

Analyzing the Impact of Multivariate Inputs on Deep Learning-Based Reservoir Level Prediction and Approaches for Mid to Long-Term Forecasting

Hyeseung Park[†] · Jongwook Yoon^{††} · Hojun Lee^{†††} · Hyunho Yang^{††††}

ABSTRACT

Local reservoirs are crucial sources for agricultural water supply, necessitating stable water level management to prepare for extreme climate conditions such as droughts. Water level prediction is significantly influenced by local climate characteristics, such as localized rainfall, as well as seasonal factors including cropping times, making it essential to understand the correlation between input and output data as much as selecting an appropriate prediction model. In this study, extensive multivariate data from over 400 reservoirs in Jeollabuk-do from 1991 to 2022 was utilized to train and validate a water level prediction model that comprehensively reflects the complex hydrological and climatological environmental factors of each reservoir, and to analyze the impact of each input feature on the prediction performance of water levels. Instead of focusing on improvements in water level performance through neural network structures, the study adopts a basic Feedforward Neural Network composed of fully connected layers, batch normalization, dropout, and activation functions, focusing on the correlation between multivariate input data and prediction performance. Additionally, most existing studies only present short-term prediction performance on a daily basis, which is not suitable for practical environments that require medium to long-term predictions, such as 10 days or a month. Therefore, this study measured the water level prediction performance up to one month ahead through a recursive method that uses daily prediction values as the next input. The experiment identified performance changes according to the prediction period and analyzed the impact of each input feature on the overall performance based on an Ablation study.

Keywords : Reservoir Water Level Prediction, Fully Connected Neural Network, Multivariate Inputs

다변량 입력이 딥러닝 기반 저수율 예측에 미치는 영향 분석과 중장기 예측 방안

박혜승[†] · 윤종욱^{††} · 이호준^{†††} · 양현호^{††††}

요약

지역 저수지들은 농업용수 공급의 중요한 수원공으로 가뭄과 같은 극단적 기후 조건을 대비하여 안정적인 저수율 관리가 필수적이다. 저수율 예측은 국지적 강우와 같은 지역적 기후 특성뿐만 아니라 작부시기를 포함하는 계절적 요인 등에 크게 영향을 받기 때문에 적절한 예측 모델을 선정하는 것만큼 입/출력 데이터 간 상관관계 파악이 무엇보다 중요하다. 이에 본 연구에서는 1991년부터 2022년까지의 전라북도 400여 개 저수지의 광범위한 다변량 데이터를 활용하여 각 저수지의 복잡한 수문학·기후학적 환경요인을 포괄적으로 반영한 저수율 예측 모델을 학습 및 검증하고, 각 입력 특성이 저수율 예측 성능에 미치는 영향력을 분석하고자 한다. 신경망 구조에 따른 저수율 예측 성능 개선이 아닌 다변량의 입력 데이터와 예측 성능 간의 상관관계에 초점을 맞추기 위하여 실험에 사용된 예측 모델로 합성곱신경망 또는 순환신경망과 같은 복잡한 형태가 아닌 완전연결계층, 배치정규화, 드롭아웃, 활성화 함수 등의 조합으로 구성된 기본적인 순방향 신경망을 채택하였다. 추가적으로 대부분의 기존 연구에서는 하루 단위의 단기 예측 성능만을 제시하고 있으며 이러한 단기 예측 방식은 10일, 한 달 단위 등 중장기적 예측이 필요한 실무환경에 적합하지 않기 때문에, 본 연구에서는 하루 단위 예측값을 다음 입력으로 사용하는 재귀적 방식을 통해 최대 한 달 뒤 저수율 예측 성능을 측정하였다. 실험을 통해 예측 기간에 따른 성능 변화 양상을 파악하였으며, Ablation study를 바탕으로 예측 모델의 각 입력 특성이 전체 성능에 끼치는 영향을 분석하였다.

키워드 : 저수율 예측, 완전연결신경망, 다변량 입력

※ 이 논문은 한국국토정보공사 공간정보연구원 산학협력R&D사업의 지원을 받아 수행된 연구임(과제명: 공간정보기반 인공지능분석을 활용한 농업용 저수지의 가뭄대비 저수율 예측, 과제번호: 2023-501).

† 정 회 원 : 협성대학교 소프트웨어공학과 조교수

†† 비 회 원 : 협성대학교 경영학과 교수

††† 정 회 원 : (주)수리이앤씨 SI사업부 이사

†††† 정 회 원 : 국립군산대학교 컴퓨터소프트웨어학부 교수

Manuscript Received : February 19, 2024

First Revision : March 19, 2024

Accepted : March 22, 2024

*Corresponding Author : Hyunho Yang(hhyang@kunsan.ac.kr)

1. 서론

대한민국 전국에서 사용되는 물은 생활용수, 공업용수, 농업용수 등 다양한 용도로 활용되고 있으며, 연간 수자원 이용량의 상당 부분, 약 63%가 농업용수로 소비된다는 사실은 농업용수의 중요성을 강조한다[1]. 특히, 지구 온난화와 이상 기후 현상으로 인해 가뭄과 같은 극단적 기상 조건이 발생함에 따라, 농업용 저수지의 수위 변화는 물 공급과 농업의 지속 가능성에 직간접적인 큰 영향을 미친다[2]. 이러한 배경에서 저수지의 저수율을 정확히 모니터링하고 예측하는 것은 농업용수 관리와 가뭄 대비 계획 수립에 있어 필수적이다.

기존 저수율 예측은 주로 저수지의 과거 저수율과 기후 데이터를 통계적으로 해석하여 미래 저수율을 예측하는 추계학적 방법에 의존해왔다. 이러한 방식은 시간의 흐름에 따른 데이터의 패턴을 분석하여 단기 예측에 유용하게 적용될 수 있으나, 장기간 데이터의 복잡한 패턴과 최근 기후 변화 추세를 포괄적으로 반영하는 데는 한계가 있다.

최근 연구에서는 LSTM(Long Short-Term Memory)과 같은 순환신경망을 포함한 딥러닝 모델을 활용하여 저수율 예측의 정확도를 높이려는 다양한 시도가 있었다. 그러나 대부분 단기 예측과 고도화된 모델 구조를 통한 성능 개선에만 초점을 맞추고 있어, 중장기 예측과 다변량 입력 데이터의 중요성을 심도 있게 다룬 연구는 상대적으로 부족하다.

