J. Korea Water Resour. Assoc. Vol. 54, No. S-1 (2021), pp. 1119-1130
doi: 10.3741/JKWRA.2021.54.S-1.1119

pISSN 1226-6280
eISSN 2287-6138

Real-time flood prediction applying random forest regression model in urban areas

Kim, Hyun Il[a] · Lee, Yeon Su[b] · Kim, Byunghyun[c]*

[a]Researcher, Nakdong River Flood Control Office, Busan, Korea
[b]Undergraduate Research Student, Department of Civil Engineering, Kyungpook National University, Daegu, Korea
[c]Assistant Professor, Department of Civil Engineering, Kyungpook National University, Daegu, Korea

Paper number: 21-086
Received: 30 September 2021; Revised: 1 November 2021; Accepted: 3 November 2021

Abstract

Urban flooding caused by localized heavy rainfall with unstable climate is constantly occurring, but a system that can predict spatial flood information with weather forecast has not been prepared yet. The worst flood situation in urban area can be occurred with difficulties of structural measures such as river levees, discharge capacity of urban sewage, storage basin of storm water, and pump facilities. However, identifying in advance the spatial flood information can have a decisive effect on minimizing flood damage. Therefore, this study presents a methodology that can predict the urban flood map in real-time by using rainfall data of the Korea Meteorological Administration (KMA), the results of two-dimensional flood analysis and random forest (RF) regression model. The Ujeong district in Ulsan metropolitan city, which the flood is frequently occurred, was selected for the study area. The RF regression model predicted the flood map corresponding to the 50 mm, 80 mm, and 110 mm rainfall events with 6-hours duration. And, the predicted results showed 63%, 80%, and 67% goodness of fit compared to the results of two-dimensional flood analysis model. It is judged that the suggested results of this study can be utilized as basic data for evacuation and response to urban flooding that occurs suddenly.

Keywords: Urban flood, Real-time flood prediction, Random forest, Machine learning

랜덤포레스트 회귀모형을 적용한 도시지역에서의 실시간 침수 예측

김현일[a] · 이연수[b] · 김병현[c]*

[a]낙동강홍수통제소 예보통제과 시설연구사, [b]경북대학교 토목공학과 학부연구원, [c]경북대학교 토목공학과 조교수

요 지

불안정한 기후와 함께 나타나는 국지적 집중호우로 인한 도시 침수는 끊임없이 발생하고 있으나, 강우량을 포함한 기상정보 현황 또는 예보정보를 활용하여 공간적인 도시홍수 예측정보를 제공할 수 있는 체계는 아직 마련되지 못한 상황이다. 공간적인 홍수정보는 하천의 제방, 도시 하수관거의 통수능, 저류지, 펌프시설과 같은 구조물적 대책에 어려움이 있을 시 발생할 수 있는 최악의 홍수상황을 미리 파악함으로써 피해를 최소화하는데 직접적인 영향을 미칠 수 있다. 이에 본 연구에서는 기상청에서 제공되는 강수량, 도시 유역에 대한 2차원 침수해석 결과, 그리고 기계학습 모형 중 하나인 랜덤포레스트 회귀모형을 활용하여 실시간으로 도시유역에 대한 침수지도를 예측할 수 있는 방법론을 제시하고자 한다. 연구유역은 내수침수가 빈번하게 발생하는 울산시 우정태화지구로 선정하였다. 지속시간 6시간의 총강우량 50 mm, 80 mm 그리고 110 mm 대한 랜덤포레스트 회귀분석 예측 침수면적과 검보정된 2차원 물리모형의 침수해석 결과 비교시 각각 63%, 80%, 그리고 67%의 적합도를 보여주어, 빠른 시간안에 발생하는 도시 침수에 대한 대응, 대피를 위한 기초자료로 활용될 수 있을 것으로 판단된다.

핵심용어: 도시홍수, 실시간 침수예측, 랜덤포레스트, 기계학습

***Corresponding Author.** Tel: +82-53-950-7819
E-mail: bhkimc@knu.ac.kr (B. Kim)

© 2021 Korea Water Resources Association. All rights reserved.

1. 서 론

도시유역은 인구가 밀집되고 불투수층과 주요시설이 다수 분포되어 있어 국지성 집중호우에 따른 홍수에 취약한 여건을 가지고 있다. 도시홍수는 도시를 인접하여 흐르는 하천의 수위가 상승하여 제방을 월류하거나 붕괴되어 발생하는 외수범람과 외수범람이 없는 상황에서 하수관거 및 저류시설 용량 부족, 저지대로의 유출량 집중 등의 원인으로 발생된다. 특히 내수범람에 의한 도시침수는 저지대를 중심으로 반복적으로 발생하는 특징을 가지고 있으며, 지난 2016년 태풍 '차바'로 인하여 심각한 침수피해를 입은 울산광역시 태화시장은 2021년 12호 태풍 '오마이스'로 인한 내수침수가 또다시 발생하였다. 2020년 발생한 7월 시간당 최대 80 mm를 넘는 집중호우는 만조시기와 겹치면서 부산시 곳곳에 침수를 발생시킨 바가 있으며, 대전광역시에는 시간당 102 mm의 폭우로 KTX선로 및 아파트단지에 침수를 발생시켰다. 2018년 8월 집중호우로 인해 광주광역시, 대전광역시, 경상남도 등의 도시지역에 침수가 발생했으며, 특히 침수피해가 컸던 광주광역시와 대전광역시는 하수관거의 설계기준을 초과하는 집중호우로 인하여 발생한 것으로 확인되었다(Kim, 2020).

예보 또는 실시간 계측 강수량과 수리학적 이론에 근거한 수치해석모형을 이용한 침수예측은 대응시간을 확보함으로써 재산·인명피해 저감을 위한 기초자료가 될 수 있다. 다만 수치해석모형의 경우, 매개변수의 조정, 자료의 수집, 입출력 자료에 대한 전·후처리에 시간이 다소 소요될 수 있는 단점이 있다. 강수량을 포함한 수문자료의 지속적인 축적과 데이터활용 기법의 발달로 인하여 최근에는 기계학습을 이용한 자료기반 홍수예측 연구가 활발히 이뤄지고 있으며, 수치해석모형 대비 비교적 홍수예측 결과를 빠르게 제공할 수 있는 장점이 있다(Mosavi *et al*., 2018). 본 연구에서는 앙상블 분류 및 회귀를 수행하는 랜덤포레스트 회귀모형을 활용한 실시간 도시침수예측을 수행하였으며, 랜덤포레스트 회귀모형은 의사결정나무의 과적합(Overfitting)으로 인한 예측결과의 편중을 보완하기 위하여 개발되었다. Albers *et al*. (2015)은 랜덤포레스트를 이용하여, 대상지역에 대하여 계절별로 큰 영향을 끼치는 하천에 대한 중요도를 산정하였다. 1955-2008년의 일 유량 관측자료를 이용하였으며, 민감도 분석, 유하시간, 중요도 분석을 실시하였다. 제시된 기법을 이용하여 잠재적인 재산피해와 경제적 손실을 저감할 것으로 판단하였다. Li *et al*. (2016)은 저수지의 수위 예측을 위하여 랜덤포레스트를 적용하였는데, 일반적인 인공신경망과 지지벡터회귀(Support vector regression, SVR) 모형의 예측결과치와 비교를 수행하였다. 4일 선행시간으로 예측 시에 가장 높은 정확도를 나타낼 수 있었으며, 수자원 관리 의사결정을 위한 정보를 제공하는데 입력 자료의 시간을 고려하여 랜덤포레스트 회귀모형을 효과적으로 사용하였다. Lee *et al*. (2017)은 랜덤 포레스트와 Boosted tree 를 이용하여 다양한 수문학적 지형자료와 홍수이력자료를 이용하여 서울시에 대한 홍수취약지구를 성공적으로 분석하였다. 랜덤포레스트 회귀 및 분류 모델과 Boosted tree의 회귀 및 분류 모델을 비교하여, 비선형성 자료 처리를 위하여 어떤 모형이 합리적인지 실험적으로 제시하였다.

