J Korean Soc Disaster Secur 15(2): 45-56, 2022
https://doi.org/10.21729/ksds.2022.15.2.45

Print ISSN: 2466-1147 Online ISSN: 2508-285X
www.dssms.org

KSDS

Original Article

LSTM기반의 자료 변동성을 고려한 하천수 회귀수량 예측 알고리즘 개발연구

이승연[1] · 유형주[2] · 이승오[3]*

[1]홍익대학교 과학기술연구소 연구원, [2]홍익대학교 토목공학과 박사과정, [3]홍익대학교 건설환경공학과 교수

Development of Return flow rate Prediction Algorithm with Data Variation based on LSTM

Seung Yeon Lee[1], Yoo, Hyung Ju[2], and Seung Oh Lee[3]*

[1]Researcher, Hongik University Research Institute of Science and Technology

[2]Ph.D Student, Dept. of Civil Engineering, Hongik University

[3]Professor, Dept. of Civil Engineering, Hongik University

요 약

가뭄 및 갈수시에 용수부족 현상이 발생하나 회귀수량을 고려한 대응이나 대책 마련이 진행되지 않고 있다. 이에 본 연구에서 자료기반의 기계학습 모형(LSTM)을 통해 회귀수량 중 하수종말처리장의 방류량을 예측하였다. 입력자료로 방류량, 유입량, 강수량, 수위를 사용하였고 예측 결과의 정확도를 개선하기 위하여 추가적으로 입력변수의 변동성 분포를 고려하였다. 방류량 자료의 변동성을 확인하기 위해서 관측값과 분포 사이의 잔차를 복합삼각함수 형태로 가정하여 이론적인 확률분포와 함께 방류량 최적의 분포 형태로 나타내었다. 변동성 분포를 고려한 입력자료를 이용한 결과와 그렇지 않는 결과를 비교한 결과, 오차정도가 감소함을 보였으며 이는 변동성 분포가 계절성을 상대적으로 잘 재현하였기 때문이라 판단된다. 따라서 본 연구에서 구축한 하수종말장처리장의 방류량 예측 모형을 활용할 경우 보다 정확한 회귀수량 예측이 가능하여 효율적인 하천수 관리 체계를 수립하는데 기초자료로 활용될 수 있을 것으로 기대된다.

핵심용어: 회귀수량, 변동성, 확률분포, LSTM

ABSTRACT

The countermeasure for the shortage of water during dry season and drought period has not been considered with return flowrate in detail. In this study, the outflow of STP was predicted through a data-based machine learning model, LSTM. As the first step, outflow, inflow, precipitation and water elevation were utilized as input data, and the distribution of variance was additionally considered to improve the accuracy of the prediction. When considering the variability of the outflow data, the residual between the observed value and the distribution was assumed to be in the form of a complex trigonometric function and presented in the form of the optimal distribution of the outflow along with the theoretical probability distribution. It was apparently found that the degree of error was reduced when compared to the case not considering where the variance distribution. Therefore, it is expected that the outflow prediction model constructed in this study can be used as basic data for establishing an efficient river management system as more accurate prediction is possible.

Keywords: Return flowrate, Variability, Probability Distribution, LSTM

***Corresponding author:** Seung Oh Lee, seungoh.lee@hongik.ac.kr

Received: 10 April 2022, **Revised:** 14 April 2022, **Accepted:** 10 June 2022

1. 서 론

지난 109년간 우리나라의 계절별 강수량 변화를 살펴본 결과, 10년당 강수량 변화율이 +15.55 mm 증가한 여름철 강수량에 비해 겨울철 강수량은 약 −0.65 mm로 큰 변화를 나타내지 않았다(KMA, 2021). 연강수량은 최저 754 mm에서 최고 1,756 mm로 변화폭이 크므로 여름에는 홍수, 겨울과 봄에는 가뭄 피해가 빈번하게 발생하고 있으며 극한 홍수 및 가뭄 발생 빈도가 전망되고 있다. 또한 기온의 상승으로 인해 지표면 증발량이 더욱 증가하여 가용 수자원 확보에 대한 대비가 필수적인 상황이다. 특히 갈수기인 겨울과 봄에 가용수량이 부족하며 지역과 유역별로 편차가 심하기 때문에 효율적인 하천수 계획 수립이 마련이 필요하다. 국내에서는 물수지분석을 수자원장기종합계획을 통해 이용 가능한 수자원량을 예측하고 있다. 그러나 용수 수요량을 위치와 무관하게 일률적으로 행정구역별 유역면적비를 적용(Jang and Moon, 2022)하는 등의 수요량 산정의 문제점과 생 · 공용수의 일관적인 회수율 적용 방식(Oh et al., 2019) 등의 공급량 산정의 문제점이 존재한다. 이 중 회귀수량은 물 수요를 충족시킨 후 다시 하천으로 회귀되는 물의 양을 의미하고 생활 · 공업용수는 65%, 농업용수는 35%의 회귀율을 보여 회귀수량의 정확한 예측을 통해 효율적으로 하천수를 관리하는 것이 중요하다(Yoo et al., 2020, MOLIT, 2016). 따라서 본 연구는 Yoo et al.(2020)의 후속 연구로 기계학습을 통해 회귀수량 중 하수종말처리장의 방류량을 중기 예측(선행적으로 1달 수행)하였고 정확도 개선을 위해 방류량의 변동성 분포를 고려하여 입력자료로 활용하였다.

