

ORIGINAL ARTICLE

호우 영향예보를 위한 머신러닝 기반의 수문학적 정량강우예측(HQPF) 최적화 방안

이한수 · 지용근* · 이영미 · 김병식¹⁾

(주)에코브레인, ¹⁾강원대학교 도시·환경방재공학전공

Optimizing Hydrological Quantitative Precipitation Forecast (HQPF) based on Machine Learning for Rainfall Impact Forecasting

Han-Su Lee, Yongkeun Jee*, Young-Mi Lee, Byung-Sik Kim¹⁾

ECOBRAIN Co. Ltd., Jeju 63309, Korea

¹⁾Department of Urban & Environmental Disaster Prevention Engineering, Kangwon National University, Samcheok 25913, Korea

Abstract

In this study, the prediction technology of Hydrological Quantitative Precipitation Forecast (HQPF) was improved by optimizing the weather predictors used as input data for machine learning. Results comparison was conducted using bias and Root Mean Square Error (RMSE), which are predictive accuracy verification indicators, based on the heavy rain case on August 21, 2021. By comparing the rainfall simulated using the improved HQPF and the observed accumulated rainfall, it was revealed that all HQPFs (conventional HQPF and improved HQPF 1 and HQPF 2) showed a decrease in rainfall as the lead time increased for the entire grid region. Hence, the difference from the observed rainfall increased. In the accumulated rainfall evaluation due to the reduction of input factors, compared to the existing HQPF, improved HQPF 1 and 2 predicted a larger accumulated rainfall. Furthermore, HQPF 2 used the lowest number of input factors and simulated more accumulated rainfall than that projected by conventional HQPF and HQPF 1. By improving the performance of conventional machine learning despite using lesser variables, the preprocessing period and model execution time can be reduced, thereby contributing to model optimization. As an additional advanced method of HQPF 1 and 2 mentioned above, a simulated analysis of the Local ENsemble prediction System (LENS) ensemble member and low pressure, one of the observed meteorological factors, was analyzed. Based on the results of this study, if we select for the positively performing ensemble members based on the heavy rain characteristics of Korea or apply additional weights differently for each ensemble member, the prediction accuracy is expected to increase.

Key words : HQPF, Machine learning, Optimization, Adjustment of rainfall forecast, LENS, Ensemble

Received 3 December, 2021; Revised 13 December, 2021;

Accepted 16 December, 2021

*Corresponding author: Yongkeun Jee, ECOBRAIN Co. Ltd., Jeju 63309, Korea
Phone : +82-70-7018-0794
E-mail : ykjee@ecobrain.net

© The Korean Environmental Sciences Society. All rights reserved.
© This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. 서론

강우는 비선형적인 현상이므로 일반적인 선형모형으로 예측하는 것에는 한계가 있다. 기존 통계 기법과 달리 머신러닝을 기반으로 한 접근법은 어떠한 가정도 필요하지 않으므로 비선형성 관계를 가진 빅데이터 기반의 자료를 분석하는 수치 모델의 예측성능을 보완하는데 기여한다. 최근에는 기상 및 기후와 관련된 자료가 방대해지고 이와 같은 자료를 보다 정확한 예측정보로 만들어내기 위한 빅데이터 기반의 다양한 방법론이 개발되고 있다(Ghada et al., 2019; Valipour et al., 2019; Zamami et al., 2019). 또한 수치예보자료의 정량적 강우량 추정 개선을 위해 인공지능망을 활용한 여러 연구가 이루어져 왔고 집중호우와 같이 예측오차가 크고 단기간에 정확한 예측이 필요한 현상에 대해 기존의 예측한계를 보완하기 위한 방법으로 머신러닝이 널리 사용되고 있다(Hong, 2008; Kang and Lee, 2011; Sumi et al., 2012).

현재 한국기상청(Korea Meteorological Administration, KMA)에서 제공하는 기상학적 정량적 강우 예보(Quantitative Precipitation Forecast, QPF)인 동네예보(Digital forecast)를 활용하여 최근 기후변화로 인한 수자원의 영향을 분석하기에는 어려움이 많다. 수자원의 수문학적 이벤트에 대한 대표적인 사례 중 침수는 주로 집중호우시기에 많이 발생한다. 집중호우는 짧은 시간에 지역적 편차가 큰 사례가 대부분이지만 그에 비해 QPF의 공간해상도와 시간해상도는 집중호우에 대비하기에는 한계가 있다. 따라서 수문학적 활용을 하기 위한 강우 예보가 필요하다. 그러기 위해서는 수치모델의 정확도 향상에 대한 노력도 필요하지만 머신러닝 기법을 적용한 향상된 강우정보에 대한 연구도 필요하다. 신속하게 생산되는 머신러닝 기반의 강우 예측결과가 집중호우시기에 도시와 지역사회 중 재해 취약성이 큰 도시에서 신속한 긴급상황관리에 매우 도움이 될 것이기 때문이다.

침수 예측과 같은 수문학적 활용을 위한 강우 정보를 생산하기 위해 다양한 강우 정보에 머신러닝 기법 중 하나인 XGBoost (eXtreme Gradient Boosting) 기법을 적용한 수문학적 강우 정보인 Hydrological Quantitative Precipitation Forecast (HQPF)를 생산한다. HQPF 생산을 위해 기상청(KMA)에서 제공하는 앙상블 수치모델 자료(Local ENsemble prediction System, LENS), 레

이다(Radar-AWS rainrates, RAR), 지점관측자료 및 동네예보(Digital forecast)의 강우예보에 대한 전처리 과정을 통하여 모든 자료의 시간 및 공간 해상도를 동일하게 변환하여 머신러닝 기상예측 인자를 도출한다. 한편, 현재 HQPF의 경우 앙상블 수치모델인 LENS의 13개 멤버별 보정강우량을 평균하여 적용하고 있어 일부 앙상블 멤버가 관측을 잘 모의하지 못하는 경우 예측성능이 떨어지는 한계를 가지고 있다.

본 연구에서는 HQPF 생산을 위해 개발한 호우예보 보정기술의 과정 중 머신러닝의 입력자료로 사용되는 기상예측 인자를 축소함으로써 머신러닝 과정을 최적화하고 HQPF의 예측기술을 고도화하고자 하였다. 이를 위해 최근 호우 사례 중 2021년 8월 21일 호우 사상을 대상으로 기존 HQPF와 기상예측 인자를 축소한 경우의 예측성능을 비교 분석하였다. 또한, LENS 앙상블 멤버 중 우리나라의 호우 특성을 잘 반영하는 멤버를 활용하여 HQPF의 예측성능을 향상시키는 방법을 모색하였다.