위 제시된 연구와 함께 Kim and Kim (2020)은 Long-Short Term Memory (LSTM), 랜덤포레스트 분류모형을 통하여 실시간으로 격자 단위의 홍수위험등급을 예측하여 랜덤포레스트 분류모형의 높은 정확도를 확인하였다. 다만 홍수위험등급 예측을 위하여 강수량에 따른 총 월류량을 예측하는 과정에 불확실성이 있으며, 일정 기준의 정확도를 확보하기 위한 학습 및 보정시간이 다소 소요되는 한계가 있었다. 이에 모형 구축과정의 간소화, 학습 소요시간의 단축에도 불구하고 높은 예측력을 확보할 수 있는 방안을 마련하고자 하였으며, 도시유출 해석을 위한 EPA-SWMM, 2차원 침수해석을 위한 FLO-2D를 활용하여 다양한 강수량조건 대비 도시침수지도를 생성하였고 강수량 정보와 도시침수 지도만을 랜덤포레스트 회귀모형에 학습시켜 도시침수예측 모형을 구축하였다.

2. 연구방법

2.1 관측 강우자료의 조사

울산시 우정태화지구에 대하여 침수를 일으킨 2016년 10월 05일 등을 포함하여 2001~2020 기간 동안의 AWS, 레이더, 그리고 ASOS 관측자료를 조사하였다. 강우 지속시간은 6시간으로 선정하였으며, 기상청(Korea Meteorological Administration, KMA) 웹사이트를 통하여 조사하였다. 울산 우정태화지구 인근에서 관측된 자료를 사용함으로써 다양한 강우 통계학적 특성에 따른 도시홍수의 양상을 검토하고자 하였다. Table 1에서는 관측소별 관측날짜와 지속시간 6시간의 총 강우량이 나타나 있다. 동일 날짜에 관측된 자료는 같은 시간대에서 관측된 강우 정보를 수집하였다. 총 79개의 관측강우 사상을 수집하였다.

2.2 1차원 및 2차원 홍수분석

앞서 조사한 관측 강우에 따른 도시유역의 월류량을 EPA-SWMM (EPA's Storm Water Management Model, 5.1 version)

Table 1. Observed rainfall data

Rainfall Stations	Observed Date	Total Rainfall (mm)
Ulsan (152) ASOS	2016-10-05	233.8
Ganjeolgot (924) AWS	2014-08-25	150.0
	2005-09-06	135.0
	2020-07-10	127.0
	2009-07-07	118.0
	2003-06-19	110.0
	2009-07-16	99.5
	2012-09-16	92.5
	2016-09-17	97.5
	2004-09-07	93.5
	2019-08-06	92.5
	2005-08-24	88.0
	2017-09-11	87.5
	2013-10-08	85.5

을 통하여 산출하였다. SWMM 모형의 실행블록 중 Runoff와 Extran을 활용하였으며, Runoff 블록은 강우사상 기반의 초기연산이 수행되는 블록으로써 배수유역 내의 유출현상, 맨홀에 대한 유입수문곡선을 계산한다. Extran블록의 경우 Runoff 블록의 출력자료를 활용하여 배수관망 시스템의 유량과 수심을 구하기 위해 사용되며, 계산 시간별 배수관망 내의 유량 및 수위를 통해 관내의 역류현상, 월류량 해석이 가능하다(Park *et al.*, 2019). 연구 대상지역에서 다양한 강우 사상에 따른 월류량 해석을 위하여 Eqs. (1) and (2)에 나타난 Saint-Venant의 연속방정식과 운동량 방정식을 활용하였다. Q는 유출량(m^3/sec), W는 소유역의 폭(m), n은 Manning의 조도계수, d는 수심(m), d_p는 지면저류 손실수심(m), S는 소유역 경사를 의미하며, A는 소유역의 지표흐름 단면적(m^2), V는 지표흐름 속도(m/sec)를 나타낸다.

$$Q = W \times \frac{1}{n}(d-d_p)^{\frac{5}{3}} S^{\frac{1}{2}} \tag{1}$$

$$\frac{\partial Q}{\partial t} + gAS_f - 2V\frac{\partial A}{\partial t} - V^2\frac{\partial A}{\partial x} + gA\frac{\partial H}{\partial x} = 0 \tag{2}$$

침수 범위에 대한 분석을 위한 2차원 홍수해석은 유한차분기법을 기반으로 하는 수리해석 프로그램을 이용하였다. 적용한 FLO-2D 모형은 FEMA (Federal Emergency Management Agency, 미국 연방재난관리청)가 승인한 하천흐름과 홍수해석을 위한 수리모형으로 유한차분법을 이용한다(Risi *et al.*, 2015). 2차원 수리학적 침수해석의 결과를 이용하여 각 강우조건에 따른 최대 침수 지도를 생성하였으며, 결과는 로지스틱 회귀의 종속 변수로 사용된다. 홍수파 해석을 위한 2차원 격자의 크기는 5 m 정방형으로 하였다.

