Prediction of Health Evaluation Indices for Aquatic Ecosystem using Extreme Gradient Boosting Tree (XGBoost) and SWAT

Chung Gil Jung

(wjd0823@konkuk.ac.kr)

So Young Woo and Seong Joon Kim

School of Civil and Environmental Engineering, College of Engineering,

Konkuk University, South Korea

KONKUK University

Earth Information Erineering Lab.

Introduction

- In South Korea, since 1990s, the people became interested in restoration of river environmental function for their life quality with water.

- Since 2007, the Ministry of Environment has monitored the aquatic ecosystem health (AEH) and evaluated the stream and river AEH.

- By the limitations of monitoring sites, we need some technique to develop AEH evaluation for the whole country streams.

- This study is to predict the AEH indices (FAI, TDI, and BMI) of ungauged streams for a standard watershed scale (about 500 km^2) using SWAT results and Ensemble Machine Learning algorithm.

Assessment Procedure

SWAT
Model set up

- Calibration and validation of **Streamflow, ET, SM, T-N, T-P** using 11 years (2005-2014) observed data

Ecological Health Monitoring Data (2008-2015)
at 320 locations in Han-river basin

- Fish Assessment Index (FAI)
- trophic diatom index (TDI)
- benthic macroinvertebrate index (BMI)

- Total P
- Phosphate (PO_4^-)
- Total N, Nitrate (NO_3^-)
- Ammonium (NH_4^+)

- water temperature (WT)
- stream discharge (SD)

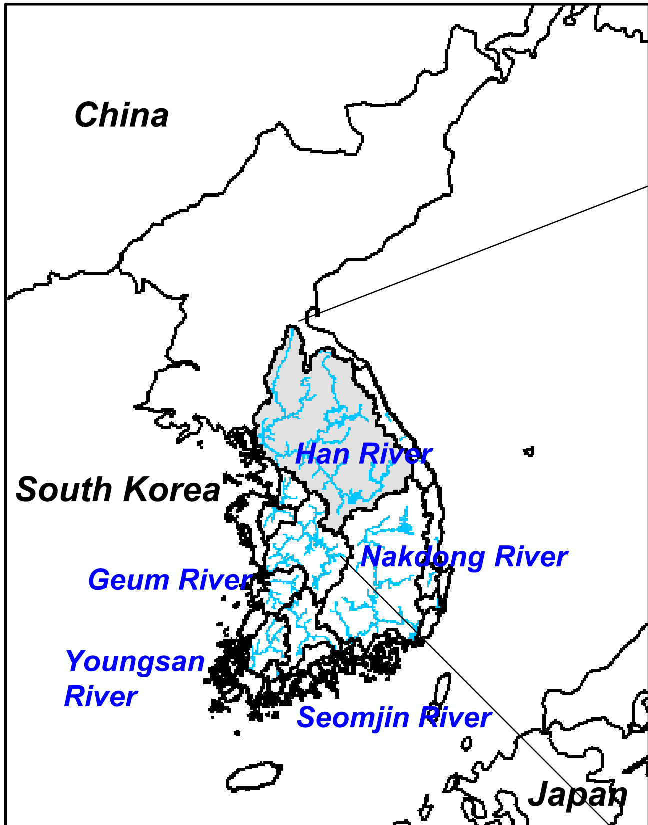
Correlation Analysis
Between data

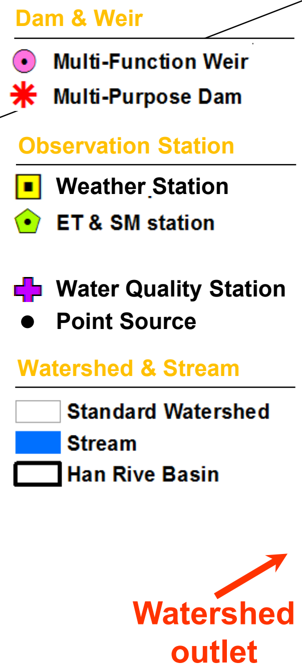
- **FAI, TDI, BMI** **versus** T-P, PO_4^-, T-N, NO_3^-, NH_4^+
- T-P, PO_4^-, T-N, NO_3^-, NH_4^+ **versus** **WT, SD**

