서울시립대학교 디지털 물산업 혁신인재 양성사업단

AI for in-stream Total Organic Carbon (TOC) prediction

Sangchul Lee, Ph.D.

Assistant professor, Division of Environmental Science & Ecological Engineering, College of Life Science, Korea university

Dec 6th, 2024

서울시립대학교 디지털 물산업 혁신인재 양성사업단

Table of Contents

01 Presenter introduction

Who is the presenter?

고려대학교

이상철 교수

2007 고려대학교 환경생태공학과 이학사
2011 고려대학교 환경생태공학과 이학석사
2017 Univ. of Maryland at College Park 이학박사
2017-2020 USDA Research Scientist
2020-2024 서울시립대학교 환경공학부 교수
2024-현재 고려대학교 환경생태공학과 교수

Korea University	Korea University	University of Maryland	USDA	University of Seoul	Korea University
(2003-2007)	(2009-2011)	(2011-2017)	(2017-2020)	(2020-2024)	(2024 ~)

B.S. to M.S.

(2003-2007)

(2009-2011)

Ph.D.

(2011-2017)

Postdoc – Returning home in glory (금의환향: 錦衣還鄕))

USDA

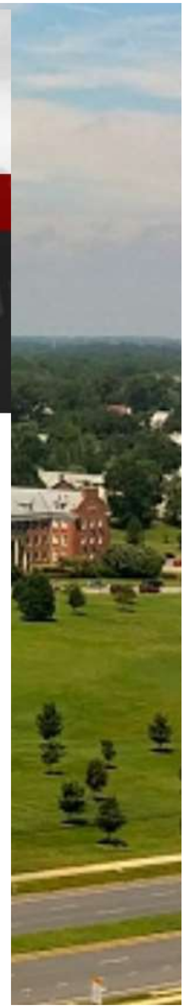
(2017-2020)

DEPARTMENT OF GEOGRAPHICAL SCIENCES

About Us · Undergraduate · Graduate · Research · GIS · Alumni · Diversity · High School Hub · Search

Dr. Sangchul Lee appointed as Assistant Professor at the University of Seoul

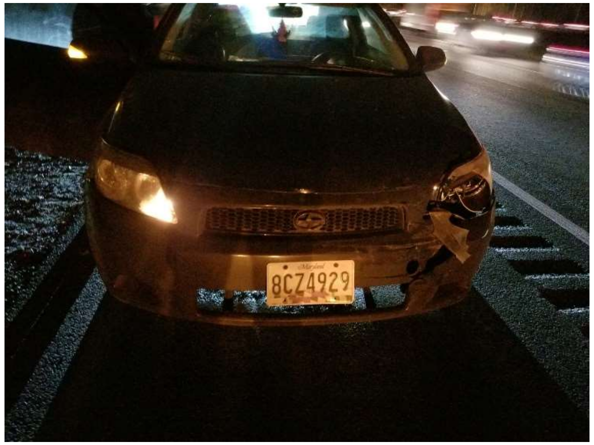
Congratulations to Dr. Sangchul Lee on his new position as an Assistant Professor in the School of Environmental Engineering at the University of Seoul in South Korea. Dr. Lee was a Graduate Research Assistant in our department from 2011- 2017. He graduated from the department in 2017 and he continued his research here as a Post-Doctoral Associate, then in AGNR as a Visiting Scientist for the USDA. Dr. Lee's new position will be focusing on environmental big data analysis at the first school for environmental engineering in South Korea. The school "aims to protect nature and humans from artificial pollutants while preserving the ecosystem. [It] teaches students environmental engineering theories through fundamental and applied courses such as environmental physics and chemistry, water and sewage supply planning, and noise and vibration control." (source) The department congratulates Dr. Lee and wishes him well in his new position and thanks him for his contributions to the department as a researcher and colleague.

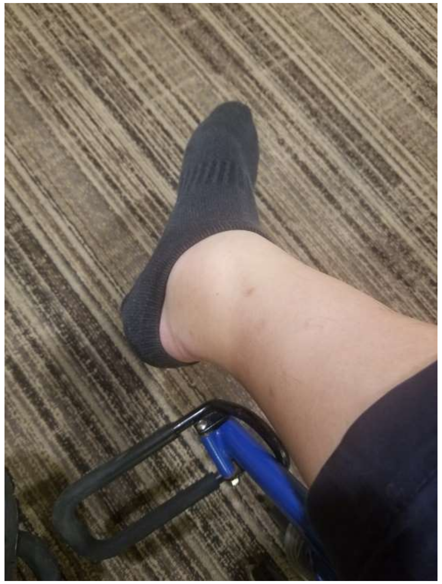


Postdoc – the dark times

USDA

(2017-2020)





Postdoc – the dark times

USDA

(2017-2020)

01_계명대학교	2021-04-30 오후 1:30	파일 폴더
02_강원대학교	2021-04-30 오후 1:30	파일 폴더
03_충남대학교	2021-04-30 오후 1:30	파일 폴더
04_서울시립대	2021-04-30 오후 1:31	파일 폴더
05_한국외대	2021-04-30 오후 1:31	파일 폴더
06_전북대학교	2021-04-30 오후 1:31	파일 폴더
07_경북대학교	2021-04-30 오후 1:31	파일 폴더
08_서울연구원	2021-04-30 오후 1:32	파일 폴더
09_명지대학교	2021-04-30 오후 1:32	파일 폴더
10_인천대학교	2021-04-30 오후 1:32	파일 폴더

22_Simon Fraser University	2021-04-30 오후 1:29	파일 폴더
23_Stevens	2021-04-30 오후 1:29	파일 폴더
24_El_Paso	2021-04-30 오후 1:29	파일 폴더
25_UC_merced	2021-04-30 오후 1:29	파일 폴더
26_Miami	2021-04-30 오후 1:29	파일 폴더
27_Alabama	2021-04-30 오후 1:29	파일 폴더
28_Delaware	2021-04-30 오후 1:29	파일 폴더
29_FloridaStateUniversity-civil-no	2021-04-30 오후 1:29	파일 폴더
30_Hawai	2021-04-30 오후 1:29	파일 폴더
31_Louisiana	2021-04-30 오후 1:29	파일 폴더
32_Wisconsin	2021-04-30 오후 1:29	파일 폴더
33_UtahState_fail	2021-04-30 오후 1:30	파일 폴더
34_UofVA	2021-04-30 오후 1:30	파일 폴더

후 1:27	파일 폴더
후 1:28	파일 폴더
후 1:28	파일 폴더

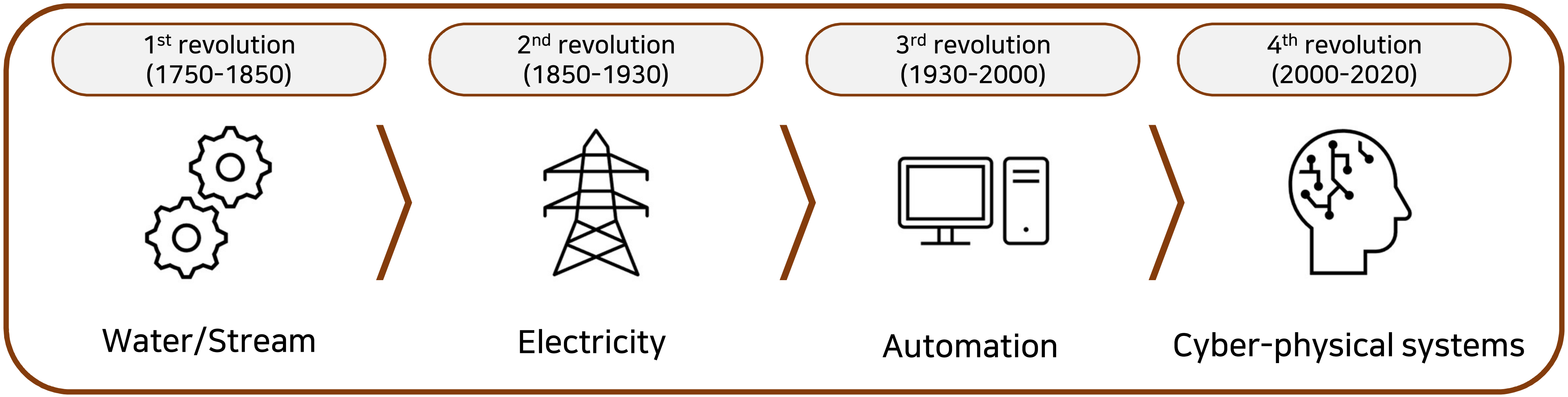
Professor

(2020-2024)

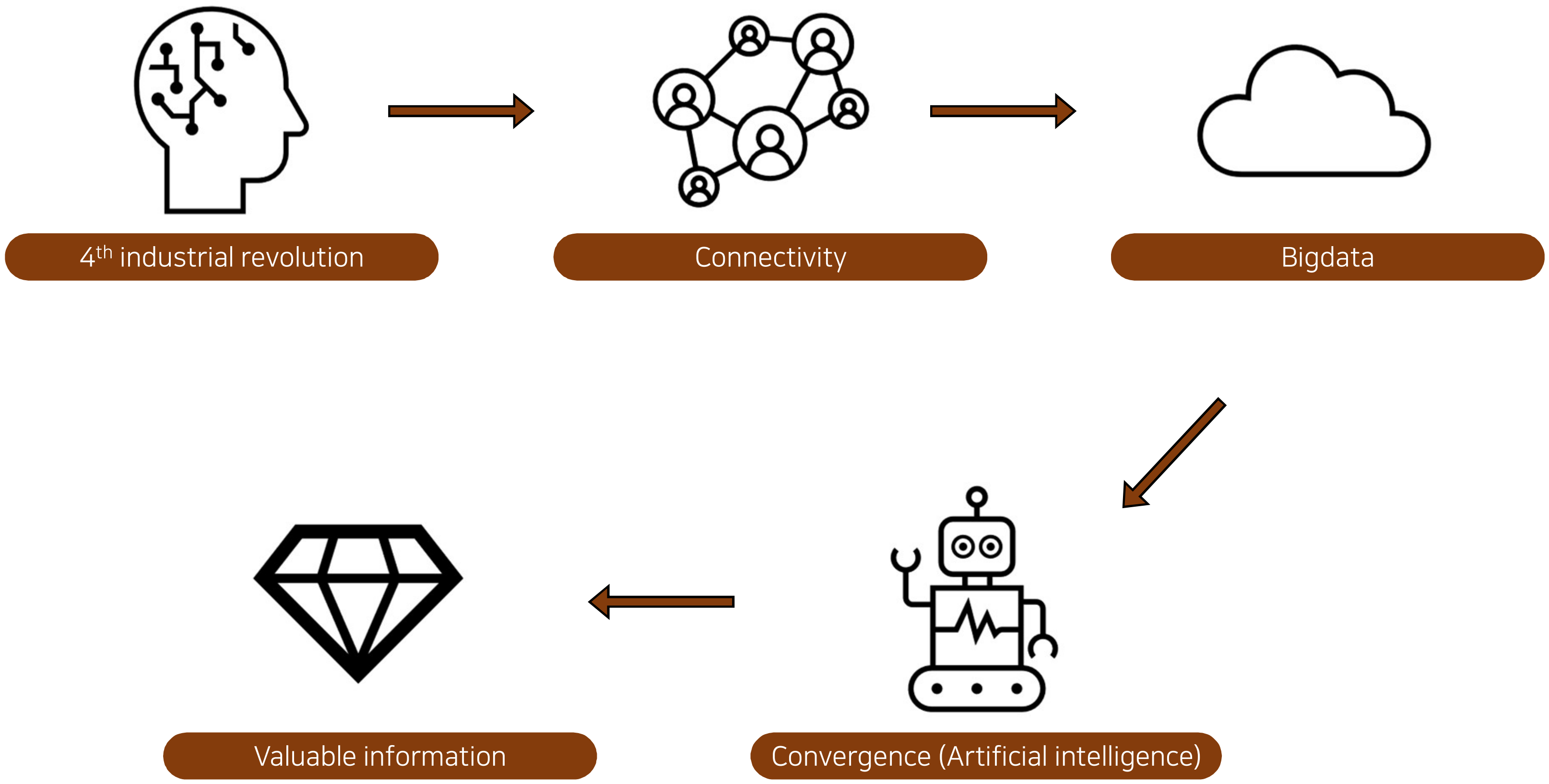
02 Background

4th Industrial revolution

The Fourth Industrial Revolution represents a fundamental change in our lives by the fusion of the physical, digital and biological worlds (Klaud shwab, 2016)

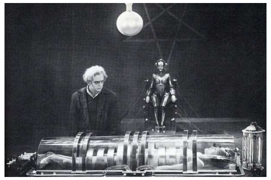


Characteristics of 4th industrial revolution

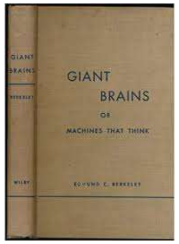


History of Artificial Intelligence (AI)

One of the topics for scientific fiction



SF film, *Metropolis* (1927)



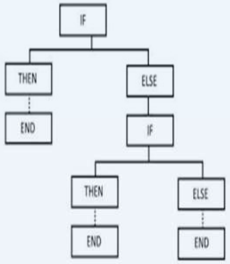
Book, *Giant Brains* (1949)

Artificial Intelligence (AI)

A machine mimic human activity such as "learning something" or "solving problem"

The scientific concept of AI was found

Logic Theorist



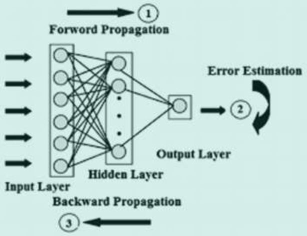
Too simple to solve complicate problems

Machine Learning

Algorithms with the ability to learn from data

The machine learning study is started with Artificial Neural Networks

Artificial neural networks



Require a long time to calculate some problems

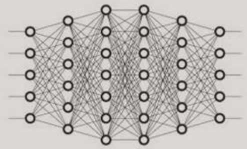
Deep Learning

The algorithms based on the deep neural networks

Application of Deep learning for industries

Big Data + Enhanced computing

Deep neural networks

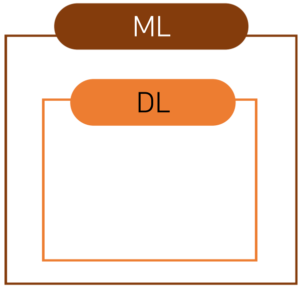


1950s

1980s

2010s

Deep Learning (DL) vs Machine Learning (ML)