기계학습을 이용한 연구가 많이 진행되어 왔으며 주식 예측(Song et al., 2017; Lee, 2017; Kim et al., 2014), 경기 결과 예측(Kim and Kim, 2021; Seo et al., 2019; Kim et al., 2015), 질병 확산 예측(Arun et al., 2020) 등 다양한 분야에서 사용되었다. 수자원분야에서도 최근 기계학습을 통한 연구가 활발하게 이루어지고 있고 홍수 피해를 예방하기 위한 하천의 수위 예측 알고리즘 개발 연구가 주를 이루고 있다(Lee et al., 2021; Yoo et al., 2019; Jung et al., 2018; Tran et al., 2016). 기계학습의 기법 중 LSTM(Long Short-Term Memory) 기법을 사용하였는데 이는 시계열자료에 특화되어 있으며 기존 RNN(Recurrent Neural Network) 기법의 가중치 소실 문제를 보완하였다. Zhang et al.(2018)은 Elman, NARX, LSTM 기법으로 하수관 시스템을 모의한 결과, LSTM 기법이 가장 우수한 성능을 보였으며 Kim et al.(2019)은 RNN 기법에 비해 LSTM 기법이 유출량 모의 성능 연구에서 있어 더 정확도가 높다고 판단하였다. 그러나 Yoo et al.(2020)은 LSTM 기법이 극값에서 과대 및 과소 산정되는 경향이 있어 보완이 필요하다는 결과를 도출하였다.

따라서 본 연구에서는 청평댐 유역에 존재하는 하수종말처리장의 1달 후 방류량을 예측하기 위해 LSTM 기법을 활용하였으며 입력자료는 방류량을 포함한 수문자료를 사용하였다. 기존 연구에서의 한계점을 보완하기 위하여 1) 최신 데이터(2019~2020년)를 추가적으로 수집하여 학습 데이터의 양을 늘렸고 2) 시계열 자료로 구성되어 있는 입력자료의 특성 중 변동성을 파악하기 위해 통계특성을 고려하여 새로운 입력인자로 구축하였다.

2. 방법론

2.1 LSTM(Long Short-Term Memory Model)

LSTM(Long Short-Term Memory) 기법은 RNN(Recurrent Neural Network) 기법 중 최적화 오류 문제를 보완한 기법으로 Hochreiter and Schmidhuber(1997)이 제안하였다. 또한 시계열 자료 처리에 특화되어 있어(Tran et al., 2016) 시계열 자료 예측에 많이 활용되고 있다. LSTM 기법은 입력 게이트(Input gate), 출력 게이트(Output gate), 망각 게이트(Forget gate)로 구성되어 있고 시간에 따른 상태를 유지하기 위한 셀의 데이터 이동을 조절한다(Fig. 1). 관련 식은 Eq. (1) ~ (6)과 같다.

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \tag{1}$$

여기서 σ는 시그모이드 활성화 함수, w_f는 망각 게이트의 가중치, h_t는 새로운 출력 값, x_t는 입력값, b_f는 망각 게이트의 기울기, 나타낸다.

$$\widetilde{C_t} = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{2}$$

$$i_t = \sigma(w_f[h_{t-1}, x_t] + b_i) \tag{3}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \widetilde{c_t} \tag{4}$$

$\widetilde{C_t}$는 활성화 함수에 의해 생성된 새로운 셀 상태를 업데이트 시 사용하는 후보 셀, W_c는 후보 셀의 가중치, b_c는 후보 셀의 기울기를 나타낸다. 다음 단계인 입력 게이트(i_t)는 입력할 값을 결정하고 새로운 셀 상태를 업데이트한다. c_t는 현재의 셀 상태로 과거의 셀 상태인 c_{t-1}과 후보 셀 $\widetilde{C_t}$로 이루어져 있다.

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{6}$$

마지막 단계인 출력 게이트(o_t)는 무엇을 출력할지 결정하고 쌍곡탄젠트 함수를 이용해 새로운 출력 값(h_t)를 도출한다.

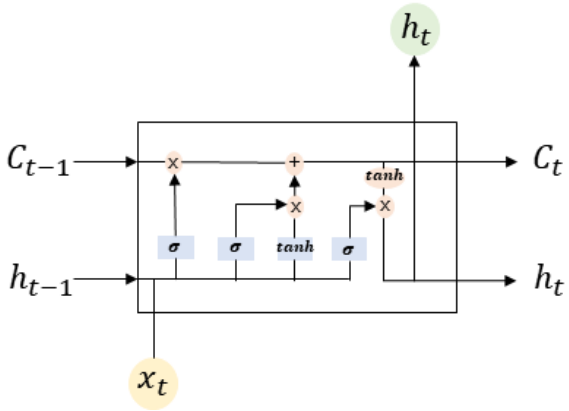

Fig. 1. LSTM structure

2.2 확률분포

2.2.1 Weibull 분포

Weibull 분포(Weibull distribution)는 연속확률 분포의 하나로, Weibull(1951)이 수명 검정 분석을 위해 고안한 분포로 고장 확률 밀도 함수를 나타내기 위해 제안하였다. Weibull 분포는 하한치가 0이고 일반적으로 오른쪽으로 왜곡되어 있는 형태를 보이므로 갈수량 또는 수명 데이터 분석에 자주 사용되고 있다. Weibull 분포의 확률밀도함수 형태는 Eq. (7)로 나타낼 수 있다. 여기서, β는 형상변수(>0), α는 축척변수(>0)이다.

