Simulation of Total Organic Carbon (TOC) Using Integration of SWAT-C Model and Machine Learning: A Case Study of the Bogangcheon Watershed

2025.06.25.

Master's Course in Regional Environmental Engineering
Kangwon National University, South Korea

Changyeom Kim

SWAT Soil & Water Assessment Tool

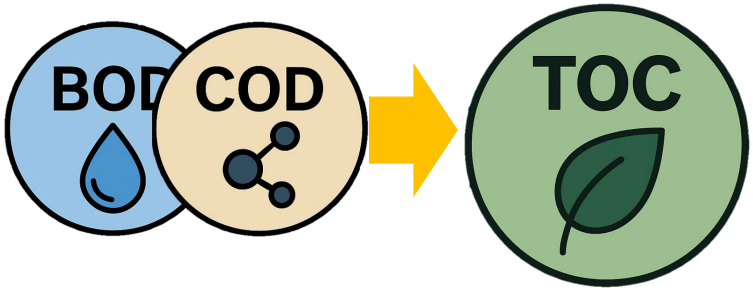
CONTENTS

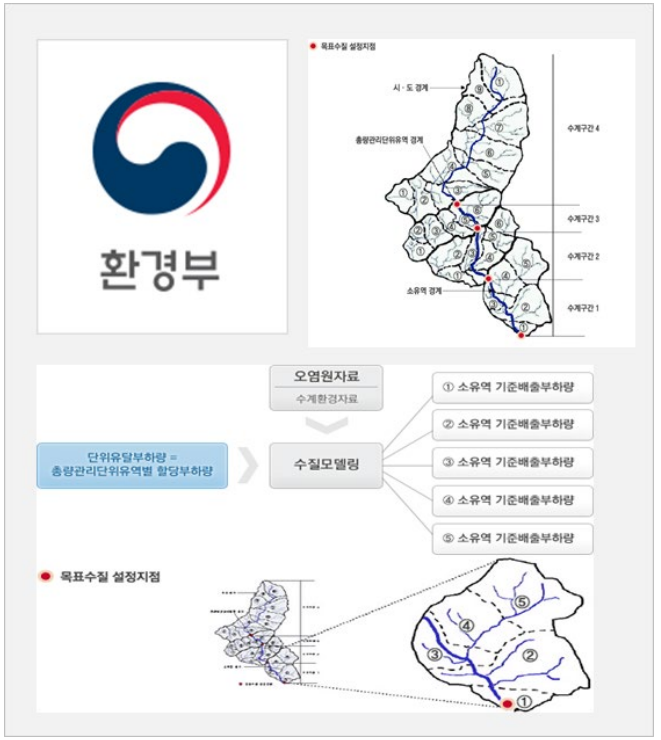
I. Introduction

❖ Research Background

<Non- point Source Pollution>

<Organic Pollutants>





- A precise water quality assessment is necessary due to the rise in non-point pollution and varied organic inputs.
- Current water quality indicators, such as BOD and COD, are insufficiently representative of diverse organic matter loads.
- The Ministry of Environment introduced Total Organic Carbon(TOC) as a new indicator for water quality monitoring and is conducting pilot projects that apply TOC through unit load management under the TMDL system.

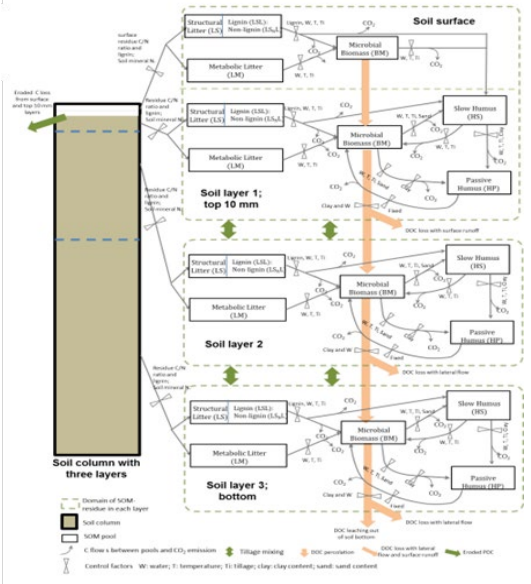
I. Introduction

❖ SWAT-C for TOC Simulation

SWAT-C(Soil and Water Assessment Tool-Carbon)

Soil & Water Assessment Tool | SWAT

<SWAT Model>



<Carbon Cycling Module>

- SWAT-C is a modified version of the SWAT (Soil and Water Assessment Tool) model that integrates a carbon cycling module.
- It enables TOC simulation by representing hydrological and biogeochemical processes at the watershed scale.

I. Introduction

❖ Previous Studies Utilizing Machine Learning for Water Quality Prediction

J. Korea Water Resour. Assoc. Vol. 54, No. 2 (2021), pp. 121-133
doi: 10.3741/JKWRA.2021.54.2.121
pISSN 1226-6280
eISSN 2287-6138

Development of suspended solid concentration measurement technique based on multi-spectral satellite imagery in Nakdong River using machine learning model

Kwon, Siyoon* · Seo, Il Won†* · Beak, Donghae*

*Ph.D Course, Department of Civil and Environmental Engineering, Seoul National University, Seoul, Korea
†Professor, Department of Civil and Environmental Engineering, Seoul National University, Seoul, Korea
*Ph.D researcher, Department of Civil and Environmental Engineering, Seoul National University, Seoul, Korea

Paper number: 20-115
Received: 18 December 2020; Revised: 15 January 2021; Accepted: 15 January 2021

Abstract

Suspended Solids (SS) generated in rivers are mainly introduced from non-point pollutants or appear naturally in the water body, and are an important water quality factor that may cause long-term water pollution by being deposited. However, the conventional method of measuring the concentration of suspended solids is labor-intensive, and it is difficult to obtain a vast amount of data via point measurement. Therefore, in this study, a model for measuring the concentration of suspended solids based on remote sensing in the Nakdong River was developed using Sentinel-2 data that provides high-resolution multi-spectral satellite images. The proposed model considers the spectral bands and band ratios of various wavelength bands using a machine learning model, Support Vector Regression (SVR), to overcome the limitation of the existing remote sensing-based regression equations. The optimal combination of variables was derived using the Recursive Feature Elimination (RFE) and weight coefficients for each variable of SVR. The results show that the 705nm band belonging to the red-edge wavelength band was estimated as the most important spectral band, and the proposed SVR model produced the most accurate measurement compared with the previous regression equations. By using the RFE, the SVR model developed in this study reduces the variable dependence compared to the existing regression equations based on the single spectral band or band ratio and provides more accurate prediction of spatial distribution of suspended solids concentration.