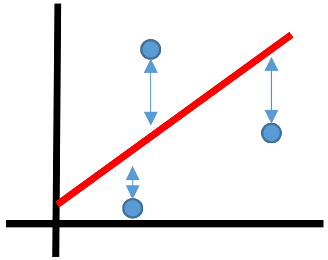
The basic of Machine learning

$H(x) = w\,x + b$

Hypothesis Weight Bias

$Error = H(x) - y$

Hypothesis True value



The purpose of machine learning is to find the optimum **w** and **b**

Chemistry

Studying time (x_1)	Score (Y)
3	40
5	50
7	70
9	90

$H(x) = w\,x + b$

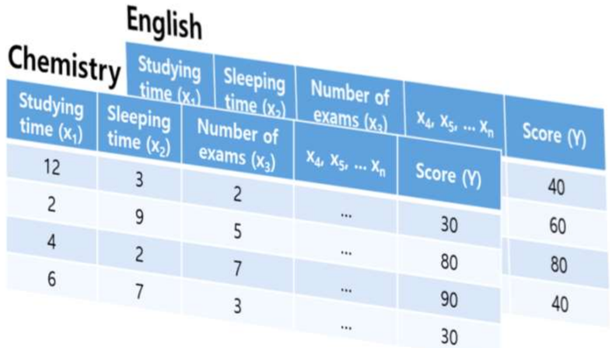
Easy

Chemistry

Studying time (x_1)	Sleeping time (x_2)	Score (Y)
3	9	40
5	3	50
7	2	90
9	4	70

$H(x) = w_1 x_1 + w_2 x_2 + b$

Easy



English

Studying time (x_1)	Sleeping time (x_2)	Number of exams (x_3)	$x_4, x_5, \dots x_n$	Score (Y)
				40
				60
				80
				40

Chemistry

Studying time (x_1)	Sleeping time (x_2)	Number of exams (x_3)	$x_4, x_5, \dots x_n$	Score (Y)
12	3	2	...	30
2	9	5	...	80
4	2	7	...	90
6	7	3	...	30

$H(x)_1 = w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_n x_n + b_1$

$H(x)_2 = w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_n x_n + b_2$

$H(x)_3 = w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_n x_n + b_3$

Difficult

Deep Learning (DL) vs Machine Learning (ML)

ML

Support Vector Machine

Self-Organizing-Map

Decision tree

Random Forest

⋮

DL

RNN

CNN

Limitations in ML

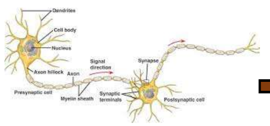
1. Impossible to handle Big data
2. Requires a long time to solve complex problems
3. Requires human helps for data preprocess
4. Each algorithm uses different approach

Advantages in DL

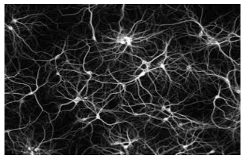
1. Possible to handle Big data
2. Provides high-quality results
3. Reduce the time for data preprocessing
4. Solve the complex problems at once

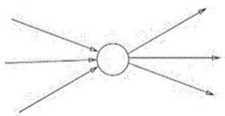
DL = Deep Neural Network

It is an algorithm which mimic human neural networks.

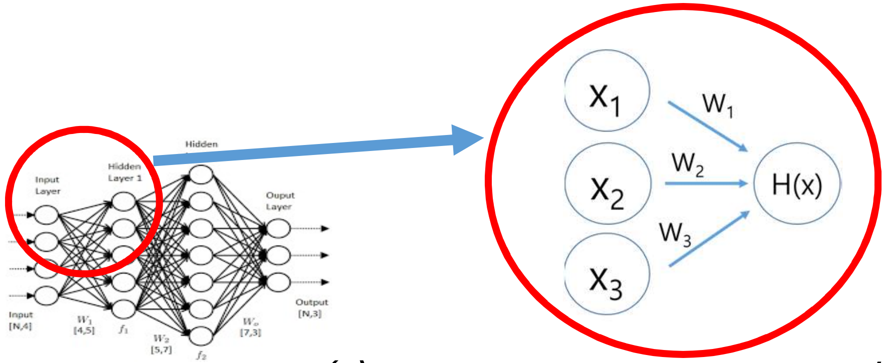


neuron





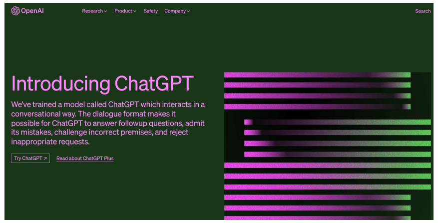
node



$$H(x) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

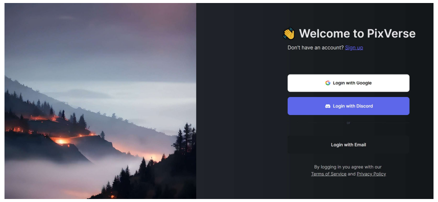
AI applications

Generative AI



ChatGPT

Text to image



Text to Video

Text to Action

Music generator

Coding AI

...

Automatic robot

Self driving car

Cleaning robot

Wearable robot

Industry

Military

5th industrial revolution (2020~)

Differences between 4th and 5th industrial revolution (Noble et al., 2022)

		4th industrial revolution	5th industrial revolution
Human-technology focus	Maximization strategy	• Maximizing the number and scope of technologies and their interconnectedness	• Maximizing the strengths of both technology and humans by understanding where each excels
	Competition vs. collaboration	• Human compete with machines for jobs	• Humans and machines dance together, metaphorically • Humans harmoniously collaborate with machines
Well-being focus	Environmental emphasis	• No environmental emphasis • Prioritization of technological progress (e.g., smart factories) • Pursuit of profits	• Well-being of all of humanity and the planet • Focus on sustainable and renewable resources • Pursuit of profits with a purpose
	Pushing the boundaries of technology	• Need to trust technology	• Humane uses of technology

Humans should collaborate closely with technology, to exploit the strengths of each and compensate for their corresponding weaknesses

Environmental and Ecological Engineering towards 5th industrial revolution

Sustainable Society

Advanced technologies (AI, DL, ML, ..)	Climate change	ESG	Sustainability

03 AI for TOC predictions

Total Organic Carbon?

The total amount of carbon that makes up organic matter present in water, and one of the representative water quality indicators that show the 'degree of water pollution

- In 2013, TOC was introduced as an organic matter management indicator in the living environment standards for rivers and lakes, and it was used alongside COD. Starting in 2016, COD was replaced with TOC.
- The adoption of TOC was driven by its advantages in speed, comprehensiveness, and precision, enabling a shift from a BOD-centric approach to a more efficient and accurate water quality assessment system.
- Because TOC has fewer years of observations compared to other water quality indicators such as TN and TP, the data size is small.

Type	Grade		TOC (mg/L)
	Very good	Ia	≤ 2
	Good	Ib	≤ 3
	Slightly good	II	≤ 4
Stream	Normal	III	≤ 5
	Slightly bad	IV	≤ 6
	Bad	V	≤ 8
	Very bad	VI	> 8

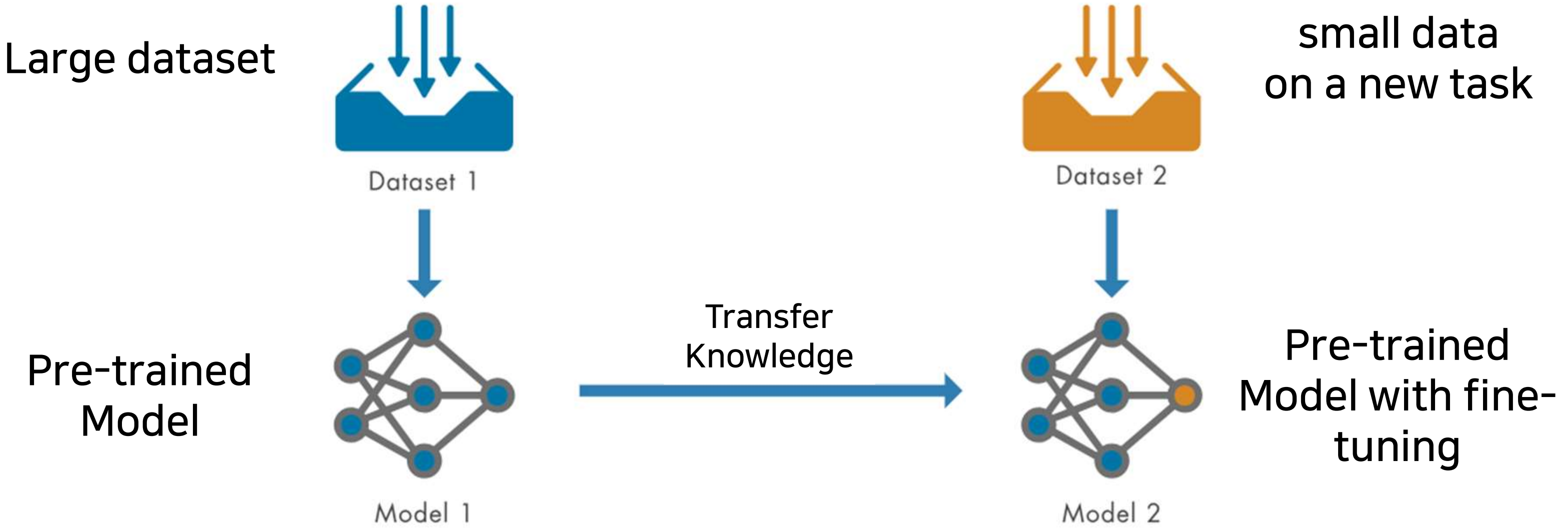
Transfer learning

Applying what you learn in one context to enhance generalization in another

Instead of training a model from scratch, knowledge of the model gained from a large dataset.

It allows to perform well on a new task with less data and computational resources.

Fine-tuning is the process of adapting a pre-trained model to a new task.

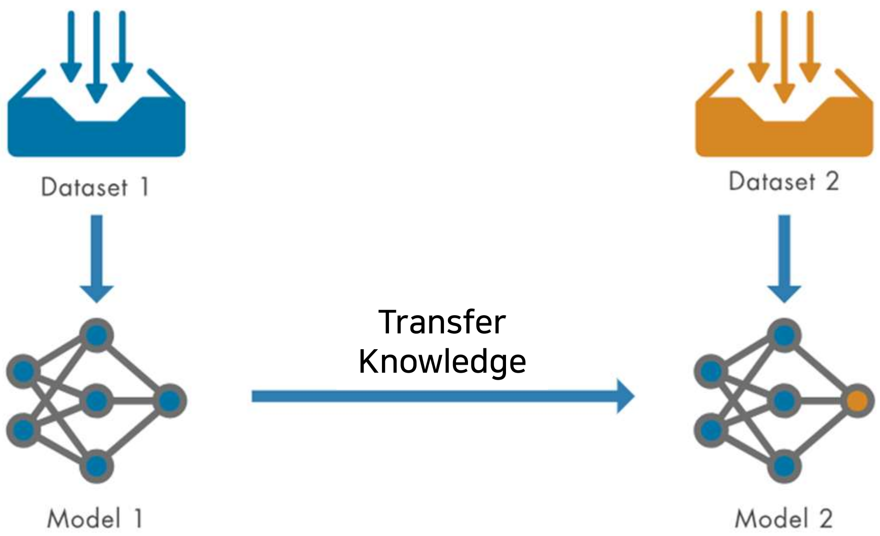


Study objective

Application of Transfer learning to predict TOC at the watershed level

Large dataset (ALL data from monitoring stations across the nation)

A watershed data



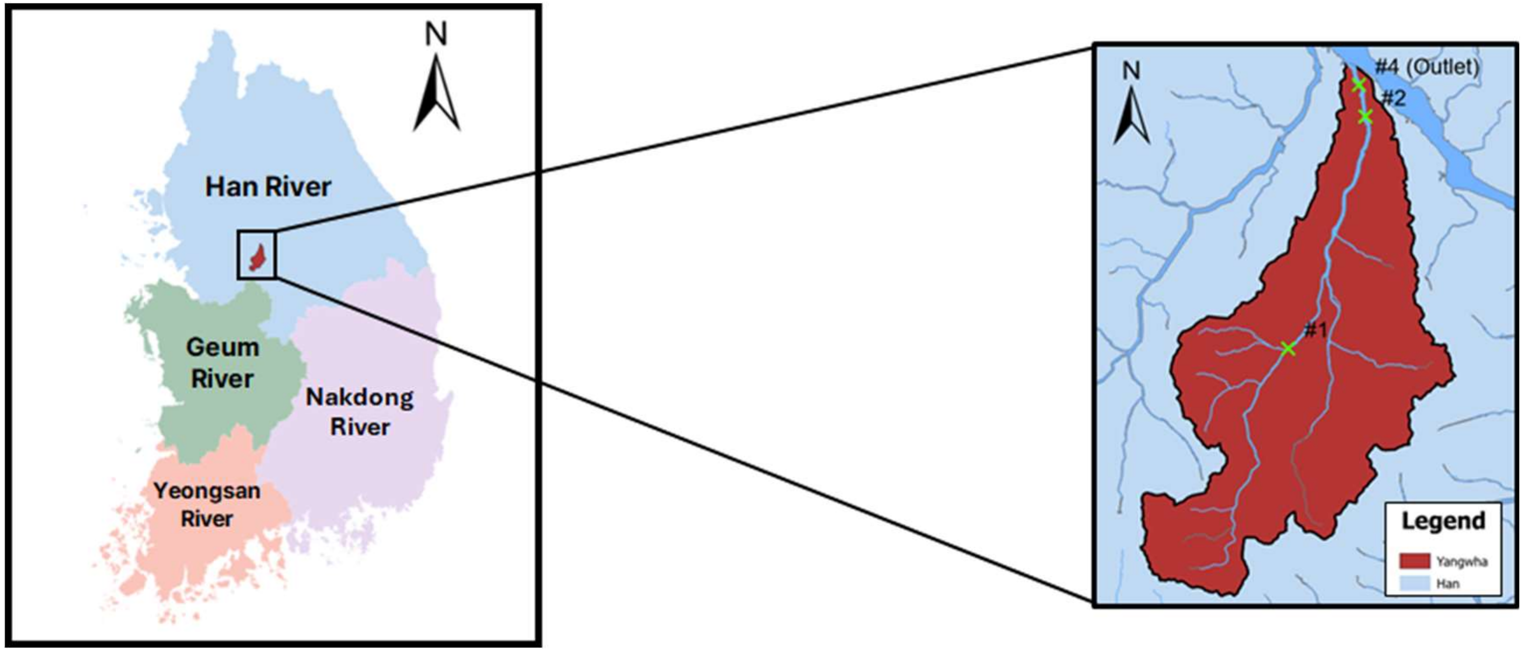
Pre-trained Model

Pre-trained Model with fine-tuning

Study Area

Target watershed: Yanghwa stream, Han River

The largest water source of Seoul is Paldang Reservoir.
Yanghwa Stream is one of the important sub-watersheds that affect the water quality of Paldang Reservoir.