본 연구는 모델 개발 자체보다는 지역적 기후 특성(예: 국지적 강우)과 계절적 요인(예: 작부시기)을 포함한 다양한 변수들이 저수율 예측 성능에 미치는 영향을 체계적으로 분석하는 데 중점을 둔다. 이를 위해 본 연구진은 실험을 위한 예측 모델로 합성곱 신경망(CNN, Convolutional Neural Network) 또는 순환 신경망(RNN, Recurrent Neural Network)과 같은 복잡한 신경망 구조가 아닌 완전연결 계층, 배치 정규화, 드롭아웃, 활성화 함수 등으로 구성된 기본적인 순방향 구조의 완전연결 신경망(FCNN, Fully-connected Neural Network)을 채택하였다.

학습을 위해 1991년부터 2022년까지 전라북도 지역의 약 400개 저수지에서 수집된 기후 및 환경 데이터를 활용한다. 이 데이터에는 강수량, 온도, 필요수량, 연 주기성이 반영된 시간 등의 다변량 정보가 포함되어 있다. 이를 활용하여 저수지별로 모델을 학습시키고, 각 저수율 예측 모델의 성능을 평가한다. 예측 기간은 1일, 3일, 5일, 10일, 20일, 그리고 월 단위로 설정하고, 이 기간 동안 재귀적 예측 방식을 통해 중장기적 예측 성능을 검증한다. 추가로, Ablation study를 통해 다변량 입력 변수가 예측 성능에 미치는 영향을 분석한다.

본 논문은 다음과 같이 구성되어 있다. 제2장에서는 저수지, 댐, 하천 등의 수위 예측에 관한 국내외 연구를 소개한다. 제3장에서는 연구 대상 저수지와 학습 데이터셋에 대해 상세히 설명하고, 사용된 저수율 예측 신경망을 다룬다. 제4장에서는 성능 평가 지표를 소개하고, 400여 개 저수지에 대한 예

측 모델의 성능을 검증한다. 마지막으로, 제5장 결론에서는 연구의 주요 발견 사항을 요약한다.

2. 관련 연구

최근 딥러닝으로 대표되는 머신러닝 기술을 활용한 저수율 예측 연구들이 활발히 진행되고 있다. 대표적인 국내의 연구 현황은 다음과 같다.

Seo 등[3]은 충주댐 유역에 대해 다양한 ANN, GRNN, ANFIS, RF와 같은 다양한 기계학습모델을 적용하여 저수율을 예측하도록 하고, 모델 성능을 제곱근평균제곱오차, 평균절대오차, 평균제곱상대오차 및 평균교차오차 지표를 활용하여 평가하였다.

Jung 등[4]은 텐서플로우를 활용하여 다중선형회귀모델과 LSTM 모델을 구현하였다. 이를 활용하여 금강 유역 내 대청댐 옥천 관측소 지점 수위 예측 연구를 수행하였다. 연구에서 제시한 모델은 하천 수위에 영향을 미치는 수많은 물리적 매개변수를 배제하고 오직 상류에서 관측된 수위 정보를 바탕으로 하류의 수위를 예측하는 것에 초점을 맞추고 있으며, 2시간 단위 단기 예측을 수행하였으며, 이를 통해 다중선형회귀모델의 한계와 LSTM[5] 모델의 우수성을 보였다.

Han 등[6]은 소양강 댐의 유입량을 예측하기 위해 시계열 데이터 예측에 높은 성능을 보이는 LSTM 기반의 Sequence-to-Sequence (s2s) 모델을 적용하였다. 입력 특성으로는 2013년부터 2019년까지 소양강 댐에서 관측된 일 유입 유량과 인제, 홍천, 춘천에서 관측된 일 강우량 데이터를 사용했다. 출력 특성은 소양강 댐의 일 유입량이며, 하루 단위 예측을 수행하였다. 해당 연구진은 상관계수, Nash-Sutcliffe 효율 계수, 평균 편차 비율 및 최대값 오차 등 다양한 통계적 지표를 사용하여 모델 예측 성능을 평가하였으며, 결과적으로 LSTM-s2s 모델은 댐 유입량 예측뿐만 아니라 단일 유량 수문학적 사건 기반 예측에서도 높은 정확도와 신뢰성을 보였다.

Yang 등[7]은 LSTM 모델을 활용하여 손항저수지, 수양저수지, 용암저수지에 대한 이상치 탐지 방안을 제안하였다. Seong 등 [8]은 유량 변동이 크고 하구에서의 조석의 영향이 거의 없는 팔당대교 관측소의 2년 7개월간 수위 및 유량 자료를 바탕으로 LSTM과 GRU(Gated Recurrent Unit)[9] 딥러닝 모델을 사용하여 수위와 유량을 10분 단위로 예측하는 연구를 다뤘다. 입력 특성으로는 저수지의 수위 자료와 인근 지역의 강수량 데이터가 사용되었으며, 출력 특성은 이상치가 탐지된 저수지 수위 데이터이다.

Joh 등[10]은 농업용 저수지의 저수율 예측을 위해 단변량 및 다변량 LSTM 모델을 활용한 연구를 수행했다. 이 연구에서는 인공지능 기술을 이용하여 농업용 저수지의 저수율을 예측한다. 단변량 LSTM 모델은 저수율만을 입력 특성으로 사용한다. 반면 다변량 LSTM 모델은 저수율 외에도 n일 누적 강수량과 연중 시기(DOY, Date of Year)를 추가적인 입력 특성