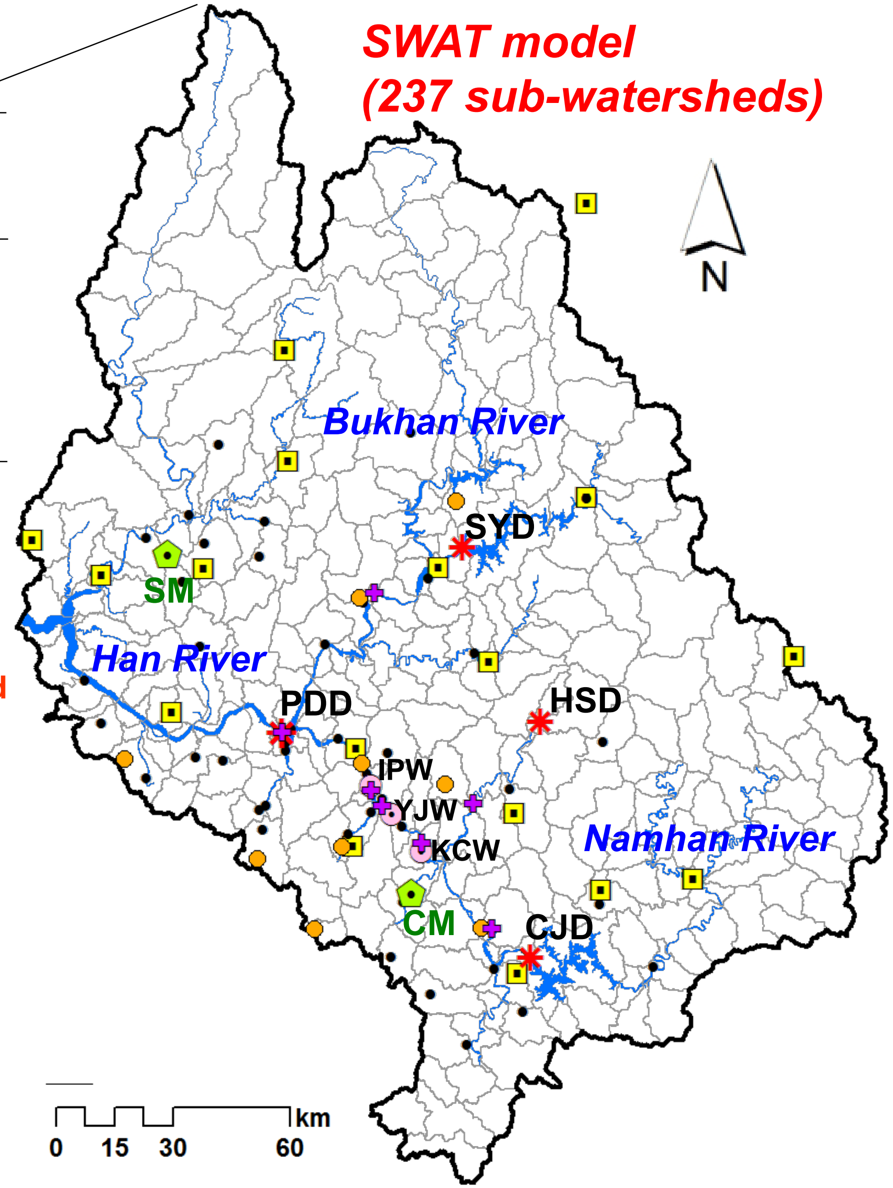
Prediction of Ecological Health Indices

- Development of XGBoost algorithm
- Input variables
  - (SWAT) **WT, SD,** T-P, PO_4^-, T-N, NO_3^-, NH_4^+
  - (monitoring) **FAI, TDI, BMI**
- Training and verification on FAI, TDI, BMI each sub-watershed

◆ Study area







- The largest river basin in South Korea (Han, Geum, Yeongsan, Seomjin, Nakdong)
- Han River basin (34,148 km^2)
  - ✓ Average precipitation 1254 mm
  - ✓ Average temperature 11.5℃