Input data preparation

Flow, in-stream TOC, rainfall, and average temperature from all monitoring stations

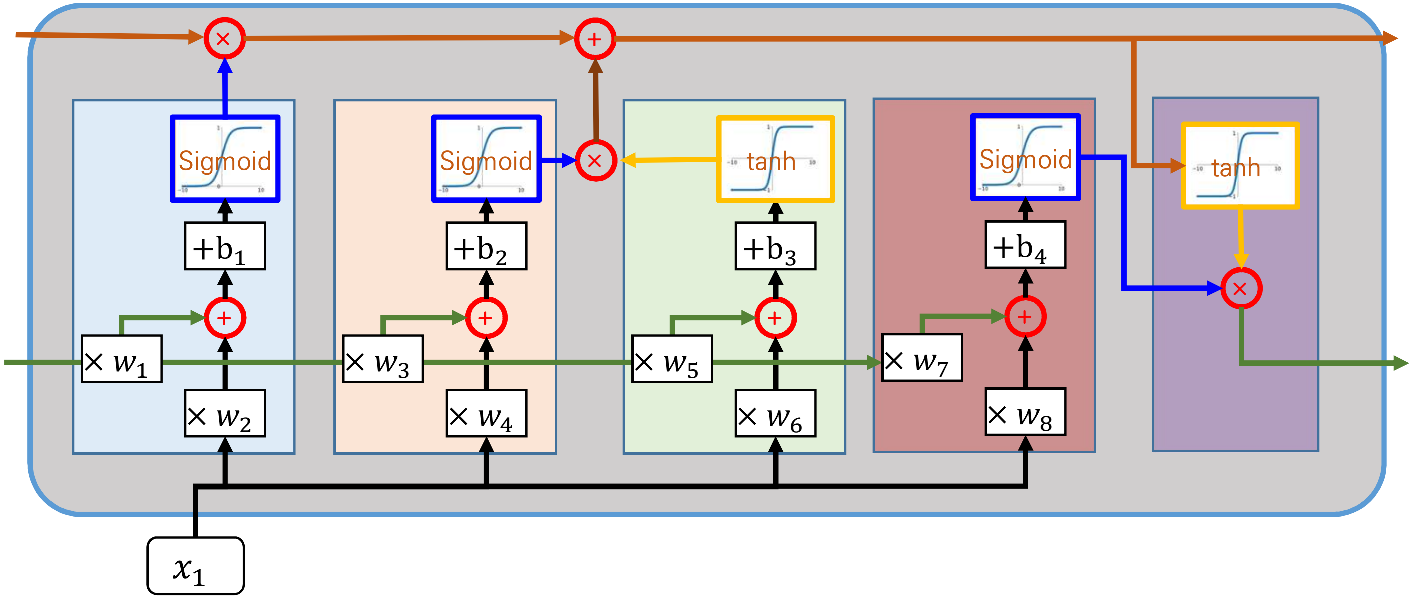
Input Variables (Weekly basis)		Target Variable
Water data	Weather data	
Flow (m^3/s) TOC (mg/L)	Rainfall (mm) Average temperature (℃)	TOC (mg/L)

Watershed	Water Quality Monitoring Station #
Han	139
Nakdong	117
Geum	69
Youngsan	67
Total	392

The climatic monitoring data closest to individual water quality monitoring stations were used as input data

LSTM

Long-Short Term Memory (LSTM) was chosen to predict TOC.

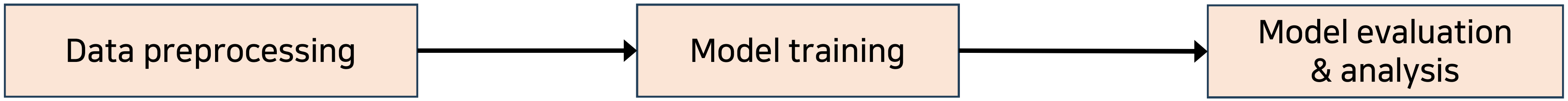


LSTM consists of 2states and 3 gates.
Two states: Cell states (Long-term memory) & hidden states (Short-term memory)
Three gates: Forget gate, input gate, and output gate
It maintains important patterns from historical data some while filtering irrelevant data.

Study design

Four different pre-trained cases were set to predict TOC from Yanghwa Stream.

Case	Watershed	WQ monitoring stations
I	Han River	139
II	Nakdong River	117
III	Han, Geum and Yeongsan River	275
IV	Han, Geum, Yeongsan, and Nakdong River	392

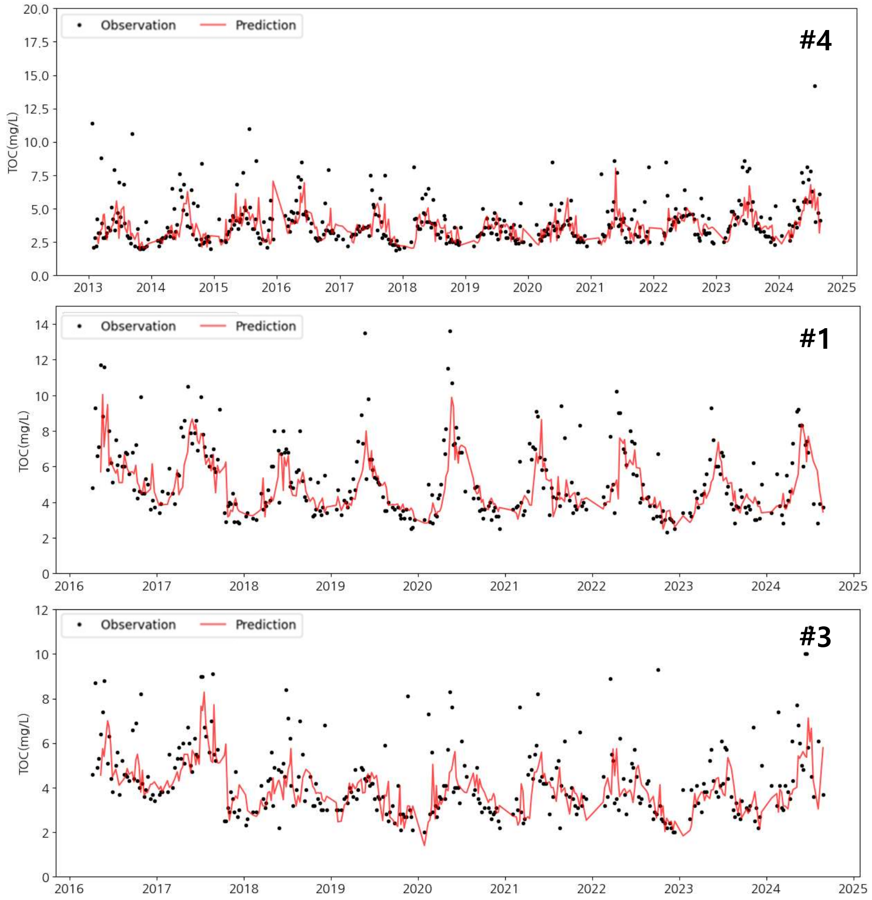


Study results (Han River)

Han River watershed data were used to predict TOC from Yanghwa Stream.

- The overall pattern is well captured, but there is difficulty in predicting the high TOC concentrations that occur mainly during the summer period.

	Train (Han)		Test (Han)	
Spot	NSE	P-bias	NSE	P-bias
#4 (outlet)			0.120	8.345
#1	0.778	0.120	0.513	8.599
#3			0.389	8.719

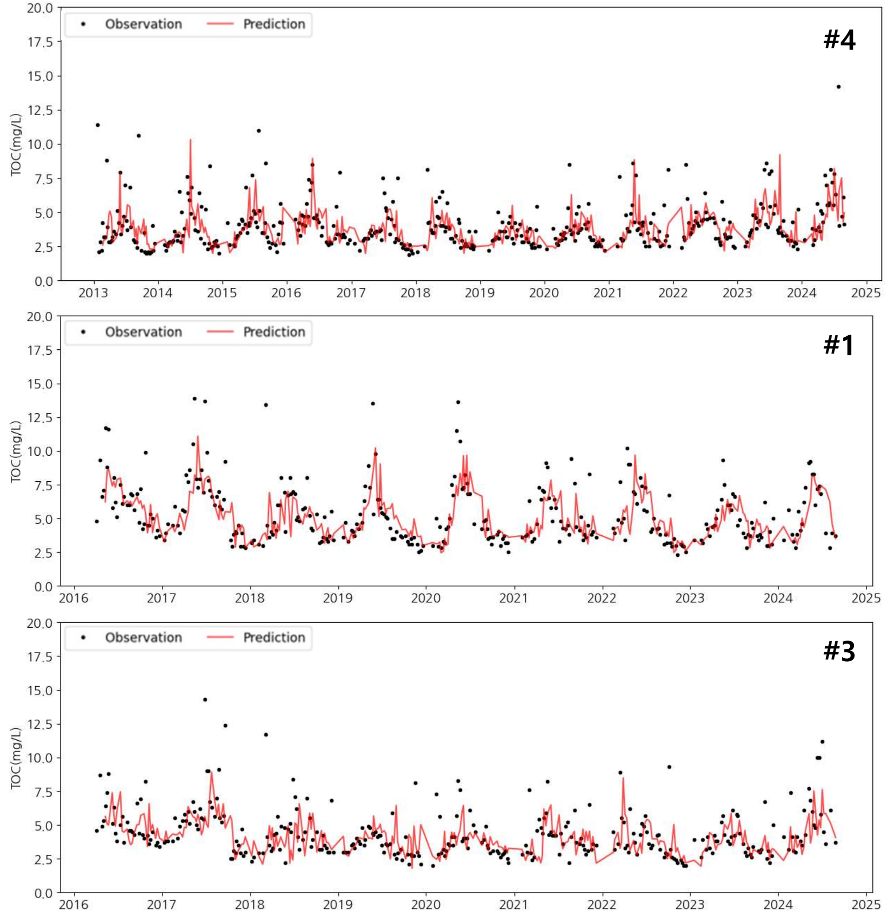


Study results (Nakdong River)

Nakdong River watershed data were used to predict TOC from Yanghwa Stream.

- It was confirmed that the test performance is lower when compared to Case I, which was pre-trained on the Han River.
- It was observed that the TOC patterns of the Nakdong River differ to some extent from those of the Han River.

	Train (Han)		Test (Han)	
Spot	NSE	P-bias	NSE	P-bias
#4 (outlet)			0.041	4.751
#1	0.811	0.042	0.390	4.121
#3			0.042	5.086

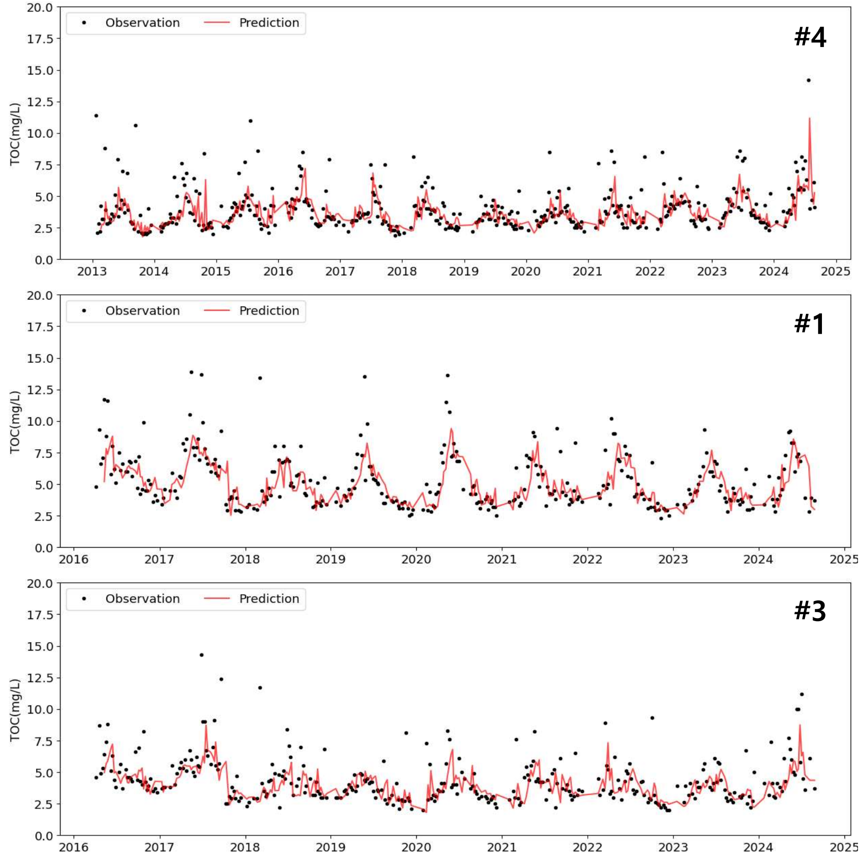


Study results (Han + Geum + Yeongsan River)

Han, Geum, and Yeongsan River watershed data were used to predict TOC from Yanghwa Stream.