Keywords: River suspended solid, Remote sensing, Multi-spectral satellite imagery, Support vector machine (SVR), Recursive feature elimination (RFE), Spatial monitoring

기계학습모형을 이용한 다분광 위성 영상 기반 낙동강 부유 물질 농도 계측 기법 개발

권시윤* · 서일원†* · 백동해*

*서울대학교 건설환경공학부 박사과정, *서울대학교 건설환경공학부 교수, *서울대학교 건설환경공학부 박사후 과정

요 지

하천에서 발생하는 부유 물질은 주로 유역으로부터 유입되거나 하천 내에서 자생으로 발생하기도 하며, 퇴적되어 중장기적인 수질 오염을 초래할 수도 있는 중요한 수질 인자이다. 하지만, 부유물질의 계측방식은 점 단위 계측이기 때문에 노동 집약적이며 방대한 양의 자료를 취득하기는 어렵다. 따라서, 본 연구에서는 고해상도 다분광 위성영상을 제공하는 Sentinel-2 위성 자료를 이용하여 낙동강 전역에 대한 원격탐사 기반 부유 물질 농도 계측 기법을 개발하였다. 개발된 기법은 기존 원격탐사 기반 회귀식들의 한계점을 개선하고 낙동강 전체 영역의 지역적 특성을 반영하기 위해 기계학습 모형인 서포트 벡터 회귀(Support Vector Regression, SVR) 모형을 이용하여 다양한 파장대의 분광 밴드들과 밴드비(band ratio)를 고려하였으며, 이를 입력 변수들의 최적 조합으로 제귀적 특징 제거법(Recursive Feature Elimination, RFE)과 SVR의 각 변수별 가중계수를 활용하여 도출하였다. 가장 중요도가 높은 분광 밴드로는 Red-edge 파장대 영역의 속하는 705 nm 밴드가 산출되었으며, 최종적으로 구축된 SVR 모형을 선행 연구들에서 제시한 회귀식들과 비교한 결과, 가장 정확한 계측 결과를 제공하는 것으로 분석했다. 본 연구에서 개발된 SVR 모형은 RFE를 통해 산출된 최적 분광 밴드 조합을 바탕으로 하기 때문에 기존 단일 분광 밴드 혹은 밴드비를 기반으로 구축된 회귀식들이 가지는 변수 의존도를 낮추는 동시에 더욱 정확한 부유물질 농도 공간분포를 제공할 수 있을 것으로 판단된다.

핵심용어: 하천 부유물질, 원격탐사, 다분광 위성 영상, 서포트 벡터 회귀(SVR), 재귀적 특징 제거법(RFE), 공간적 모니터링

*Corresponding Author. Tel: +82-2-880-7345
E-mail: seoilwon@snu.ac.kr (I. W. Seo)

J. Korea Water Resour. Assoc. Vol. 51, No. 10 (2018), pp. 863-874
doi: 10.3741/JKWRA.2018.51.10.863
pISSN 1226-6280
eISSN 2287-6138

Assessment of climate change impact on aquatic ecology health indices in Han river basin using SWAT and random forest

Woo, So Young* · Jung, Chung Gil†* · Kim, Jin Uk* · Kim, Seong Joon*

*Department of Civil, Environmental and Plant Engineering, Konkuk University

Paper number: 18-061
Received: 2 August 2018; Revised: 4 September 2018; Accepted: 4 September 2018

Abstract

The purpose of this study is to evaluate the future climate change impact on stream aquatic ecology health of Han River watershed (34,148 km²) using SWAT (Soil and Water Assessment Tool) and random forest. The 8 years (2008~2015) spring (April to June) Aquatic ecology Health Indices (AHI) such as Trophic Diatom Index (TDI), Benthic Macroinvertebrate Index (BMI) and Fish Assessment Index (FAI) scored (0~100) and graded (A~E) by NIER (National Institute of Environmental Research) were used. The 8 years NIER indices with the water quality (T-N, NH_4-N, NO_3-N, T-P, PO_4) showed that the deviation of AHI score is large when the concentration of water quality is low, and AHI score had negative correlation when the concentration is high. By using random forest, one of the Machine Learning techniques for classification analysis, the classification results for the 3 indices grade showed that all of precision, recall, and f1-score were above 0.81. The future SWAT hydrology and water quality results under HadGEM3-RA RCP 4.5 and 8.5 scenarios of Korea Meteorological Administration (KMA) showed that the future nitrogen-related water quality in watershed average increased up to 43.2% by the baseflow increase effect and the phosphorus-related water quality decreased up to 18.9% by the surface runoff decrease effect. The future FAI and BMI showed a little better Index grade while the future TDI showed a little worse index grade. We can infer that the future TDI is more sensitive to nitrogen-related water quality and the future FAI and BMI are responded to phosphorus-related water quality.

Keywords: Aquatic ecology health, Climate change, Water quality, SWAT, Random forest

SWAT 및 random forest를 이용한 기후변화에 따른 한강유역의 수생태계 건강성 지수 영향 평가

우소영* · 정충길†* · 김진욱* · 김성준*

*건국대학교 공과대학 사회환경플랜트공학과

요 지

본 연구에서는 SWAT 모형과 random forest를 이용하여 미래 기후변화에 따른 한강유역(34,148 km^2)의 수생태계 건강성을 평가하였다. 국립환경과학원에서 8년간(2008~2015년) 봄철(4~6월)에 모니터링한 부착돌말류 지수(TDI), 저서형 대형무척추동물지수(BMI), 어류평가지수(FAI)는 0~100점, A~E등급으로 평가되며, 이를 본 연구에서 사용하였다. 수생태 건강성에 영향을 미치는 변수로는 수질(T-N, NH_4-N, NO_3-N, T-P, PO_4)과 수문을 선정하였으며, 수질 오염도가 낮은 경우에는 수생태계 건강성 점수가 불안정하게 분포되지만 수질 오염도가 높은 경우 수생태계 건강성 점수가 낮아지는 역상관관계를 확인하였다. 기계학습의 분류 분석 기법 중 하나인 random forest 모델을 이용한 세 개의 수생태 건강성 지수 등급 분류 결과 정밀도, 재현율, f1-score 모두 0.81 이상의 매우 정확도를 나타내었다. 기상청의 HadGEM3-RA RCP 4.5와 8.5 시나리오를 적용한 미래 SWAT 수문, 수질 결과 기저유출의 증가로 인해 질소 계열 수질 농도는 기준년도 대비 최대 43.2% 증가하였고, 지표유출 감소로 인해 인 계열 수질 오염도는 최대 18.9% 감소하는 것으로 분석되었다. 미래 FAI, BMI의 등급은 개선되는 경향을 보이지만 TDI는 등급이 악화되는 것으로 나타났다. 이를 통해 TDI는 질소 계열 수질에 민감하고 FAI, BMI는 인 계열 수질에 더 민감하다고 판단하였다.