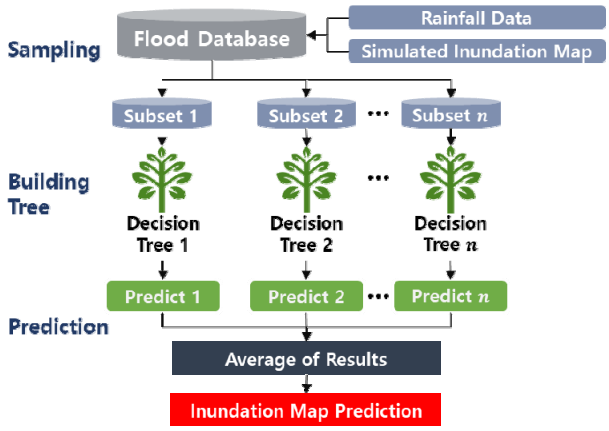


Fig. 1. Structure of random forest regression model

2.3 랜덤 포레스트

랜덤포레스트는 특정 사건 발생에 대한 분류 및 회귀를 수행하는데 다수의 의사결정나무를 구성하는 앙상블 학습을 사용하는 모형이다(Fig. 1). Zhou *et al.* (2018)이 시도한 서로 다른 종류의 인공신경망 간의 앙상블 학습을 이용하여 원하는 수문 자료를 예측할 수도 있지만, 본 연구에서 적용하는 랜덤포레스트는 랜덤 샘플링을 통하여 서로 다른 자료를 학습한 다양한 의사결정나무들의 결과를 종합하는 모형이다. Tyralis *et al.* (2019)에 따르면 랜덤 포레스트는 단순하지만 자연현상에 대한 해석을 실시하는데 높은 예측력을 나타내고 있다. 학습과정에서 다양한 의사결정나무를 구성하고 최종적인 평균값을 출력하는 방식으로 작동되며, 주요 매개변수는 max_features, bootstrap의 사용 여부, 그리고 n_estimator으로 볼 수 있다. max_features 매개변수는 각 노드에서 사용할 특성의 최대 개수를 결정한다. Bootstrap은 각각의 분류 모델을 위한 학습자료 샘플링 시 중복을 허용하는가에 대한 옵션이다. n_estimator의 경우 랜덤포레스트에서 만들 의사결정나무 개수를 결정하며, 랜덤 포레스트의 알고리즘은 다음 네 단계로 요약될 수 있다.

1) n개의 임의의 bootstrap 샘플을 추출한다.
2) bootstrap 샘플에서 의사결정 나무를 학습하는데, 각 노드에서는 중복을 허용하지 않고 랜덤하게 d개의 특성을 선택한다. 정보 이득과 같은 목적 함수를 기준으로 최선의 분할을 만드는 특성을 사용해서 노드를 분할한다.

3) 단계 1), 2)를 k번 반복한다.
4) 각 의사결정나무 회귀결과를 평균하여 최종적인 결과를 도출한다.

가장 정보가 풍부한 특성으로 노드를 나누기 위해 최적화할 목적 함수를 정의한다. 랜덤포레스트에서 사용될 수 있는 목적 함수는 각 분할에서 정보 이득을 최대화한다. 정보 이득은 Eq. (3)와 같이 정의할 수 있다.

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \tag{3}$$

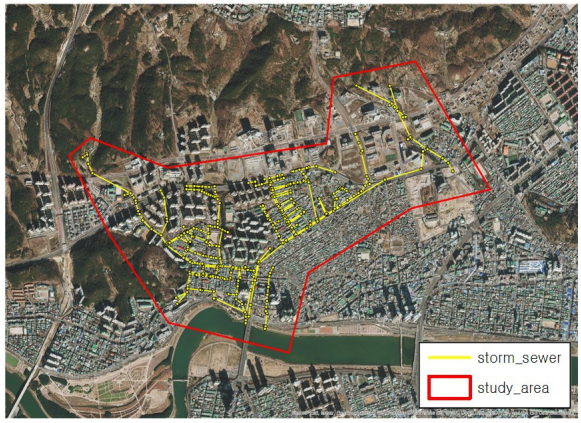
여기서 f는 분할에 사용할 특성이며, D_p와 D_j는 부모와 j번째 자식 노드의 데이터 세트이다. I는 불순도(impurity) 지표이다. N_p는 부모 노드에 있는 전체 샘플 개수이며, N_j는 j번째 자식 노드에 있는 샘플 개수를 나타낸다. 정보 이득은 단순히 부모 노드의 불순도와 자식 노드의 불순도 합의 차이이다. 자식 노드의 불순도가 낮을수록 정보이득이 더 커지게 된다.

3. 연구유역

3.1 울산광역시 우정태화지구

랜덤포레스트 회귀모델을 적용하기 위해 2016년 태풍 '차바' 때의 집중호우로 인해 침수피해를 입은 울산광역시 우정태화지구를 연구유역으로 선정하였다(Fig. 2). 연구유역은 태풍 '차바'로 인해 울산 태화강 둔치, 태화시장 일대 등이 침수되었으며, 2021년 8월 집중호우 및 두 차례의 태풍으로 인한 침수가 발생한 이력이 있으며, 태화강 인접지역, 상대적인 저지대로 인한 매년 침수피해 위험성이 있다. 2016년 태풍 '차바'로 인한 피해상황을 Fig. 3에 나타내었다.

해당 연구유역 인근의 방재기상관측소(AWS)는 매곡(943),



(a) Study area and storm sewer

(b) Rainfall Station

Fig. 2. (a) Study area and (b) Rainfall station

(a)

(b)

Fig. 3. Photo of damaged area (a) Riverside of Taehwa river, (b) Taehwa market (Source: (a) Chosunilbo, (b) Ulsanilbo)

정자(949), 장생포(898), 을기(901), 온산(954), 간절곶(924) 등이 있는데, 장생포(898) 관측소는 관측개시일이 2019년 12월 24일이고, 온산(954) 관측소는 관측개시일이 2008년 6월 27일로 비교적 최근이기 때문에 강우자료가 충분하지 않았고, 을기(901) 관측소는 폭이 넓은 하천의 하류부 건너에 있는 관측소이기 때문에 강우량에 영향을 준다고 판단하였다. 매곡(943)과 정자(949) 관측소는 간절곶(924) 관측소에 비해 상대적으로 거리가 먼 곳에 위치하므로 간절곶(924) 관측소에서 강우자료를 수집하기로 결정하였다. 연구유역 근처의 종관기상관측소(ASOS)는 울산(152) 관측소 한 곳이 존재했고, 연구유역과 멀지 않은 곳에 위치하여 강우자료 수집에 활용하였다.

4. 모형의 적용

4.1 모형의 검증

2016년 연구유역에 큰 피해를 줬던 태풍 차바를 적용하여 1차원 및 2차원 연계 도시홍수해석 모형인 EPA-SWMM과 FLO-2D 모형을 검증하였다. 2016년 10월 5일 태풍 차바가 연구유역에 내린 강우량은 Fig. 4에 나타내었으며, 총 강우량은 233.8 mm이다. 1차원 및 2차원 연계모형의 검증을 위해, 이 강우량을 SWMM에 적용하고 배수관망 해석을 수행하여 맨홀에서의 월류량을 산정하였다. 산정한 맨홀 월류량은 Fig. 5에 나타내었다. 또한, 맨홀에서의 월류량을 다시 FLO-2D에 입력하여 연구유역에 대한 2차원 홍수해석을 수행하였다. 그리고 나서 FLO-2D 모의 결과를 당시에 주민 신고 및 현장 조사된 National Disaster Management System (NDMS) 자료와 비교하였다. 비교결과는 Fig. 6에 나타내었다. Fig. 6에서 보여주듯이, 큰 피해가 발생하였던 태화시장 일대에서는 NDMS와 모의 결과가 비교적 잘 일치하나, 피해가 크지 않았던 일부 지역에서는 1-2차원 연계모형의 결과와 NDMS 자료간 차이가 발생하는데, 이는 주민신고가 적극적으로 이루어지지 않는 등 NDMS 자료가 실제 침수양상을 재현하지 못하는 것도 하나의 원인일 수 있다.