- With more than twice as much data trained as in Scenario I and II, NSE performance is improved.
- Prediction performance has improved, especially at the outlet, but at the expense of relatively poor prediction of TOC at high concentrations.

	Train (Han)		Test (Han)	
Spot	NSE	P-bias	NSE	P-bias
#4 (outlet)	0.777	1.861	0.443	7.795
#1			0.503	5.338
#3			0.380	7.487

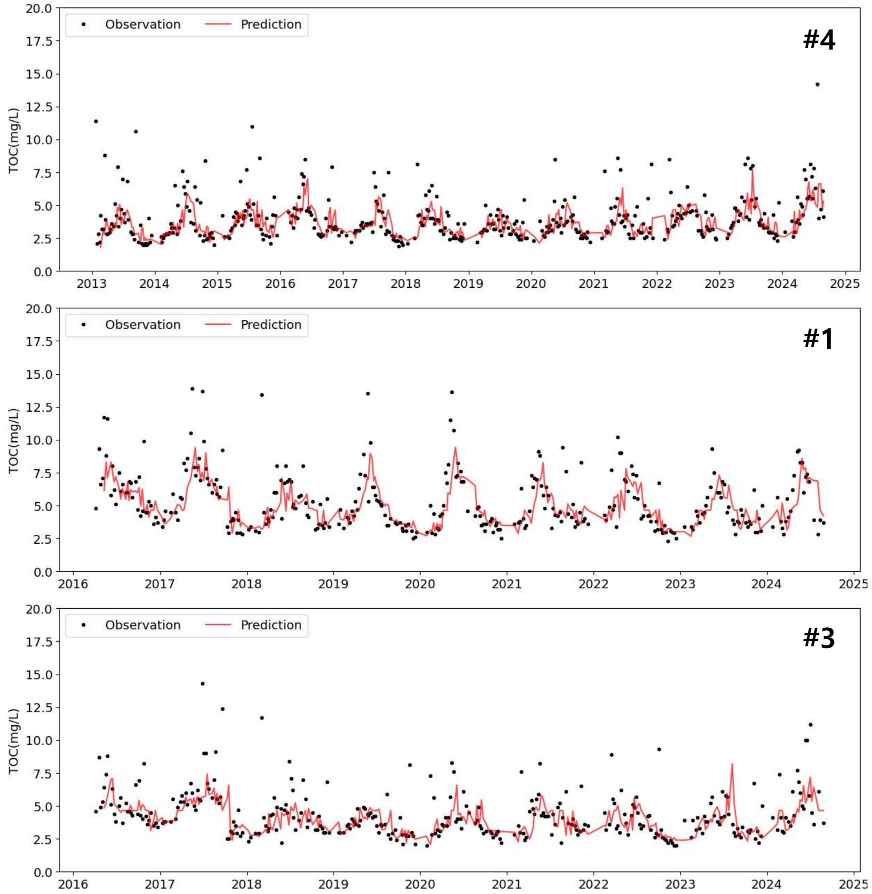


Study results (ALL)

Han, Geum, Yeongsan, and Nakdong River watershed data were used to predict TOC from Yanghwa Stream.

- When data from the Nakdong River basin is added to the training data of the pre-trained model, the NSE performance in the Yanghwa Stream basin significantly decreases.

	Train (Han)		Test (Han)	
Spot	NSE	P-bias	NSE	P-bias
#4 (outlet)			0.218	7.512
#1	0.707	2.939	0.449	5.759
#3			0.222	7.787



Conclusion

Transfer learning enhanced the prediction accuracy of in-stream TOC.

- Interestingly, the inclusion of Nakdong data reduced the prediction capacity.
- This is likely due to the path of Typhoon and densely distribution of waste water treatment plants in Nakdong River.

Case	Watershed	WQ monitoring stations	Train	Test
I	Han River	139	0.778	0.120
II	Nakdong River	117	0.811	0.041
III	Han, Geum and Yeongsan River	275	0.777	0.443
IV	Han, Geum, Yeongsan, and Nakdong River	392	0.707	0.218
	Single Model	1	0.39	-0.59

04 AI applications on Environmental Engineering

04 Case study (I):
Mosquito population predictions

Background

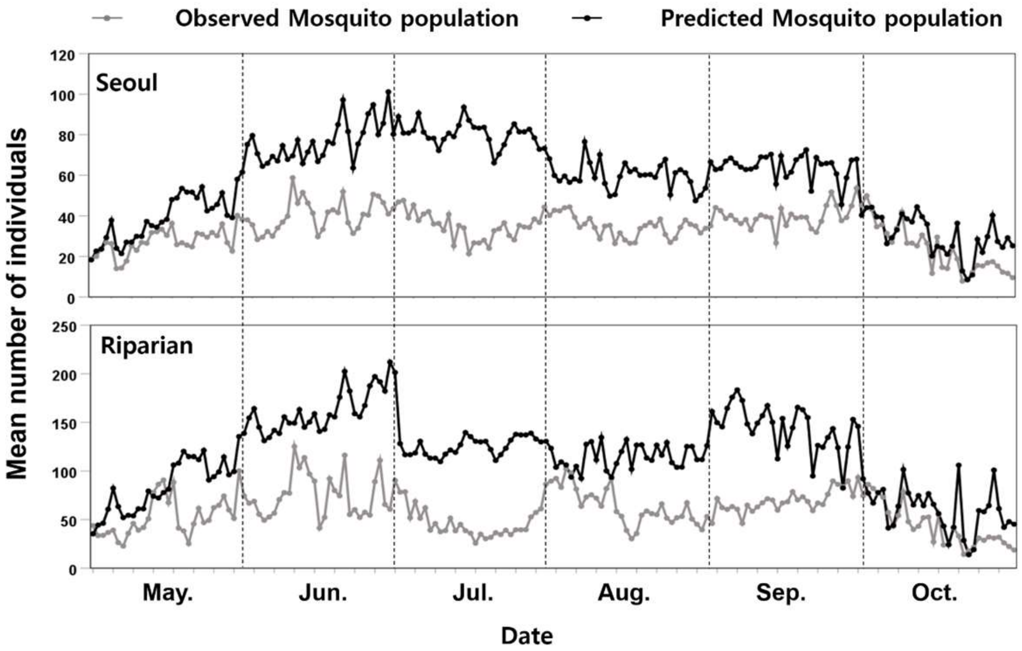
Study site

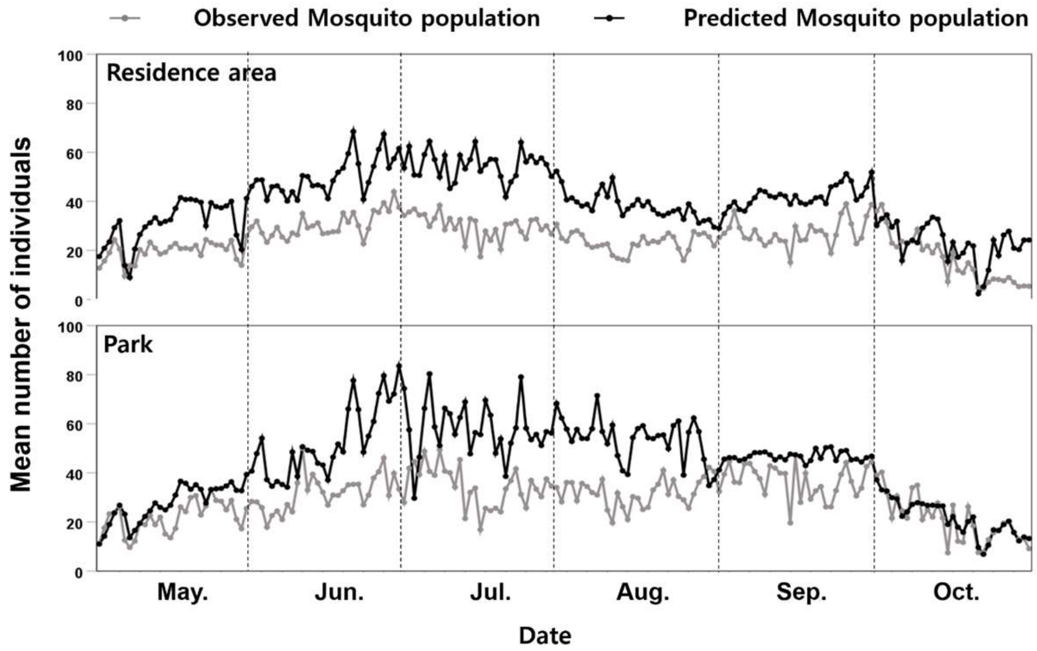
Prediction equations

월	모기 개체수 예측산식	r^2	AIC
5	(평균기온*3.462)+(최저기온*-3.027)+(강수 계속시간*-1.105)+(가조시간*24.246)+(평균 지면온도*-1.356) +(최저 초상온도*2.639)+{ (DMS개방도*-52.669)/50}+394.057	0.49	0.61
6	(최고기온*2.094)+(강수 계속시간*-1.678)+(합계 일사량*-0.452)+(평균 지면온도*-1.798) +(최저 초상온도*2.753)+{ (가로등*504.297)/50}-2810.323	0.44	601.31
7	(평균기온*-13.122)+(최고기온*3.822)+(강수 계속시간*-1.673)+(일강수량*0.102)+(합계 일사량*-0.991)+(평균 지면온도*5.226)+(최저 초상온도*5.82)+{ (가로등*750.55)/50}-4163.259	0.48	814.90
8	(평균기온*-8.067)+(최고기온*2.232)+(강수 계속시간*-1.337)+(평균 이슬점온도*1.238)+(가조시간*7.71) +(합계 일사량*-0.875)+(평균 지면온도*3.125)+{ (가로등*226.917)/50}-1273.444	0.28	835.32
9	(최저기온*-4.375)+(최고기온*2.271)+(강수 계속시간*-1.116)+(가조시간*-7.253)+(합계 일사량*-0.894) +(최저 초상온도*1.861)+{ (DMS개방도*-203.186)/50}+2912.897	0.28	458.80
10	(최고기온*2.722)+(일강수량*-0.206)+(최저 초상온도*1.676)-37.636	0.85	292.95

Background

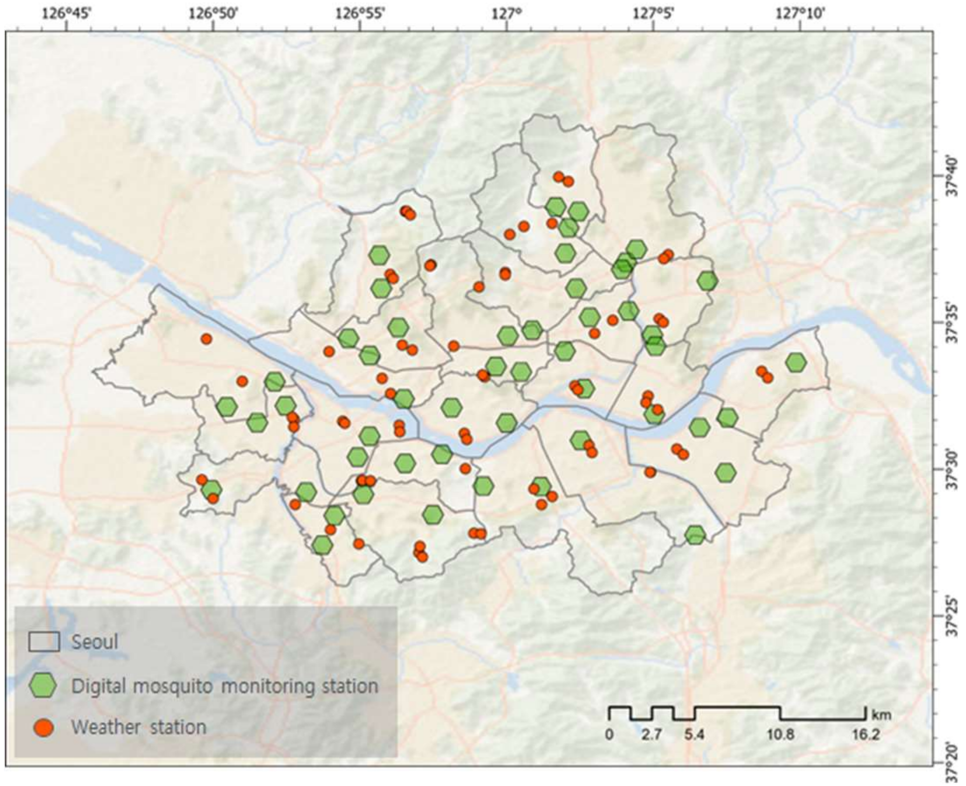
Low Accuracy





Study site & data collection

Study site



Seoul

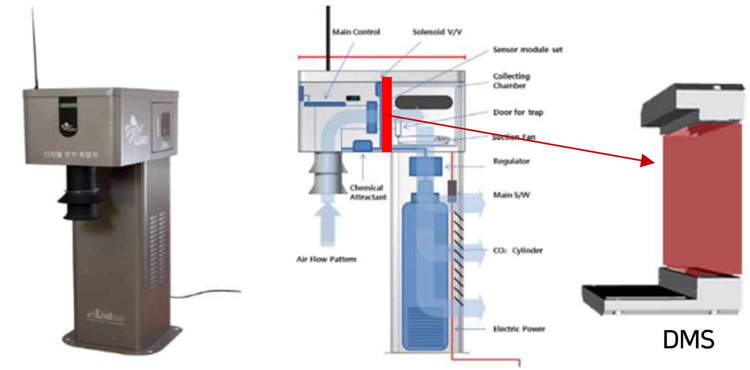
Area: 605,243,985㎡

Population: approximately 9,635,445

Digital smart Mosquito monitoring System (DMS)