핵심용어: 수생태 건강성, 기후변화, 수질, SWAT, Random forest

*Corresponding Author. Tel: +82-2-444-0186
E-mail: wjd0823@konkuk.ac.kr (C. G. Jung)

Journal of the Korean Society of Agricultural Engineers
DOI : https://doi.org/10.5389/KSAE.2017.59.4.097
Vol. 59, No. 4, pp. 97~107, July 2017
ISSN 1738-3692 | eISSN 2093-7709

머신러닝 기법을 활용한 유황별 LOADEST 모형의 적정 회귀식 선정 연구: 낙동강 수계를 중심으로

Study of Selection of Regression Equation for Flow-conditions using Machine-learning Method: Focusing on Nakdonggang Waterbody

김종건* · 박윤식** · 이서로* · 신용철*** · 임경재* · 김기성*,†

Kim, Jonggun · Park, Youn Shik · Lee, Seoro · Shin, Yongchul · Lim, Kyoung Jae · Kim, Ki-sung

Abstract

This study is to determine the coefficients of regression equations and to select the optimal regression equation in the LOADEST model after classifying the whole study period into 5 flow conditions for 16 watersheds located in the Nakdonggang waterbody. The optimized coefficients of regression equations were derived using the gradient descent method as a learning method in Tensorflow which is the engine of machine-learning method. In South Korea, the variability of streamflow is relatively high, and rainfall is concentrated in summer that can significantly affect the characteristic analysis of pollutant loads. Thus, unlike the previous application of the LOADEST model (adjusting whole study period), the study period was classified into 5 flow conditions to estimate the optimized coefficients and regression equations in the LOADEST model. As shown in the results, the equation #9 which has 7 coefficients related to flow and seasonal characteristics was selected for each flow condition in the study watersheds. When compared the simulated load (SS) to observed load, the simulation showed a similar pattern to the observation for the high flow condition due to the flow parameters related to precipitation directly. On the other hand, although the simulated load showed a similar pattern to observation in several watersheds, most of study watersheds showed large differences for the low flow conditions. This is because the pollutant load during low flow conditions might be significantly affected by baseflow or point-source pollutant load. Thus, based on the results of this study, it can be found that to estimate the continuous pollutant load properly the regression equations need to be determined with proper coefficients based on various flow conditions in watersheds. Furthermore, the machine-learning method can be useful to estimate the coefficients of regression equations in the LOADEST model.

Keywords: LOADEST model, Machine-learning, Pollutant load, Flow conditions, Regression equation

Ⅰ. 서 론

하천에는 청천시 연중 일정량의 기저유량과 하수처리장과 같은 시설물에서의 배출수가 유입되고 있으며 이로 인한 기저유출에 대한 비점오염원과 시설물에서의 점오염원이 유입되고 있다. 강우시 강우 유출수에 의한 비점오염물질이 하천으로 유입되어 수질오염원의 주요 원인이 되고 있다 (Lee et al., 2014). 또한, 기후변화로 인한 이상 기후에 따른 홍수 및 가뭄의 발생은 하천의 수량과 오염물질 발생에 영향을 미치고 있다. 현재 국내에서는 수질오염총량관리제를 시행하여서, 군 구간별로 목표수질을 설정하고, 이를 달성하기 위해 허용부하량을 산정해 오염물질 배출량을 규제하는 데 노력을 기울이고 있다 (ME, 2004). 4대강 사업 이후 악화된 수질을 관리하기 위해 낙동강 수계에 대해서는 수계 내 지류를 대상으로 지류총량제를 적용하여 오염물질의 근본 원인을 파악하고 이를 저감할 수 있는 방안을 수립하고 있다. 지류총량제를 비롯한 수질오염총량관리제에서는 유역 내 오염부하량을 통해 수질오염 기준을 마련하고 수질 관리 대책을 수립하고 있다. 오염부하량은 유역 내 배출되는 유량과 그 유량 속에 포함되어 있는 오염물질 농도의 곱으로 산정되며, 오염부하 특성 분석을 통해 주요 오염원 파악 및 오염원의 기여도 분석이 가능하다. 이러한 오염부하 특성은 토지이용도, 유역 크기 및 지형 등 유역의 특성에 따라 다양하게 나타날 수 있다. 뿐만 아니라, 유황별 오염부하 특성을 통해 점오염원 및 비점오염원 등의 주요 오염원 파악이 가능하다. 그러하여 오염부하량의 정확한 산정은 유역 내 하천의 수질오염 분석에 있어 중요한 부

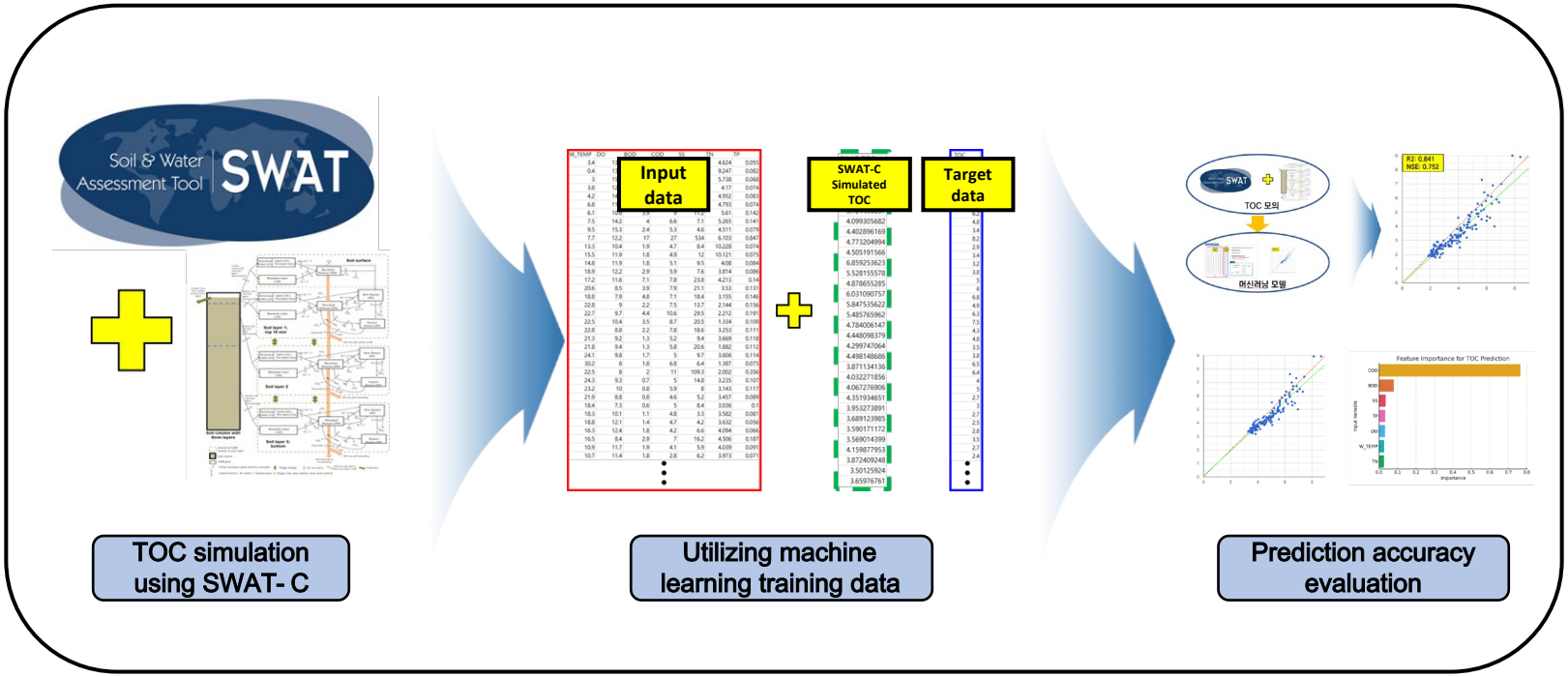
• Department of Regional Infrastructures Engineering, Kangwon National University
•• Department of Rural Construction Engineering, Kongju National University
••• School of Agricultural Civil and Bio-Industrial Engineering, Kyungpook National University
† Corresponding author
Tel.: +82-33-250-6466 Fax: +82-33-259-5560
E-mail: kskim@kangwon.ac.kr