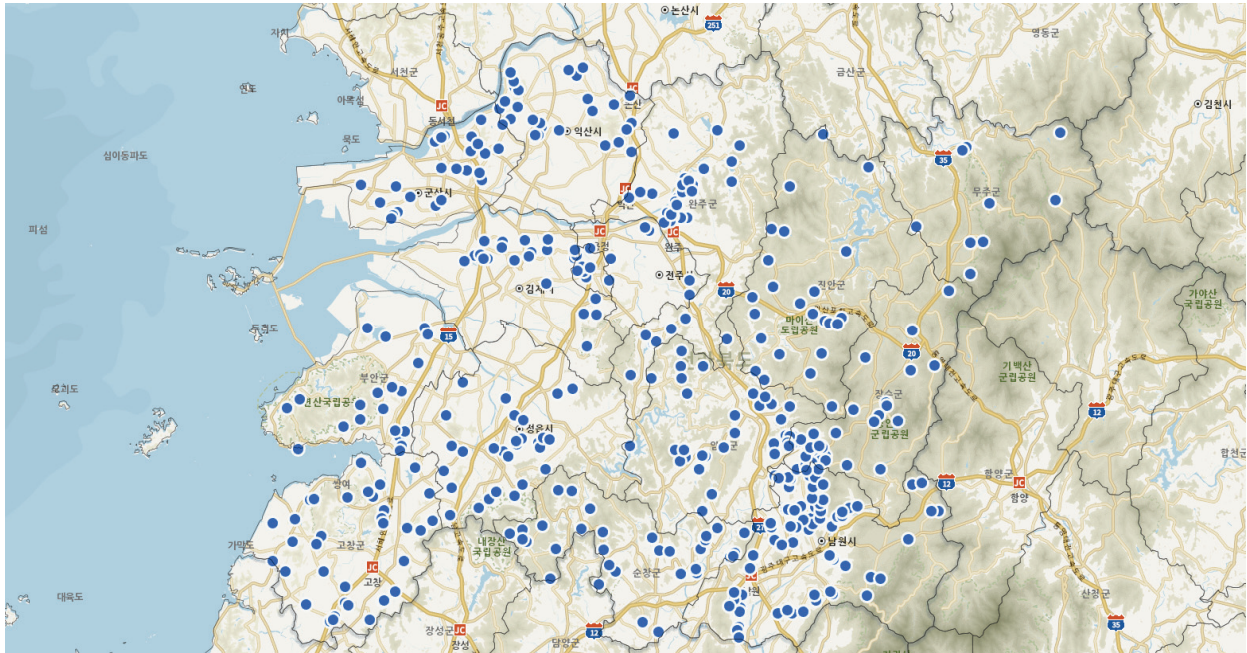


Fig. 1. Research Subject Reservoirs: Approximately 400 Reservoirs in Jeollabuk-do

으로 포함하였다. 이 모델들은 이동저수지의 2013년부터 2020년까지 8년간의 데이터를 학습하여 2021년의 일일 저수율을 예측하는 데 사용되었다. 1일, 3일, 5일에 대한 단기적 저수율 변화를 예측하였으며, 실험을 통해 다변량 모델에서 저수율 외에 DOY, 일일 및 5일 누적 강수량이 중요한 변수로 확인되었다.

Das 등[11]은 저수지의 수위를 예측하기 위해 베이지안 네트워크 기반의 확률론적 비선형 접근 방식을 사용한다. 모델의 입력 특성으로는 과거의 수위 데이터가 사용되며, 출력 특성은 예측되는 저수지의 일일 수위이다. 인도의 마유라크시 저수지 과거 22년간 데이터를 기반으로 하루 단위 예측을 수행하였다.

위 연구들은 딥러닝 기반 모델이 효과적으로 수위 예측이 가능함을 보였다. 그러나 주로 10분에서 5일 이내의 단기 예측 성능만을 평가 및 검증하였던 한계를 가지며, 단기 예측은 기존 추계학적 방법론에서도 낮지 않은 성능을 보여왔기 때문에 중장기적인 예측 모델 연구 및 성능 검증에 대한 필요성은 여전히 존재한다. 또한 대부분의 연구에서 LSTM이나 GRU와 같은 고도화된 순환신경망 구조를 기반으로 한 예측 모델의 우수성을 보였으나, 순환신경망이 갖는 시간적 지연 현상 문제에 대한 고려와 예측 성능에 영향을 줄 수 있는 다변량 입력 특성에 대한 연구가 필요하다.

3. 연구자료 및 방법

3.1 연구 대상 저수지 및 학습용 데이터셋 구성

Fig. 1은 본 연구 대상 저수지의 위치를 보여준다. 전주시 은석 저수지, 백석 저수지, 군산시 원우 저수지, 옥구 저수지, 익

산시 도촌 저수지 등과 같은 전라북도 소재 400여 개 저수지들을 실험 대상으로 선정하였으며, 농촌용수종합정보시스템[12], 기상정보개방포털[13] 등에서 제공하는 1990년부터 2022년까지의 저수지별 일 단위 저수율(Reservoir Water Level), ASOS 기반 강수량(precipitation), 온도(Temperature) 데이터를 신경망 학습 데이터로 사용하였다. 필요수량(Water Demand)과 날짜(Day) 데이터는 본문에서 제시하는 특정 공식에 따라 계산하여 학습에 활용하였다.

본 연구에서는 효율적인 학습, 검증, 및 테스트를 위해 저수지별 데이터셋을 1990년부터 2018년, 2019년부터 2020년, 2021년부터 2022년까지의 구간으로 분할하였다. 검증 데이터는 모델의 하이퍼파라미터 최적화, 테스트 데이터는 모델의 일반화 성능을 검증하기 위해 사용된다.

실제 저수지별 예측 모델의 입력으로 사용되는 다중 독립 변수로는 각 저수지의 일별 강수량, 일별 온도, 일별 필요수량, 날짜, 하루 전 저수율(Reservoir Water Level of the Previous Day)을 사용하였다. 모델이 예측하는 단일 종속 변수는 저수율로 설정하였다. 이외 저수지 유역면적, 수혜면적, 깊이 등과 같은 정적인 특성을 고려할 수 있으나, 이는 고정값으로 하나의 저수지에 대한 예측 모델을 구성함에 있어 입력 특성에서 배제하였다. 이와 같은 변수는 일반화된 모델, 즉 다수의 서로 다른 저수지 데이터를 하나의 모델로 학습시킬 때 의미있는 특성이 될 것이다.

필요수량 데이터는 작물의 정상적인 생육 하에 논과 밭에서 소비되는 수량으로 수로손실, 증발산량, 그리고 침투량으로 구성되며, Penmann 공식[14]에 기반하여 다음과 같이 산정한다.

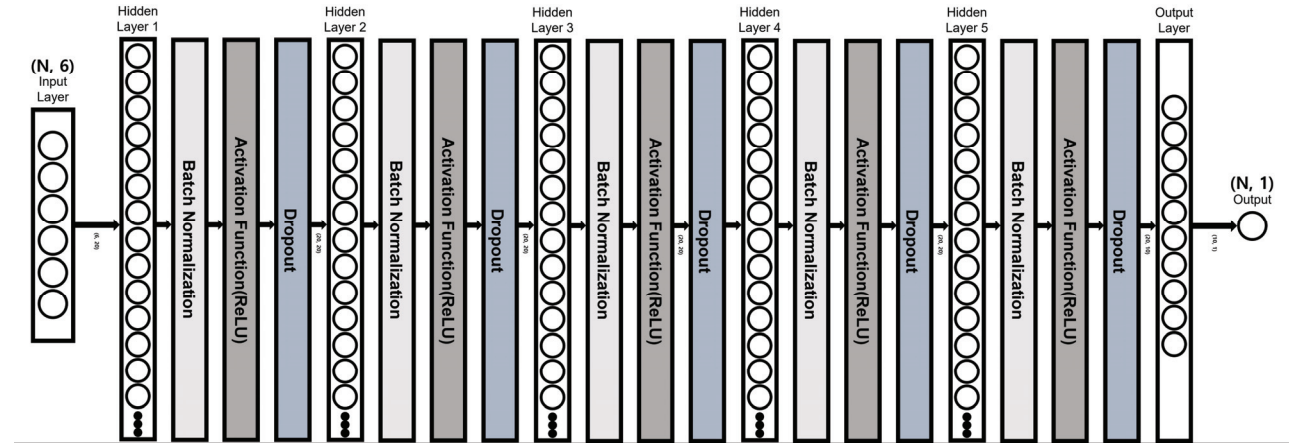


Fig. 2. Architecture of the Fully-connected Neural Network for Predicting Water Level of a Reservoir

