디지털 물산업 혁신인재 양성사업

2차년도 해외 네트워크 탐방 프로그램 지원

논문 리뷰 및 연구 내용 적용방안 제시

기상학적 요인이 원수 수질에 미치는 영향 확인 및 ML 기반 모델을 활용한 수질 항목 변화 예측을 중심으로

Subin Jang, Nahyun Lee
Dept. of Environmental Engineering, University of Seoul

CONTENTS

Introducing **Our Team**

장수빈

환경공학부 22학번
(3-2 재학)

2022학년도 서울시립대학교 환경공학부 입학
제4기 서울시 도시미래인재양성프로젝트 선발
서울시립대학교 디지털 물산업 혁신인재 양성사업 **참여('23.08 ~)**
2023 사회적경제 공학아이디어 경진대회 **장려상 수상**
2023 환경독성·보건 분야 프로젝트 공모전 **환경보건센터장상(금) 수상**
2024 1학기 서울시립대학교 학생포상 **학술상 수상**
2024 The 17th National University Student Social Practice and Science Contest on Energy Saving & Emission Reduction **동상 수상**
2024 환경독성 · 보건분야 과학소통 아이디어 공모전 **환경독성보건학회장상(은) 수상**
디지털 물산업 마이크로 디그리 **수료(예정)**

이나현

환경공학부 22학번
(3-2 재학)

2022학년도 서울시립대학교 환경공학부 입학
서울시립대학교 디지털 물산업 혁신인재 양성사업 **참여('23.08 ~)**
상하수도시스템 연구실 미래설계학기 연구인턴십 트랙 **참여('24.09~)**
디지털 물산업 마이크로 디그리 **수료(예정)**

활동 목표

Goal of Participation

Introduction & Purpose

조사 논문 Christian Ortiz-Lopez, et al*., 『**Ensemble machine learning using hydrometeorological information to improve modeling of quality parameter of raw water supplying treatment plants**』, Journal of Environmental Management, Vol. 362, 2024.

하천, 호소수 등에서 **기상학적 사건(강우 등)은 수질에 영향**을 줌
→ 이의 예측을 도와주는 도구 필요

조기 경보 시스템(Early Warning Systems)은 이러한 상황 대응에 유용
→ EWSs에는 단일 알고리즘 ML 기법이 사용되어 성능 제한적

원수수질 변동을 감지하여 **효과적인 정수장 운영**에도 필요
→ ex) 원수의 탁도, 유기물 농도에 따라 약품 주입량 산정

기상 및 수문학적 데이터를 활용하여
음용수 처리 시설에 공급되는 원수 수질을
결정하는 **매개변수를 모델링**하는 기법 제안

단일 알고리즘이 아닌 RF, GB, XBG**의 앙상블
기법 적용하여 정확성 향상하는 것이 핵심

*저자: Christian Ortiz-Lopez[A], Christian Bouchard[A], Maneul J. Rodriguez[B]
[A] Center Research Planning and Development, Laval University, Quebec, Canada
[B] School of Regional Planning and Development(ESAD), Laval University, Quebec, Canada