2. 연구자료 및 방법

2.1. 연구자료

2.1.1. 강우 관측 자료

본 연구에서 사용한 관측자료로는 Automated Surface Observing System (ASOS)과 Automatic Weather Station (AWS)이 있다. 한국 기상청에서는 전국 95개소의 ASOS와 494개의 AWS를 사용하여 한반도의 강수, 기온 등에 대해 관측업무를 수행하고 국내 전역의 강수특성을 정량적으로 파악하는데 효과적으로 활용할 수 있다. 본 연구에서는 사례기간 동안의 ASOS 자료를 이용하여 보정강우량을 오차검증 하였다(Fig. 1).

Radar-AWS Rainrate (RAR)는 레이더와 AWS를 활용하여 실시간으로 Z-R계수를 산출 및 적용하는 강우 강도 산출시스템이다. 또한 이 시스템은 한반도 영역의 고해상도 정량적 강우량 자료를 산출한다. 해상도는 1 km의 공간해상도와 10분 주기의 시간해상도로 제공된다.

2.1.2. 강우 예측 자료

본 연구에서는 예측 강우보정 알고리즘 개발을 위해 기상청에서 제공하는 Local ENsemble prediction System (LENS)과 동네예보를 사용하였다. LENS 자료는 한반도 위험기상 확률 예측을 위해 13개의 앙상블 섭동

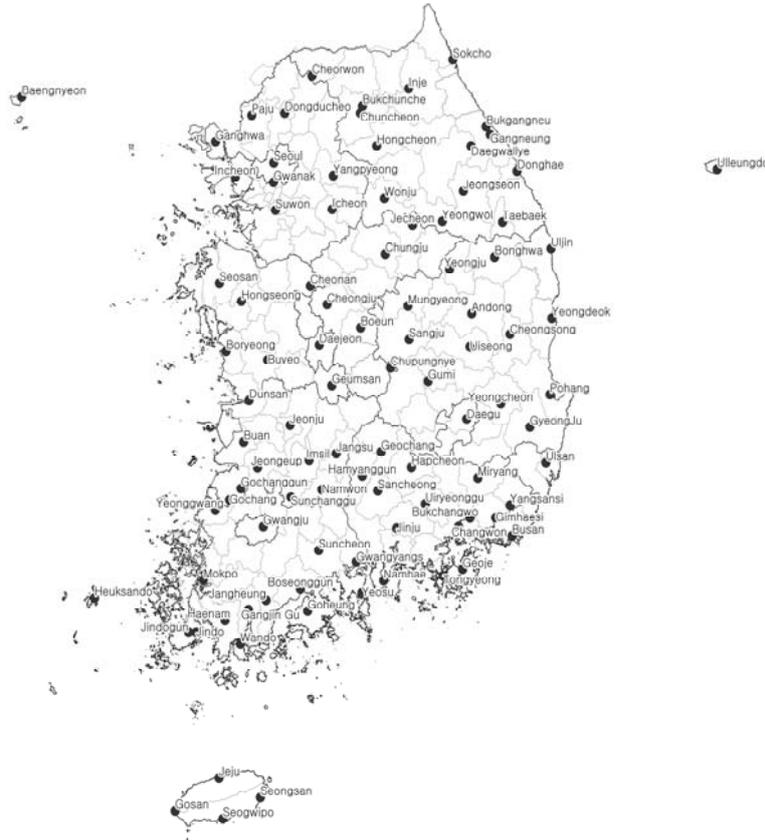


Fig. 1. ASOS observation points (95 station).

멤버로 72시간까지 예측정보를 제공하는 Unified Model (UM)기반 국지 앙상을 예측시스템이고 2.2 km의 공간 해상도와 12시간 주기의 시간 해상도로 제공된다.

동네예보자료는 한국기상청에서 2008년부터 전국 읍, 면, 동 단위로 상세하게 제공하고 있다. 이 예보자료는 한반도 전체 영역에 대해서 5 km 격자 간격으로 기온, 풍향, 강수확률, 강수 형태 등의 정량적 예보를 제공하고 있다. 본 연구에서는 동네예보 자료에서 6시간 누적 강우자료를 추출하여 사용하였다.

2.2. 머신러닝

2.2.1. 기상예측인자

HQPF를 생산하는데 필요한 입력자료는 Ko et al.(2020)에서 제시한 기상예측 인자를 기준으로 하였다. 기존에는 LENS 자료에서 3차원 공간자료를 활용하여

층후(Thickness), 상층제트, 하층제트, 연직바람쉬어, 습수, 가강수량, 불안정지수, 연직속도, 지상기온, 지상풍속, 해면기압, 강우량을 추출하였고 ASOS, AWS, RAR, 동네예보자료에서는 강우량 자료를 활용하였다. 본 연구에서는 머신러닝 과정의 최적화를 위하여 입력되는 기상예측 인자를 달리하여 모델을 구축하고 결과를 비교하였다. Table 1에서 기존의 HQPF 생산에 사용된 예측인자와 변경된 예측인자를 나타내었다. 기존 15개 예측인자를 이용한 경우를 HQPF라 하였으며, 상대습도와 기온 등 8개의 예측인자를 이용한 경우를 Improved HQPF 1, LENS의 지표면 풍속과 예측자료들의 강수량 변수만 사용한 경우를 Improved HQPF 2라고 정의하였다. 수집된 자료들은 머신러닝의 입력자료로 사용하기 위해 호우영향모델의 격자를 기준으로 처리되었다. ASOS와 AWS는 지점자료이므로 호우영향모델 격자

Table 1. Predictor variables used machine learning

Variables	Source	Selected variables		
		HQPF	Improved HQPF 1	Improved HQPF 2
Atmospheric thickness 500 hPa ~ 1000 hPa [m]	LENS	○		
Wind speed at 200/500/850 hPa [m/s]	LENS	○		
Vertical wind shear 200 hPa ~ 850 hPa [m/s]	LENS	○		
Dew point temperature at 700 hPa [°C]	LENS	○		
K-Index (Instability Index)	LENS	○		
Sea level pressure [hPa]	LENS	○		
Vertical velocity at 700 hPa [hPa/hour]	LENS	○		
Relative humidity at 850 hPa [%]	LENS	○	○	
Temperature at surface [°C]	LENS	○	○	
Precipitable water [mm]	LENS	○	○	
Wind speed at surface [m/s]	LENS	○	○	○
Precipitation [mm]	LENS	○	○	○
Precipitation [mm]	ASOS, AWS	○	○	○
Precipitation [mm]	RAR	○	○	○
Precipitation [mm]	Digital forecast	○	○	○

중에서 1 km 이내 자료를 활용하고 그 외의 격자는 RAR 자료로 내삽되었다. 동네예보자료는 5 km에서 1 km의 공간해상도로 변환하기 위해 통계학적인 방법으로 내삽하는 크리깅 기법(kriging method)을 적용하였다. 그리고 동네예보자료의 6시간 누적강우량을 1시간으로 상세화하기 위해 LENS 앙상블 평균자료를 이용하였다. LENS 자료는 동네예보 자료와 동일한 시점의 6시간 누적강우량을 계산하였고, 동네예보에 대한 비율을 산출하여 1시간 간격의 LENS 강우량 자료에 곱하였다. 상세화된 1시간 간격의 동네예보 강우자료는 6시간 누적강우량과 동일하면서, LENS 자료의 시간 변화를 따른다.