$$f(x)=\frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1}\exp\left[-\left(\frac{x}{\alpha}\right)^{\beta}\right], \qquad x>0 \tag{7}$$

2.2.2 Gumbel 분포

Gumbel 분포(Gumbel distribution)도 연속확률 분포의 하나로, Gumbel(1935)이 자료의 극치 중에서도 최대치에 해당하는 자료에 대한 분포를 표시하기 위해 발표하였다. Gumbel 분포는 연 최대홍수량 및 강우량자료의 분석에 많이 사용되고 있으며, Gumbel 분포의 확률밀도함수 형태는 Eq. (8)로 나타낼 수 있다. 여기서, α는 축척변수(>0)이고, x_0은 위치변수로 최빈값(mode)의 위치를 나타낸다.

$$f(x)=\frac{1}{\alpha}\exp\left[-\frac{x-x_0}{\alpha}-\exp\left(-\frac{x-x_0}{\alpha}\right)\right], \qquad -\infty<x<\infty \tag{8}$$

2.2.3 Nomal 분포

Normal 분포(Normal distribution)는 Gaussian 분포 또는 표준오차곡선(normal error curve)이라 하며, 확률 및 통계분야에서 가장 중요한 분포이다. 일반적인 조건에서 독립확률변수들의 합이 커질수록 그 합의 분포는 Normal 분포에 가까워진다는 중심극한정리(central limit theorem) 때문이다. Normal 분포는 가설검정, 품질관리 등과 같은 통계분야 뿐만 아니라 수문분야에서도 많이 적용하는 분포이다. Normal 분포의 확률밀도함수는 Eq. (9)로 나타낼 수 있다. 여기서, μ는 평균이고, σ는 표준편차, x는 자료 값이다.

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \qquad -\infty<x<\infty \tag{9}$$

2.2.4 Generalized Extreme Value 분포

GEV 분포는 Gumbel 분포, Frechet 분포, Weibull 분포를 결합한 분포로 자료의 최대치 또는 최소치계열을 분석하는 경우에 많이 쓰이므로 수문분야에서는 홍수, 가뭄 등의 분석에 활용된다. 극치분포의 확률밀도함수는 Eq. (10)과 같다. 여기서, α는 축척변수, β는 형상변수, x_0은 위치변수를 나타낸다.

$$f(x)=\exp\left[-\left(1-\frac{\beta(x-x_0)}{\alpha}\right)^{\frac{1}{\beta}}\right] \tag{10}$$

2.2.5 Gamma 분포

Gamma 분포는 왼쪽에서 경계를 갖고 오른쪽으로(양) 왜곡되어 있어 수문자료특성과 유사성을 가지고 있다. 따라서 연최대홍수량, 갈수량 및 극대강수량 등의 확률분포를 나타내는데 사용되고 있다. Gamma 분포의 확률밀도함수는 Eq. (11)~(13)로 나타낼 수 있다. 여기서, β는 형상변수(>0), α는 축척변수(>0)이고 a_1~a_8은 상수 값이다.

$$f(x)=\frac{1}{\alpha\Gamma(\beta)}\left(\frac{x}{\alpha}\right)^{\beta-1}\exp\left(-\frac{x}{\alpha}\right) \tag{11}$$

$$\Gamma(\beta) = \Gamma(\beta+1)/\beta \tag{12}$$

$$\Gamma(\beta+1) = 1 + a_1\beta + a_2\beta^2 + \cdots + a_8\beta^8\,,\ \ 0 \le \beta \le 1 \tag{13}$$

3. 자료 구축

3.1 연구대상지

본 연구의 대상지는 청평댐 유역으로 북위 37°32'~37°52', 동경 127°15'~127°26'에 위치한다(Fig. 2). 2019년 기준 청평댐 유역의 총 인구는 180,970명으로 대부분 주거(53.84 km^2), 농업(79.08 km^2)용지로 사용하고 농업용수(53.7%), 생활용수(41.0%), 공업용수(5.3%) 순으로 용수를 이용하고 있다.