Real-time remote mosquito monitoring device using carbon dioxide as an attractant

- Data collected from 19:00 the previous day to 5:00 on the same day
- Averaged and converted into daily data



Data collection

Weather data

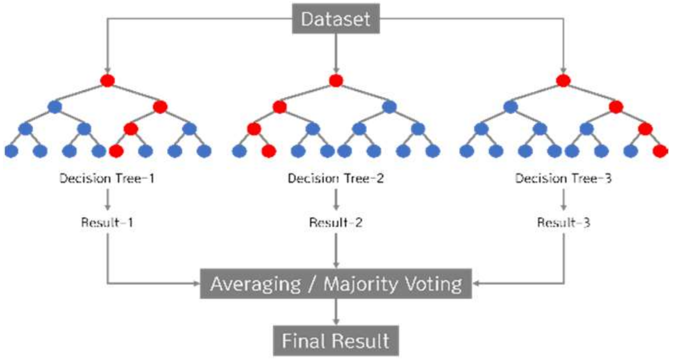
- **Identifying the weather stations nearest to individual DMSs**
- **Collected data includes**
  Maximum daily temperature,Minimum daily temperature
  Wind speed, Precipitation,Humidity
- **Additional temperature metrics**
  Monthly mean temperature, Yearly mean temperature

Landscape data

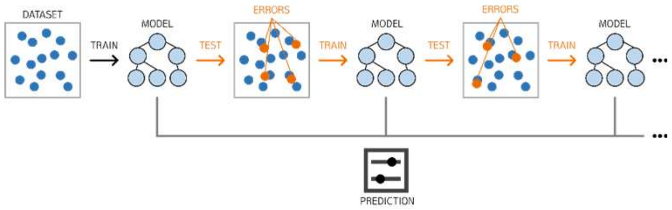
- **Environmental factors and surface data collected within specific radius (1m, 3m, 5m, and 10m) around each DMS**
  DMS Openness [%]
  Environmental factors
  Natural, artificial features and surface types

Machine learning

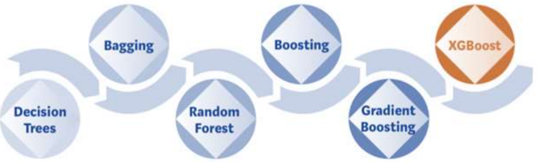
Random forest



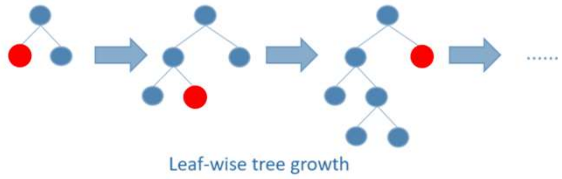
Gradient Boosting



eXtreme Gradient Boosting

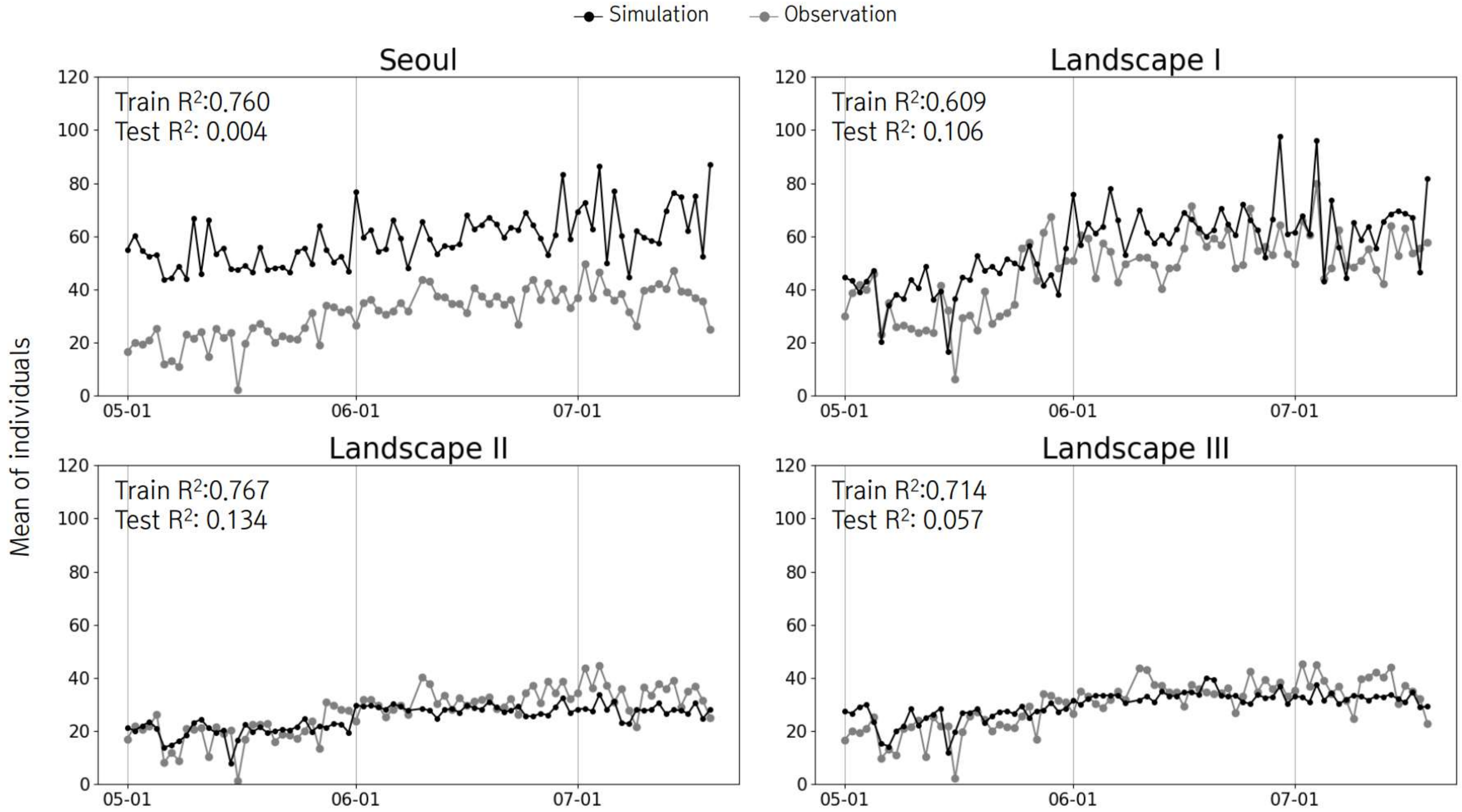


Light Gradient Boosting



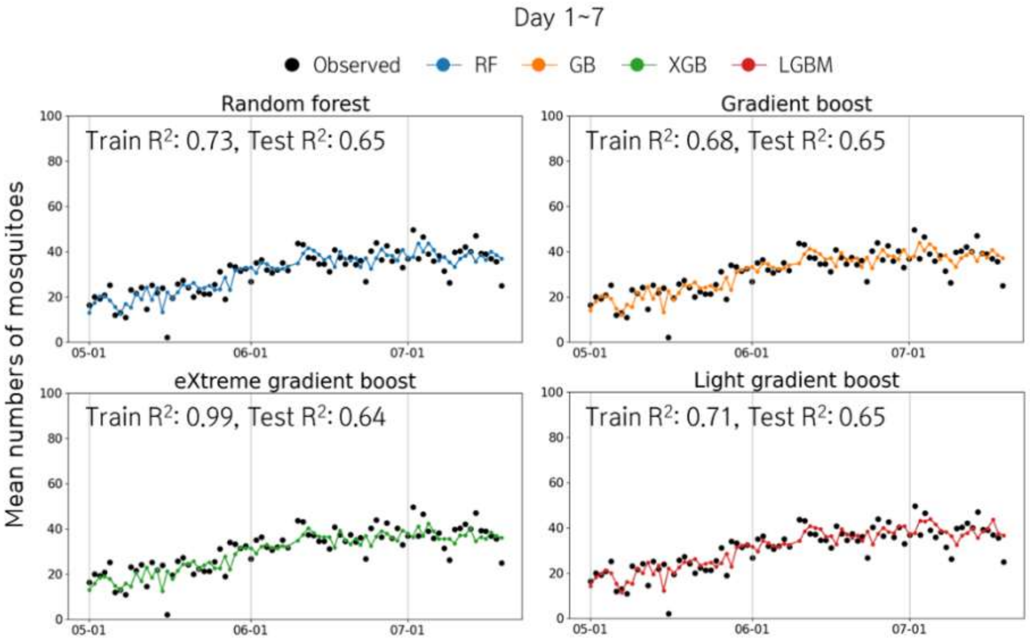
Results

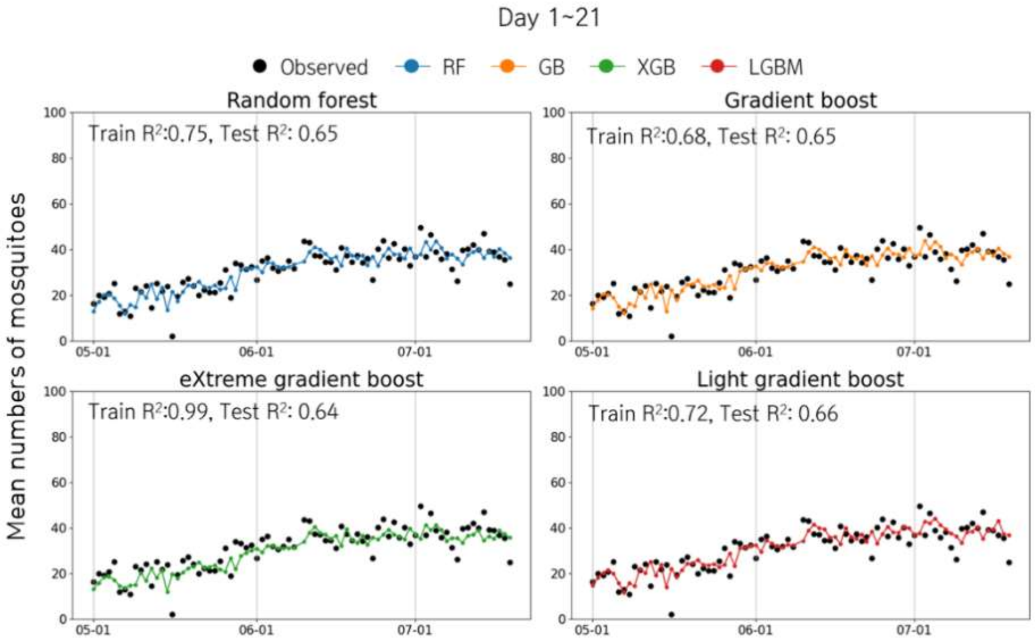
Model performance measures WITHOUT considering mosquito characteristics



Results

Model performance measures WITH considering mosquito characteristics





04 Case study (Ⅱ):
Groundwater quality monitoring

Introduction

Groundwater degradation in Jeju island

질산성질소 위협...제주 지하수 오염 '심각'

김정은 기자 | 승인 2022.06.07

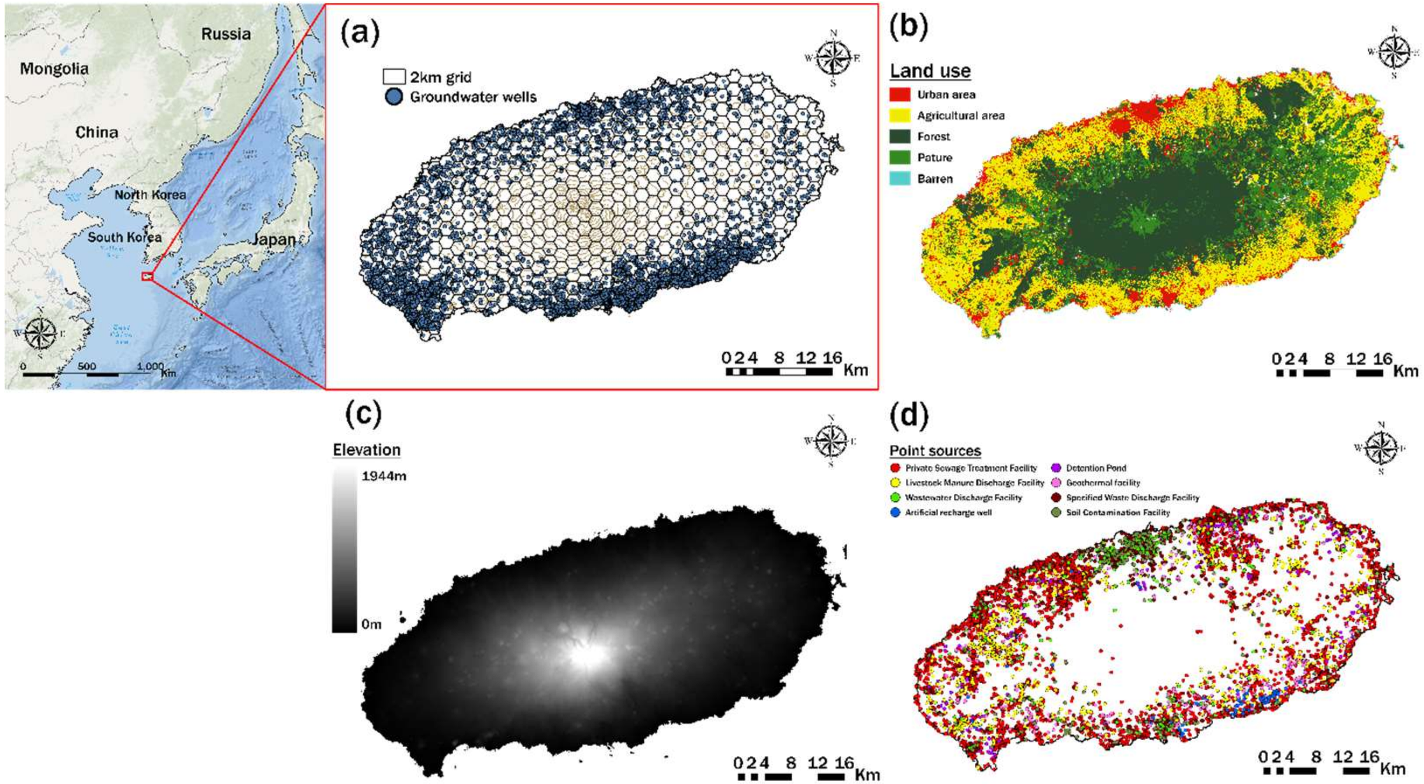
> 제주도 보건환경연구원 조사
> 한림읍 농업용 수질기준 초과
> 한경면 사설 관정 오염 심각
> 대정읍 오염도 지수 높아