Received: April 24, 2017
Revised: July 19, 2017
Accepted: July 20, 2017

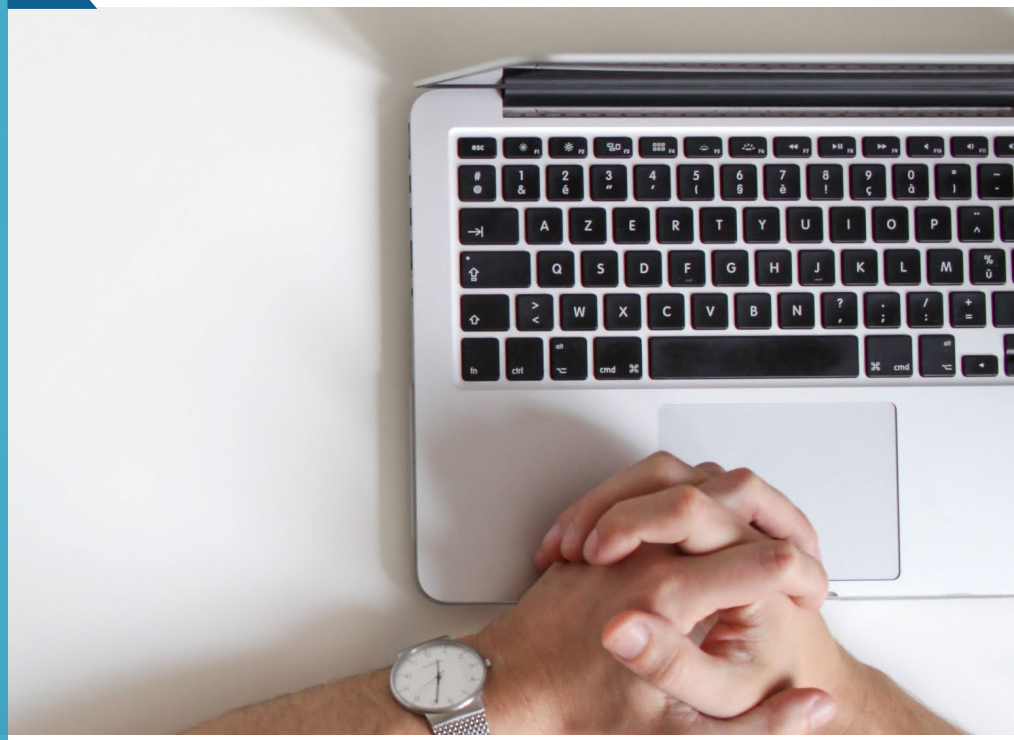
- Machine learning and deep learning techniques have been actively applied to water quality prediction worldwide.
- Data-driven machine learning models are highly dependent on the quality and quantity of training data, and may show reduced predictive performance in data-scarce environments.
- To address these limitations, studies have explored integrating machine learning with process-based hydrological models.

❖ The Goal of This Study

- ✓ Improve TOC simulation accuracy by integrating the SWAT-C model with a machine learning model
- ✓ Evaluate prediction accuracy by comparing with observed TOC load data