2.2.2. XGBoost (eXtreme Gradient Boosting)

기상예측인자들 사이에서 비선형 관계를 찾고 가중치를 계산하기 위해 머신러닝이 사용되었다. 머신러닝의 앙상블 학습에는 배깅과 부스팅 방식이 있다. 배깅 방식은 학습데이터를 무작위로 샘플링하여 분할하고, 약한 예측 모델들을 결합하여 최종 결과를 산출한다. 부스팅 방식은 이전 모델들이 예측하지 못한 오차 데이터에 가중치를 부여하여 강한 예측모형을 만드는 알고리즘으로, 이전 모델의 오차를 고려한다는 점에서 배깅 방식과 다

르다. 배깅 방식에는 Random forest, 부스팅 방식에는 Adaptive Boosting (AdaBoost), Gradient Boosting (GBM), eXtreme Gradient Boosting (XGBoost)이 있다. Adaboost는 각 모델별로 cost function을 통해 가중치를 고려한 예측모델이고, GBM 방식은 Adaboost와 기본개념은 동일하고, 가중치를 계산하는 방식에서 gradient descent를 이용한다. 여기에서 활용한 XGBoost 방식은 GBM 대비 분산 및 병렬 처리를 통해 성능이 향상되었다는 점에서 우수하다. 일반적으로 XGBoost의 속도는 GBM 보다 10배 이상 빠르다. 그리고 이 기법은 효율적이고 확장성 있는 구현이 가능하므로 여러 선행연구에서 사용되어왔다(Friedman et al., 2000; Ke et al., 2017; Chen and He, 2021).

XGBoost는 여러 개의 분류 회귀나무(classification and regression tree)를 묶어 오류값을 낮추는 부스팅 기법으로 다음과 같은 방식으로 구성된다.

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (1)$$

Table 2. XGBoost regression that model hyperparameters from the grid search

Parameter	Range	Optimum Value
Max_depth	2 ~ 10	5
Eta	0.1 ~ 1	0.5
Nrounds	10 ~ 100	50

식(1)은 예측모델을 나타내고, K는 tree 개수, F는 CART의 모든 regression tree 세트, f는 F공간의 함수를 나타낸다.

$$L(\phi) = \sum_i (y_i, \hat{y}_i) + \sum_k \Omega(f_k),$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (2)$$

식(2)는 예측모델의 최적화를 위해 우변의 첫 번째 항은 학습데이터가 모델에 잘 피팅되었는지를 측정하고, 두 번째 항은 정규화를 통해 모델 복잡도가 단순화되었는지를 측정한다.

$$L^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (3)$$

식(3)은 여러 개의 tree 결과를 반영하여 단계별로 트리의 loss를 계산한다.

$$L^{(t)} \approx \sum_{i=1}^n [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \quad (4)$$

식(4)는 Taylor expansion을 이용하여 2차 다항식으로 단순화하고, 다양한 loss function을 대입하여 t단계의 식을 도출하면 새로운 트리에 대한 학습을 최적화할 수 있다.

$$\bar{L}^{(t)} = \sum_{i=1}^n [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2$$

$$= \sum_{j=1}^T [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \quad (5)$$

식(5)는 정규화를 활용하여 모델의 복잡도를 조정한다.

다. 같은 leaf에 속한 모든 데이터는 같은 score를 가지므로, 합계를 구하는 식으로 변경한다.

$$\bar{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^T \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \quad (6)$$

식(6)에서 tree의 leaf에 최소의 objective value (score)를 계산하고, score 값을 이용하여 모델(트리 구조)을 평가한다. score 값이 작을수록 좋은 트리 구조를 나타낸다.

2.2.3. 하이퍼 파라미터와 분석틀

XGBoost 분석을 수행하려면 모델의 하이퍼 파라미터를 조정해야 한다. grid search로 추출된 하이퍼 파라미터 범위는 Table 2에 제시하였다. 이 논문에서 사용한 머신러닝은 자료 전처리, 계산, 머신러닝 학습 등의 과정을 거쳤으며 이 과정은 R 언어의 package를 사용하여 코드화되어있다.

2.2.4. 머신러닝 학습(Training)

머신러닝은 2016년부터 2019년까지 수집된 기상예측인자를 활용하여 학습하였다. 기상예측인자들은 머신러닝의 XGBoost 기법을 통해 강우량에 대한 가중치가 계산되었고 이렇게 생산된 앙상블 결과를 평균하여 HQPF를 산출하였다(Fig. 2)

2.2.5. HQPF 알고리즘 디자인

예측 강우보정 알고리즘은 관측자료와 수치예보를 사용하여 전처리, 머신러닝, 후처리의 과정을 통해 개발되었다(Fig. 3). 디자인 과정은 다음과 같다. 첫번째 전처리 과정은 수치예보자료에서 머신러닝의 입력자료로 사용되는 기상예측 인자들을 추출하기 위한 과정이다. 이 단계에서는 기상관측자료와 기상예측자료에서 기상예측인자들을 추출하고, 호우영향모델의 격자로 내삽하는 과정이 포함되어있으며 2.2.1.에 설명하였다. 두번째 머신

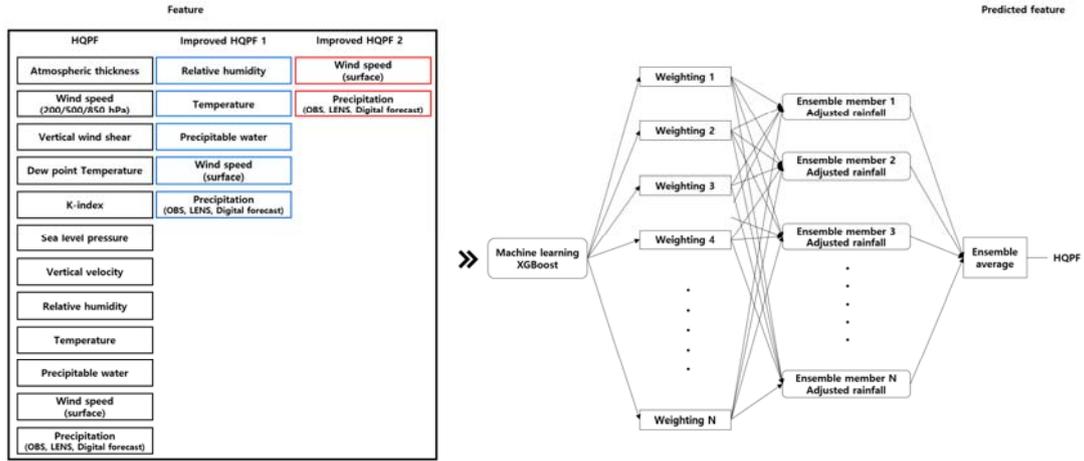


Fig. 2. Mimetic diagram for machine learning training.