4.2 랜덤포레스트 예측모형의 입력자료

2차원 도시홍수 예측을 위한 머신러닝 적용을 위해 먼저, 강우자료를 수집하였다. 강우자료는 앞서 설명하였던 연구유역인 울산시 우정태화지구 인근에 위치한 간절곶(924) AWS 관측소와 울산(152) ASOS 관측소 두 곳의 자료를 활용하였다. 이 두 관측소에서 2001년에서 2020년 사이의 강우자료를 수집하였다. 강우자료의 시간은 모형의 혼동을 방지하기 위해서 6시간 강우에 대한 자료를 수집하였다. 그 방법으로는 각 강우에 대한 시간으로부터 6시간 동안의 값을 모두 합하여 시간별 6시간 강우를 산정한 다음 내림차순으로 정렬하면, 지속기간 6시간의 강우를 총강우량이 많은 순서대로 얻을 수 있다. 이를 이용해 6시간 동안 많은 비가 내린 날짜를 조사하고, 다시 해당 날짜에 대한 분 자료를 수집하여 10개 단위로 묶으면 10분 간격의 강우를 얻을 수 있다. 이러한 방법으로 총강우량 233.8 mm에서 30 mm 사이의 지속기간 6시간, 10분 간격의 강우 79개를 수집하였다. 수집한 79개의 강우 중에서 76개는 학습에 사용하고, 나머지 3개는(총강우량 50 mm, 80 mm, 그리고 110 mm) 예측을 위해 사용하였다. Fig. 7은 학습에 사용한 76개의 강우자료 중에서 태풍 차바(238 mm)를 제외하고 가장 강우량이 많은 4개의 사상에 대해 앞서 설명한 방법으로 정리한 예를 보여준다.

NDMS 자료와 비교를 통해 검증된 1-2차원 연계모형 즉, SWMM과 FLO-2D 모형에 수집한 79개의 강우를 적용하여

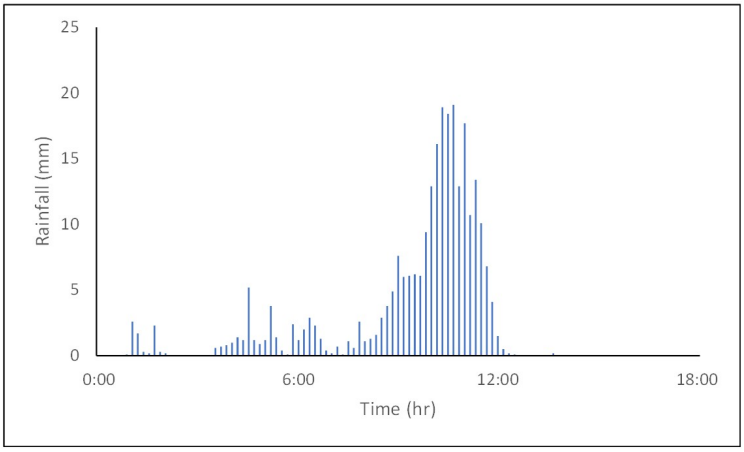
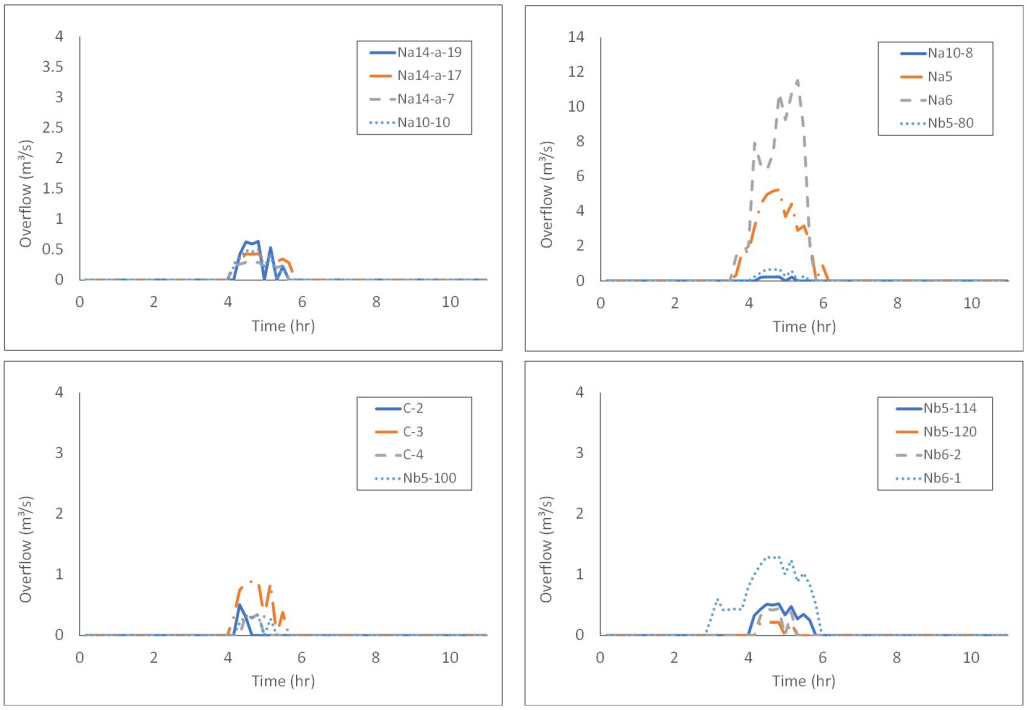
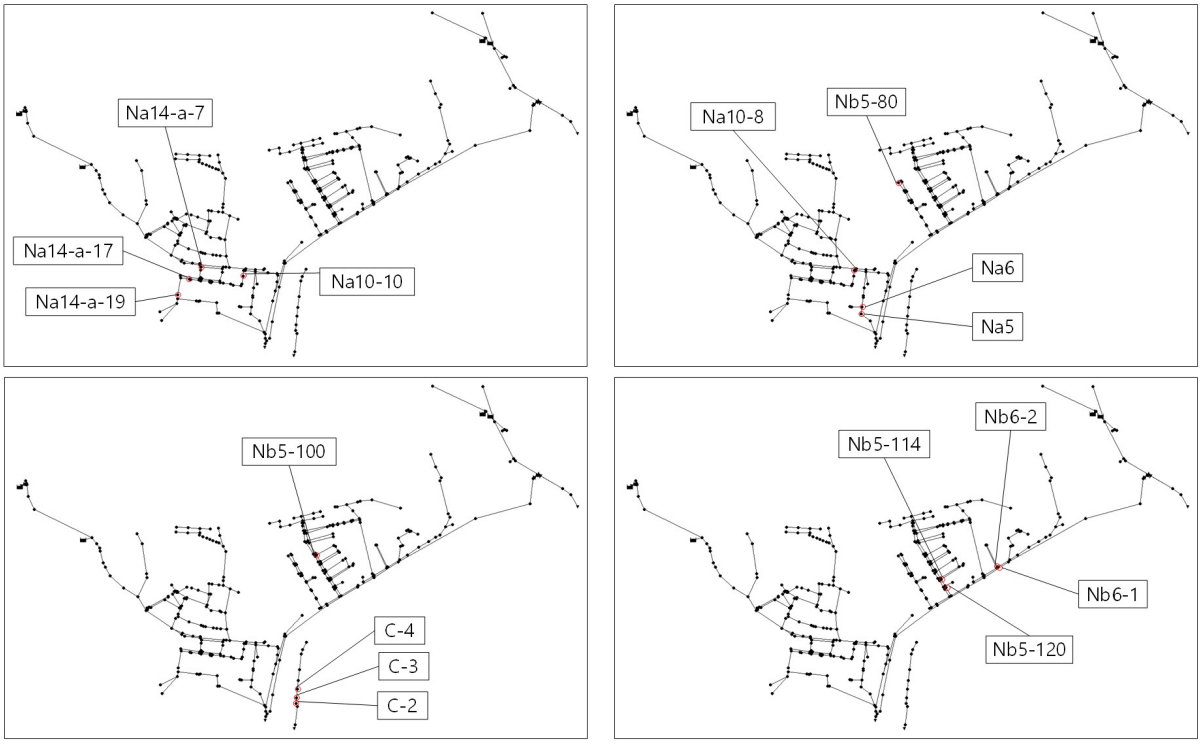