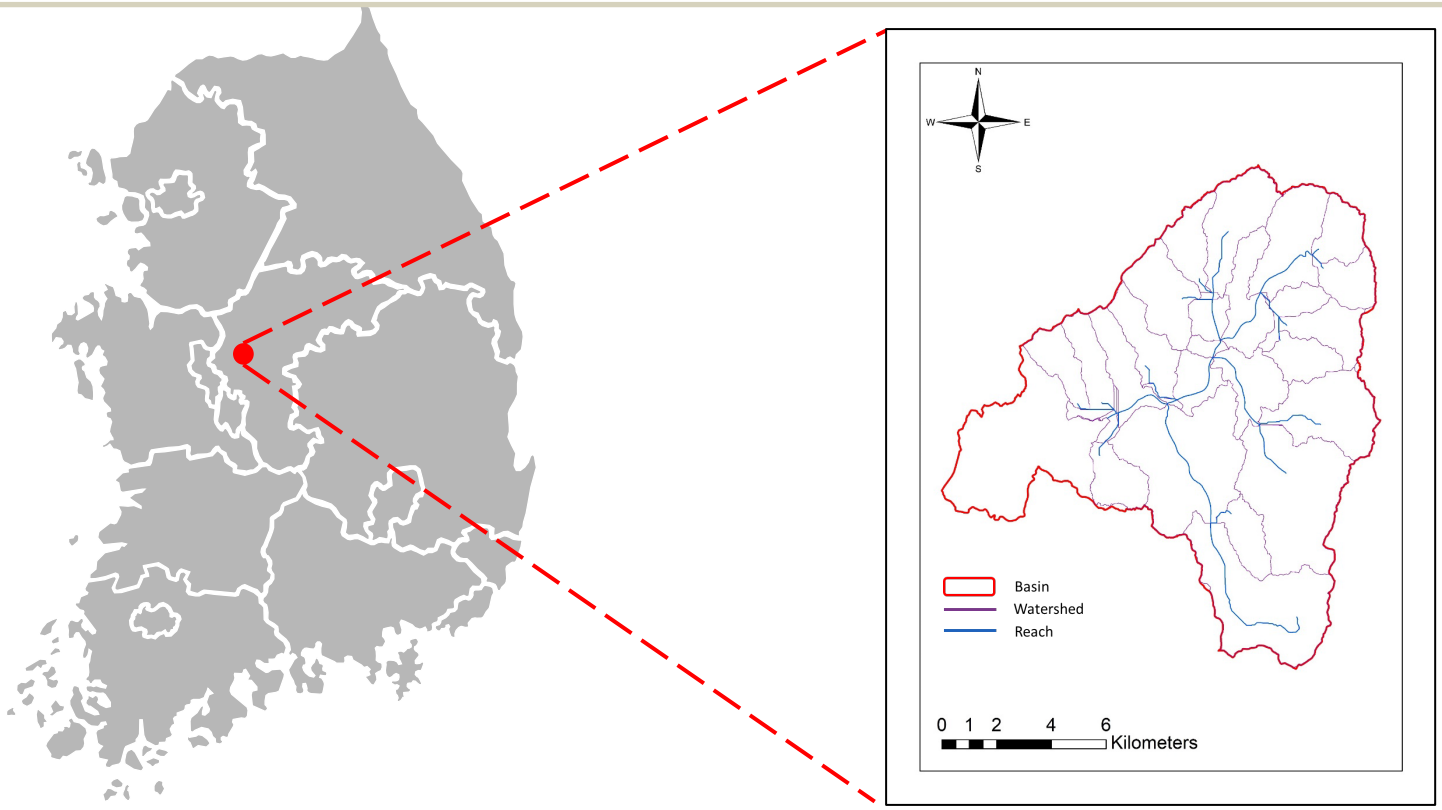


II. Methodology



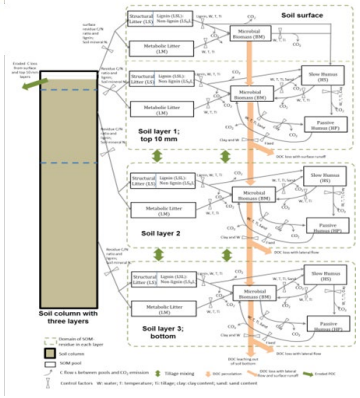
SWAT Soil & Water Assessment Tool

❖ Study Area

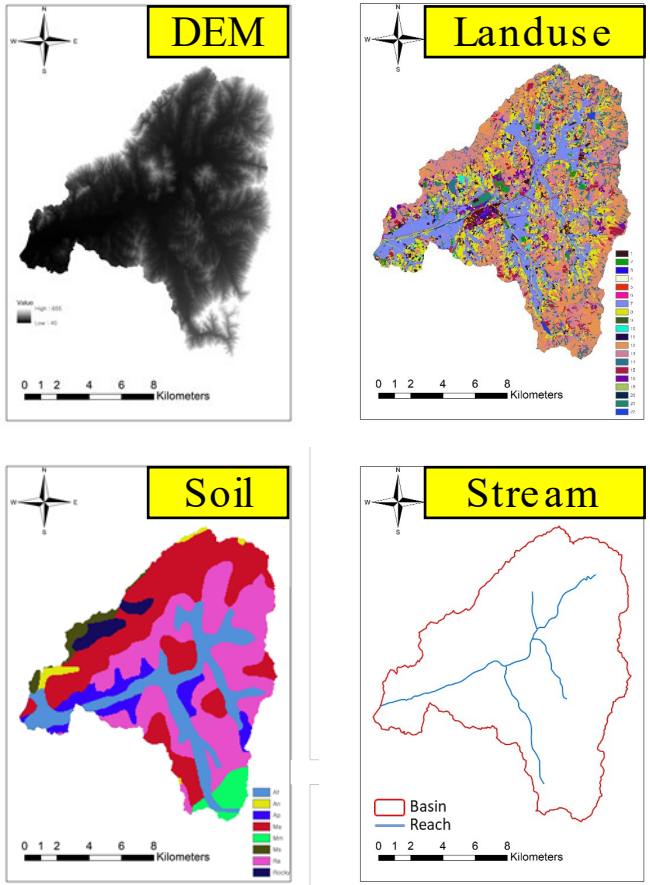


- ✓ Study Site: Bogangcheon watershed, located in the upper region of the Mihocheon River Basin
- ✓ Watershed Area: 149.22 km²

❖ Overview of SWAT-C and Input Data Configuration



- SWAT- C is a modified version of the original SWAT (Soil and Water Assessment Tool) model, incorporating a carbon cycling module for TOC simulation.
- The model requires both spatial and meteorological inputs as primary input data.

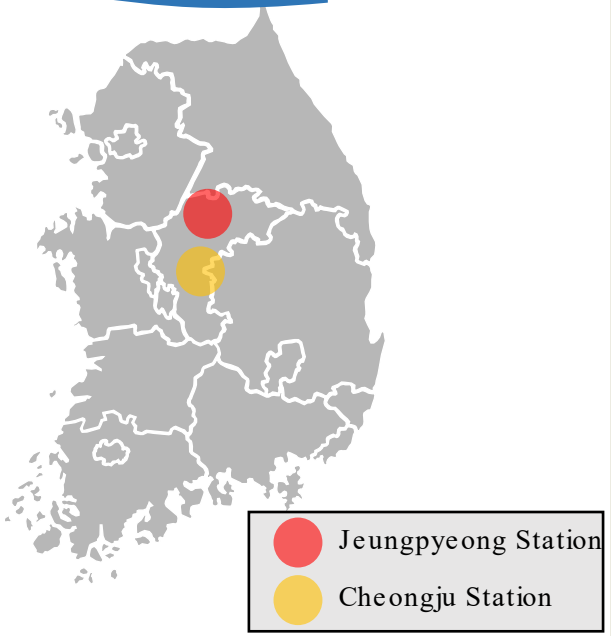


<SWAT- C Spatial Data>

SWAT Soil & Water Assessment Tool

❖ SWAT-C Input Data Configuration

Weather Stations

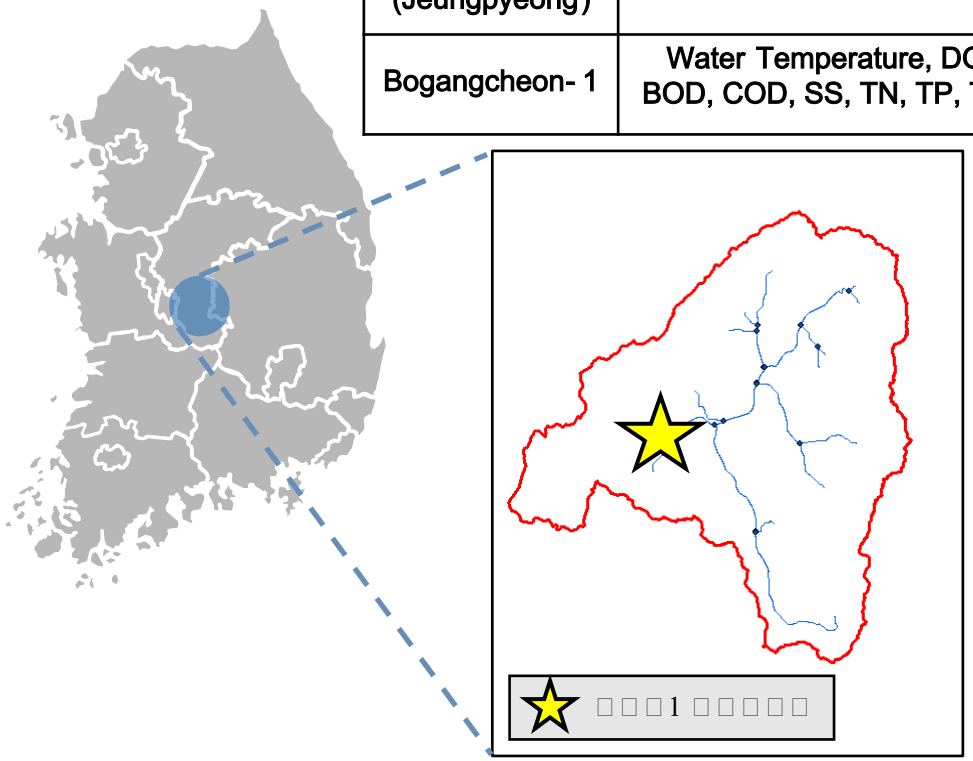


Weather Stations	Weather Data (Daily)
Jeungpyeong (AWS)	Precipitation, Max/Min Temperature, Wind Speed
Cheongju (ASOS)	Precipitation, Solar Radiation, Relative Humidity, Max/Min Temperature, Wind Speed