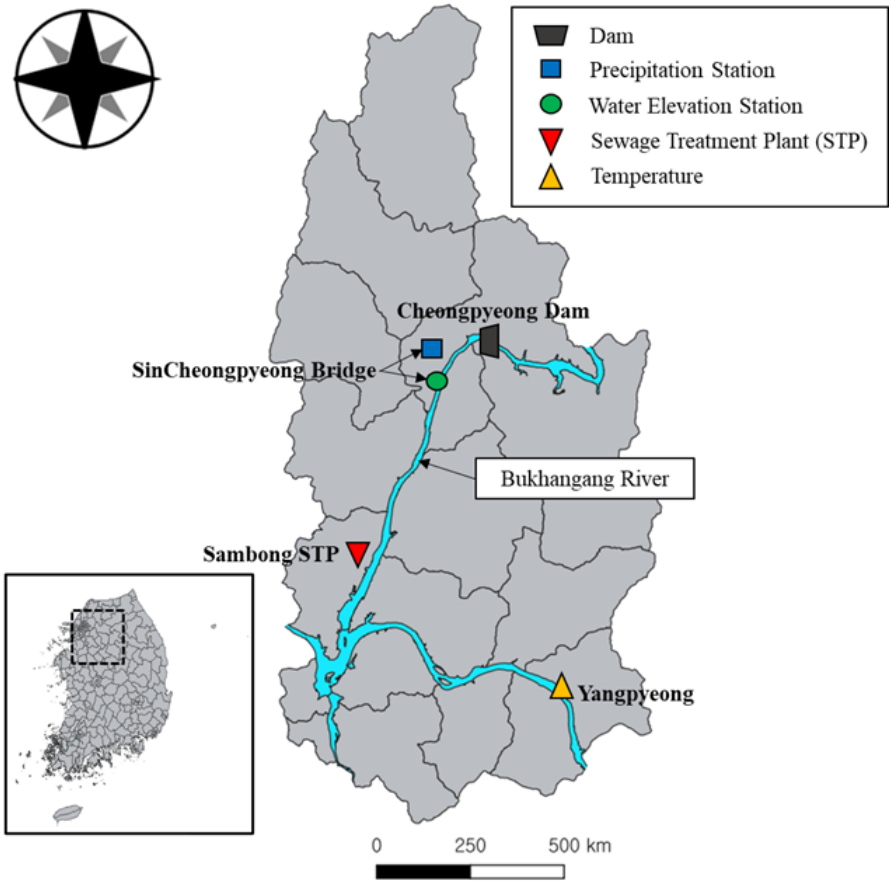

Fig. 2. Study area

3.2 입력인자

방류량 예측 알고리즘에 사용된 기본 입력인자는 강수량(mm), 수위(EL.m), 유입량(m^3/s), 방류량(m^3/s), 기온(℃)으로 총 5가지이며 각각 한강홍수통제소, 국립환경과학원, 기상청에서 수집하였다. 입력인자는 모두 시계열 자료의 1일 단위이며 2012년~2020년(9년)의 자료를 연구기간으로 활용하였다. 입력인자의 지점, 단위, 보유 기관, 기간에 대해서는 Table 1에 나타내었다.

Table 1. Information about input data

Input Data	Stations	Unit	Reference	Period
Precipitation (mm)	SinCheongpyeong Bridge	1 day	Han River Flood Control Office (HRFCO)	9 years (2012~ 2020)
Water Elevation (EL.m)	SinCheongpyeong Bridge			
Inflow (m^3/s)	Sambong STP		National Institute of Environmental Research (NIER)	
Outflow (m^3/s)	Sambong STP			
Temperature (°C)	Yangpyeong		Korea Meteorological Administration (KMA)	

3.3 자료 전처리

수집한 입력인자를 방류량 예측 알고리즘에 적용하기 위해서 t-test 및 p-value를 통해 입력인자의 유효성을 검증하였다. t-test는 2개의 집단에서 평균의 차이를 통해 통계적으로 유의미한 지 파악하기 위한 보편적인 방법으로 3가지 가정사항을 충족시켜야한다. 먼저, 수집된 데이터는 모두 같은 간격의 연속형 수치(identical interval and continuity)이어야 하며 둘째, 2개의 집단은 서로 독립적(independent)이어야 한다. 마지막으로 데이터의 수치는 정규성을 보여야 한다(normality). 또한 p-value가 0.05 이하인 경우 유의미한 인자로 판단하여(Ruxton, 2006) 본 연구의 입력자료인 강수량, 수위, 유입량, 방류량, 기온을 대상으로 t-test와 p-value를 도출한 결과, 0.0 이하의 값이 도출되어 사용 가능함을 알 수 있었다(Table 2).

Table 2. Results of t-test

Input Data	P-value	Usage status
Precipitation (mm)	~ 0.00	Available
Water Elevation (EL.m)		
Inflow (m^3/s)		
Outflow (m^3/s)		
Temperature (°C)		

또한, 방류량과 각 입력인자 간 상관관계를 판별하기 위해 상관성분석을 실행하였다. 일반적으로 상관성분석에서 많이 사용되고 연속형 변수의 상관관계를 측정하는 Pearson 상관계수(r)를 사용하였으며 Eq. (14)로 나타낼 수 있다. 여기서 $\overline{x_i}$, $\overline{y_i}$는 x_i, y_i의 평균을 의미한다.

$$r = \frac{\sum(x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum(x_i - \overline{x_i})^2}\sqrt{\sum(y_i - \overline{y_i})^2}} \tag{14}$$

입력 자료 중에 방류량과 관련이 없는 경우, LSTM 모형이 과거 자료로부터 학습할 때 교란 요인이 될 수 있기 때문에 입력인자로 사용할 수 없으며 분석결과는 Table 3와 같다. Pearson 상관계수(r)는 −1 ~ +1 사이의 값으로 일반적으로 0.1~0.3은 약한 양적 선형관계, 0.3~0.7은 뚜렷한 양적 선형관계, 0.7~1.0은 강한 양적 선형관계를 나타낸다. 강수량과 수위는 약한 양의 상관성, 유입량은 매우 강한 상관성, 기온은 강한 상관성을 나타내는 것을 알 수 있어 모두 입력인자로 사용하였다.

Table 3. Results of outflow correlation analysis

Input Data	Correlation Coefficient (r)	Correlation
Precipitation (mm)	0.20	Weak
Water Elevation (EL.m)	0.27	Weak
Inflow (m^3/s)	0.98	Very Strong
Temperature (°C)	0.34	Strong

4. 결과 및 분석

4.1 민감도 분석

방류량 예측 알고리즘의 실제 모의 수행에 앞서 최적의 매개변수 값을 적용하기 위해 민감도 분석을 실시하였다. 선정한 매개변수는 총 4가지로 시퀀스 길이(Sequence Length), 반복횟수(Iteration), 은닉층(Hidden Layer), 학습률(Learning Rate)이다. Table 4에는 매개변수별 고정값을 표기하였으며 예측자료의 정확도를 판단하기 위해 오차지표인 RMSE, MAE, IOA, R^2 로 평가하였다. 입력자료는 9년간의 수문자료를 사용하였으며 예측자료는 1달 후의 방류량으로 설정하였다.