Fig. 4. Rainfall in typhoon Chaba on october 5, 2016



(a) Overflow at Manhole



(b) Location of Manhole

Fig. 5. Overflow at manhole and location of manhole

도시침수해석을 수행하였다. 학습을 위해 사용된 76개의 강우에 대한 FLO-2D 모의결과 침수심은 랜덤포레스트 입력자료의 최대침수심, 평균침수심, 침수횟수, 목표침수심에 사용하고, 예측을 위한 3개의 강우에 대한 FLO-2D 모의결과 침수

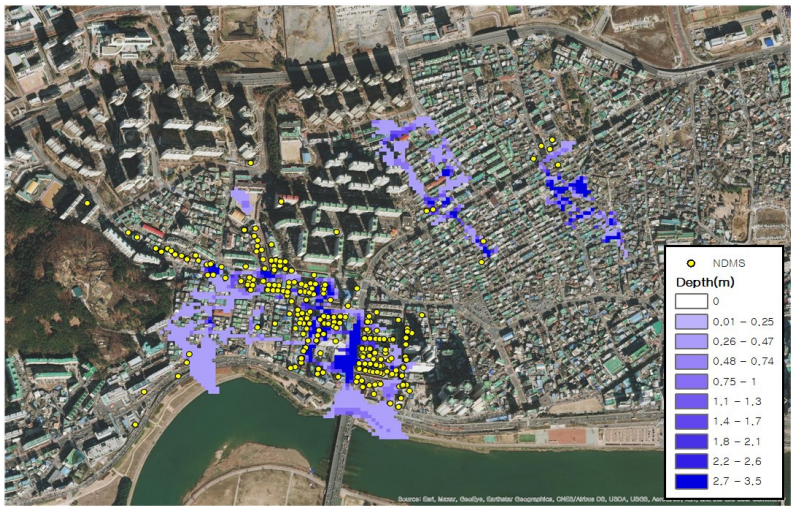


Fig. 6. Comparison of FLO-2D result with NDMS

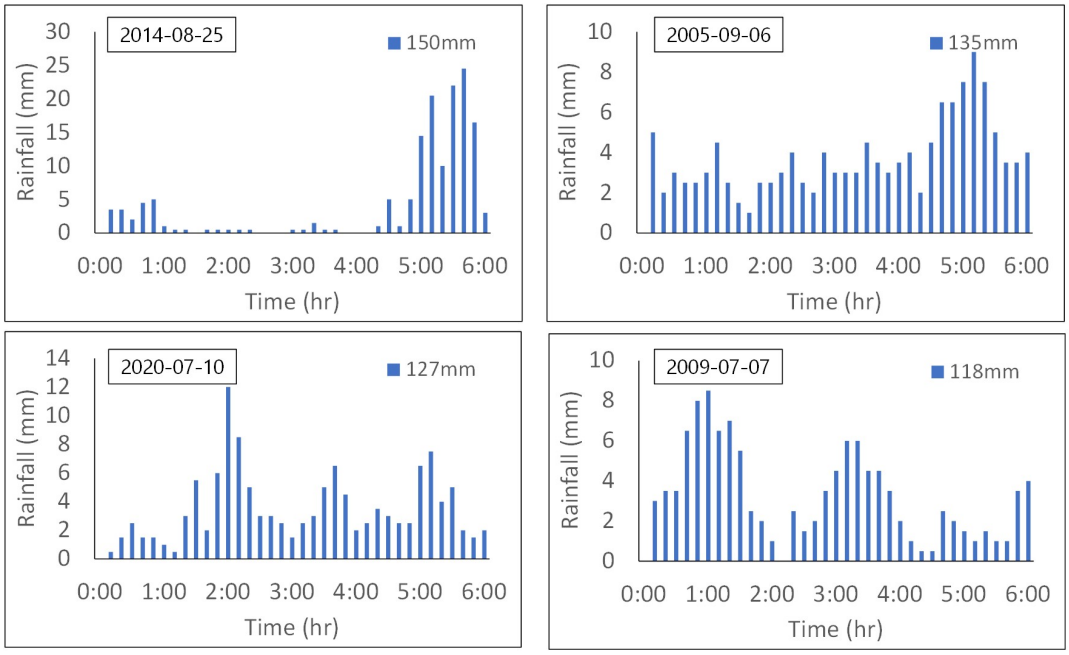


Fig. 7. Four representative rainfall events used for training

심은 랜덤포레스트로 예측한 침수심과 비교하는 데 사용하였다. 학습자료 76개의 FLO-2D 모의결과에 대한 침수심을 격자별로 최대값, 평균값, 침수발생횟수를 산정하여 격자 기본정보로 사용하였다. 격자의 X, Y좌표와 표고, 최대침수심, 평균침수심, 침수횟수가 격자 기본정보를 구성하고, 격자 기본정보와 총강우량을 학습자료로, 2차원 해석을 통해 얻은 침수심을 목표자료로 사용하였다 Table 3은 랜덤포레스트를 이용한 학습을 위한 자료의 구성 일부를 보여준다. 격자수(41,908개) × 강우자료 수(76개)로 총 3,185,008개의 입력자료로 모델을 학습하였다. 학습자료와 동일한 격자 기본정보와 예측강우의 총강우량을 사용하여 예측자료를 생성하였다. Table 2는 랜덤포레스트를 이용하여 침수심을 예측하기 위해 구성한 자료의 일부를 보여준다.