제주 서부 지하수 오염수준 '최악'...청정 제주는 옛말

서부 지역 관정 중 심각 수준 59% 넘어
8백미터 이상 관정도 오염정도 심각
대대적인 오염원 관리 대책 시급

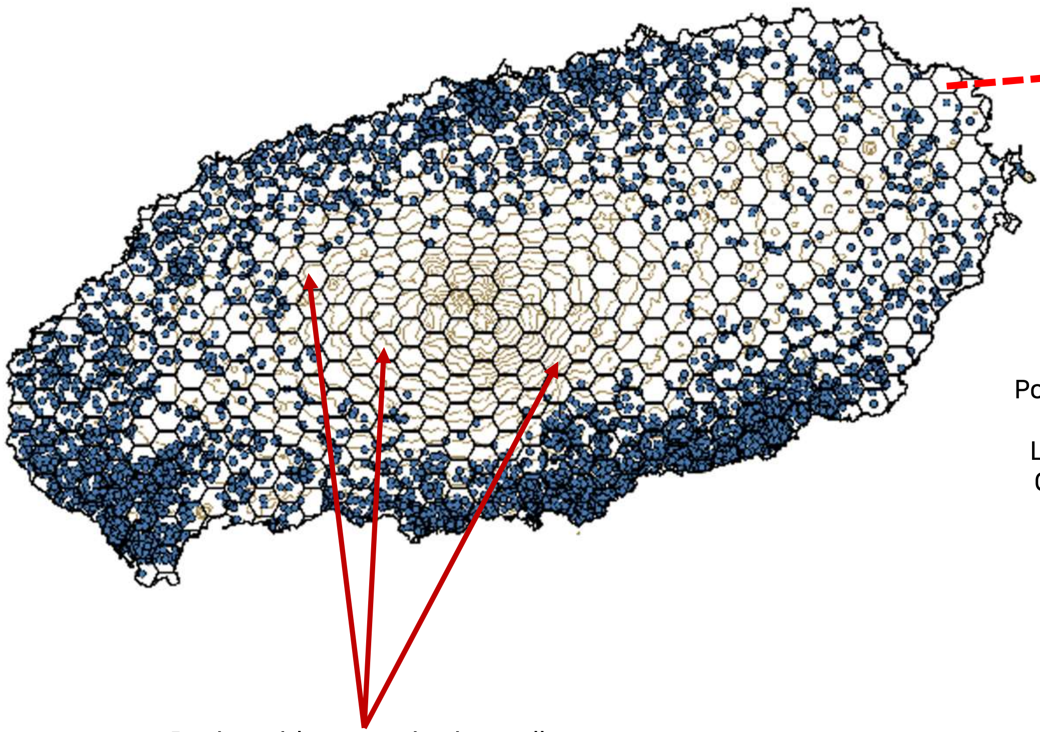
Methods

Site discretization for predictions



Methods

Data preparation



25 input
Point sources
Soil
Land cover
Geological

Grid	X1	X2	X3	…	Xk
1	X1,1	X1,2	X1,3	…	X1,k
2	X2,1	X2,2	X2,3	…	X2,k
…	…	…	…	…	…
n	Xn,1	Xn,2	Xn,3	…	Xn,k

Target
Mean NO3-N

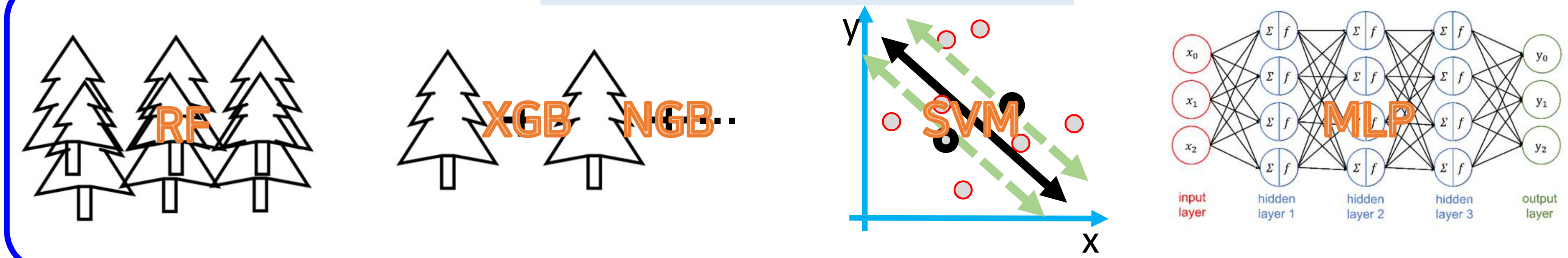
NO3-N
0.5
1.4
…
0.75

Train : Test = 70% : 30% (n = 440)
5-fold cross validation

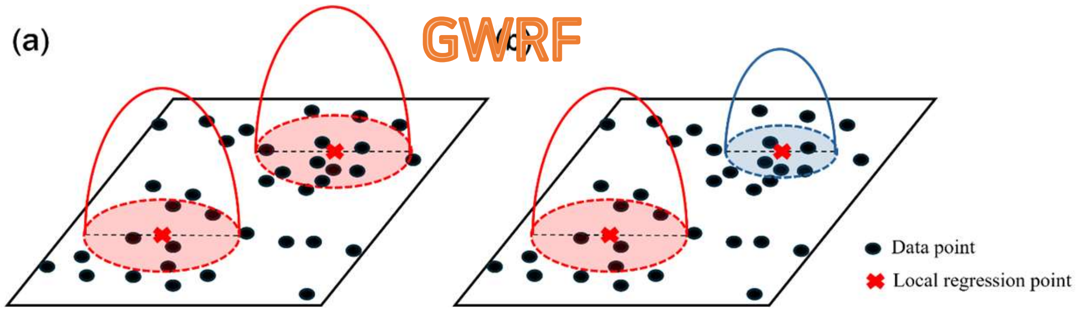
Methods

Machine learning models (MLMs)

Conventional MLMs (CMLMs)



Geographically Weighted MLMs (GWMLMs)



Results

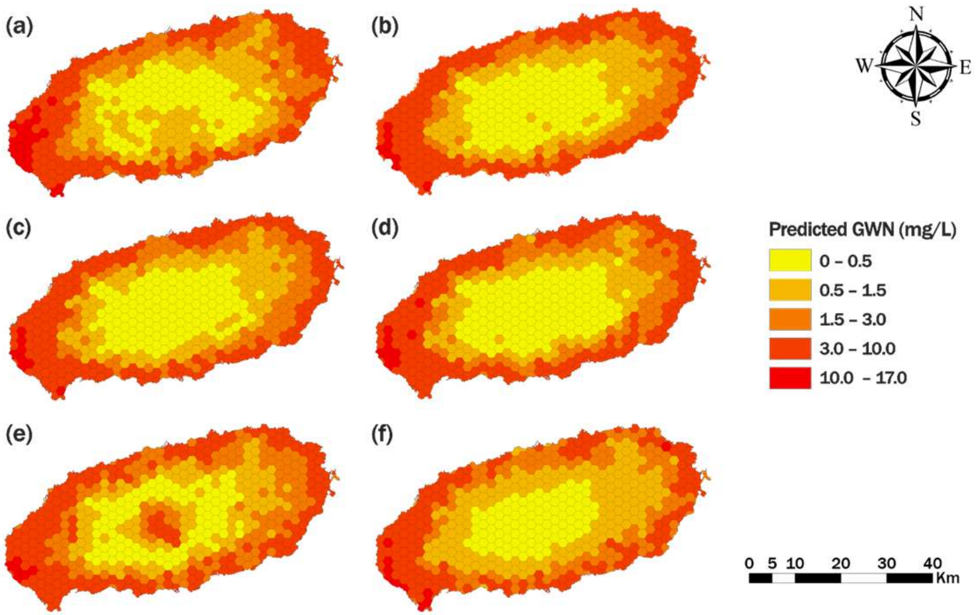
Performance measures

	GWMLMs (GWRF)	CMLMs				
		RF	XGB	NGB	SVM	MLP
Train R2	0.960	0.848	0.840	0.807	0.705	0.853
Test R2	0.728	0.571	0.619	0.636	0.473	0.513
Train RMSE	0.685	1.333	1.370	1.502	1.860	1.311
Test RMSE	1.497	1.879	1.772	1.732	2.083	2.003

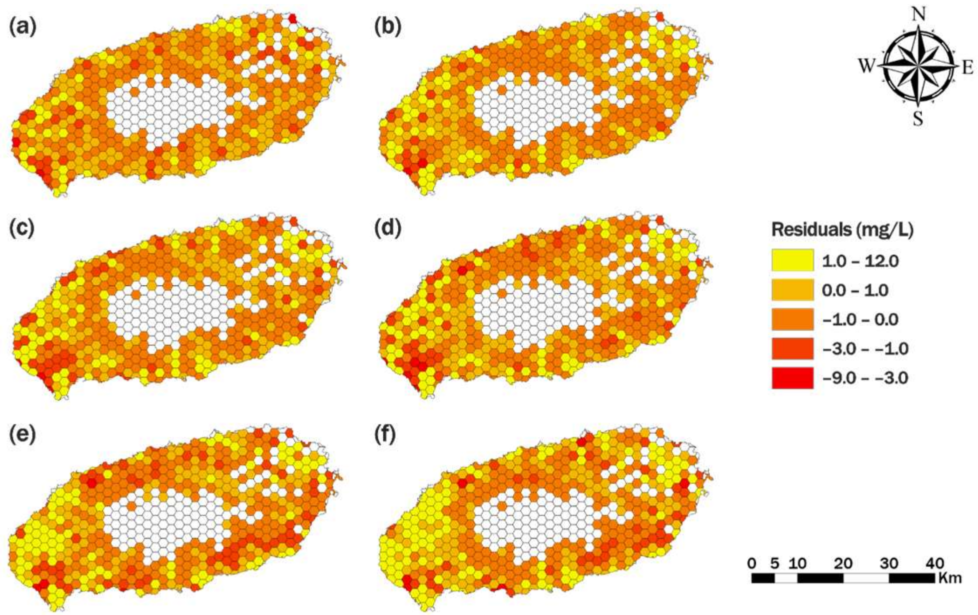
RF: Random Forest; XGB: Extreme Gradient Boosting; NGB: Natural Gradient Boosting; SVM: Support Vector Machine; MLP: Multi-Layer Perceptron

Results

Mapping results



Residual Map



(a) GWRF, (b) RF, (c) XGB, (d) NGB, (e) SVM, and (f) MLP.

04 Case studies (III):
Surface water quality monitoring

Introduction

Algal blooms

'녹조라떼' 명명 10년, 환경부는 문제없다지만 원수관리 취약

윤수은 기자 | 입력 2022.08.09 17:08 | 댓글 0

녹조로 뒤덮인 낙동강. 사진=뉴시스

인기뉴스
1 메이플
2 SVB 파
3 암호화
4 전두환
5 순다 피
6 GDC 2
7 육아 휴
8 '포켓몬
9 세계의
10 SVB발

[이코리아] 계속된 폭염과 가뭄이 길어지면서 낙동강의 녹조 현상이 심상치 않자 환경부가 직접 조사에 착수해 '문제없다'는 결론을 내렸다. 하지만 녹조 해소에 대해서는 가뭄 때문에 어렵다며 뚜렷한 해결책을 내놓지 못했다. 환경단체들은 시민 건강과 직결된 사안인 만큼 시민 우려를 해소하는 투명하고 적극적인 환경당국의 대응이 필요하다는 지적이다.

8일 환경부는 '녹조 현황 및 대책' 자료를 통해 대구·부산·경남지역 정수장 5곳의 수돗물을 대상으로, 녹조 독성 물질인 마이크로시스틴이 검출되지 않았다고 밝혔다. 이를 위해 환경부 고시에 규정된 마이크로시스틴 분석법인 LC-MS/MS법과 환경단체가 사용한 ELISA법 두 가지 방법 모두 사용했다.

다시 돌아온 '녹조라떼'
– 낙동강 녹조의 주범은 4대강사업이 만든 초대형 보

2012년 4대강사업이 준공되자 나타난 낙동강 녹조현상이 올해도 어김없이 '녹조라떼'로 돌아왔다.