**RF: Random Forest, GB: Gradient Boosting, XBG: Extreme Gradient Boosting

<Methods> 1. Study Area

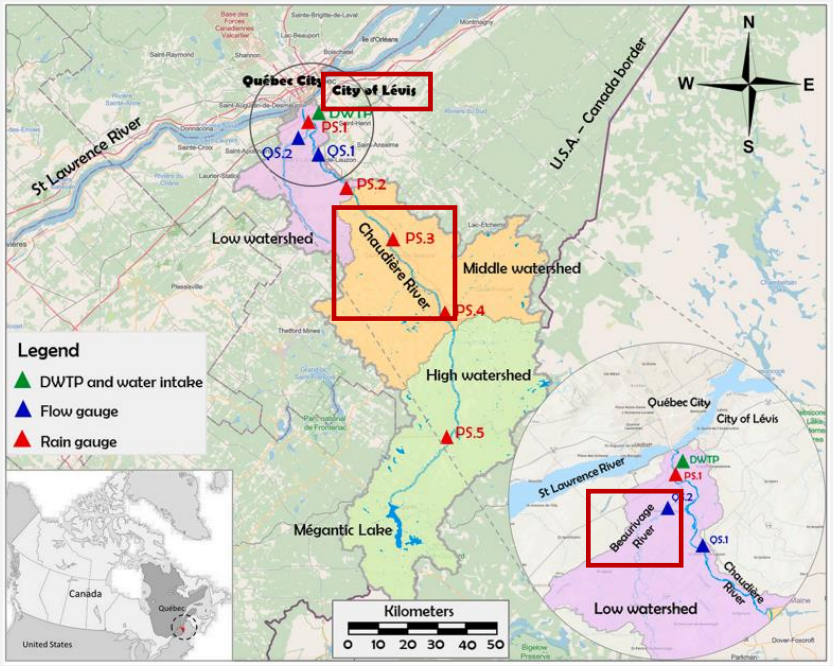

Fig 1. Chaudiere 강 유역 모식도

- 캐나다 레비스 시의 음용수 공급망을 연구 대상으로 설정
- Chaudiere 강물을 끌어오는 음용수 처리 시설 포함
- 유역은 숲 67%, 농경지 20%, 도심 4%로 구성
- 연간 평균 수온 2.9°~4.6°
- Beaurivage 강은 원수 수질에 중요한 영향을 미치는 하위 유역

2. Data Selection

- 강우계(기상 정보)와 유량계(유속), 음용수 처리 시설(수질)에서 수집된 데이터 사용
- **강우계**: '17.05.~'17.10. 시간 단위 강수량 정보 수집(PS.1~PS.5)
- **유량계**: Chaudiere, Beaurivage 두 강의 수류 속도 측정(QS.1~QS.2)
- **음용수 처리 시설**: 원수의 **탁도와 UV 흡광도** 측정, 실시간 분석기와 분석 광도계로 측정
- 누락 데이터는 Kalman smoothing, ARIMA 모델로 처리
- 탁도가 UV254 측정에 미치는 간섭은 보정 기능으로 제거

3. Methodology

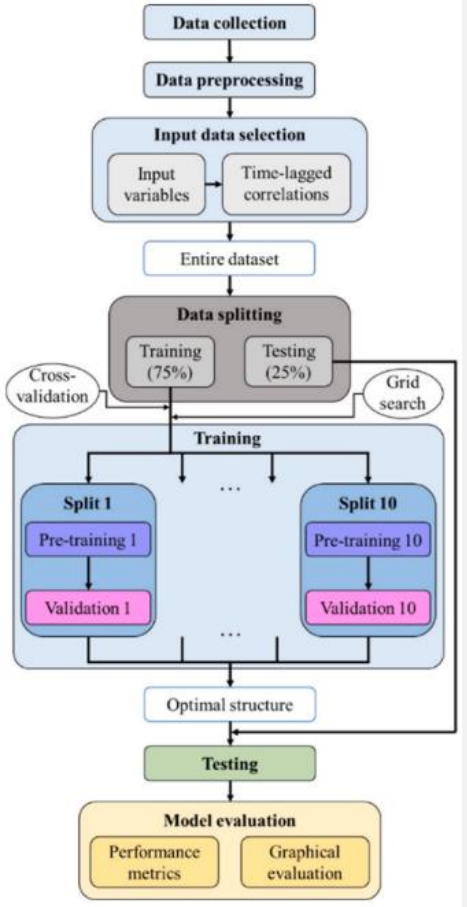

Fig 2. Framework of methodology

4.Input Data Selection

1. 강수량과 하천 유량을 입력 변수로 선택

- 하천 유량은 강수에 대한 유역의 반응을 통합한 변수로 간주되어 원수 수질 예측에 사용
- 5개 강우계와 2개의 유량계 데이터를 사용하여 각각 강수량과 하천 유량 표현

2. 선택한 입력 변수를 음용수 처리 시설에서 측정한 출력 변수(탁도, UV254)와 연결

- 연구에 따르면 강수 후 탁도와 UV254의 피크가 발생하며, 일정 시간 후 수치가 기준 수준으로 돌아옴
- 즉 강수량과 하천 유량이 원수 품질에 시간 지연을 두고 영향을 미치는 것
- Ortiz-Lopez et al.(2023)의 방법론을 사용하여 입력 변수와 출력 변수 간의 지연 시간을 계산
- 약 4,300개의 시간별 관측치로 구성된 데이터셋 분석
- Spearman의 순위 상관 계수를 사용하여 1시간의 시간 지연으로 변수 간의 상관관계를 계산