4.3 모형의 학습

Table 2에서 제시된 바와 같이 훈련자료(Training data)와 목표값(Target data)로 이뤄진 자료를 활용하여 랜덤포레스트를 학습하였으며, Table 3에서 제시된 자료를 활용하여 학습

모형에 대한 검증을 수행하였다. 본 연구에서는 10개의 의사결정나무가 하나의 랜덤포레스트를 구성하는데, 그 중 하나의 의사결정나무를 Fig. 8에 나타내었다. 각 단계에서 X[i], mse, samples, value 총 4개의 인자로 구성되어 있으며, X[i]는 현재 노드를 두 개의 노드로 분할할 때 사용한 분할 기준값이다. i값은 1에서 7까지로, i값이 의미하는 자료를 Table 4에 나타내었다. mse는 평균제곱오차를 의미하고, 노드 분할 시에 평균제곱오차가 최소가 되는 분할 기준을 자동으로 선정하게 된다. 노드에서 평균제곱오차가 감소하였다는 것은 노드의 순도가 올라갔다는 것을 의미하며, samples는 해당 노드에 포함된 자료의 개수를 의미한다. 노드 분할 시에 이 자료들을 분할하는 것이므로, 분할된 두 노드의 samples 값을 더하면 분할되기 전의 samples 값과 일치한다. value는 목표값을 의미하고, 해당

Table 2. Random forest training data

Training data							Target
Grid basic information						Total rainfall	Target depth
X	Y	Elev.	Max. depth	Avg. depth	Num. of flooding		
227923	328732	6.47	3.52	0.54	18	233.8	3.43
227933	328732	6.82	3.29	0.48	17	233.8	3.08
228643	329132	9.22	3.17	0.53	38	233.8	3.17
⋮							
227923	328732	6.47	3.52	0.54	18	99.5	2.69
227933	328732	6.82	3.29	0.48	17	99.5	2.47
228643	329132	9.22	3.17	0.53	38	99.5	1.49
⋮							
227923	328732	6.47	3.52	0.54	18	71	0
227933	328732	6.82	3.29	0.48	17	71	0
228643	329132	9.22	3.17	0.53	38	71	0.09
⋮							

Table 3. Random forest prediction data

Prediction data						
Grid basic information						Total rainfall
X	Y	Elev.	Max. depth	Avg. depth	Num. of flooding	
227923	328732	6.47	3.52	0.54	18	50
227933	328732	6.82	3.29	0.48	17	50
228643	329132	9.22	3.17	0.53	38	50
228633	329132	9.12	3.08	0.57	38	50
228633	329142	9.12	3.05	0.57	38	50
228643	329142	9.36	2.89	0.45	35	50
227923	328722	7.11	2.87	0.39	18	50
227923	328742	7.14	2.85	0.38	16	50
227933	328722	7.18	2.81	0.37	16	50
227803	328882	7.55	2.71	0.11	9	50
⋮						

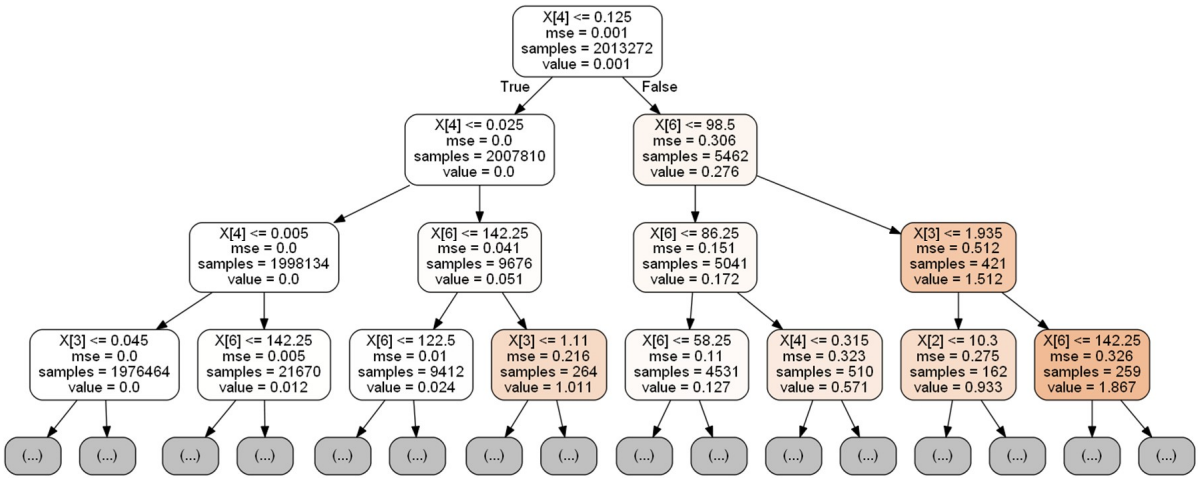


Fig. 8. Random forest tree

Table 4. Split citeria for i.

i	1	2	3	4	5	6	7
Split Criteria	x coord.	y coord.	Elev.	Max. depth	Avg. depth	Num. of flooding	Total rainfall

노드에 포함된 자료들의 평균 목표침수심이다. 예측 단계에서 학습시에 생성된 의사결정나무의 노드를 따라서 끝 노드로 도달하게 되면, 그 끝 노드의 value값으로 침수심을 예측하게 된다. 각 의사결정나무에서 예측된 값을 평균하여 최종적인 예측 침수심을 좌표와 함께 제시한다. Tables 2 and 3 자료를 활용하여 학습을 마친 랜덤포레스트의 실질적인 예측력을 확인하기 위해 활용된 강우사상은 Table 5와 같다.

4.4 침수예측 평가

랜덤포레스트 회귀모형을 통하여 총강우량 50 mm, 80 mm, 110 mm 실측강우에 대해 침수예측을 실시하였다. 50 mm 강우는 비교적 한쪽으로 치우친 분포를 나타내고, 80 mm 와 110 mm 강우는 비교적 균등한 분포를 나타낸다. 침수심 예측을 위해 적용한 세 개의 강우는 Fig. 9에 나타내었다.

예측을 위해 사용된 강우에 대한 FLO-2D 침수해석 결과와 랜덤포레스트 회귀분석으로 예측한 결과를 비교하여, Fig. 10에 나타내었다. 여기서, FLO-2D 모의를 위해서는 6시간 강우 모두를 사용하지만, 학습된 랜덤포레스트 모형에는 총 강우량만을 입력하여 침수지도를 예측하였다. 비교적 한쪽으로 치우친 분포를 나타내는 50 mm 강우에 대해서는 과소예측을, 비교적 균등한 분포를 나타내는 80 mm 와 110 mm 강우에 대해서는 과대예측을 수행하였다. 예측을 위한 입력자료에서 격자 기본정보를 제외하고 총 강우량만을 적용하였기에, 입력 강우자료의 분포특성이 예측결과의 과대 및 과소예측 여부에 적지 않은 영향을 주었음을 확인할 수 있었다.

2차원 수리해석 결과와 랜덤포레스트 회귀분석 예측결과에 대한 오차분석을 실시하였다. 면적에 대한 적합도, 침수심에 대한 평균제곱근오차(Root mean square error, RMSE), 그리고 상관계수를 각각 산정하였으며, 면적에 대한 적합도는 Eq. (4)을 적용하여 계산하였다.

$$goodness\ \ of\ \ fit(\%) = \frac{g_b}{g_f + g_p - g_b} \times 100 \tag{4}$$

g_b는 FLO-2D와 랜덤포레스트 모두에서 침수가 일어난 격자의 수, g_f는 FLO-2D에서만 침수가 일어난 격자의 수, g_p는 랜덤포레스트 예측에서만 침수가 일어난 격자의 수를 의미한다.