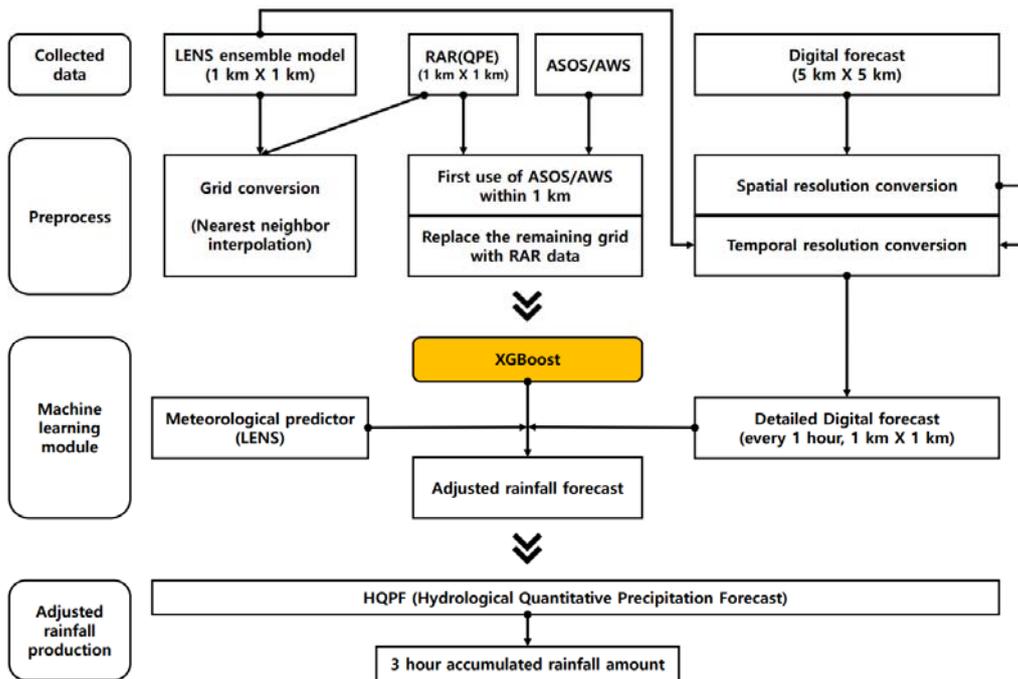


Fig. 3. HQPF calculation process.

러닝 과정은 2016년부터 2019년까지 수집된 기상예측 인자를 활용하여 학습하였다. 기상예측인자들은 머신러닝의 XGBoost 기법을 통해 LENS 예보자료와 동네예

보자료를 입력하여 1시간 간격의 앙상블 멤버별 보정강우를 산출하였다. 마지막으로 후처리 과정은 머신러닝 과정을 통해 산출된 앙상블 멤버별 보정 강우를 평균하여

Table 3. Classification of the pressure patterns causing precipitation over Korean peninsula

Patterns (Symbol)	Description
Low pattern (Lo)	Extra tropical low pressure system passing through the Korean Peninsula
Front pattern (Fr)	Front extended from the low pressure system over northern Manchuria or Aleusian
Col pattern (Co)	Col between two high pressure systems or weak trough from the north or the south
Changma Pattern (Ch)	Stationary Changma front in summer monsoon
Typhoon direct effect (Ty)	Direct effect of typhoon
Typhoon indirect effect (Ti)	Indirect effect of typhoon or direct effect of tropical depression
Typhoon transformed effect (Tt)	Extra tropical cyclones transformed from tropical cyclone after landfall
Convergence of tropical air (Ct)	The strong tropical air convergence. Generally it transport high moist air to Korean peninsula.
Local instability (Li)	Local instability due to strong surface heating. Generally precipitation is shower types
Orography (Or)	Orographic forcing

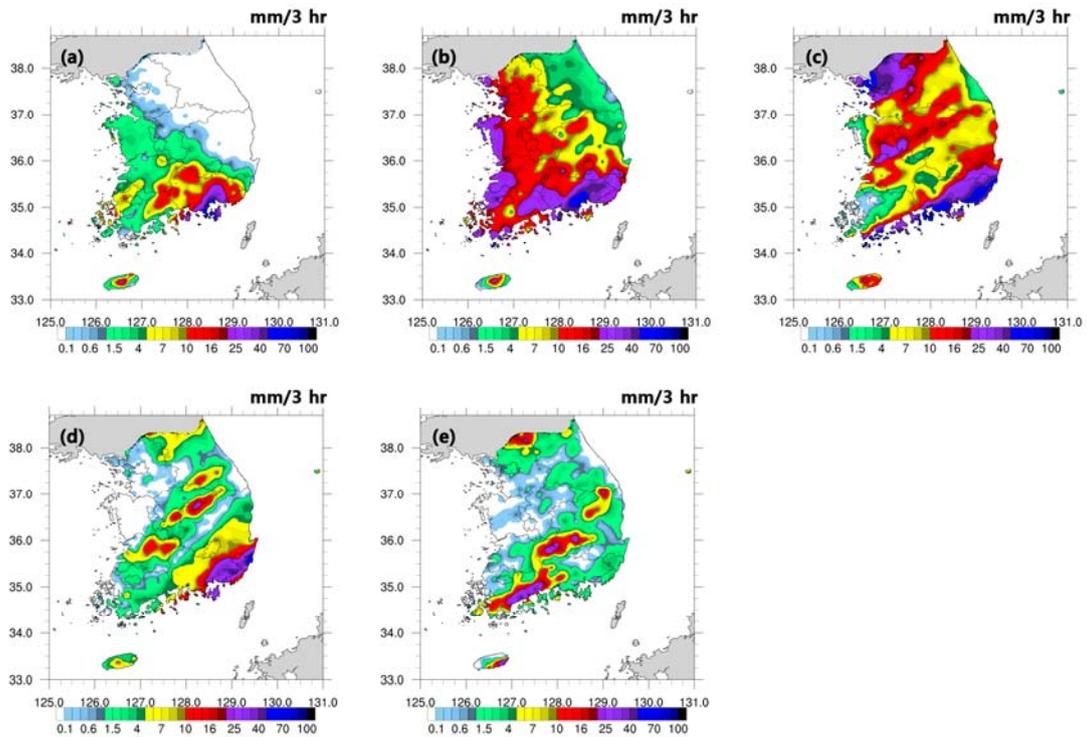


Fig. 4. The horizontal distribution of 3 hr accumulated rainfall at (a) 0600, (b) 0900, (c) 1200, (d) 1500, (e) 1800 LST in August 21, 2021 (Observation).

최종으로 3시간 누적강우의 보정강우(Hydrological Quantitative Precipitation Forecast, HQPF)를 생산하였다.

2.3. 호우 사례 및 연구지역 선정

본 연구에서는 전국 95개 ASOS 지점자료 중 56개 지점에서 일강수량이 30 mm를 초과하고, 12개 지점에서 80 mm를 초과하는 2021년 8월 21일 사례를 선정하였다.