5. Modeling Techniques

유역의 강수량과 하천 유량 시계열을 입력 변수로, 원수 수질의 두 매개변수(탁도, UV254)를 모델링

- 세 가지의 decision tree based 앙상블 기계학습 기법(RF, GB, XGB) 사용

Decision Tree는 데이터를 순차적으로 분할하는 알고리즘

- 데이터 해석에 용이하나, 보지 못한 데이터에 대해서는 높은 분산을 보이는 단점
- 이의 해결을 위해 앙상블 기법을 사용하여 여러 단순 모델을 집계하고 결합

*배깅(Bootstrap aggregating): 재샘플링을 통해 여러 하위 집합을 생성하고, 각 모델이 예측을 평균화하여 최종 예측을 수행

*부스팅: 약한 모델을 반복적으로 결합하여 강력한 예측 모델을 생성하며, 편향과 분산을 정상화하는 데 초점을 맞춤

5. Modeling Techniques(이어서)

Random Forest (RF)	$\hat{f}_{bag}(x) = \frac{1}{M}\sum_{m=1}^{M} DT_{(x)}$	• RF는 여러 결정 트리를 사용하여 회귀 및 분류 문제를 해결하는 배깅 기법 • 부트스트랩 샘플링을 통해 무작위로 데이터 집합을 생성하고, 각 트리에서 예측 변수를 무작위로 선택하여 분할 수행하여 예측 간 상관관계를 줄이고 분산 감소
Gradient Boosting (GB)	$\hat{f}(x) = ip + \sum_{m=1}^{M} \lambda \hat{f}^{m}(x)$	• GB는 수정된 초기 데이터셋을 기반으로 DT를 순차적으로 구축하는 부스팅 기법 • 모델은 최소제곱오차의 합을 최소화하는 초기 상수 값으로 시작, 각 단계에서 잔차를 계산하여 회귀 DT를 적합시키며, 이전 트리의 정보를 사용하여 예측값 업데이트
eXtreme Gradient Boosting (XGB)	• XGB는 최적화된 GB 라이브러리로, 초기 예측 후 잔차를 계산하여 DT를 구성 • 정규화 매개변수로 과적합을 방지하고, 이득을 계산하여 다양한 임계값을 평가하며, 최소손실감소량을 기반으로 가지치기 수행	

Data splitting & hyperparameter tuning

모델 적용 전, 데이터셋의 관측치는 무작위로 **75%는 훈련용, 25%는 테스트용**으로 나뉨

- RF: 후보 변수의 수(m)는 탁도 모델에서 5, UV254 모델에서 6으로 설정
- GB: 나무의 수(M)=6,000, 학습률(λ)=0.30, 최대 트리 깊이(db)=7, 최소 관측치 수(notn)=5로 설정
- XGB: 최대 깊이(md), 최소 손실 감소(γ), L2 및 L1 정규화 매개변수가 각각 탁도 모델에서는 md=10, γ=5, L2=2, L1=4로, UV254 모델에서는 md=12, γ=5, L2=4, L1=3으로 설정

Model evaluation

- R square
- NSE Coefficient
- RMSE
- MAE

Result 1
: 원수 수질 매개변수의 시계열 분석

연구 기간동안 원수의 탁도와 UV254, 강우 누적량, 하천 유량에서 큰 변동이 관찰됨.

하천 유량 변화를 기준으로 계절적 시기를 두 갈래로 구분함.

- **눈 녹는 시기(4월 중~5월 초)**: 이 시기에는 최대 하천 유량이 487m^3/h에 도달함. 이러한 높은 하천 유량은 눈과 얼음이 녹는 현상과 연관됨.
- **따뜻한 시기(5월 말~10월 말_시계열 종료 시점)**: 이 시기에는 주로 강우와 지하수 기저 유량으로 인해 하천 유량의 급증 발생.