Water Quality and Discharge Data

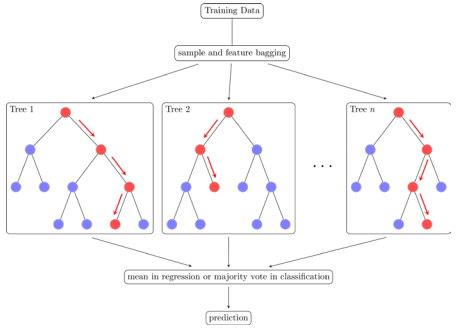
물환경정보시스템 Water Environment Information System

Monitoring Site	Observed Data (8- Day intervals, 2012~2024)
Bantan Bridge (Jeungpyeong)	Flow Data
Bogangcheon- 1	Water Temperature, DO, BOD, COD, SS, TN, TP, TOC



❖ Input Data for Machine Learning

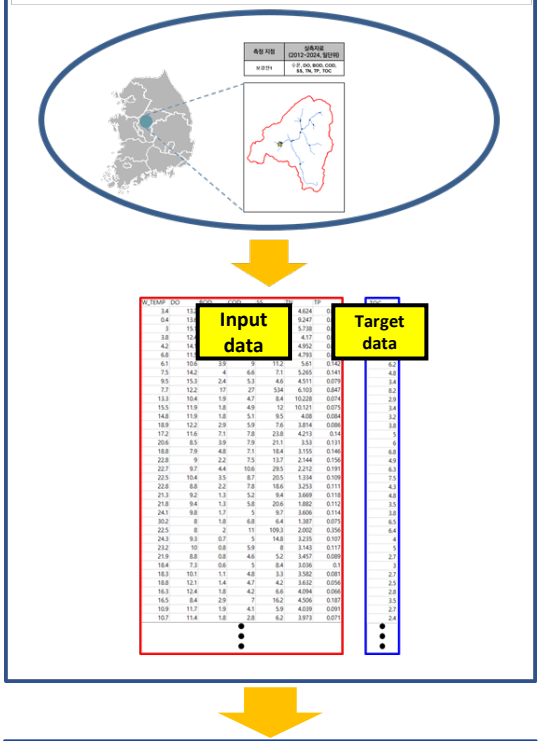
- Algorithm Used: RandomForest
  - Random Forest generates multiple decision trees and aggregates their results for prediction.
  - A supervised learning algorithm suitable for regression tasks.