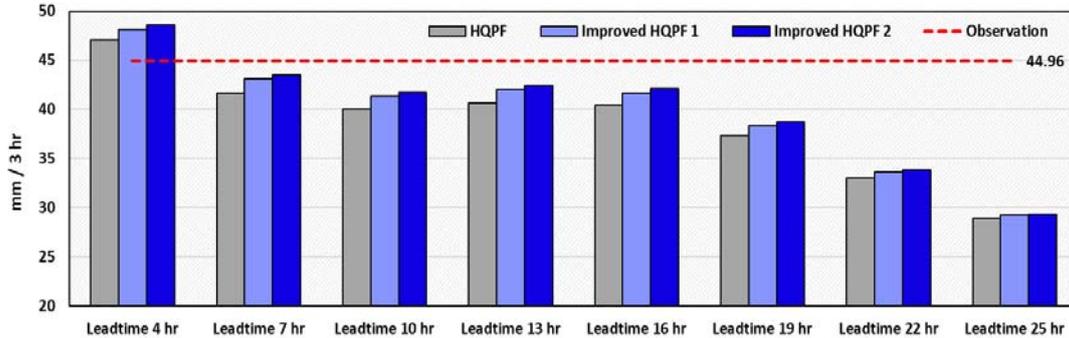


Fig. 5. All region accumulated precipitation.

일반적으로 일강수량이 80 mm 이상의 경우 집중호우로 정의한다. 분석 기간은 8월 21일 06시 부터 8월 21일 18시 까지이며 연구지역은 전국이다. 선정된 호우 사례를 Rha et al.(2005)의 강우특성 분류기준(Table 3)으로 분류했을 때 중부지역은 저기압형, 남부지방은 장마형으로 두 가지 강우특성이 나타났다. 저기압형은 온대저기압이 우리나라를 통과하거나 직접적으로 영향을 미쳐서 내린 강우 유형으로, 저기압의 중심과 연결된 강우 구역이 나타나는 경우이다. 장마형은 정체전선에 의하여 우리나라에서 강수가 발생하는 유형으로, Fig. 4에서는 중부지방은 저기압의 중심이 지나감에 따라 강우가 이동하는 패턴을 보이고 남부지방은 장마전선으로 인한 강한 강우 패턴이 보인다.

2.4. 검증지표 및 방법

본 연구에서는 XGBoost 보정강우량에 대한 정확도를 평가하기 위해 오차검증 방법인 bias와 Root Mean Square Error (RMSE)를 활용하였다. bias는 편향을 나타내는 지표로 예측에 대한 편향이 크면 과소 적합(under-fitting)을 의미한다. RMSE는 관측강우량 대비 보정강우량의 오차정도를 파악할 수 있는 지표이며 bias와 RMSE는 아래와 같은 식을 활용한다.

$$Bias = P_i - O_i \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (P_i - O_i)^2}{n}} \quad (8)$$

여기에서 n 은 자료 개수, P_i 는 보정강우량, O_i 는 관측강우량을 의미한다.

선정된 호우 사례에 대해 XGBoost로 보정된 결과의 검증지표 중 전국의 RMSE 결과를 선행시간별로 제시하고 관측강우량과의 bias 공간장을 선행시간별로 나타내었다. RMSE의 경우 호우에 대한 예측정확도를 확인하기 위하여 시간대별로 관측강우량이 10 mm를 초과한 경우 계산하였다. 이를 통해 선행시간의 증가에 따른 보정강우량의 예측정확도 변화 및 예측강우량과 보정강우량과의 공간패턴의 차이를 확인하였다. 또한, HQPF, Improved HQPF 1, Improved HQPF 2의 RMSE 결과 비교를 통해 최적의 예측인자를 파악하였다.

3. 결과 및 고찰

3.1. 예측인자 변경에 따른 누적 강우량 및 통계적 오차 분석

Fig. 5는 관측 및 XGBoost 보정결과의 3시간 누적강우량에 대한 전국평균 결과이다. 관측기반 전국평균 누적강우량은 점선으로 나타내었으며, 44.96 mm이다. HQPF, Improved HQPF 1, Improved HQPF 2의 XGBoost 보정결과에서는 선행시간이 +4 hr에서 +25 hr로 증가함에 따라 누적강우량이 감소하는 경향을 보이며 +25 hr일 때가 +4 hr일 때 보다 18.08, 18.86, 19.26 mm 감소하였다. 선행시간의 증가에 따른 보정강우량의 감소로 관측값과의 차이가 발생하게 되며, 예측정확도에 영향을 미친다. XGBoost 입력변수 변경에 따른 누적강우량의 차이를 비교했을 때, 변수를 줄인 Improved