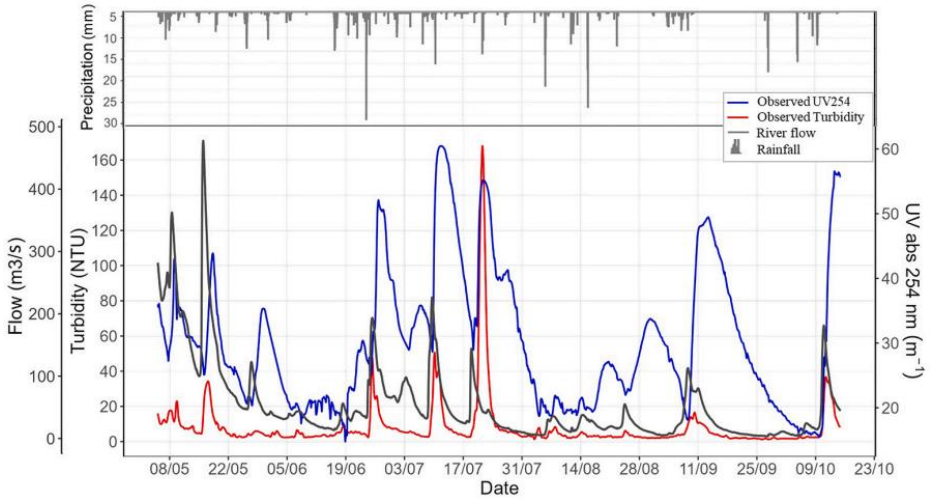

Fig 3. Hourly variation for raw water turbidity and UV254 measured at the DWTP water intake, hourly river flow measured at QS.1 near the water intake and hourly rainfall accumulation measured at PS.3 in the middle watershed.

Result 2
: 원수 탁도와 UV254 모델링

RF, GB, XGB 세 가지의 앙상블 머신러닝 기법을 사용함.

모든 모델은 hyperparameter 튜닝을 통해 최적화됨.

UV254 예측은 탁도 예측보다 더 나은 성능을 보였는데, 이는 UV 데이터가 탁도 데이터에 비해 극단적인 값이 적고 균일했기 때문.

• 탁도 예측

RF는 낮은 값부터 높은 값까지 정확히 예측하며 가장 좋은 일반화 능력을 보임.

일부 매우 낮거나 높은 값을 예측하는 것에 한해서 오차가 발생.

GB, XGB도 비슷한 수준. XGB는 RF보다 나은 탁도 peak 예측 성능을 보임.

지표로 따졌을 때, **RF는** 가장 높은 R^2값과 낮은 RMSE, MAE 값을 기록하며 **탁도 예측에서 가장 우수**했음.

• UV254 예측

RF는 UV254예측에서도 가장 우수한 성능을 보였음.

지표로 따졌을 때, 가장 높은 R^2값과 낮은 RMSE, MAE 값을 기록.

GB는 RMSE와 MAE 값이 가장 높아 **UV254 예측 성능이 상대적으로 낮았음**.

XGB는 RF와 유사 수준의 성능을 보였고, 높은 R^2값과 낮은 RMSE, MAE 값 기록.

Table 1. Performance metrics of ensemble techniques for raw water turbidity modeling on the test and validation datasets.

Ensemble technique	R^2_{cal}	R^2_{test}	NSE_{cal}	NSE_{test}	$RMSE_{cal}$ [NTU]	$RMSE_{test}$ [NTU]	MAE_{cal} [NTU]	MAE_{test} [NTU]
RF	0.97	0.87	0.94	0.75	3.64	9.11	0.99	2.45
GB	0.99	0.80	0.99	0.64	0.35	10.91	0.22	3.47
XGB	0.99	0.81	0.99	0.65	1.56	10.71	0.89	2.62

Table 2. Performance metrics of ensemble techniques for raw water UV254 modeling.

Ensemble technique	R^2_{cal}	R^2_{test}	NSE_{cal}	NSE_{test}	$RMSE_{cal}$ [m^{-1}]	$RMSE_{test}$ [m^{-1}]	MAE_{cal} [m^{-1}]	MAE_{test} [m^{-1}]
RF	0.98	0.89	0.96	0.80	2.17	5.00	1.25	2.94
GB	0.99	0.85	0.99	0.73	0.88	5.80	0.64	3.80
XGB	0.98	0.88	0.96	0.77	2.09	5.32	1.33	3.31

Discussion

모델 성능 및 한계

연구에서 사용된 앙상블 모델들은 원수 수질항목 예측에서 우수한 성능 보임
탁도와 UV254 예측에서, low peak, mid peak 모두에서 높은 정확도 보임
RF가 가장 우수했음. 하지만 high peak에 대한 예측 정확도는 낮았는데, 사례 부족을 원인으로 추정

기존 연구와의 비교

선행 연구에서 많이 사용된 단일 ML 기법 대비 앙상블 모델은 높은 R^2값을 기록하고, 일반화 능력이 좋았으며, 데이터 변환 기법 없이도 우수한 성능을 발휘함
본 연구는 시간 단위 데이터와 기상 및 수문학적 데이터를 사용하여, 기존의 일별/월별 데이터에 의존한 연구보다 더욱 정밀한 결과 도출