$$PET = c \left[\frac{\Delta}{\Delta + \gamma} R_n + \frac{\gamma}{\Delta + \gamma} f(u_2)(e_{sa} - e_a) \right] \quad (1)$$

$$R_n = R_{ns} - R_{nl} \quad (2)$$

$$f(u_2) = 0.27 \left(1 + \frac{u_2}{100} \right) \quad (3)$$

Equation (1), (2), (3)에서 PET , c , Δ , R_n , $f(u_2)$, e_{sa} , e_a , R_{ns} , R_{nl} , u_2 는 각각 잠재 증발산량(mm/day), 조정계수, 온도~포화 증기압곡선의 기울기(mb/°C), 건습계정수(기온 15°C에서 0.66mb/°C), 순복사량(mm/day), 풍속에 관계하는 함수, 포화수증기압(mb), 공기 중의 수증기압(mb), 순단파 복사량(mm/day), 순장파 복사량(mm/day), 지상 2m에서의 일평균풍속(km/d)를 의미한다.

날짜 데이터는 년, 월, 일을 있는 그대로 사용하지 않고, 작부시기와 같은 계절별 주기적 특성을 반영하기 위하여 다음과 같은 공식으로 산출된 값 두 가지를 사용한다.

$$Day_{\sin} = \sin(2\pi \times doy)/t \quad (4)$$

$$Day_{\cos} = \cos(2\pi \times doy)/t \quad (5)$$

Equation (4), (5)에서 doy (Day of Year)는 특정 날짜가 위치한 순번을 의미한다. 예를 들어 1월 1일은 1, 2월 1일은 32로 치환된다. 는 전체 날짜의 개수로 윤년의 경우 366, 그렇지 않은 경우에는 365를 의미한다. 위 식을 통해 12월 31일과 1월 1일이 공간적으로 가깝게 표현되어, 계절별 주기성을 효과적으로 학습시킬 수 있을 것으로 기대한다.

학습 과정에서 하루 전 저수율 데이터를 입력으로 사용하는 이유는 이전 시점의 저수율이 다음 저수율과 밀접한 상관관계, 즉 자기상관성(self-correlation)을 가질 수 있기 때문이며, 이는 예측 정확성 향상에 도움이 될 것이다. 학습 이후 검

증 또는 테스트, 즉 추론 과정에서는 한 달 단위의 재귀적 예측을 가정했을 때 매월 1일을 제외한 나머지 일에 대해서는 실측값이 아닌 예측값이 사용된다. 매월 1일 예측에 대해서는 하루 전 실측값을 그대로 사용하고, 나머지에 대해서는 invalid로 전처리하여 직전 저수율 예측치로 덮어쓰도록 하였다. 이와 같은 방식은 3일, 5일, 10일, 20일 단위 예측에도 동일하게 적용되었다.

3.2 저수율 예측 신경망

저수율 예측에는 LSTM, GRU와 같은 시계열 데이터 처리에 최적화된 모델 또는 CNN과 같이 위성 사진 등의 이미지 데이터 처리에 유리한 모델이 사용될 수 있다. 그러나 이러한 고도화된 모델들은 과적합이나 기울기 소실 같은 문제에 직면할 수 있다. 특히, 순환신경망(RNN)에서는 모델이 현재 데이터보다 과거 데이터에 더 크게 의존하여 예측을 수행하는 시간적 지연(Lag)이 발생하기 쉬우며, 예측 결과가 이전 실측값을 일정 시점 뒤에 따라가는 경향을 보인다 [15]. 본 연구에서는 이러한 문제에 대해 고려하기보다는 전통적인 다중퍼셉트론(Multi-Layer Perceptron)에 기반한 기본적인 구조의 신경망을 구성하여 효과적으로 다변량의 독립 변수들과 종속 변수 간의 상관관계를 파악하고자 하였다.

본 연구에서는 저수지별 저수율을 예측하기 위해 Fig. 2에 나타난 기본적인 순방향 구조의 완전연결 신경망을 설계 및 구현하였다. 이 신경망은 예측하고자 하는 저수지의 다음 날 강수량, 온도, 필요수량, 날짜의 순환적 특성을 반영한 Day_{\sin} 와 Day_{\cos} , 그리고 하루 전 저수율로 구성된 총 6개의 입력 채널을 사용한다. 이러한 입력 변수들은 저수율에 직접적인 영향을 미칠 것으로 예상되는 환경적 및 시간적 요소들을 포함한다. 모델은 다음 날의 저수율을 예측하며, 따라서 출력 채널은 1로 설정된다. 효율적인 학습을 위해, 모든 특성에 대해 Min-Max 정규화를 수행하여 서로 다른 특성 간의 상관관계를 동일한 스케일에서 분석하고자 하였다.

제안된 모델은 5개의 완전연결 계층(Fully Connected Layer)으로 이루어진 은닉층(Hidden Layer)을 포함한다. 각 은닉층은 동일한 20개의 노드를 포함하며, 복잡한 데이터 패턴을 학습할 수 있도록 ReLU(Rectified Linear Unit) 비선형 활성화 함수를 사용한다. 학습 효율성을 높이고 과적합을 방지하기 위해, 은닉층에는 배치 정규화 계층(Batch Normalization Layer)과 드롭아웃 계층(Dropout Layer)을 추가하였다. 최종 출력층 직전의 계층(Output Layer)은 10개의 노드로 구성되어 있으며, 이는 복잡한 특성 추출 후 저수율 예측을 위한 정보를 축적하는 역할을 한다. 실제 최종 출력층은 1개의 노드로 구성되어 있어, 다음 날의 저수율을 예측한다. 이러한 구조와 데이터 처리 방법을 통해, 본 연구는 환경적 및 시간적 요소들이 저수율에 미치는 영향을 효과적으로 분석하고 예측할 수 있는 모델을 제안한다.