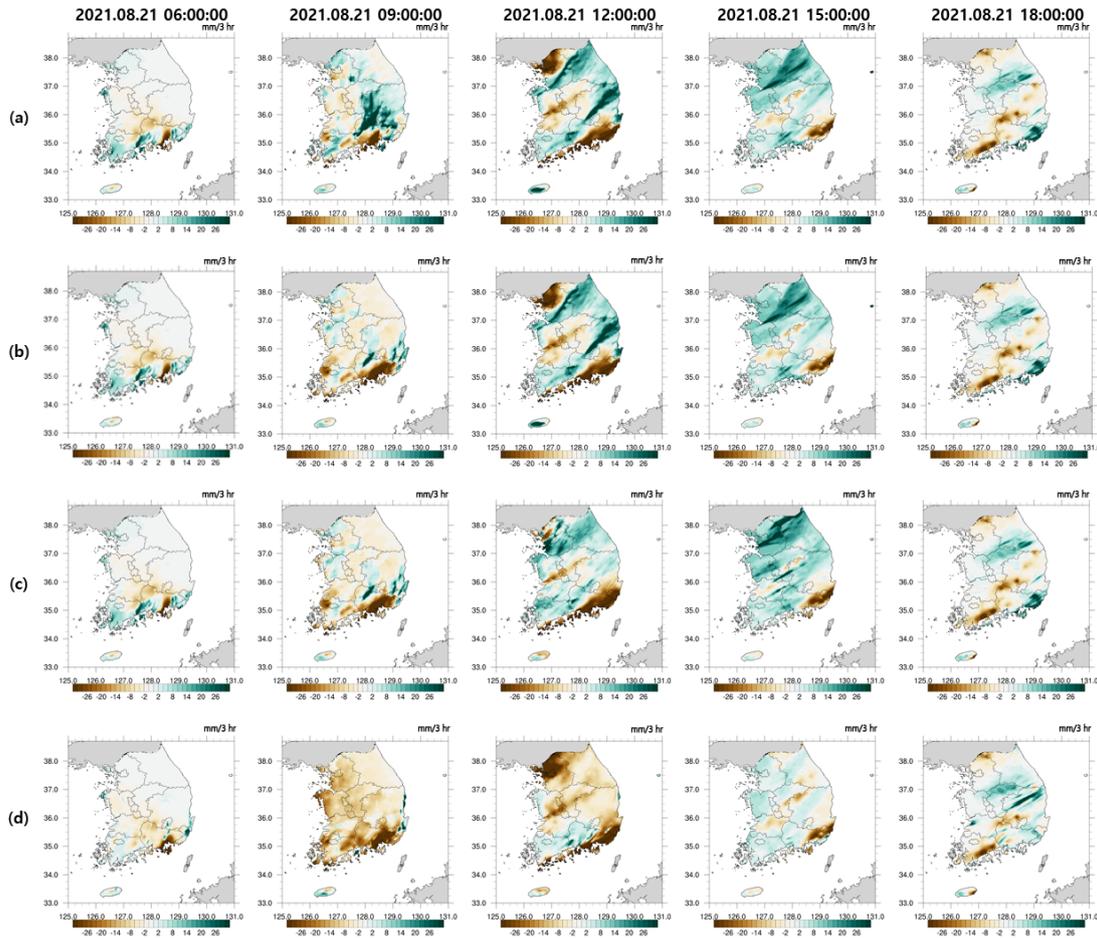


Fig. 6. The horizontal distribution of 3 hr accumulated rainfall bias for improved HQPF 2 and observation (a) lead time 4 hr, (b) lead time 7 hr, (c) lead time 13 hr, (d) lead time 25 hr.

HQPF 2가 더 높은 예측정확도를 나타내었다.

Fig. 6은 Improved HQPF 2의 3시간 누적강우량 예측결과와 관측강우량의 bias 공간 분포도이다. 선행시간 증가에 따라 보정강우량이 감소하여 선행시간 +25 hr의 09시와 12시에는 대부분 지역에서 관측강우량이 보정강우량보다 많은 것을 볼 수 있다.

Table 4와 Fig. 7은 호우 사례의 관측강우량과 보정강우량에 대한 3시간 누적강우량의 오차검증 결과이다. 지점별 관측강우량이 10 mm를 초과하였을 때에 대한 RMSE로 Table 4 및 Fig. 7에서 선행시간의 증가에 따라 RMSE가 증가하며, 이는 Fig. 5와 Fig. 6의 결과에서

선행시간에 따라 보정강우량이 감소는 결과와 동일하다. 또한, Improved HQPF 2의 RMSE가 HQPF와 Improved HQPF 1의 결과보다 작은 것을 볼 수 있다. 이는 3.1.의 결과와 같이 변수를 줄인 Improved HQPF 2의 성능이 향상되었음을 나타낸다.

변수의 수에 따른 예측성능 변화에 대한 원인으로, 머신러닝 학습에 들어가는 많은 예측인자가 호우 예측에 필요한 주요 인자의 가중치 적용에 영향을 미쳐 예측정확도를 하락시킬 수 있다. 불필요한 예측인자를 제외하는 것으로도 예측정확도가 향상될 수 있으며, XGBoost의 입력자료 전처리 및 데이터 처리 등의 소요시간 감소로

Table 4. RMSE for observed rainfall and adjusted rainfall

20210821 - RMSE								
	Lead time 4 hr	Lead time 7 hr	Lead time 10 hr	Lead time 13 hr	Lead time 16 hr	Lead time 19 hr	Lead time 22 hr	Lead time 25 hr
HQPF	19.321	19.619	20.368	20.690	20.796	21.176	21.526	21.781
Improved HQPF 1	18.966	19.343	20.049	20.359	20.398	20.836	21.382	21.671
Improved HQPF 2	18.850	19.220	19.921	20.222	20.268	20.735	21.349	21.659

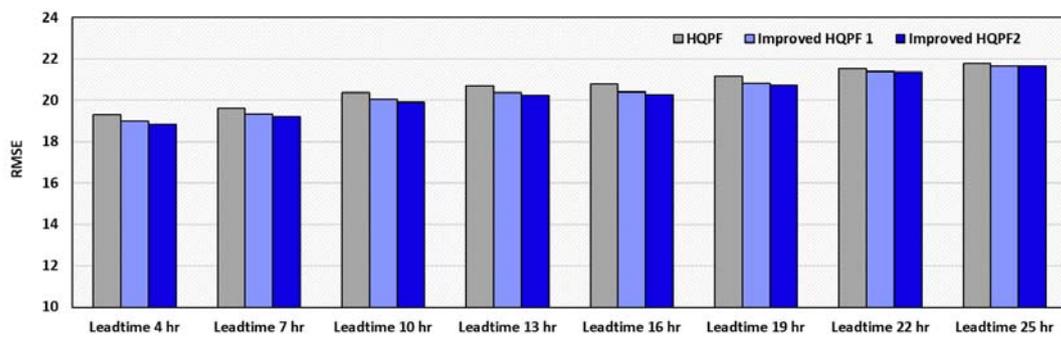


Fig. 7. All region over 10mm precipitation of RMSE.

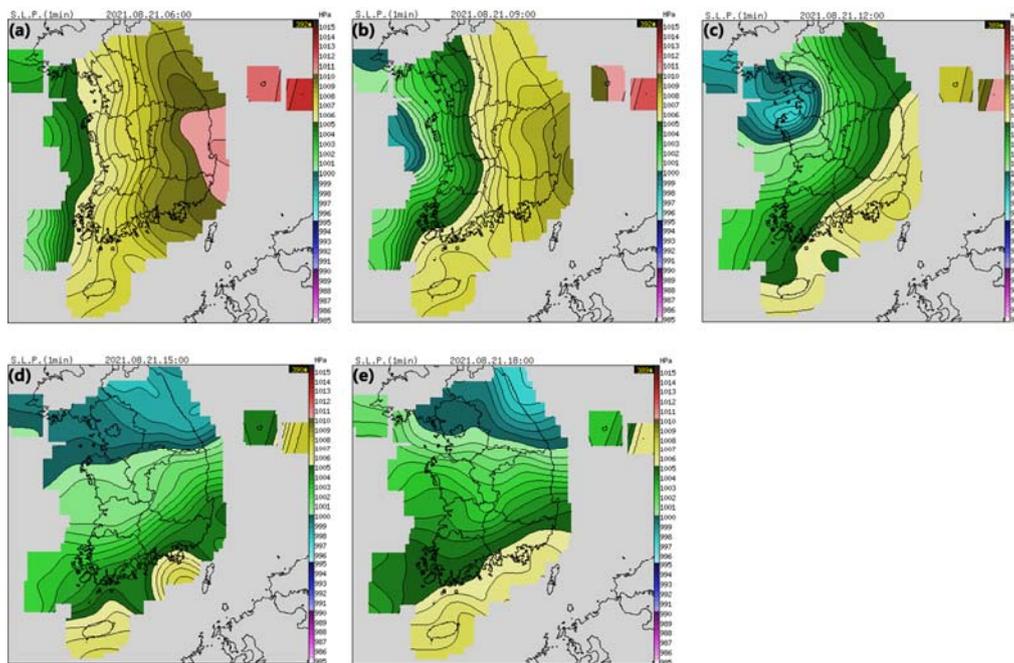


Fig. 8. The horizontal distribution of sea level pressure at (a) 0600, (b) 0900, (c) 1200, (d) 1500, (e) 1800 LST in August 21, 2021 (Observation).