모델의 실무적 시사점

정수처리장의 조기 경보 시스템을 위한 모델로의 활용 가능성 기대
강우 후 수질항목의 급격한 변화를 사전 탐지, 처리 약품 투입량을 조정하는데 기여 가능
이는 정수처리장 운영 효율 향상과 이어짐

추가 개선 방향

다음 요소(선행 강우량, 토양 상태, 유역 내 토지 이용 현황 등)를 추가로 고려한 연구가 필요할 것.
특정 시점의 수질을 보는 정적 모델링에서 나아가, n시간 후의 수질을 예측하는 동적 모델링을 통해 조기 경보 시스템의 실효성을 강화할 수 있을 것

동일 분야 선행 연구는 어땠을까?

국내 기준 유사 연구 현황은 어떻고, 무엇을 다루었는지 그 개요를 살펴보자.

입력자료 군집화에 따른 앙상블 머신러닝 모형의 수질 예측 특성 연구

The Effect of Input Variables Clustering on the Characteristics of Ensemble Machine Learning Model for Water Quality Prediction(박정수, 2021)

KMC(K-Means Clustering)을 이용하여 입력자료의 특성에 따른 군집화를 수행한 후 앙상블 머신러닝 모형인 GBDT(Gradient Boosting Decision Tree) 알고리즘 중 하나인 XGB를 이용하여 하천의 SSC(suspended sediment concentration) 를 예측하는 모형을 구축.

- 독립변수: 하천 유량
- 종속변수: SSC
- RMSE 및 RSR을 이용하여 모델 성능 평가
- 입력자료의 특성을 고려하지 않은 model 2, 군집화를 한 model 1을 구분하여 RSR 평과 결과, model 1에서 개선된 성능을 보임

→ 입력자료의 특성을 고려한 접근을 통해 머신러닝 모델의 성능 개선이 가능함을 확인

데이터 불균형 개선에 따른 탁도 예측 앙상블 머신러닝 모형 성능 특성

Performance Characteristics of an Ensemble Machine Learning Model for Turbidity Prediction With Improved Data Imbalance(양현석, 박정수, 2023)

앙상블 머신러닝 알고리즘 중 하나인 LightGBM (light gradient boosting machine)을 이용하여 탁도를 예측하는 다중 분류 모형을 구축.

- 독립변수: 수온, pH, 전기전도도, 용존산소량, 유량
- 종속변수: 탁도 (class 1 : 10NTU 미만 / class 2 : 30NTU 미만 / class 3 : 100NTU 미만 / class 4 : 100NTU 이상)
- 입력자료를 일정한 비율로 SMOTE를 적용하여 모형의 예측성능을 비교 (다수 class와 소수 class의 데이터 수 차이로 인한 불균형을 해소)
- 모든 class에 동일하게 SMOTE를 적용하는 것보다 일정한 비율로 적용하는 것이 더 향상된 모델의 성능을 보임

→ 데이터불균형의 해소를 통한 모형성능의 향상이 가능할 것

해당 연구를 어떻게 현장 적용할 수 있을까?

기상 사건 등의 외부 요인이 원수 수질에 미치는 영향의 근거를 **확인**하고,
다루어진 유형의 모델이 국내에서는 어떻게 활용될 수 있을지 고안하였다.

수질 항목의 시공간적 변동 경향성

- BOD5: 단발성 강우 후 급증, 연속 강우 후 급감하는 경향(수위 변화와 Chl-a 증가가 증가 요인으로 작용)
- CODMn: BOD5와 유사한 변동 양상
- T-N: 여름철에 낮고, 겨울철에 높은 경향
- T-P: 이의 변동은 강수량과 유입량에 양의 상관관계
- SRP: 강수량 및 방류량과 밀접한 관계
- T, N: 질소는 강수량이 적었던 시기에 큰 변동, 인은 강수량이 많았던 시기에 뚜렷한 변화
- TSS: 강수량과 양의 상관관계

하천, 호소수 등에 대한 기상 · 수문학적 영향(팔당호 예시)

- 온대 몬순 지역에 위치
- 여름에는 장마와 태풍으로 인한 큰 강도의 강우가 특징
- 집중 호우 발생 빈도 증가
- 겨울에는 강풍에 의한 건조가 특징

✓기후적 요인은 하천과 저수지 수질에 유의한 영향을 미침을 확인

✓강수량의 급격한 증가는 오염물질 유입에 기여

✓수온과 유량 변화에 따른 큰 수질 변동이 있을 수 있음(성층화, 영양염 펄스 등)

기상 · 수문학적 이해는 효과적인 수질 관리에 필수적이며, 이는 국내 환경에서도 유효함.