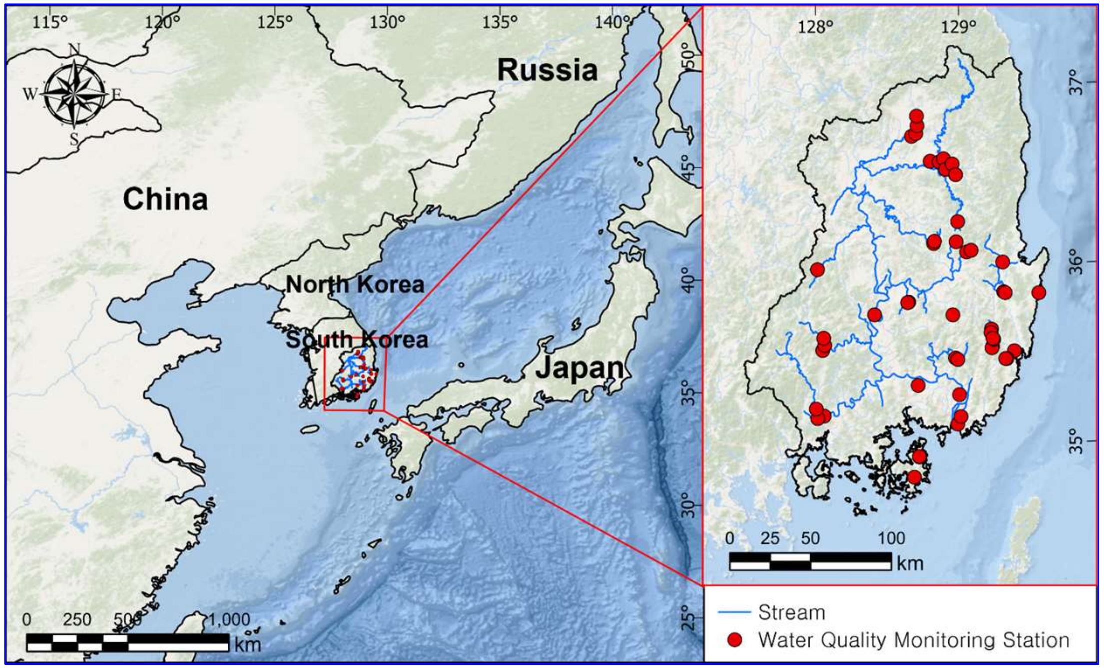
낙동강 중류 고령과 대구 인근에서 6월 초에 발견되는 녹조라떼 현상은 지난해 8월 초순 보고되던 녹조현상에 비해 두달 정도나 더 빠른 것으로, 이것은 녹조현상이 4대강사업으로 인한 결과란 것을 다시 한번 증명한다.

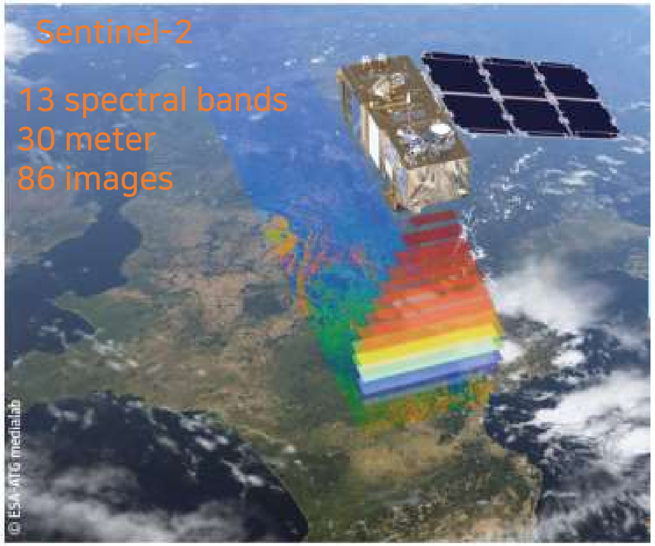
지난해 녹조가 대발생하자 그 원인을 두고 환경단체와 전문가들에서는 "이전에 없던 녹조현상이 4대강에서 발생한 것은 4대강사업으로 초대형 보로 막힌 강에서 강물이 정체되었기 때문"이라 확신했다. 그러나 정부와 환경부는 "4대강사업 때문임이 아니다"는 주장으로 일관했고, 당시 내세운 논리가 이상고온 현상이었다.

▲ 낙동강 고령 우곡교 아래서 다시 만난 '녹조라떼' ⓒ대구환경연합 정수근

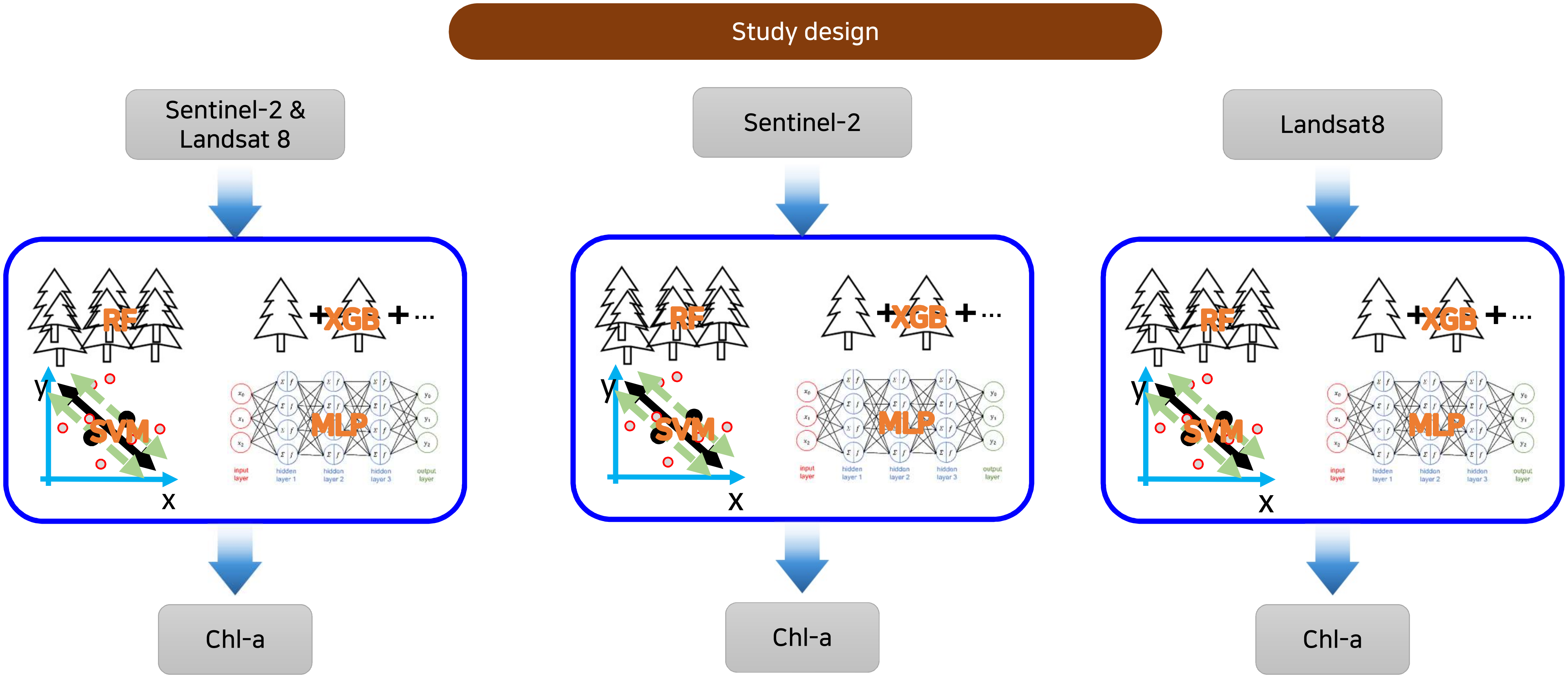
Methods

Study site and dataset



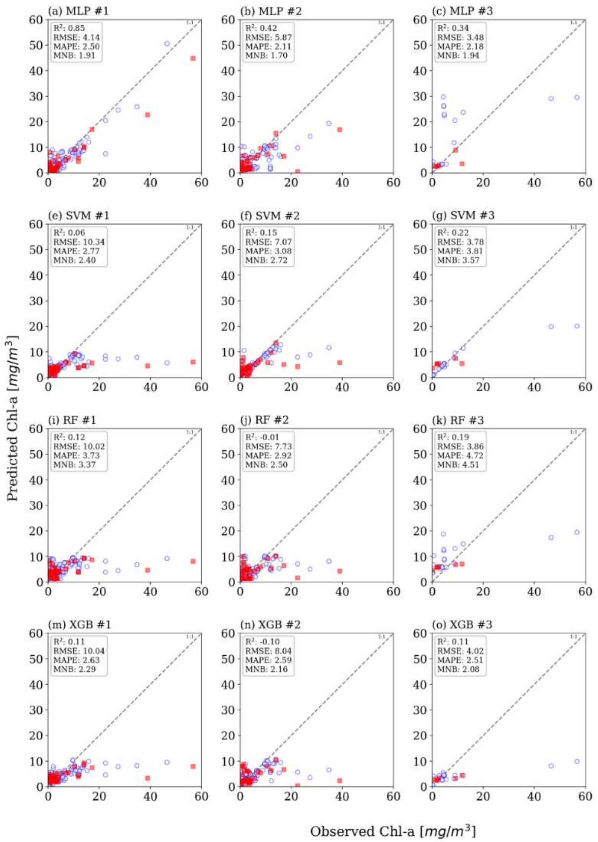


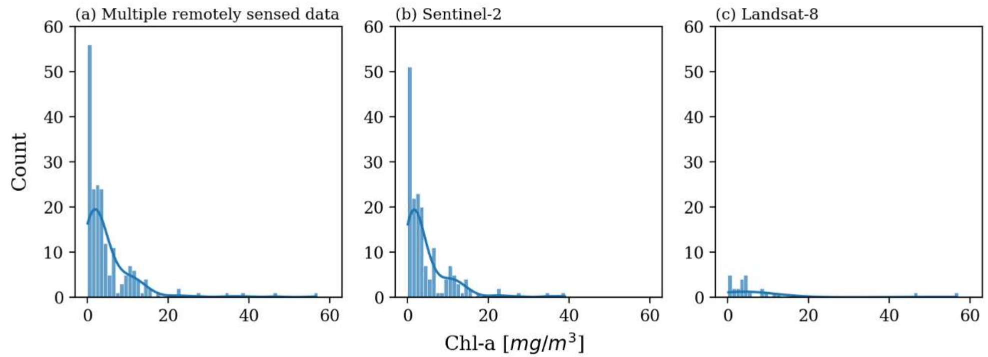
Methods


Study design
Sentinel-2 &
Landsat 8
Sentinel-2
Landsat8
RF
+XGB +…
y
SVM
x
MLP
input layer
hidden layer 1
hidden layer 2
hidden layer 3
output layer
Chl-a
Chl-a
Chl-a

Results

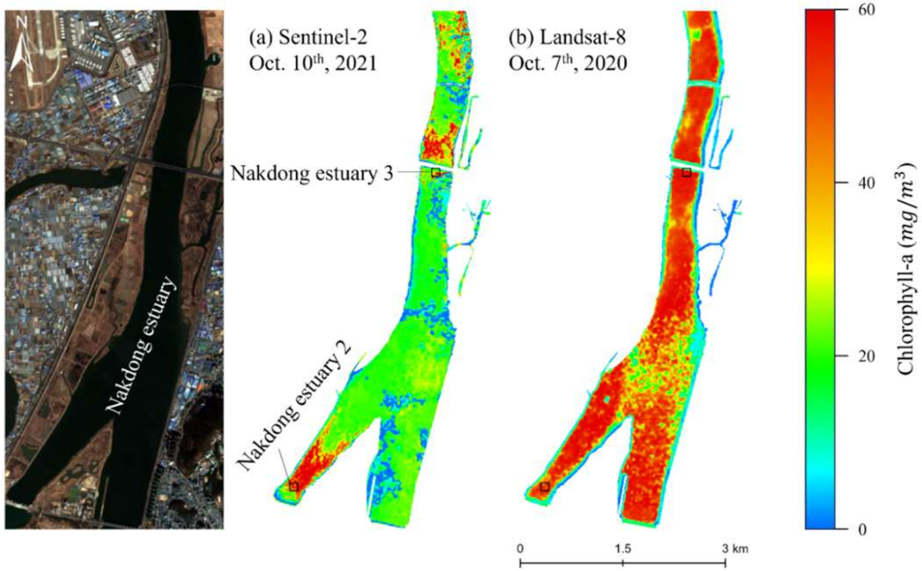
Performance measures & Data bias

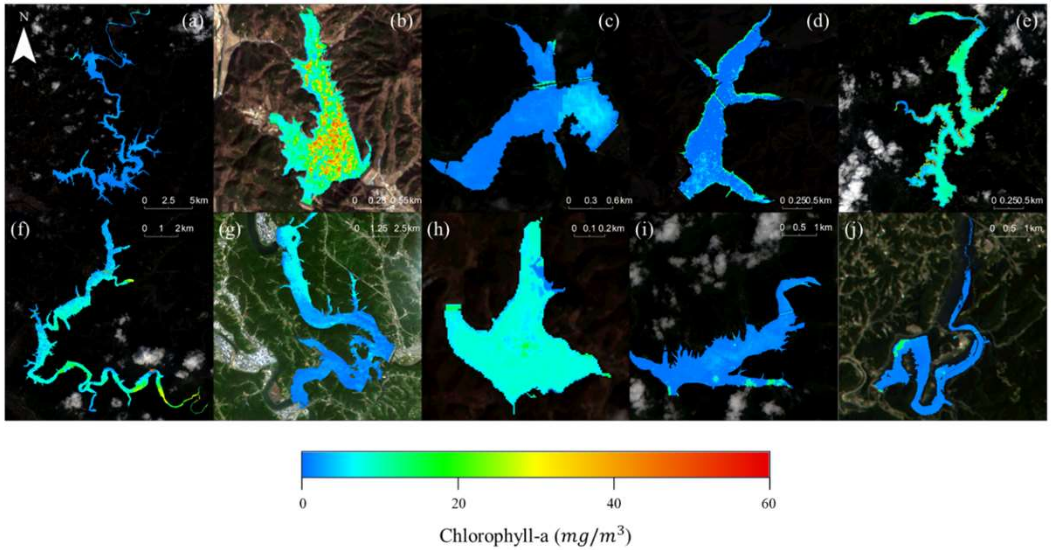




Results

Prediction map





04 Case studies (Ⅳ):
Machine learning applications on soil heavy metals

Introduction

Soil degradation around abandoned mining sites

중금속 기준치 초과 폐광산 56곳 중 ‘13곳’

신석주 기자 | 승인 2020.10.20 13:21 | 댓글 0

파킨슨, 심근경색 등 중대질환 유발하는 철·망간·아연 중금속 기준치 초과
황운하 의원, “폐광산, 수질오염뿐 아니라 토양오염에 대한 조사도 병행돼야”

[에너지신문] 한국광해관리공단이 관리하고 있는 폐광산에서 기준치 이상의 중금속이 함유된 유출수가 방류되고 있는 것으로 나타났다. 특히 기후변화로 인한 폭우, 잦은 태풍 등으로 폐광이 무너지는 경우가 많아 유출수 관리에 엄격해야 한다는 지적이다.