50 mm 강우에 대한 면적 적합도는 63%, 80 mm 강우에 대

Table 5. Rainfall information for flood prediction

Total rainfall	Rain station	Date
50 mm	Ganjeolgot (924) AWS	2002/08/06
80 mm	Ganjeolgot (924) AWS	2004/08/22
110 mm	Ganjeolgot (924) AWS	2003/06/19

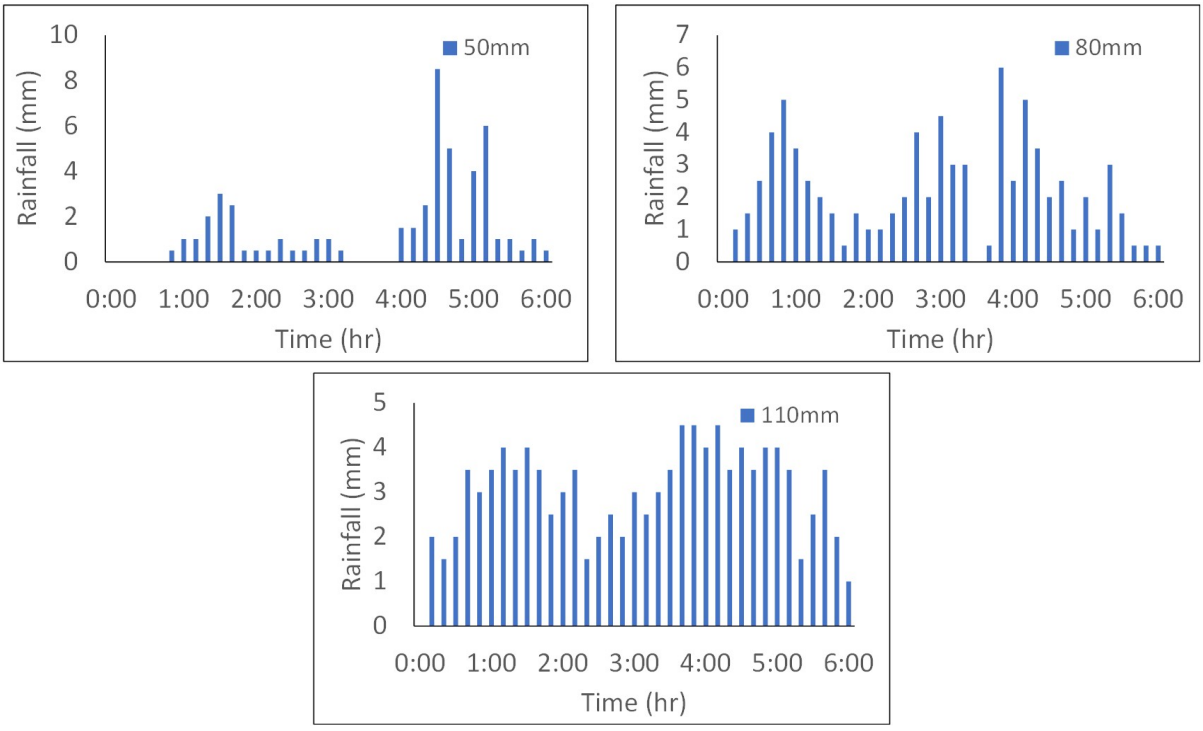


Fig. 9. Rainfall used to predict flooding

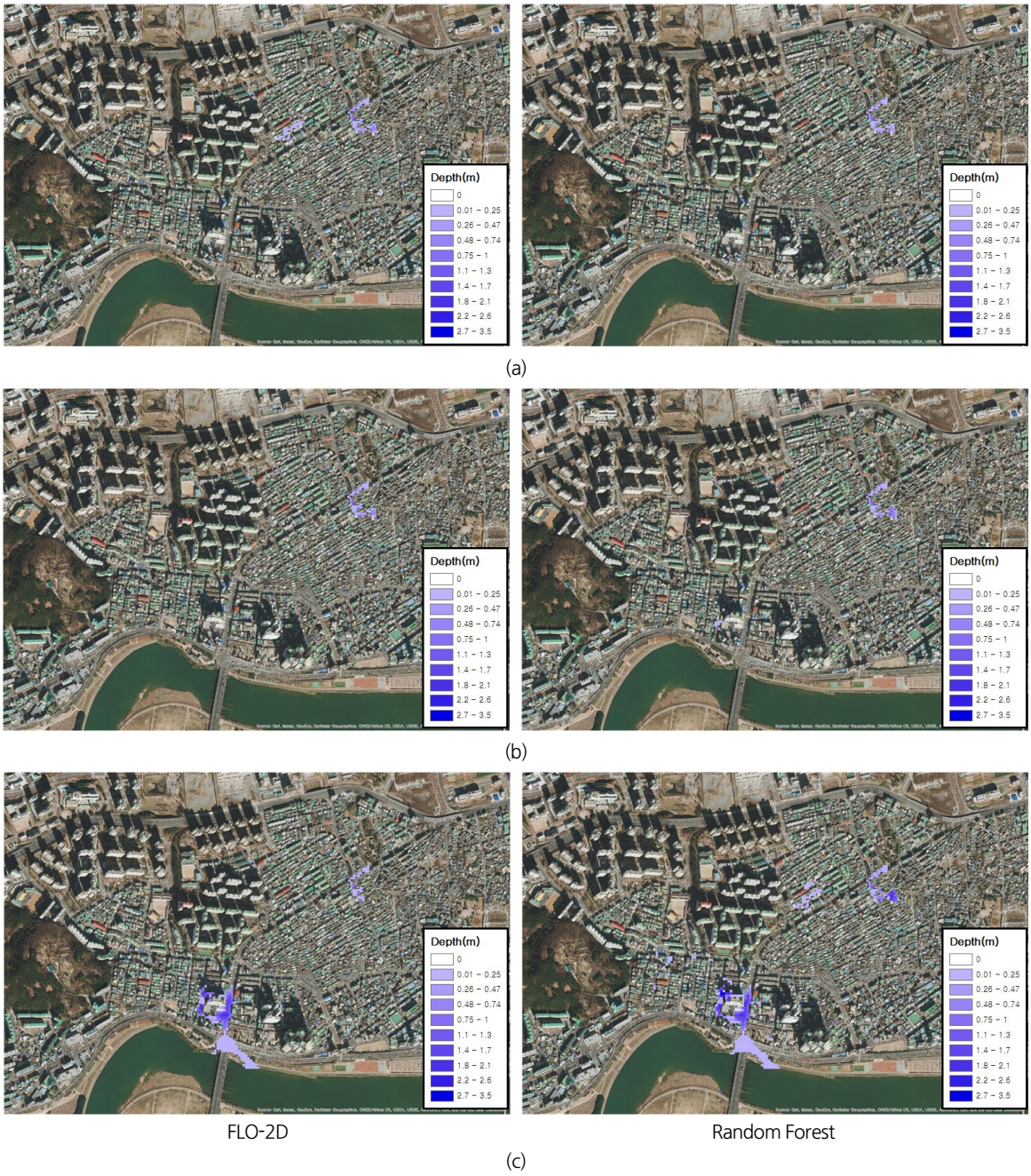


Fig. 10. Comparison of flood depth between Flo-2D and random forest regression analysis; (a) 50 mm, (b) 80 mm and (c)110 mm