Table 4. Test cases for sensitivity analysis

Parameter	Setting Value	Evaluation
Sequence Length	1, 5, 10*, 20	RMSE, MAE, IOA, R^2 Comparison
Iteration	100, 1000*, 10000, 50000	
Hidden Layer	1, 2*, 5, 10	
Learning Rate	0.005, 0.01*, 0.05, 0.1	

*Selected value.

각 매개변수별로 민감도분석을 실행한 결과 오차지표 비교를 통해 시퀀스 길이 5일, 반복횟수 10000번, 은닉층 5개, 학습률 0.05가 최적값으로 선정되었다. 해당 결과를 Table 5에 표기하였으며 향후 방류량 예측 알고리즘 수행 시에 해당 값으로 적용하였다.

Table 5. Results of Sensitivity Analysis

Parameter	Setting Value	MAE	RMSE	IOA	R^2
Sequence Length	1	128.71	140.36	0.43	0.20
	5	**115.48**	**132.33**	**0.49**	**0.25**
	10	119.11	135.97	0.44	0.21
	20	129.50	142.44	0.43	0.15
Iteration	10	170.74	230.14	0.32	0.06
	100	135.60	148.34	0.36	0.10
	1000	119.11	135.97	0.44	0.21
	10000	**112.96**	**134.55**	**0.62**	**0.20**
Hidden Layer	1	127.31	141.04	0.41	0.17
	2	119.11	135.97	0.44	0.21
	5	**112.21**	**124.28**	**0.60**	**0.30**
	10	113.82	135.87	0.61	0.17
Learning Rate	0.005	122.94	137.48	0.44	0.20
	0.01	119.11	135.97	0.44	0.21
	0.05	**114.07**	**130.28**	**0.56**	**0.28**
	0.1	114.37	131.96	0.52	0.23

4.2 최적 분포

본 연구에서 사용한 입력인자는 시계열 자료로 변동성을 가지는 특징이 있다. 방류량 자료의 변동성을 확인하고 새로운 입력인자로 사용하기 위해 변동성분에 대한 분포를 Eq. (15)으로 도출하였다.

$$Q(t) = Distribution(Q) + Q^{'}(t) \tag{15}$$

$$Q^{'}(t) = \alpha_1 \sin(\alpha_2 x) + \alpha_3 \cos(\alpha_4 x) \tag{16}$$

여기서, $Q(t)$는 새로운 입력인자로 사용하기 위한 최적분포 값이며 $Distribution(Q)$는 일반적으로 확률분포로 많이 사용되는 Normal 분포, Gumbel 분포, Gamma 분포, Weibull 분포, Generalized Extreme Value(GEV) 분포를 활용하였다. $Q^{'}(t)$는 Matlab(version, R2020b)을 이용하여 방류량 자료의 최적 분포형태를 비선형 회귀 모델 피팅(fitnlm) 라이브러리 및 카이제곱 검정(chi-squared test)을 통하여 선정하였으며 관측값과 확률 분포 간 잔차(residual)의 경우, 실제 자료 분석을 통하여 복합삼각함수 형태로 도출하였다(α_1~α_4는 상수, Eq. (16)). Table 6은 확률분포를 고려하지 않은 경우와 5가지의 확률분포와 잔차를 고려한 경우의 오차정도에 대해 비교하였다. 그 결과, 변동성 분포를 고려한 경우가 고려하지 않은 경우에 비해 정확도가 높게 나타났으며 정규분포 형태의 오차지표가 가장 최소로 도출되었다.

Table 6. Comparison of Distribution

Distribution	MAE	RMSE	IOA	R^2
No Distribution	113.15	132.39	0.57	0.22
Normal Distribution	**97.35**	**125.44**	**0.66**	**0.32**
Gumbel Distribution	103.77	141.23	0.61	0.04
Gamma Distribution	98.12	125.95	0.69	0.27
Weibull Distribution	110.20	138.92	0.56	0.11
GEV Distribution	101.42	140.51	0.61	0.07