국회 산업통상자원중소벤처기업위원회 황운하 의원(대전 중구, 더불어민주당)이 한국광해관리공단으로부터 제출받은 자료에 따르면, 공단이 관리하고 있는 폐광 56곳 중 13곳(23%)의 유출수에 기준치 이상의 아연, 망간, 구리 등의 중금속이 함유돼 있었다.

광해공단은 강원, 전남, 경북 등 지사를 통해 전국 56개 폐광의 수질정화시설을 운영·관리하고 있다. 이중 기준치 초과 중금속을 포함한 유출수가 방류되고 있는 곳은 강원 7곳, 경북 3곳, 대구 1곳, 부산 1곳, 경남 1곳이다.

이중에는 환경부 장관이 ‘매우좋음’ 등급 정도의 수질을 보전해야 한다고 지정한 ‘청정지역’이 6곳이나 포함되어 있다.

경북 문경에 위치한 석봉탄광의 경우 청정지역 수준의 수질을 유지해야 하나, 파킨슨병을 유발할 수 있는 망간이 기준치의 4.5배 이상 검출됐다. 이 광산은 낙동강으로 합류되는 하천과 불과 1.2km에 위치해 있다.

비소·카드뮴··· 섬뜩한 ‘중금속 범벅’ 폐광산, 부산에 더 있다

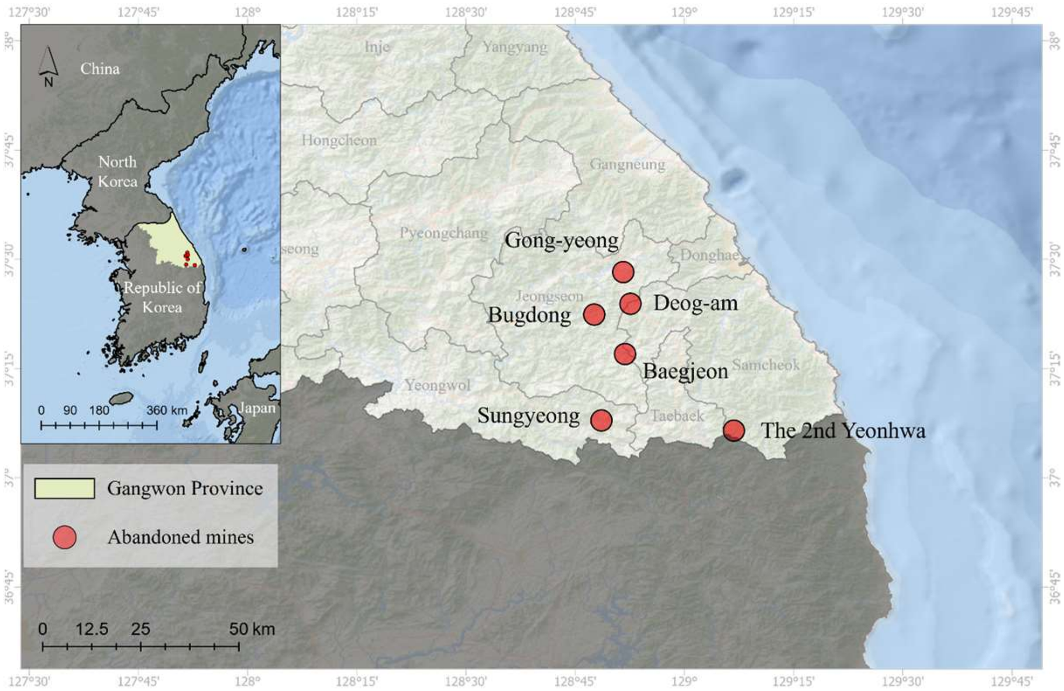
이승훈 기자 lee88@busan.com, 남형욱 기자 thoth@busan.com

입력 : 2021-05-05 17:00:00 수정 : 2021-06-01 14:45:49 게재 : 2021-05-05 14:57:49 (03면)

부산 기장군 철마면 임기납석광산에서 유출된 황철석 등 성분으로 인해 인근 하천 주변이 붉게 산화돼 있다. 남형욱 기자 thoth@

Methods

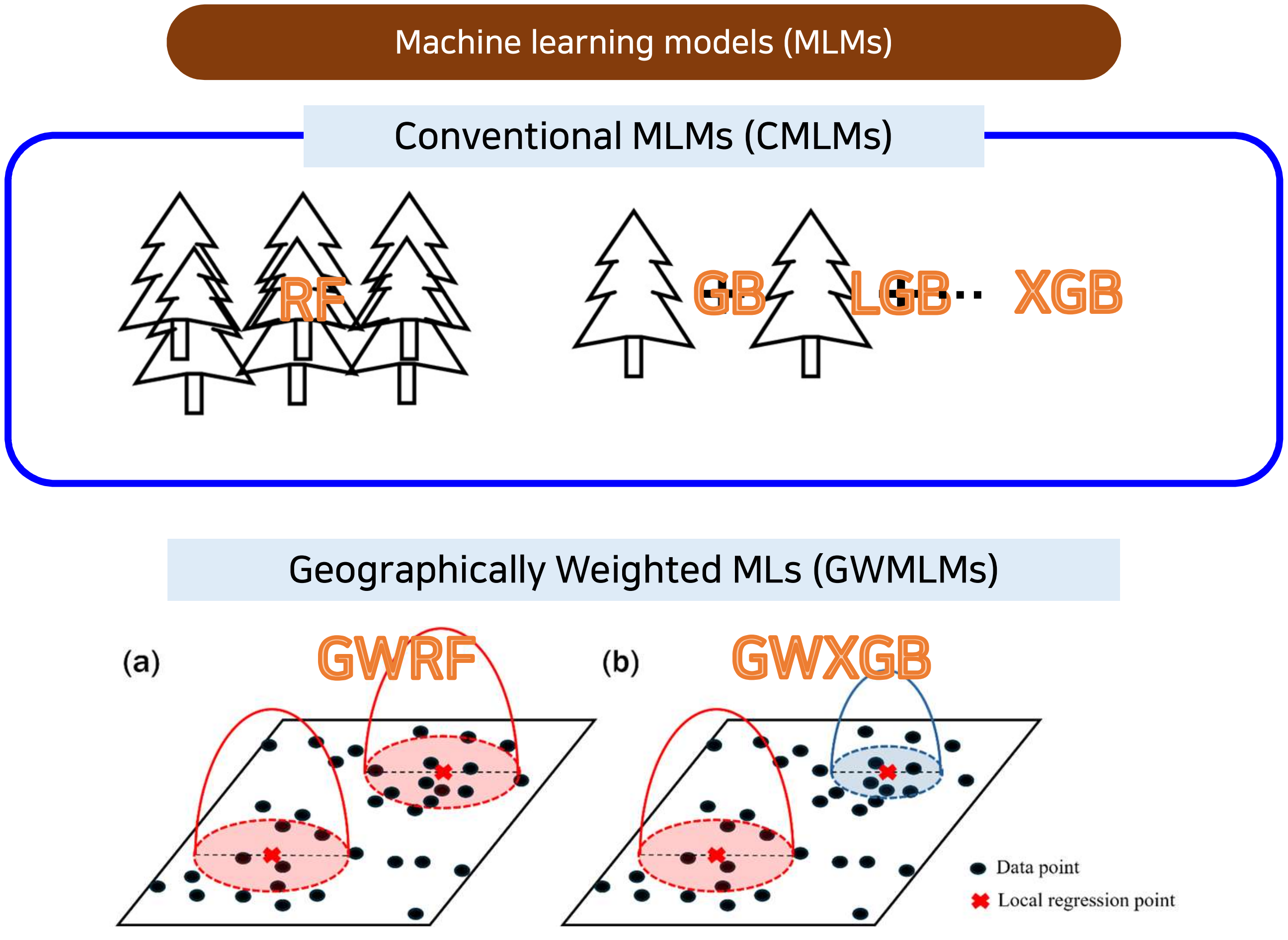
Study sites and input data



	Variables
Input Geographic (4) & Soil (13)	1) Distance from mining site to sampling point, 2) Slope, 3) Elevation, and 4) Distinction between topsoil and subsoil
	1) soil gravel content, 2) soil HSGs, 3) clay, 4) silt, 5) soil, 6) pH, 7) SOC, 8) CEC, 9) Ca, 10) Mg, 11) K, 12) Na, and 13) BSP
Target	Cadmium [Cd] (mg/kg)
	Lead (Pb) (mg/kg)

Train : Test = 70% : 30% (n = 401)
5-fold cross validation

Methods


Machine learning models (MLMs)
Conventional MLMs (CMLMs)
RF
GB
LGB
…
XGB
Geographically Weighted MLs (GWMLMs)
(a)
GWRF
(b)
GWXGB
Data point
Local regression point

Results

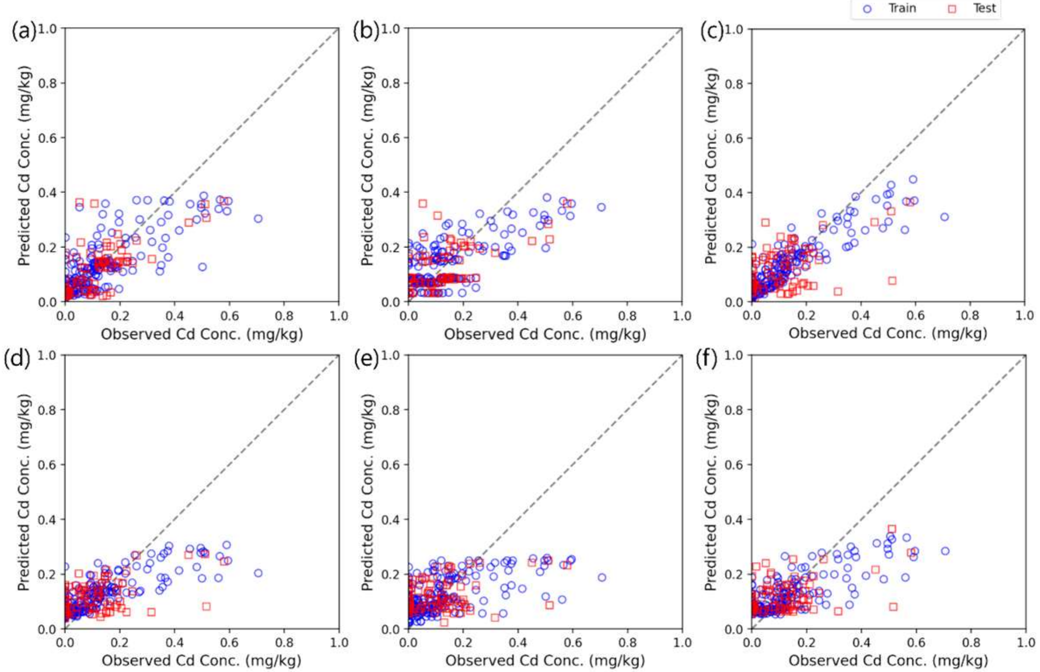
Performance measures

	GWMLMs		CMLMs			
	GWRF	GWXGB	RF	GB	LGB	XGB
Train R^2 for Cd	0.6	0.56	0.76	0.52	0.33	0.52
Test R^2 for Cd	0.5	0.37	0.21	0.22	0.07	0.19
Train R^2 for Pb	0.73	0.87	0.59	0.65	0.90	0.76
Test R^2 for Pb	0.50	0.52	0.36	0.35	0.26	0.38

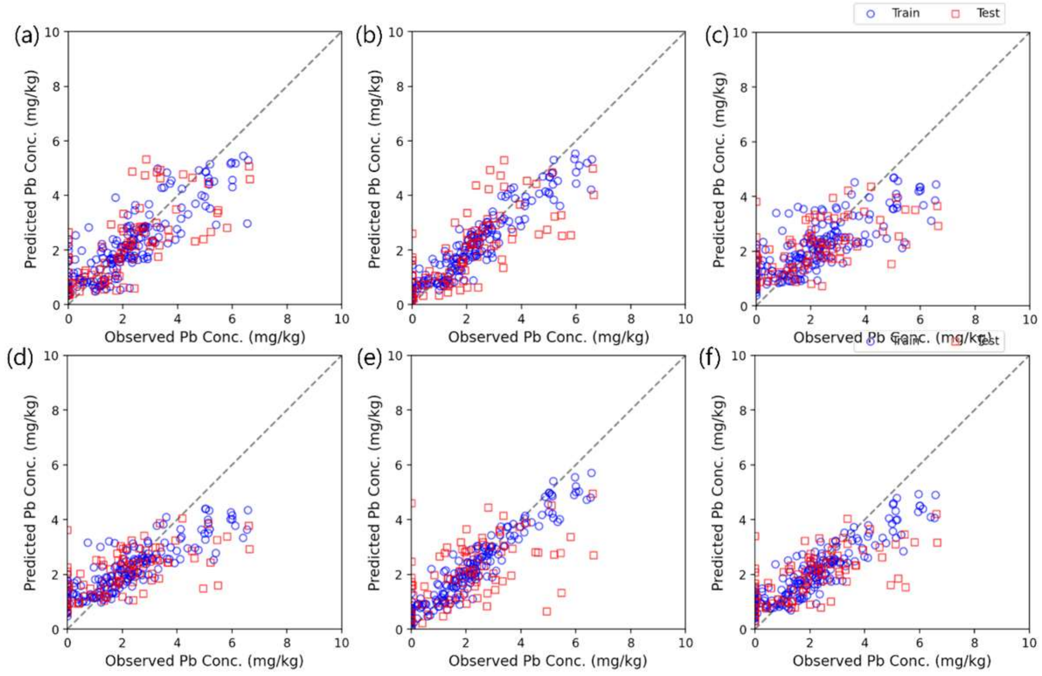
RF: Random Forest; GB: Gradient Boosting; LGB: Light Gradient Boosting; XGB: Extreme Gradient Boosting

Results

Cd



Pb



(a) GWRF; (b) GWXGB; (c) RF; (d) GB; (e) LGB; (f) XGB

Thank you