한 면적 적합도는 79%, 그리고 110 mm 강우에 대한 면적 적합도는 67%로 나타났다. 침수심에 대한 평균제곱근오차를 구하는 식을 Eq. (5)에 나타내었다. 여기서 n은 총 격자의 개수, x_1은 격자별 2차원 결과 침수심, x_2는 격자별 예측결과 침수심을 의미한다.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_{1,i} - x_{2,i})^2}{n}} \tag{5}$$

50 mm 강우에 대한 평균제곱근오차는 0.139 m, 80 mm 강우에 대한 평균제곱근오차는 0.140 m, 그리고 110 mm 강우에 대한 평균제곱근오차는 0.343 m로 나타났다. 50 mm 강우에

Table 6. The results of goodness of fit

Total Rainfall	Goodness of fit to area (%)	Ratio of overprediction (%)	Ratio of underprediction (%)	Max. flood depth (m)	RMSE for flood depth (m)	Correlation coefficient
50 mm	63%	-	37%	1.109	0.139	0.920
80 mm	79%	21%	-	1.104	0.140	0.939
110 mm	67%	33%	-	1.961	0.343	0.897

대한 최대침수심은 1.109 m, 80 mm 강우에 대한 최대침수심은 1.104 m, 110 mm 강우에 대한 최대침수심은 1.961m로 나타났다. 이러한 이유로 110 mm 강우에 대한 평균제곱근오차가 50 mm, 80 mm 강우에 대한 평균제곱근오차에 비해 크게 나타났다. 상관계수를 구하는 식을 Eq. (6)에 나타내었다. 여기서 n은 총 격자의 개수, x는 FLO-2D 모의 결과 격자별 침수심, y는 랜덤포레스트 회귀모형으로 예측한 격자별 침수심에 대한 자료이다. s는 표준편차를 의미한다.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y} \tag{6}$$

50 mm 강우에 대한 상관계수는 0.920, 80 mm 강우에 대한 상관계수는 0.939, 110 mm 강우에 대한 상관계수는 0.897로 나타났다. 면적에 대한 적합도, 평균제곱근오차, 그리고 상관계수 산정 결과를 Table 6에 나타내었다.

4. 결 론

격자 기본정보와 총 강우량만을 이용해 인공지능으로 침수심을 예측하는 것은 기존 수치해석 모형의 계산에 비해 속도면에서 큰 장점을 가지고 있다. EPA-SWMM 와 FLO-2D 모의결과는 짧게는 수 분, 길게는 수 시간 소요되는 반면, 본 연구에서 제시된 랜덤포레스트 회귀모형은 일단 학습이 완료되면 즉각적인 침수예측이 가능하다. EPA-SWMM 와, FLO-2D 모의결과를 랜덤포레스트 회귀모형의 학습자료로서 활용하였으며, 사전에 수집된 강수량 정보, 계산된 침수지도를 활용하여 즉각적인 침수지도 예측이 가능하도록 하였다.

예측속도에서 위와 같은 장점을 가지고 있으며, 입력자료의 보완을 통하여 부족한 침수면적 적합도를 개선할 수 있을 것으로 보인다. 그러나 본 연구에서는 보다 빠르고 효율적인 예측을 위해 격자 기본정보를 추출하고, 총 강우량만을 랜덤포레스트 회귀분석 입력자료로 이용하여 예측이 가능하도록 하였다. 강수정보에서 격자별 침수심 예측을 바로 도출할 수 있는 방법론을 제시하였음에 의의를 두고 있다. 세 가지 사상에 대한 예측결과를 검증된 FLO-2D 모의결과와 비교시 62.5%, 79.3% 그리고 66.7%의 적합도를 나타내었다. Kim and Han (2020)의 다양한 도시유역에 대한 예측결과를 참고하였을 때에 60% 이상의 적합도를 나타낸 본 연구결과는 준수한 예측력으로 판단되며, 단일 사상에 대한 과적합 문제를 나타내지 않기에 침수가 빈번하게 나타나는 도시유역의 홍수대응에 실질적으로 활용될 수 있을 것으로 보인다.

References

Albers, S.J., Dery, S.J., Petticrew, E.L. (2015). “Flooding in the Nechako river basin of Canada: A random forest modeling approach to flood analysis in a regulated reservoir system.” *Canadian Water Resources Journal*, Vol. 41, No. 1-2, pp. 250-260.

Kim, J.S. (2020). *Urban flood countermeasures for response of climate change*, Legislation and Policy Report, Vol. 62, National Assembly Research Service.

Kim, H.I., and Han, K.Y. (2020). “Data-driven approach for the rapid simulation of urban flood prediction.” *KSCE Journal of Civil Engineering*, Vol. 24, pp. 1932-1943.

Kim, H.I., and Kim, B.H. (2020). “Flood hazard rating prediction for urban areas using random forest and LSTM.” *KSCE Journal of Civil Engineering*, Vol. 24, pp. 3884-3896.

Lee, S.M., Kim, J.C., Jung, H.S., Lee, M.J., and Lee, S.R. (2017). “Prediction of flood susceptibility using random-forest and boosted-tree models in Seoul Metropolitan City, Korea.” *Geomatics, Natural Hazards and Risk,* Vol. 3, No. 2, pp. 1185-1203.

Li, B., Yang, G., Wan, R., Dai, X., and Zhang, Y. (2016). “Comparison of random forests and other statistical methods for the prediction of lake water level: A case study of the Poyang Lake in China.” *Hydrology Research*, Vol. 47, No. S1, pp. 69-83.

Mosavi, A., Ozturk, P., and Chaw, K.K. (2018). “Flood prediction using machine learning models: Literature review.” *Water*, Vol. 10. doi: 10.3390/w10111536

Park, J.H., Kim, S.H., and Bae, D.H. (2019). “Evaluating appropriateness of the design methodology for urban sewer system.” *Journal of Korea Water Resource Association*, Vol. 52, No. 6, pp.411-420. (in Korean)

Risi, R.D., Jalayer, F., and Paola, F.D. (2015). "Meso-scale hazard zoning of potentially flood prone areas." *Journal of Hydrology*, Vol. 527, pp. 316-325.

Tyralis, H., Papacharalmpous, G., and Langousis, A. (2019). "A brief review of random forests for water scientists and practitioners and their recent history in water resources." *Water*, Vol. 11, No. 5. doi: 10.3390/w11050910

Zhou, J.Z., Peng, T., Zhang, C., and Sun, N. (2018). "Data pre-analysis and ensemble of varios artificial neural networks for monthly streamflow forecasting." *Water*, Vol. 10, No. 5. doi: 10.3390/w10050628