학습은 NVIDIA GeForce RTX 3090 그래픽 카드 2개를 사용하여 수행되었다. 이는 고성능 컴퓨팅 환경에서 대량의 데이터를 효율적으로 처리할 수 있게 해준다. 모델 학습을 위한 하이퍼파라미터는 배치 크기(Batch size) 128, 학습률(Learning rate) 1×10^{-4} , 그리고 총 300 에폭(Epoch)이 설정되었으며, 검증(Validation) 과정을 거쳐 충분한 학습이 이루어지면서도 과적합을 방지하기 위해 선정되었다. 손실 함수로는 L1 손실을 사용하였다. 이는 모델이 실측값과 예측값 간의 절대 차이를 최소화하도록 유도한다. 최적화 알고리즘으로는 효율성과 빠른 수렴 속도로 인해 많은 딥러닝 모델에서 널리 사용되는 아담 최적화 함수(Adam Optimizer)[16]를 사용한다. 제안된 모델은 Python 기반의 딥러닝 프레임워크인 PyTorch[17]를 사용하여 구현되었다. PyTorch는 유연성과 사용 편의성으로 인해 복잡한 신경망 모델을 구축하고 실험하는 데 적합한 도구이다. 실험에서 저수지별 학습 시간은 평균적으로 5분 내외로 소요되었다.

3.3 재귀적 증장기 예측

증장기적 예측을 위해 본 연구진은 재귀적 예측 방법을 활용하였다. 재귀적 예측 방법은 시계열 데이터 예측에 널리 사용되는 기법으로, 하나의 예측값이 다음 시점의 입력 데이터로 사용되는 방식이다. 증장기적으로 저수율을 직접 예측하는 방법(예를 들면, 오늘을 기준으로 30일 이후의 당일 저수율)과 재귀적 방법(예를 들면, 오늘을 기준으로 내일 저수율을 예측하고, 예측된 내일 저수율을 바탕으로 모레 저수율을 예측하는 방식을 반복) 사이에는 몇 가지 차이점과 장단점이 존재한다. 재귀적 예측 방법의 장점을 증점적으로 살펴보자면, 먼저 재귀적 예측 방법은 매일 또는 매주와 같은 짧은 기간의 예측을 반복하여 장기 예측을 구성한다. 새로운 데이터가 수집됨에 따라 모델을 지속적으로 업데이트하고, 최근 정보를 반영하여 예측의 정확도를 개선할 수 있는 유연성을 제공한다. 두 번째로 재귀적 방법은 단기 예측을 연속적으로 수행하기 때문에, 단기적인 변동성이나 이상 현상을 더 잘 포착하고 반영할

수 있다. 세 번째로 장기 예측을 위해 복잡한 구조를 갖는 대신, 단기 예측을 반복함으로써 모델의 복잡성을 관리할 수 있다. 이는 컴퓨팅 리소스 사용 측면에서도 더 효율적이라고 할 수 있다. 재귀적 예측 방법은 여러 시점에 걸쳐 반복되며, 각 시점의 예측값은 이전 예측의 결과에 영향을 받게 된다. 따라서 연속적인 데이터 포인트 간의 관계를 모델링할 수 있으나, 시간이 지남에 따라 오차가 누적되어, 최종 장기 예측의 오차를 증가시킬 수 있다는 단점이 있다.

4. 모델 적용 및 실험 결과 분석

4.1 평가지표

본 연구에서는 모델의 예측 성능을 평가하기 위한 통계적 평가지표로 평균절대오차(MAE, Mean Absolute Error), 평균 제곱오차(MSE, Mean Squared Error), 제곱근평균제곱오차(RMSE, Root Mean Squared Error)를 사용하였다. MAE는 예측값과 실측값 간의 절대 차이 평균, MSE는 예측값과 실측값 간 차이 제곱의 평균을 의미한다. RMSE는 MSE의 제곱근으로 예측 오차의 표준편차와 유사하며, 오차의 크기를 원래 단위로 해석 가능하도록 한다. 이러한 지표들은 모두 예측 모델이 얼마나 잘 또는 잘못 예측하고 있는지를 수치적으로 나타내는데 사용된다. MSE와 RMSE는 큰 오류에 더 많은 가중치를 부여하는 반면, MAE는 모든 오류를 동일하게 취급한다는 특징을 갖는다. 각각의 공식은 다음과 같다.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (6)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (8)$$

여기서 n , y_i , \hat{y}_i 은 각각 데이터 포인트의 총 개수, 실측값, 모델 예측값을 의미한다.

4.2 저수지별 학습 결과 및 성능 평가

본 연구진은 400여 개 저수지 각각에 대해 학습된 저수율 예측 모델의 성능을 (Fig. 3)에 나타난 바와 같이 MAE, MSE, 그리고 RMSE의 분포를 통해 분석하였다. 히스토그램과 커널 밀도 추정(KDE, Kernel Density Estimation) 곡선을 통해 400여 개 저수지에 대한 각 오차 지표의 분포를 시각화하였다. KDE는 각 데이터 포인트 주위에 연속적인 밀도 함수를 배치하여 전체 데이터 세트의 확률 밀도 함수를 추정하는 방법으로 모델의 예측 성능을 직관적으로 이해하는 데 도움을 준다. Table 1은 1일부터 한 달 단위까지의 400여 개 저수지 개별 학습모델들의 예측 결과를 토대로 계산한 예측 성능 평균을