✓ Train/Test: 70%:30%

✓ Evaluation: R^2, NSE

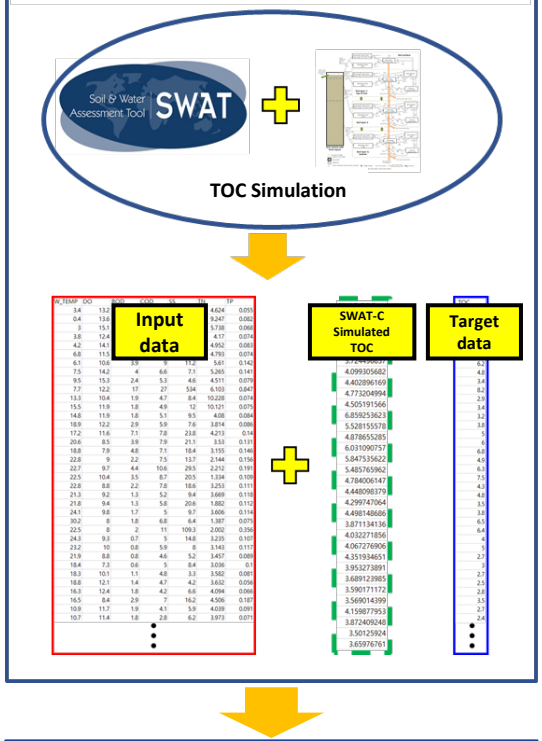
Model Using Only Observed Data



✓ Input: Water Temperature, DO, BOD, COD, SS, TN, TP

✓ Target: TOC

Model Using Observed Data + SWAT-C Output

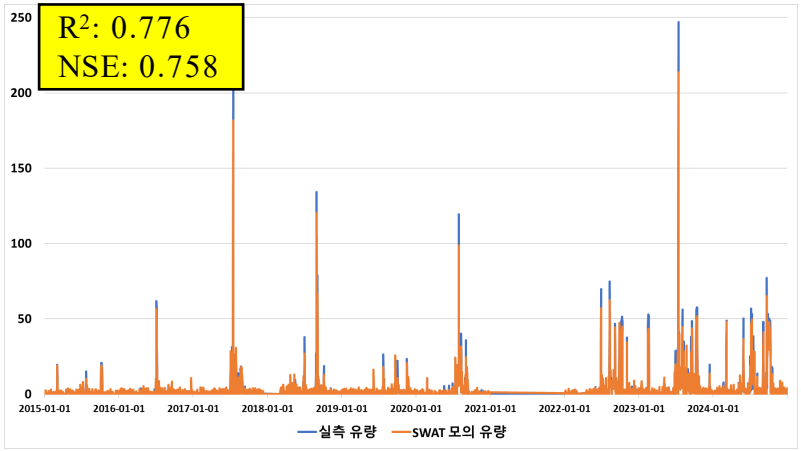


✓ Input: Water Temperature, DO, BOD, COD, SS, TN, TP, SWAT-C Simulated TOC

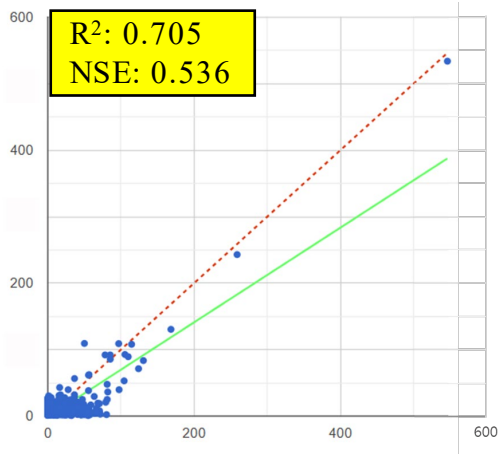
✓ Target: TOC

III. Result

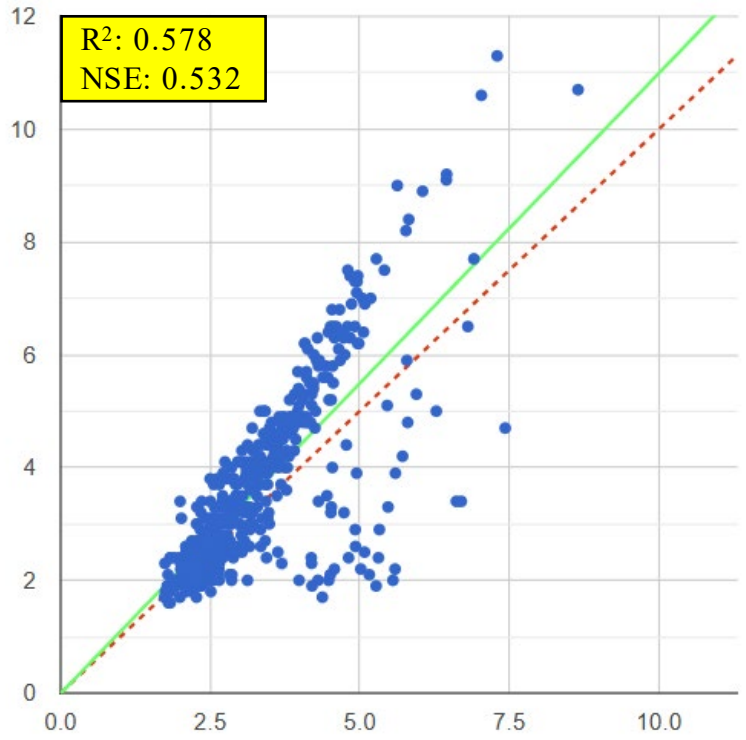
❖ SWAT- C Simulation Results



<Flow Simulation>



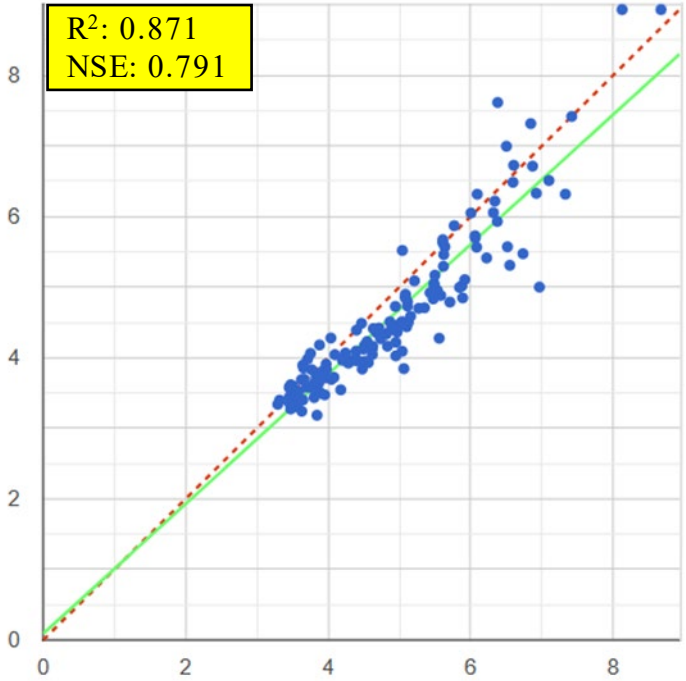
<SS (Suspended Solids) Simulation>

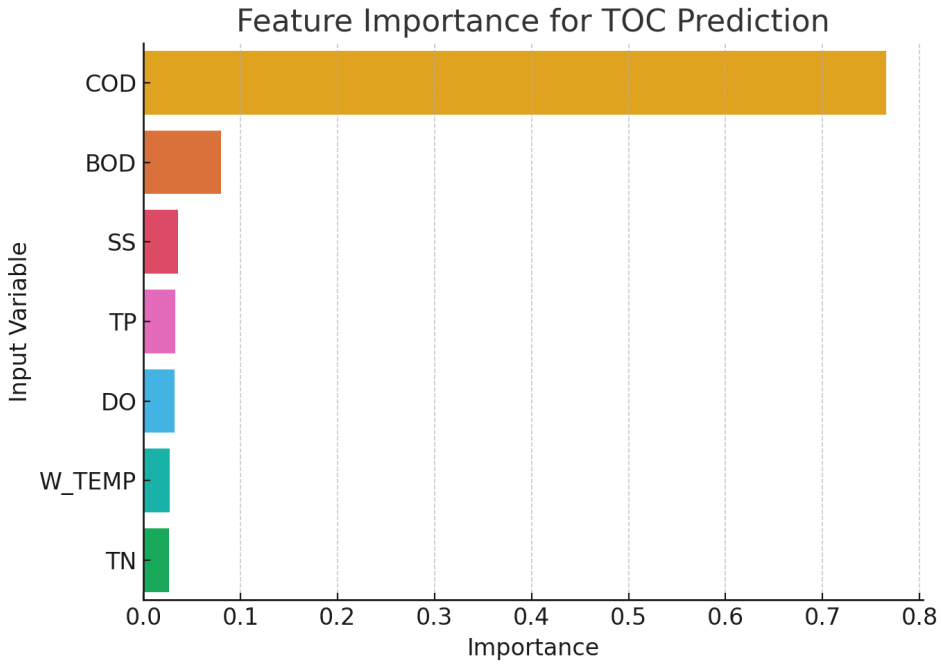


<TOC (Total Organic Carbon) Simulation>

- The TOC simulation shows acceptable performance, exceeding the general threshold of NSE > 0.5.
- This indicates that the carbon module in SWAT- C is effective in capturing TOC dynamics

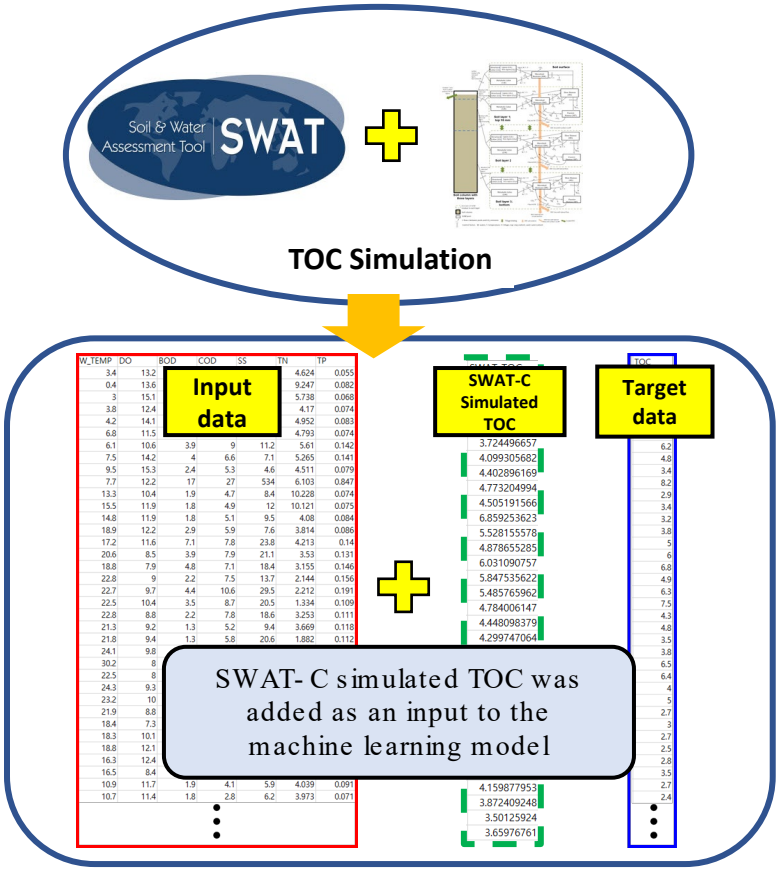
❖ Machine Learning Results Using Observed Water Quality Data