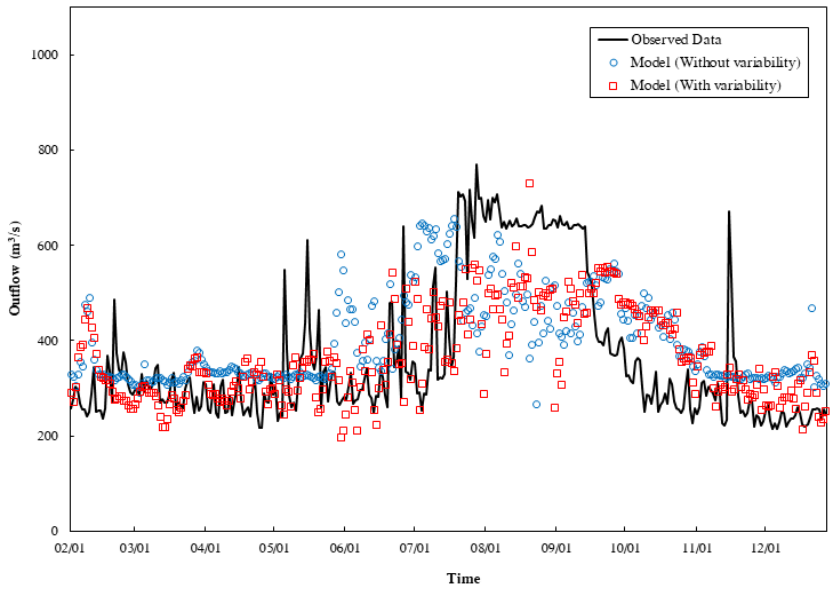

Fig. 3. Comparison of the prediction results

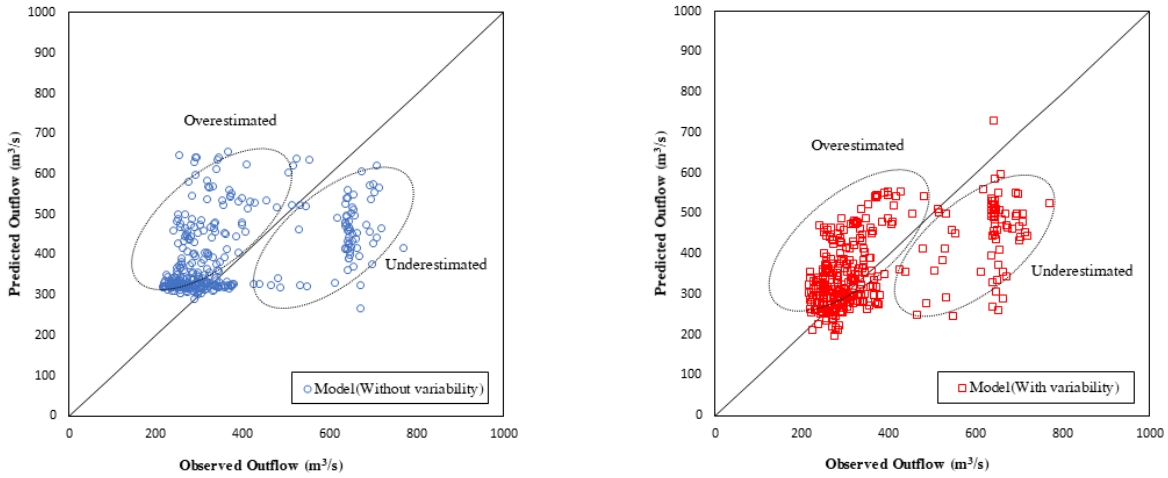

Fig. 4. Dispersion results according to variability

관측값과 변동성분포 중 정규분포를 입력자료로 고려한 경우와 고려하지 않은 경우의 모형 예측값을 Fig. 3에 제시하였다. 변동성을 입력자료로 활용하였을 때 방류량이 적은 경우가 많은 경우에 비해 정확도가 높음을 알 수 있었다. 변동성 고려 유무에 따라 산포도를 비교해본 결과(Fig. 4), 변동성을 고려하지 않은 예측값이 관측값에 비해 72.2%(239개), 변동성을 고려한 예측값이 관측값에 비해 57.7%(191개)가 과대 산정된 것을 파악할 수 있다. 변동성을 고려한 경우가 고려하지 않은 경우에 비해 적게 과대 산정되는 원인으로는 정규분포의 형태가 계절성을 간접적으로 재현했기 때문이라고 판단된다. 따라서 시계열 데이터의 자기상관성을 파악하기 위한 함수인 자기상관함수를 통해 정규분포의 고려 유무에 따른 방류량 예측 결과를 비교해보았다. Fig. 5는 변동성 고려 유무에 따른 예측된 방류량의 자기상관성을 나타내며 모두 시계열 데이터이므로 시차가 증가할 때 ACF 값이 감소하는 경향을 보인다. 또한 정규분포를 고려하여 입력인자로 사용한 경우가 그렇지 않은 경우에 비해 계절에 따라 고정된 주기가 변화하여 나타나는 Scalloped shape이 보다 뚜렷한 것으로 판단된다(Hyndman and Athanasopoulos, 2018). 따라서 정규분포의 형태를 고려한 경우 계절성의 영향을 더 받는 것을 확인할 수 있다.

회귀수량의 정확한 예측은 가뭄이나 갈수기에 가용수량을 파악하기 위해 수행되기 때문에 방류량의 값이 과대산정이 되면 재이용수로 사용되는 양에 영향을 미치게 되어 용수부족 현상이 발생할 수 있다. 따라서 변동성 분포를 입력자료로 함께 고려할 경우, 그렇지 않은 경우에 비해 정확도가 향상이 된 것으로 보아 효율적인 하천수 계획 수립에 있어 필요하다고 판단된다.