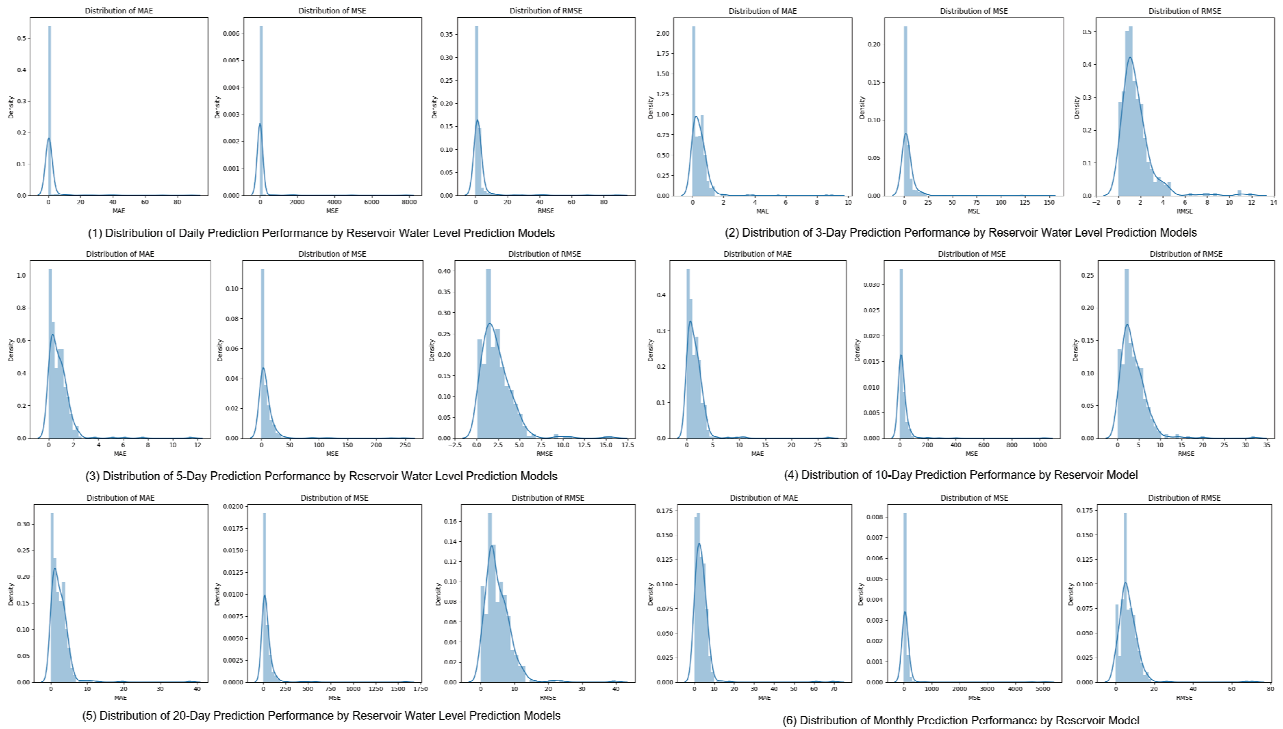


Fig. 3. Kernel Density Estimation and Histogram for the Distribution of Model Performance Across Different Reservoirs

MAE, MSE, RMSE 지표로 보여준다.

Fig. 3과 Table 1에 대한 분석 결과, MAE 값들은 주로 낮은 범위에 집중되어 있어, 대부분의 모델들이 실제 저수율과 최소한으로 벗어나는 예측을 함으로써 일반적으로 예측이 정확하였음을 시사한다. KDE 곡선의 표측한 정점은 강한 중심 경향성을 나타내며, 이는 대부분의 모델이 좁은 오차 범위를 갖는다고 볼 수 있다. 반면, MSE의 분포는 이상치에 민감하며, 특히 큰 오차 값을 가진 이상치 모델들이 존재함을 나타낸다. RMSE는 MSE의 제곱근으로, 큰 오차의 영향을 줄여 실제 오차 크기를 보다 현실적으로 나타낸다. 이로 인해 RMSE 분포는 MAE 및 MSE의 경향을 반영하면서도, 큰 오차의 영향을 완화시킨 형태를 보여준다.

더욱이, 예측 기간이 1일에서 30일로 증가함에 따라 재귀적 오차 누적이 관찰되었다. 즉, 예측 기간이 길어질수록 오차가 누적되어 전체적인 예측 오차가 증가하는 경향이 나타난다. 이러한 경향은 특히 장기 예측에서 더욱 뚜렷하게 나타나, 장기 예측의 정확도를 향상시키기 위한 연구 및 모델 최적화의 필요성을 시사한다.

Fig. 4는 2021년부터 2022년까지 전라북도 전주시 기지저수지와 군산시 옥녀저수지의 실측 저수율 값과 3일, 5일, 10일, 20일, 한 달 단위로 증가하는 재귀적 예측 기간에 따른 예측값을 비교한 것이다. 두 저수지는 측정된 전체 400여 개 저수지에 대한 MSE 값 분포에서 중위값 근처에 위치한 저수지이기 때문에 선정되었으며, 이는 중간 수준의 오차를 가진 저수지 모델을 대표적으로 분석하기 위함이다. 실험에 사용된

Table 1. Analysis of Recursive Prediction Performance for Over 400 Reservoirs on an n-day Basis

Prediction Period	MAE(Avg.)	MSE(Avg.)	RMSE(Avg.)
1	0.5525	4.5800	1.6248
3	0.4933	4.6212	1.5979
5	0.8225	8.9252	2.3512
10	1.5873	22.8555	3.7346
20	2.4535	40.7927	5.1207
30	3.3156	55.7065	6.3676

저수율 예측 모델들은 예측 기간이 짧은 단기 예측에서 실측 값과의 높은 일치도를 보였지만, 예측 기간이 길어짐에 따라 오차의 누적으로 인해 예측값의 변동성이 증가하고, 실측값과의 차이가 확대되는 경향이 명확하게 관찰되었다. 이는 장기 예측에서는 다양한 외부 요인과 복잡한 수문학적 과정이 예측 오차에 더 큰 영향을 미칠 수 있음을 의미한다.

4.3 Ablation Study

저수율 예측 모델의 각 입력 특성이 예측 정확도에 미치는 기여도를 정량적으로 평가하기 위해 본 연구진은 포괄적인 Ablation Study를 수행하였다. 이 연구는 베이스라인 모델을 여러 번 재학습하면서 매번 다른 특성을 생략하고, MAE, MSE, RMSE 측면에서 성능 저하를 평가한다. 베이스라인 모델은 강수량, 온도, 필요수량, 계절적 주기성이 반영된 시간, 그리고