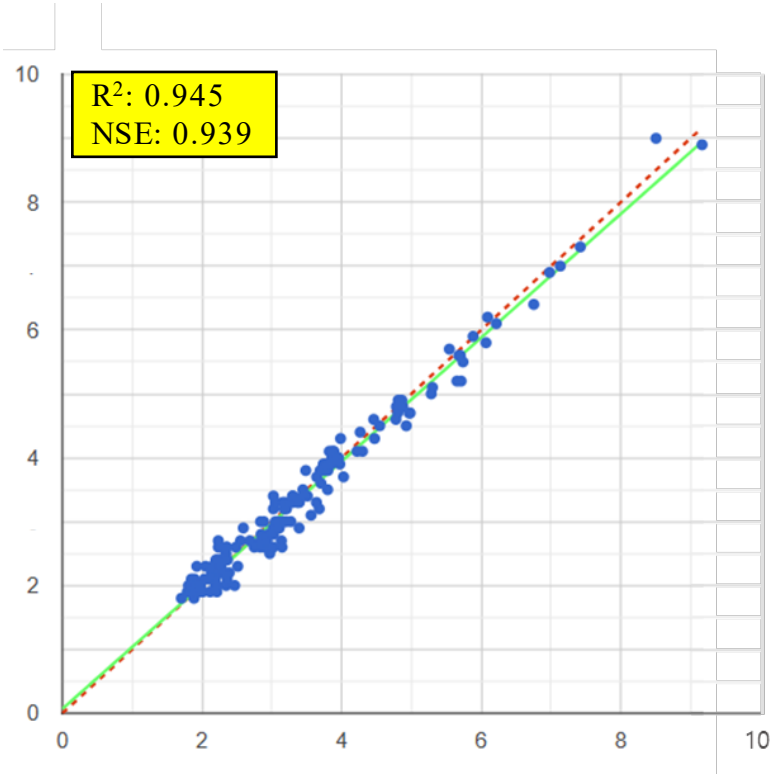




- The machine learning model (Random Forest) shows enhanced TOC prediction performance with high R^2 and NSE values.
- High importance scores for COD and BOD imply their strong correlation with TOC, due to shared chemical characteristics.
- This approach demonstrates that machine learning models can effectively reflect dominant influencing variables, improving model reliability for TOC estimation.

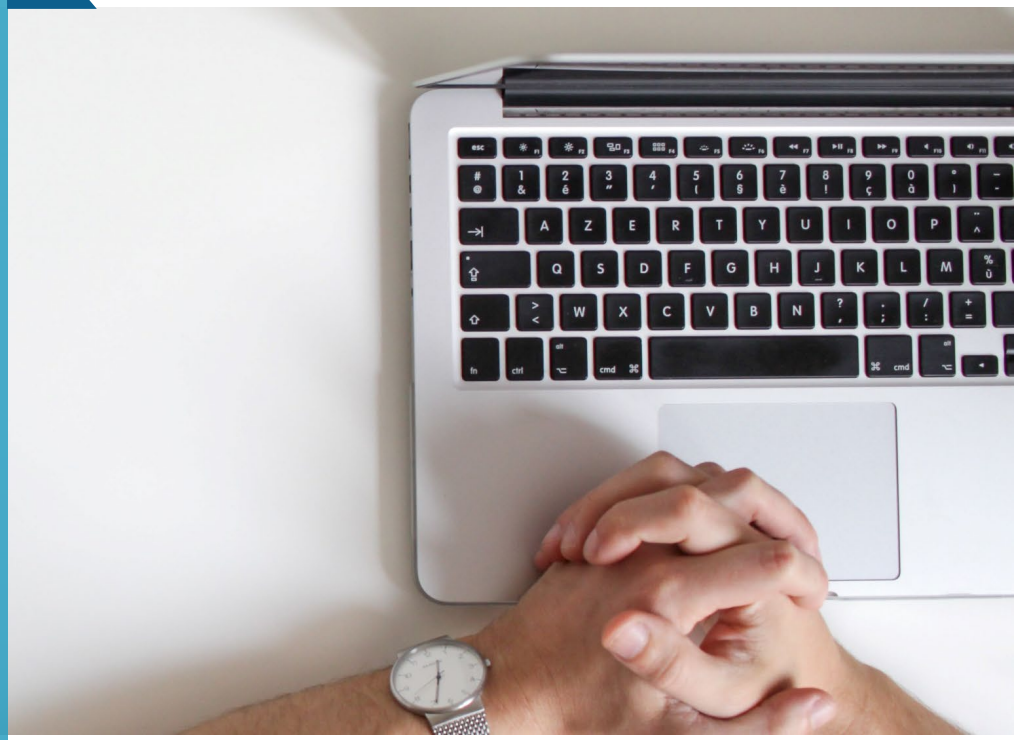
❖ Machine Learning Results Using SWAT- C Simulated TOC





- Incorporating SWAT- C simulated TOC as an input significantly enhanced the predictive performance of the machine learning model.
- This suggests that integrating physically- based model outputs with data- driven models can improve the estimation of observed TOC concentrations.

IV. Conclusion and Limitations



IV. Conclusion and Limitations

❖ Conclusion

- ✓ The SWAT-C model demonstrated acceptable performance in simulating TOC, meeting or exceeding standard accuracy thresholds.
- ✓ Machine learning models using observed input variables achieved high accuracy for TOC prediction, particularly under complex hydrological conditions.
- ✓ When SWAT-C simulated TOC was used as an additional input, the prediction performance further improved, confirming the effectiveness of hybrid modeling approaches.

❖ Limitations

- ✓ SWAT-C is a physically-based model, while machine learning is data-driven; thus, logical interpretation across the modeling chain is challenging.
- ✓ The Random Forest algorithm does not fully capture the temporal dynamics of TOC, making it less effective for time-series predictions or variable lag analysis.

❖ Future Research

- ✓ To better reflect temporal patterns in TOC, deep learning models such as LSTM (Long Short-Term Memory) should be explored.
- ✓ Evaluating model stability under various watershed conditions and hydrologic regimes is necessary.
- ✓ Expanding simulations to include DOC and POC, and utilizing SWAT-C to estimate both forms can enhance water quality management strategies, especially for inflow pollution control.

Thank You!

E-mail:
teddy00102933@gmail.com