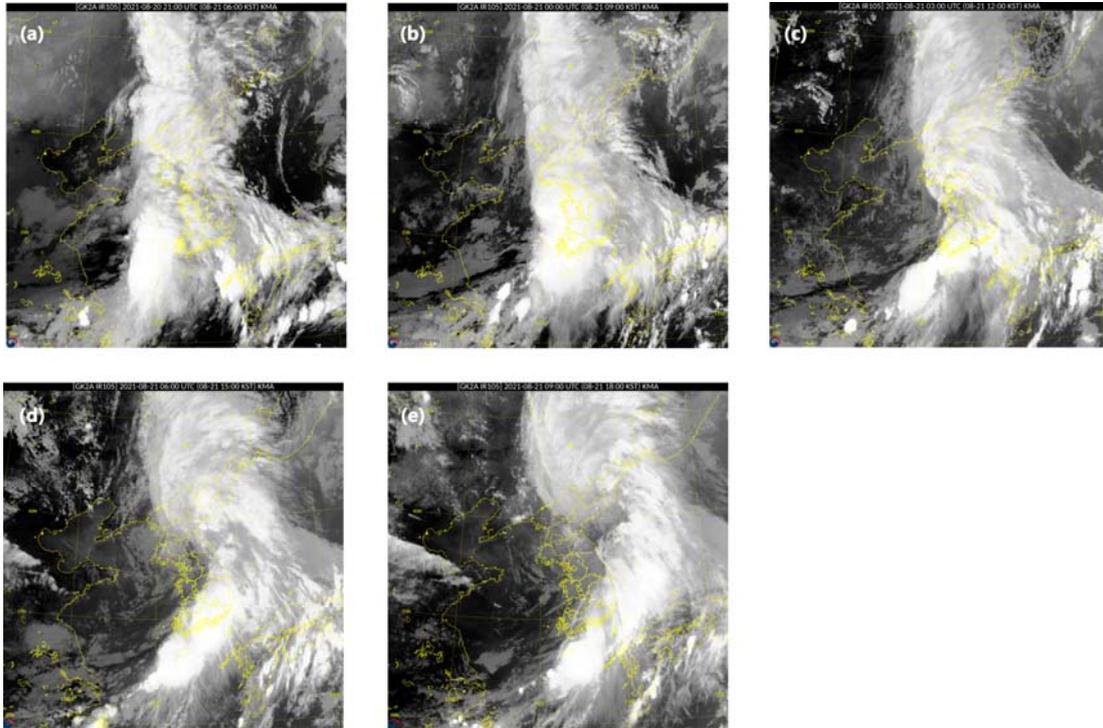


Fig. 9. The horizontal distribution of cloud at (a) 0600, (b) 0900, (c) 1200, (d) 1500, (e) 1800 LST in August 21, 2021 (Observation).

인해 최적화에 도움이 될 수 있을 것으로 판단된다.

3.2. LENS 앙상블별 기상특성 파악

Fig. 8은 8월 21일 호우 기간 중 관측의 해면기압을 공간분포로 나타낸 것이며 Fig. 9는 동일한 시간의 천리안 적외선 영상이다. Fig. 8에서 시간에 따라 저기압이 중부지역을 통과하는 것을 볼 수 있으며, 동시간 대에 호우를 유발하는 구름 영역의 이동을 Fig. 9에서 볼 수 있다. Fig. 10은 8월 21일 오전 6시의 관측 해면기압과 LENS 앙상블별(ENS_01 ~ 13) 해면기압 결과를 나타낸 것이다. ENS_01 및 ENS_13과 같이 관측 해면기압 분포와 유사하게 모의하는 앙상블 모델이 존재하는 것을 볼 수 있다.

현재의 보정강우는 XGBoost 이후 나온 앙상블별 결과를 평균하여 사용하기 때문에, 앙상블 모델별 강우의 공간장이 상이할 경우 값이 평균되어 호우와 같은 극한 값의 예측에서는 예측성능이 떨어질 수 있다.

4. 결론

본 연구에서는 HQPF 생산을 위해 개발한 호우예보 보정기술의 과정 중 머신러닝의 훈련 및 입력자료로 들어가는 변수 조합의 최적화를 통해 HQPF의 예측기술을 고도화하였다. 예측기술 고도화를 통해 수문학적 활용을 위한 강우 및 정확도가 향상된 극한 강우 정보를 생산하는 것에 그 목적이 있다.

기존에는 HQPF 생산을 위해 기상청에서 제공하는 앙상블 수치모델자료, 레이더자료, 지점관측자료 및 동네예보의 강우예보를 활용하였다. 이 자료들은 전처리 과정 및 머신러닝 과정을 통해 앙상블 멤버별 보정 강우로 산출되고 해당 앙상블 자료를 평균하여 최종으로 보정된 강우량을 생산하는 후처리 과정을 거치게 된다. 머신러닝 모델의 구축 및 보정강우 예측의 입력자료로 사용되는 변수는 LENS 앙상블 모델의 변수가 포함되어있는데, 기존의 입력변수를 변경하여 머신러닝 모델을

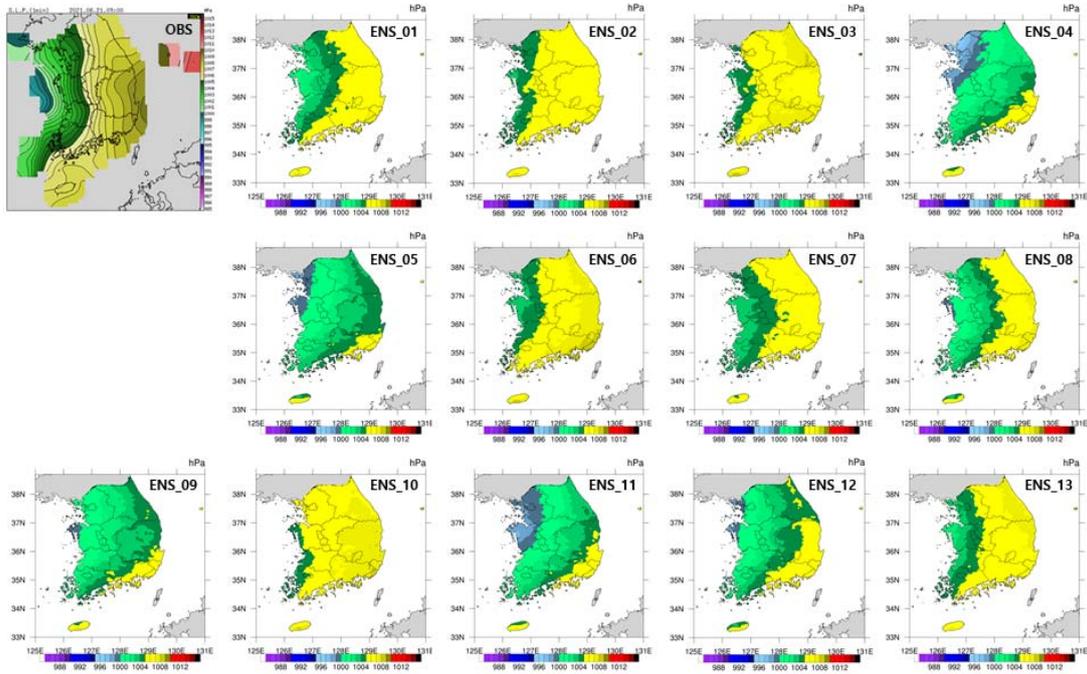


Fig. 10. The horizontal distribution of observation and ensemble sea level pressure at 0600 LST in August 21, 2021.