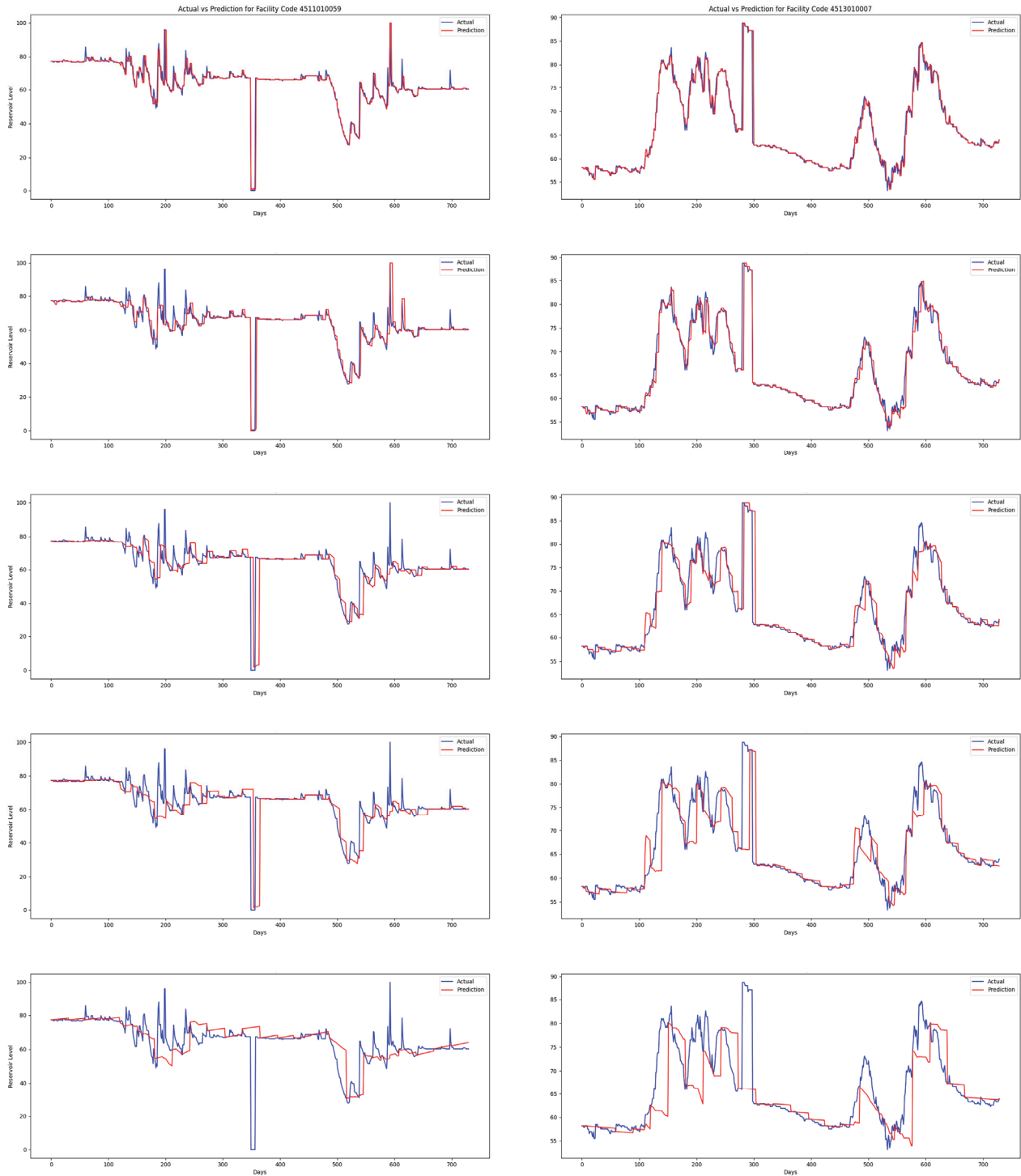


Fig. 4. Comparison of Actual and Predicted Reservoir Water Levels for Giji Reservoir (left) in Jeonju-si and Oknyeo Reservoir (right) in Gunsan-si, Conducted in 3, 5, 10, 20-Day, and Monthly Intervals from 2021 to 2022.

하루 전 저수율을 입력으로 사용한다. 각 특성의 개별적 영향력을 확인하기 위해 베이스라인 모델로부터 하나의 특성을 제거하고 그에 따른 성능 지표를 Table 2와 같이 기록하였다.

Ablation Study 결과를 살펴보면, 모든 특성이 성능 향상에 긍정적 영향을 미치고 있다. 먼저, 필요수량 특성을 제거했을

때 MAE, RMSE, MSE 평균이 각각 16%, 22.49%, 7.6%씩 증가하는 것을 알 수 있다. 이는 필요수량이 저수율 예측에 있어 가장 중요한 시기인 작부시기와 연관성이 높기 때문에 분석된다. 연별 순환성이 반영된 날짜(Day_{sin} , Day_{cos}) 특성을 제거했을 때 MAE, RMSE, MSE 평균이 각각 13.4%, 22%, 6.3%

Table 2. Impact of Input Variable Removal on Performance

Precipitation	Temperature	Water Demand	Day	Water Level of Prev. Day	MAE (Avg.)	MSE (Avg.)	RMSE (Avg.)
O	O	O	O	O	3.3156	55.7065	6.3676
X	O	O	O	O	3.7284	64.8973	6.7550
					12.5%	16.5%	6.1%
O	X	O	O	O	3.6220	61.6319	6.6314
					9.2%	10.6%	4.1%
O	O	X	O	O	3.8461	68.2359	6.8517
					16%	22.49%	7.6%
O	O	O	X	O	3.7604	67.9805	6.7712
					13.4%	22%	6.3%
O	O	O	O	X	15.6664	444.858	19.1087
					372.5%	698.6%	200%

씩 증가하는 것을 알 수 있다. 저수지의 수위 변화는 날씨, 계절, 농업용수의 수요가 물리는 작부시기 등과 같은 시간에 따라 변화하는 동적 요소들의 영향을 받는다. 날씨 특성을 통해 예측 모델은 이러한 주기적인 변화를 학습하고 예측 성능을 향상시킬 수 있다고 볼 수 있다. 당연하게도 저수지 수위에 직접적인 영향을 주는 강수량과 시기 파악에 도움을 줄 수 있는 온도 특성 역시 전체 성능 향상에 유의미한 역할을 수행함을 확인할 수 있다. 저수율 예측에 가장 큰 영향을 주는 것은 하루 전 저수율이며, 이러한 결과는 다음과 같은 몇 가지 주요 요인으로 설명할 수 있다. 먼저 저수율은 시간적 연속성을 갖기 때문이다. 직전 시점의 저수율은 당연하게도 다음 시점의 저수율을 예측하는 데 매우 중요한 정보를 제공한다. 이는 짧은 기간 내 발생하는 저수율의 변동성을 감지하기 위해서도 필수적이다. 또한 저수율은 강수량, 온도, 필요수량 등 다른 환경변수들과 그 자체로 복잡한 상관관계를 갖는다. 하루 전 저수율은 이러한 변수들의 영향을 종합적으로 반영하는 지표 역할을 수행한다.