적용 가능성 확인

한국은 강우량, 유량, 그리고 수질 데이터를 상세히 측정하고 있으며, 이 데이터는 본 연구에서 다뤄진 모델이 추가적으로 필요로 하는 것과 일치함.

특히, 강우량 및 강 유량 데이터는 이미 환경부와 각 지방 자치단체에서 운영하는 수질 및 유량 모니터링 시스템에 의해 수집되고 있어 유사한 모델을 향상시켜 구현하기에 용이.

- 이는 한국의 주요 상수원인 한강, 낙동강 등의 유역에서 발생하는 수질 변화에 대한 파악 및 조기 대응에도 유용하게 활용될 수 있을 것으로 기대.
- 모델의 조건을 바꾸어, 하천 외의 정적인 수원에 대해서도 적용할 수 있을 것.
- 기후변화로 인한 기상 이벤트에 따른 대응을 강화할 수 있을 것.

적용 방안 제안

조기 경보 시스템 구축

- 한국의 상수원 데이터를 활용한 앙상블 머신러닝 모델 개발
- 정수처리장 운영 시스템과 연계하여 실시간 예측 데이터 제공
- 집중호우 발생 시 탁도 예측을 통해 약품 투입량 조정 자동화

한국 맞춤형 모델 최적화 예시

- 토양 조건: 한국의 지역별 토양 특성을 모델에 반영
- 선행 강우 조건: 강우 전 유역의 포화 상태를 추가 변수로 포함
- 유입 오염원: 특정 유역에서 발생할 수 있는 산업 기인 오염원 데이터 통합
- 기대 효과: 수질 안정성 확보, 정수장 운영 효율성 향상

이밖에도 ….

- 생태계 변화 예측 등에도 사용될 수 있을 것
- 탁수의 변화, 수생 생물의 서식지와 생태계에 직접적인 영향을 미침
- 탁도와 영양소 농도 간의 관계를 분석함으로써, 특정 지역에서의 부영양화 현상을 조기에 발견하고 대응

Reference

1. Christian Ortiz-Lopez, Christian Bouchard, Manuel J. Rodriguez, 『Ensemble machine learning using hydrometeorological information to improve modeling of quality parameter of raw water supplying treatment plants』, Journal of Environmental Management, Vol. 362, 2024.
2. 권용수, 배미정, 김준수 외 3인, 『우리나라 주요 호소의 수질 변동 경향성 분석 및 유형화』, The Korean Society of Limnology, Vol.47, No.3, pp 146-159, 2014.
3. 박정수, 『입력자료 군집화에 따른 앙상블 머신러닝 모형의 수질예측 특성 연구』, Journal of Korean Society on Water Environment, Vol.37, No.5, pp 335-343, 2021.
4. 양현석, 박정수, 『데이터 불균형 개선에 따른 탁도 예측 앙상블 머신러닝 모형의 성능 특성』, Ecology and Resilient Infrastructure, Vol.10, No.4, pp 107-115, 2023.
5. 유호식, 『한강수질에 영향을 끼치는 요인들의 통계분석』, Journal of Korean Society of Environmental Engineers, Vol.24, No.12, pp 2139-2150, 2002.
6. 황순진, 김건희, 박채홍 외 7인, 『남・북한가오가 경안천 합류 수역 팔당호의 수질 변동성에 대한 기상・수문학적 영향』, The Korean Society of Limnology, Vol.49, No.4, pp 354-374, 2016.
7. 황순진, 심연보, 최봉근 외 6인, 『북한강 의암호의 수질 변동성에 대한 강우・수문학적 비교 분석』, The Korean Society of Limnology, Vol.50, No.1, pp 29-45, 2017.

디지털 물산업 혁신인재 양성사업

2차년도 해외 네트워크 탐방 프로그램 지원

Thank you for your participation!

Subin Jang, Nahyun Lee
Dept. of Environmental Engineering, University of Seoul