fig. 13. The distribution of selected model parameters: (a) ANN, (b) SVM.

반복예측 오차가 크다고 해서 반드시 모델의 효용성이 떨어진다고 할 수는 없다. 반복예측 시에는 초기 직접예측에 의해 계산된 지하수위 자료가 반복적으로 입력자료로 사용되기 때문에 이에 의한 오차누적의 문제점이 제기되어 왔다. 따라서 시계열 예측 모델을 이용한 강우에 대한 지하수위 모의를 위해서는 이러한 오차누적에 대해 보다 안정적인 모델 구축이 보다 중요할 수 있다. ANN 및 SVM의 반복예측 모델 안정성 평가를 위해 각 관측소 적용 결과에 대한 RPR-DPR Ratio를 분석한 결과 ANN 모델의 평균값이 SVM 보다 약 62% 크고 편차가 심한 것으로 나타나 본 연구 자료의 적용에 있어 SVM 모델이 반복예측에 대해 보다 안정적인 것으로 판단된다. 또한 본 연구에서는 국내 지하수위 자료 적용에 적합한 ANN 및 SVM 모델 파라미터를 평가하기 위해 각 관측소 자료에 대해 동일한 구간에서의 파라미터 탐색을 수행하고 적정 파라미터 구간을 제시하였다. 본 연구에서는 자료 기간 및 품질을 고려하여 국가지하수관측망 216개소에 대해서만 시계열 예측 모델을 적용하였다. 이후 보다 최신 자료까지를 고려한 연구 및 국가지하수관측망 외 농촌지하수관측망 및 보조관측망 등에 대한 적용 연구를 수행할 필요가 있다고 판단된다. 본 연구에서 수행한 방법 및 연구 결과는 지하수 관측망의 운영의 효율성을 증대하고 지하수 함양률 산정 등의 연구에 유용하게 활용될 수 있을 것으로 기대한다.

사 사

본 연구 논문은 국토교통부가 출연하고 국토교통과학기술진흥원에서 위탁 시행한 물관리연구사업(11기술혁신C05)에 의한 '수변지하수활용고도화' 연구단의 연구비 지원에 의해 수행되었습니다.

REFERENCES

Behzad, M., Asghari, K. and Coppola Jr., E.A., 2010, Comparative study of SVMs and ANNs in aquifer water level prediction. Journal of Computing in Civil Engineering, 24(5), 408-413.

Box, G.E.P. and Jenkins, G.M., 1976, Time Series Analysis-Forecasting and Control. Holden-Day, San Francisco, California, USA, 575 p.

Coppola, E., Rana, A.J., Poulton, M.M., Szidarovszky, F. and Uhl, V.V., 2005, A neural network model for predicting aquifer water level elevations, Ground Water, 43(2), 231-241.

Coulibaly, P., Anctil, F., Aravena, R. and Bobee, B., 2001, Artificial neural network modeling of water table depth fluctuations. Water Resources Research, 37(4), 885-896.

Gill, M.K., Asefa, T., Kaheil, Y. and McKee, M., 2007, Effect of missing data on performance of learning algorithms for hydrologic predictions: implications to an imputation technique. Water Resources Research 43, W07416. doi:10.1029/2006WR005298.

Herrera, L.J., Pomares, H., Rojas, I., Guillen, A., Prieto, A. and Valenzuela, O., 2007, Recursive prediction for long term time series forecasting using advanced models. Neurocomputing, 70(16-18), 2870-2880.

Ji, Y., Hao, J., Reyhani, N. and Lendasse, A., 2005, Direct and recursive prediction of time series using mutual information selection. Lecture Notes in Computer Science, 3512, 1010-1017.

Knotters, M. and Bierkens, M.F.P., 2000, Physical basis of time series models for water table depths. Water Resources Research, 36(1), 181-188.

Koo, M.H., Kim, T.K., Kim, S.S., Chung, S.R., Kang, I.O., Lee, C.J. and Kim, Y., 2013, Estimating groundwater recharge using the water-table fluctuation method: Effect of stream-aquifer interactions. Journal of Soil and Groundwater Environment, 18(5), 65-76 (in Korean with English abstract).

Platt, J.C., 1999, Fast training of support vector machines using sequential minimal optimization. In: Scholkopf, B., Burges, C.J.C., Smolar, A.J. (Eds.), Advances in Kernel Methods-Support Vector Learning, MIT Press, Cambridge, Massachusetts, USA, 376 p.

Rai, S.N. and Singh, R.N., 1995, Two-dimensional modelling of water table fluctuation in response to localized transient recharge. Journal of Hydrology, 167, 167-174.

Rumelhart, D.E., McClelland, J.L. and The PDP Research Group, 1986, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, Massachusetts, USA, 516 p.

Scholkopf, B. and Smola, A.J., 2002, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, Massachusetts, USA, 656 p.

Tankersley, C.D., Graham, W.D. and Hatfield, K., 1993, Comparison of univariate and transfer function models of groundwater fluctuations. Water Resources Research, 29, 3517-3533.

Vapnik, V.N., 1995, The Nature of Statistical Learning Theory. Springer-Verlag, New York, USA, 314 p.

Yi, M.J., Kim, G.B., Sohn, Y.C., Lee, J.Y. and Lee, K.K., 2004, Time series analyses for the groundwater level in the National Groundwater Monitoring Network. Journal of Geological Society of Korea, 50(2), 293-307 (in Korean with English abstract).

Yi, M.J. and Lee, K.K., 2004, Transfer function-noise modeling of irregularly observed groundwater heads using precipitation data. Journal of Hydrology, 288, 272-287.

Yoon, H., Hyun, Y., Ha, K., Lee, K.K. and Kim, G.B., 2016, A method to improve the stability and accuracy of ANN- and SVM- based time series models for long-term groundwater level predictions. Computers and Geosciences, 90, 144-155.

Yoon, H., Jun, S.C., Hyun, Y., Bae, G.O. and Lee, K.K., 2011, A comparative study of artificial neural network and support vector machines for predicting groundwater levels in a coastal aquifer. Journal of Hydrology, 396, 128-138.

Yoon, H., Kim, Y., Ha, K. and Kim, G.B., 2013, Application of groundwater-level prediction model using data-based learning algorithms to National Groundwater Monitoring Network data. The Journal of Engineering Geology, 23(2), 137-147 (in Korean with English abstract).

Yoon, H., Park, E., Kim, G.B., Ha, K., Yoon, P. and Lee, S.H., 2015, A method to filter out the effect of river stage fluctuations using time series model for forecasting groundwater level and its application to groundwater recharge estimation. Journal of Soil and Groundwater Environment, 20(3), 74-82 (in Korean with English abstract).

Yoon, P., Yoon, H., Kim, Y. and Kim, G.B., 2014, A comparative study on forecasting groundwater level fluctuations of National Groundwater Monitoring Networks using TFNM, ANN, and ANFIS. Journal of Soil and Groundwater Environment, 19(3), 123-133 (in Korean with English abstract).

Received : May 27, 2016
Revised : June 22, 2016
Accepted : June 22, 2016