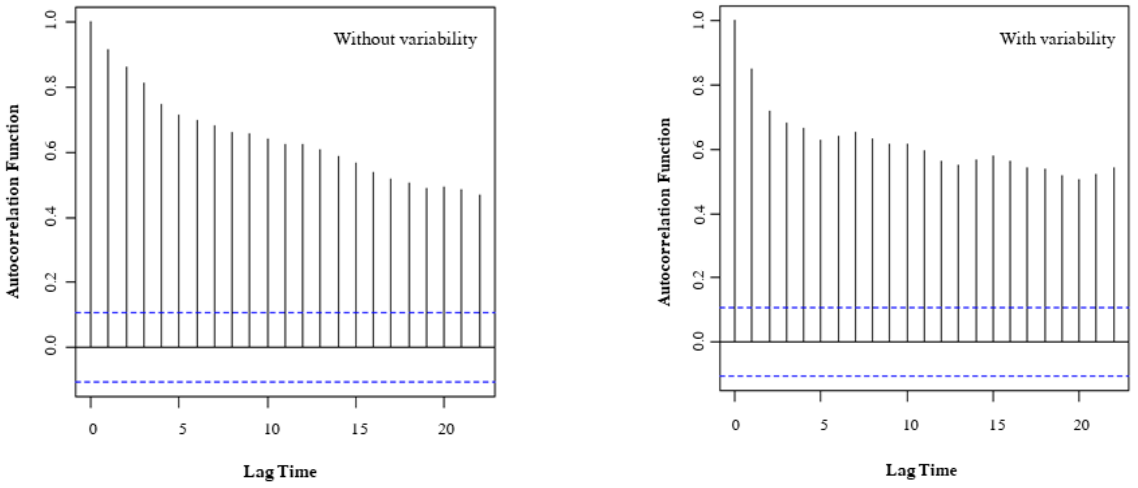

Fig. 5. Autocorrelation Function according to variability

5. 결 론

본 연구는 가뭄 등으로 인한 용수부족 현상을 대비하기 위하여 기계학습을 통해 회귀수량 중 하수종말처리장 방류량을 예측하고자 하였다. 연구대상지는 남양주시에 위치한 삼봉하수종말처리장으로 사용한 입력인자는 방류량, 유입량, 강수량, 수위이며 모두 1일 단위이다. 학습기간은 2012년~2019년, 수행기간은 2020년, 예측 결과는 1달 후 방류량으로 설정하였다. 기계학습 모형도 시계열 자료에 특화되어 있는 LSTM 모형을 활용하였고 입력인자를 입력자료로 활용하기 위해 t-test, 상관성 분석으로 검증하였으며 민감도분석을 통해 매개변수의 최적값을 도출하는 등의 모형을 구축하였다.

추가적으로 보다 정확한 예측 결과를 위해 방류량 자료의 변동성 분포도 입력인자로 고려하였다. 변동성 분포를 도출하기 위해 5가지의 분포(Normal 분포, Gumbel 분포, Gamma 분포, Weibull 분포, GEV 분포) 중 최적의 분포형태와 복합삼각함수 형태의 관측값과 분포 사이 잔차를 합한 식을 가정하였다. 오차정도를 통해 비교한 결과, Normal 분포가 가장 최적의 분포로 판단되었고 최종적으로 변동성 분포를 입력인자로 고려한 경우가 그렇지 않은 경우에 비해 오차지표인 MAE가 15.8 m^3/s, RMSE가 6.95 m^3/s 만큼 감소하였다. 특히 가용수량이 부족한 봄과 겨울에 변동성 분포를 고려하지 않은 경우가 더 과대산정되는 경향을 보였고 회귀수량을 재이용하기 위한 하천수 계획 수립에 있어 변동성 분포를 고려하는 경우가 보다 합리적인 관리체계를 수립할 것으로 예상된다.

그러나 극값 주변에서는 과소산정되는 결과를 확인하였다. 이는 LSTM 기법의 구조적인 문제와 급격한 강수량 변화로 인해 발생되었다고 판단된다. 이러한 문제를 해결하기 위해서 1) 용수이용량의 검토, 2)입력인자의 단위 축소, 3) 입력자료 전처리 고도화, 4) 변동성의 세분화 등을 고려하면 예측 방류량의 정확도를 개선시키고 최종적으로는 하천수 이용 계획을 수립하기 위한 자료로 활용될 것으로 기대된다.

Acknowledgment

본 연구는 환경부의 재원으로 한국환경산업기술원의 물관리연구사업(127572)에 의해 수행되었습니다.

References

Arun, S. S. and Iyer, G. N. (2020). On the Analysis of COVID19-novel Corona Viral Disease Pandemic Spread Data Using Machine Learning Techniques. In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE. pp.1222-1227.

Gumbel, E. J. (1935). Les Valeurs Extremes Des Distributions Statis-tiques. Annales l'institut Henri Poincar'e. 5(2): 115-158.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-term Memory. Neural Computation. 9(8): 1735-1780.

Hyndman, R. J. and Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

Jang, O. J. and Moon, Y. I. (2022). Predicting the Amount of Water Shortage during Dry Seasons Using Deep Neural Network with Data from RCP Scenarios. Journal of Korea Water Resources Association. 55(2): 121-133.

Jung, S. H., Lee, D. E., and Lee, K. S. (2018). Prediction of River Water Level Using Deep-learning Open Library. Journal of the Korean Society of Hazard Mitigation. 18(1): 1-11.

Kim, D., Park, J., and Choi, J. (2014). A Comparative Study Between Stock Price Prediction Models Using Sentiment Analysis and Machine Learning Based on SNS and News Articles. Journal of Information Technology Services. 13(3): 221-233.

Kim, J. H., Kim, K. T., and Han, J. K. (2015). Big Data Analysis based on Deep Learning for Baseball Game Data. Journal of Korea Institute of Communication Sciences. 2015(11): 262-265.

Kim, J., Kang, M. S., and Kim, S. H. (2019). Comparing the Performance of Artificial Neural Networks and Long

Short-Term Memory Networks for Rainfall-runoff Analysis. In Proceedings of the Korea Water Resources Association Conference. Korea Water Resources Association. pp.320-320.

Kim, Y. and Kim, Y. M. (2021). Predicting Game Results using Machine Learning and Deriving Strategic Direction from Variable Importance. Journal of Korea Game Society. 21(4): 3-12.