5. 결 론

본 연구는 전라북도 지역의 약 400개 저수지 데이터를 활용하여 다변량 데이터를 포함한 완전연결신경망 모델로 저수지별 저수율을 예측하는 데 중점을 두었다. 이 모델은 강수량, 온도, 필요수량, 날씨의 주기적 특성 및 하루 전 저수율을 입력 변수로 통합하여 저수율 예측의 정확도를 높이는 것을 목표로 한다.

본 연구의 핵심 발견은 중장기 예측 성능에 다변량 데이터의 활용이 중요하다는 것이다. 재귀적 예측 방법을 통해 최대한 달 단위까지의 미래 저수율 예측의 정확도를 검증하였으며, 이 방법이 예측 기간이 길어질수록 누적되는 오차에도 불구하고 단기 예측에서 변동성을 잘 포착하고 정확도를 향상시킬 수 있음을 보여준다. 또한, Ablation Study를 통해 각 입력

변수가 예측 성능에 미치는 구체적인 영향을 분석함으로써, 향후 모델의 성능 개선 방향에 대한 인사이트를 제공한다.

이러한 결과는 저수율 예측을 위한 기계학습 모델의 개발 뿐만 아니라, 저수지 관리와 농업용수 공급 계획에 있어 유용한 의사결정 도구로서의 가능성을 보여준다. 더욱이, 중장기 예측 결과는 이 분야의 향후 연구와 모델 개발에 기초자료로 활용될 수 있다.

향후 연구에서는 저수지별 학습 모델이 아닌 일반화된 통합 모델 개발이 필요하며, 이를 위해 개별 저수지 학습에서 사용되지 않은 유역면적, 수해면적, 깊이 등의 고정값을 추가적으로 활용할 수 있다. 또한 예측 모델의 고도화를 통한 성능 개선 방안과 중장기 재귀적 예측 성능을 더욱 향상시킬 방안을 모색해야 하며, 이러한 연구들은 향후 농업용수 관리와 가뭄 대비에 실질적인 도움을 줄 수 있을 것이다.

References

- [1] H. Lim, H. An, G. Choi, J. Lee, and J. Do, "Prediction of pollution loads in agricultural reservoirs using LSTM algorithm: Case study of reservoirs in Nonsan City," *Korean Journal of Agricultural Science*, Vol.49, No.2, pp.193-202, 2022.
- [2] J. Jung, S. Kim, D. Kim, and S. Yang, "A study on the use of geospatial information-based simulation for preemptive response to water disasters in agricultural land," *Smart Media Journal*, Vol.11, No.7, pp.52-60, 2022.
- [3] Y. Seo, E. Choi, and W. Yeo, "Reservoir water level forecasting using machine learning models," *Journal of the Korean Society of Agricultural Engineers*, Vol.59, No.3, pp.97-110, 2017.
- [4] S. Jung, D. Lee, and K. Lee, "Prediction of river water level using deep-learning open library," *Journal of the Korean Society of Hazard Mitigation*, Vol.18, No.1, pp.1-11, 2018.

[5] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," 2014.

[6] H. Han, C. Choi, J. Jung, and H. Kim, "Application of sequence to sequence learning based LSTM model (LSTM-s2s) for forecasting dam inflow," *Journal of Korea Water Resources Association*, Vol.54, No.3, pp.157-166, 2021.

[7] M. Yang, W. Nam, H. Kim, T. Kim, A. Shin, and M. Kang, "Anomaly detection in reservoir water level data using the LSTM model based on deep learning," *Journal of the Korean Society of Hazard Mitigation*, Vol.21, No.1, pp.71-81, 2021.

[8] Y. Seong, K. Park, and Y. Jung, "Flow rate prediction at Paldang Bridge using deep learning models," *Journal of Korea Water Resources Association*, Vol.55, No.8, pp.565-575, 2022.

[9] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[10] Joh. Sunguk, and Lee, Yangwon. "Prediction of Water Storage Rate for Agricultural Reservoirs Using Univariate and Multivariate LSTM Models," *Korean Journal of Remote Sensing* 39.5 (2023): 1125-1134.

[11] M. Das, S. K. Ghosh, V. M. Chowdary, A. Saikrishnaveni, and R. K. Sharma, "A probabilistic nonlinear model for forecasting daily water level in reservoir," *Water Resources Management*, Vol.30, pp.3107-3122, 2016.

[12] RAWRIS [Internet], <https://rawris.ekr.or.kr>

[13] KMA Weather Data Service [Internet], <https://data.kma.go.kr>

[14] H. L. Penman, "Natural evaporation from open water, bare soil and grass," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, Vol.193, No.1032, pp.120-145, 1948.

[15] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent advances in recurrent neural networks," *arXiv preprint arXiv:1801.01078*, 2017.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[17] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, Vol.32, 2019.



박혜승

<https://orcid.org/0000-0003-3564-2479>
 e-mail : hs2000park@omail.uhs.ac.kr
 2012년 한양대학교 컴퓨터전공(학사)
 2014년 한양대학교 전자컴퓨터통신공학과 (석사)
 2021년 한국기술교육대학교 컴퓨터공학과 (박사)

2021년 ~ 현재 협성대학교 소프트웨어공학과 조교수
 관심분야 : 인공지능, 컴퓨터비전 등



윤종욱

<https://orcid.org/0009-0007-2601-4012>
 e-mail : biziq@naver.com
 1988년 인하대학교 산업공학과(학사)
 1991년 한국외국어대학교 경영과학과 (석사)
 1997년 인하대학교 경영학과(박사)

1995년 ~ 현재 협성대학교 경영학과 교수
 관심분야 : 빅데이터, 인공지능, 경영혁신, 정보기술전략 등



이호준

<https://orcid.org/0009-0006-5771-628X>
 e-mail : eagle_ff@hotmail.com
 1995년 중앙대학교 제어계측공학과(학사)
 2011년 (주)한국공간정보통신 정보사업실 부장
 2023년 네이버시스템(주) 통합관계플랫폼팀 팀장

2023년 ~ 현재 (주)수리이엔씨 SI사업부 이사
 관심분야 : 농업가뭄관리, 물수지관리, GIS



양현호

<https://orcid.org/0000-0002-8884-9935>
 e-mail : hhyang@kunsan.ac.kr
 1986년 광운대학교 전자공학과(학사)
 1990년 광운대학교 전자공학과(석사)
 2003년 광주과학기술원(GIST) 정보통신공학과(박사)

2005년 ~ 현재 국립군산대학교 컴퓨터소프트웨어학부 교수
 관심분야 : 기계학습, 인공지능, 빅데이터 등