구축하고 그 결과를 비교하였다. 결과 비교는 예측정확도 검증지표인 bias와 RMSE를 이용하여 진행하였다. 본 연구의 결론을 요약하면 다음과 같다.

첫째, 보정강우량과 관측강우량의 누적강우량을 비교한 결과, 선행시간이 증가할수록 전체 영역에서 보정강우량을 낮게 모의하여 관측강우량과의 차이가 증가하였다. 이는 예측시간이 증가할수록 호우 강도를 예측하지 못한 것으로 판단된다. 변수 변경에 따른 누적강우량의 차이를 보면 기존의 HQPF 보다 Improved HQPF 1, 2가 누적강우량을 더 크게 예측하였다. 또한, 모든 선행시간에서 변수가 가장 적은 Improved HQPF 2의 결과가 HQPF 및 Improved HQPF 1 보다 누적강우량을 더 많이 모의하였다.

둘째, 선행시간별 관측 누적강우량과 보정된 누적강우량의 bias 공간장 결과도 앞선 누적강우량의 비교결과와 동일하게 나타났다. 선행시간이 +4 hr일 때 9시와 12시의 bias 분포도에서는 보정강우량이 관측보다 많은 지

역이 나타났으나 +25 hr일 때 bias 분포도에서는 거의 모든 지역에서 관측강우량이 보정강우량보다 많았다. 변수 변경에 따른 누적강우량의 RMSE를 보면 기존의 HQPF보다 Improved HQPF 1, 2의 RMSE가 더 낮게 나타나 예측성능이 향상되었음을 확인하였다. 이는 호우 강도를 약화시킬 수 있는 변수를 제거함으로써 예측정확도가 향상되었을 것으로 판단된다. 또한, 변수의 감소로 인해 기존 머신러닝 대비 전처리 시간 및 모델 수행시간이 감소하여 모델 최적화에 기여할 수 있을 것으로 판단된다.

셋째, 보정강우량 생산에 입력자료로 들어가는 자료 중 LENS 앙상블 멤버별 기상요소의 모의정도를 비교하였다. 일부 앙상블 멤버는 해당 사례의 저기압 이동을 시간대별로 관측과 유사하게 모의하였다. 저기압형 강우 특성을 가지는 호우 사례에서 앞서 저기압의 이동을 관측과 유사하게 모의하는 앙상블 모델은 다른 앙상블 모델에 비해 강우 강도의 시공간적 정확도가 높을 것으로

판단된다. 현재의 보정강우량 생산 시에는 각 멤버별 보정강우량을 평균한 공간장에 후처리 과정이 적용되고 있다. 모든 앙상블 멤버가 관측과 유사한 결과를 모의한다면 예측성능이 높겠지만, 일부의 앙상블 멤버만 관측과 유사한 결과를 모의한다면 평균 되어 예측성능이 떨어질 것이다. 우리나라의 호우 특성별로 모의를 잘하는 앙상블 멤버를 선정하여 해당 멤버의 결과만 사용하거나 추가적인 기중치를 적용한다면 현재보다 예측정확도가 높아질 것으로 판단된다.

감사의 글

본 연구는 한국기상산업기술원 호우 분야 재해영향모델을 위한 예측강우 생산기술 고도화(KMI2021-00311) 연구사업의 연구비 지원에 의해 수행되었습니다.

REFERENCES

- Chen, T., He, T., 2021, The comprehensive R archive network, <https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf>.
- Friedman, J. H., Hastie, T., Tibshirani, R., 2000, Additive logistic regression: a statistical view of boosting, *Ann. Stat.*, 28(2), 337-374.
- Ghada, W., Eastrella, N., Meanzel, A., 2019, Machine learning approach to classify rain type based on this disdrometers and cloud observations, *Atmosphere*, 10(5), 251-268.
- Hong, W. C., 2008, Rainfall forecasting by technological machine learning models, *AMC*, 200, 41-57.
- Kang, B. S., Lee, B. K., 2011, Application of artificial neural network to improve quantitative precipitation, *J. Korea Water Resour. Assoc.*, 44(2), 97-107.
- Ke, G., Meng, Q., Finely, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. Y., 2017, LightGBM: A highly efficient gradient boosting decision tree, 31st conference on neural information processing systems, Long beach, CA, USA, 3149-3157.
- Ko, C. M., Jeong, Y. Y., Lee, Y. M., Kim, B. S., 2020, The development of a Quantitative Precipitation Forecast correction technique based on machine learning for hydrological applications, *Atmosphere*, 11(1), 111-129.
- Rha, D. K., Kwak, C. H., Suh, M. S., Hong, Y., 2005, Analysis of the characteristics of precipitation over South Korea in terms of the associated synoptic patterns: a 30 years climatology (1973~2002), *The Journal of The Korean Earth Science Society*, 26(7), 732-743.
- Sumi, S. M., Zaman, M. F., Hirose, H., 2012, A Rainfall forecasting method using machine learning models and its application to the Fukuoka city case, *Int. J. Appl. Math. Comput. Sci.*, 22(4), 841 - 854.
- Valipour, M., Sefidkouhi, G., Ali, M., Raeni-Sarjaz, M., Guzman, S. M., 2019, A Hybrid data-driven machine learning technique for evapotranspiration modeling various climates, *Atmosphere*, 10(6), 311-325.
- Zamami J. M., Cao, C., Ni, X., Bashir, B., Talebiesfandarani, S., 2019, PM_{2.5} Prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data, *Atmosphere*, 10(7), 373-391.

-
- Manager. Han-Su Lee
ECOBRAIN Co. Ltd.
lhs@ecobrain.net
 - Principal Research Engineer. Yong-Keun Jee
ECOBRAIN Co. Ltd.
ykJee@ecobrain.net
 - CEO. Yonug-Mi Lee
ECOBRAIN Co. Ltd.
leeym@ecobrain.net
 - Professor. Byung-Sik Kim
Department of Urban & Environmental Disaster Prevention
Engineering, Kangwon National University
hydrokbs@kangwon.ac.kr