Korea Meteorological Administration (KMA) (2021). Korea Climate Change Assessment Report 2021. Seoul: KMA.

Lee, S. Y., Yoo, H. J., and Lee, S. O. (2021). Role of Unstructured Data on Water Surface Elevation Prediction with LSTM: Case Study on Jamsu Bridge, Korea. Journal of Korea Water Resources Association. 54(spc1): 1195-1204.

Lee, W. (2017). A Deep Learning Analysis of the KOSPI's Directions. Journal of the Korean Data and Information Science Society. 28(2): 287-295.

Ministry of Land Infrastructure and Transport (MOLIT) (2016). National Water Resources Plan (2011~2020) (3rd revision). Sejong: MOLIT.

Oh, J. H., Ryu, K. S., Bok, J. S., Jang, Y. S., Bae, Y. D., and Lee, B. G. (2019). Water Supply-and-Demand Analysis Considering the Actual Water-Use System in the East Basin of Han River. Journal of the Korean Society of Hazard Mitigation. 19(7): 529-543.

Ruxton, G. D. (2006). The Unequal Variance T-Test is an Underused Alternative to Student's T-Test and the Mann-Whitney U Test. Behavioral Ecology. 17(4): 688-690.

Seo, Y. J., Moon, H. W., and Woo, Y. T. (2019). A Win/Lose Prediction Model of Korean Professional Baseball Using Machine Learning Technique. Journal of the Korea Society of Computer and Information. 24(2): 17-24.

Song, Y. J., Lee, J. W., and Lee, J. W. (2017). A Design and Implementation of Deep Learning Model for Stock Prediction Using Tensorflow. KIISE Transactions on Computing Practices. 23(11): 799-801.

Tran, Q. T., Hao, L., and Trinh, Q. K. (2016). A Novel Procedure to Model and Forecast Mobile Communication Traffic by ARIMA/GARCH Combination Models. In 2016 International Conference on Modeling. Simulation and Optimization Technologies and Applications (MSOTA2016). Atlantis Press.

Weibull, W. (1951). A Statistical Distribution Function of wide Applicability. Journal of Applied Mechanics.

Yoo, H. J., Lee, S. O., Choi, S. H., and Park, M. H. (2020). Development of a Data-Driven Model for Forecasting Outflow to Establish a Reasonable River Water Management System. Journal of Korean Society of Disaster and Security. 13(4): 75-92.

Yoo, H., Lee, S. O., Choi, S., and Park, M. (2019). A Study on the Data Driven Neural Network Model for the Prediction of Time Series Data: Application of Water Surface Elevation Forecasting in Hangang River Bridge. Journal of Korean Society of Disaster and Security. 12(2): 73-82.

Zhang, D., Martinez, N., Lindholm, G., and Ratnaweera, H. (2018). Manage Sewer In-line Storage Control Using Hydraulic Model and Recurrent Neural Network. Water Resources Management. 32(6): 2079-2098.

Korean References Translated from the English

국토교통부 (2016). 수자원장기종합계획(2001~2020) 제3차 수정계획. 세종: 국토교통부.

기상청 (2021). 한국 기후변화 평가보고서 2021 -기후변화 과학적 근거-. 서울: 기상청.

김동영, 박제원, 최재현 (2014). SNS 와 뉴스기사의 감성분석과 기계학습을 이용한 주가예측 모형 비교 연구. 한국 IT 서비스 학회지. 13(3): 221-233.

김용우, 김영민 (2021). 기계학습을 활용한 게임승패 예측 및 변수중요도 산출을 통한 전략방향 도출. 한국게임학회 논문지. 21(4): 3-12.

김종훈, 김경태, 한종기 (2015). Deep Learning 기반 기계학습 알고리즘을 이용한 야구 경기 Big Data 분석. 한국통신학회 학술대회논문집. 2015(11): 262-265.

서영진, 문형우, 우용태 (2019). 기계학습 기법을 이용한 한국프로야구 승패 예측 모델. 한국컴퓨터정보학회논문지. 24(2): 17-24.

송유정, 이종우 (2017). 텐서플로우를 이용한 주가 변동 예측 딥러닝 모델 설계 및 개발. 한국정보과학회 학술발표논문집. pp.799-801.

오지환, 류경식, 복정수, 장연석, 배영대, 이봉국 (2019). 실제 물이용 체계를 고려한 한강 동해 권역 물수급 평가. 한국방재학회 논문집. 19(7): 529-543.

유형주, 이승오, 최서혜, 박문형 (2019). 시계열 자료의 예측을 위한 자료 기반 신경망 모델에 관한 연구: 한강대교 수위예측 적용. 한국방재안전학회논문집. 12(2): 73-82.

유형주, 이승오, 최서혜, 박문형 (2020). 합리적인 하천수 관리체계 구축을 위한 자료기반 방류량 예측모형 개발. 한국방재안전학회 논문집. 13(4): 75-92.

이승연, 유형주, 이승오 (2021). LSTM 기법을 활용한 수위 예측 알고리즘 개발 시 비정형자료의 역할에 관한 연구: 잠수교 사례. 한국수자원학회 논문집. 54: 1195-1204.

이우식 (2017). 딥러닝분석과 기술적 분석 지표를 이용한 한국 코스피주가지수 방향성 예측. 한국데이터정보과학회지. 28(2): 287-295.

정성호, 이대업, 이경상 (2018). 딥러닝 오픈 라이브러리를 이용한 하천수위 예측. 한국방재학회논문집. 18(1): 1-11.