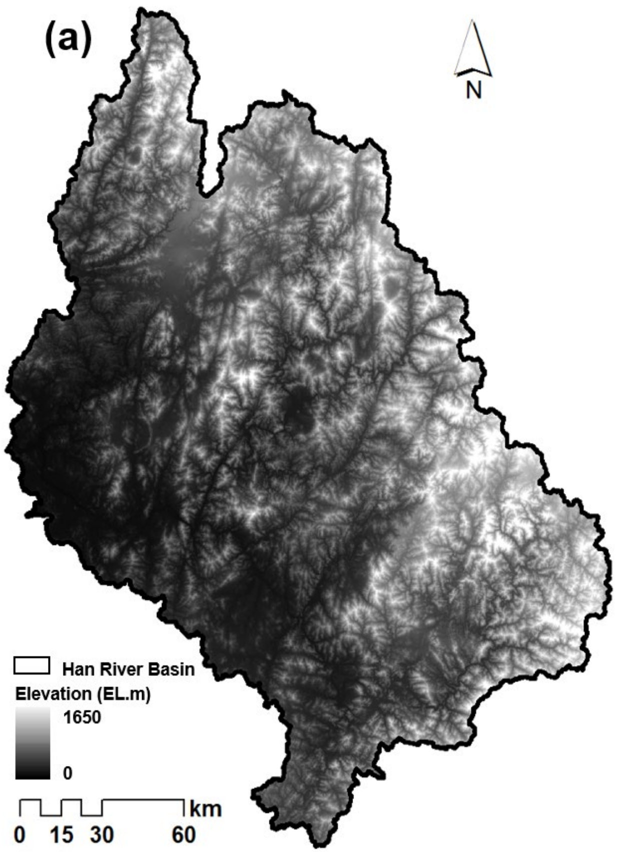
◆ Data for SWAT model evaluation

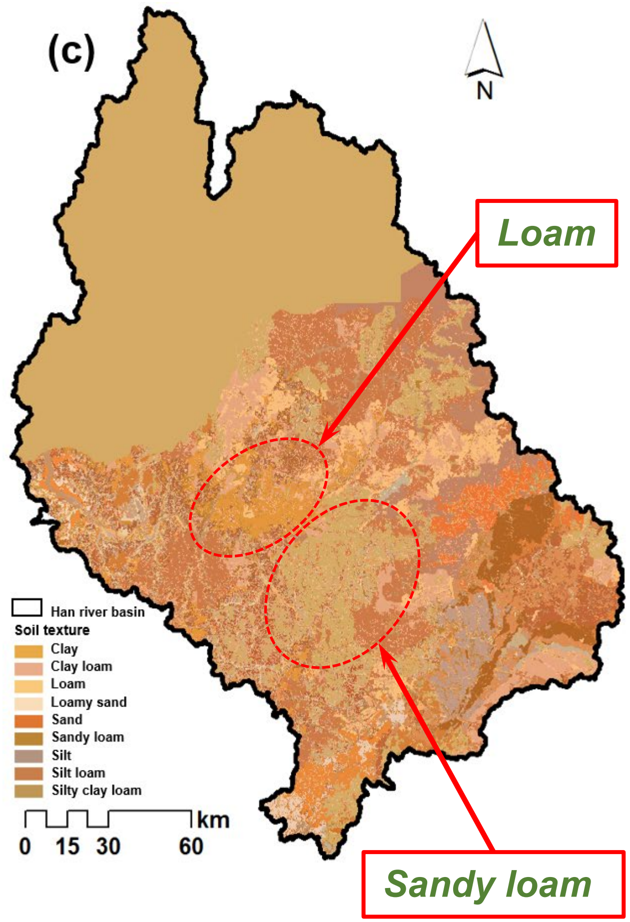
GIS data

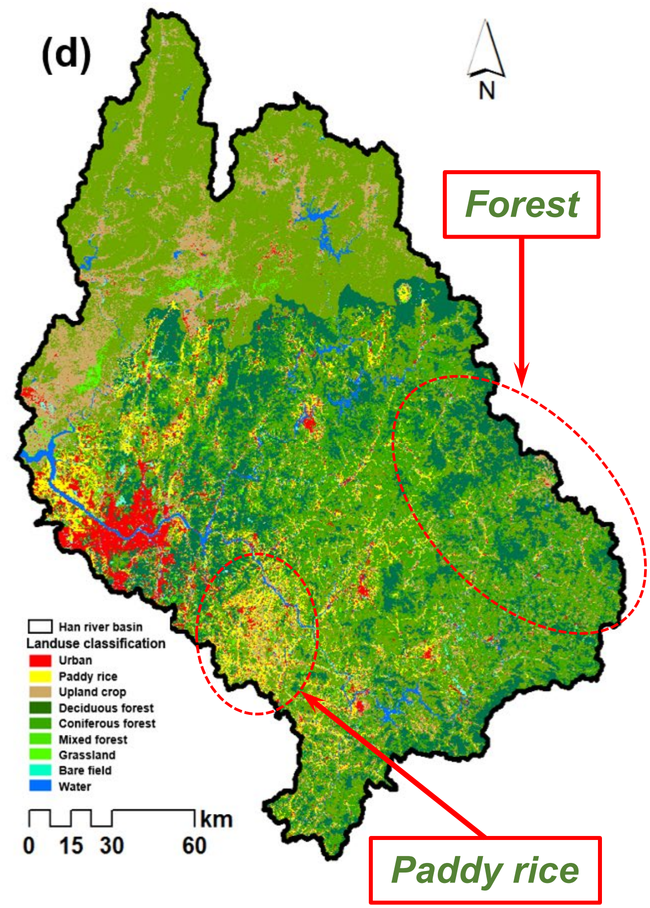
Elevation : 0 - 1650m (SRTM 90m grid size)

Soil : Loam (24%) and sandy loam (58%)

Land cover (2008) : Forest (73%) and paddy rice (6%)

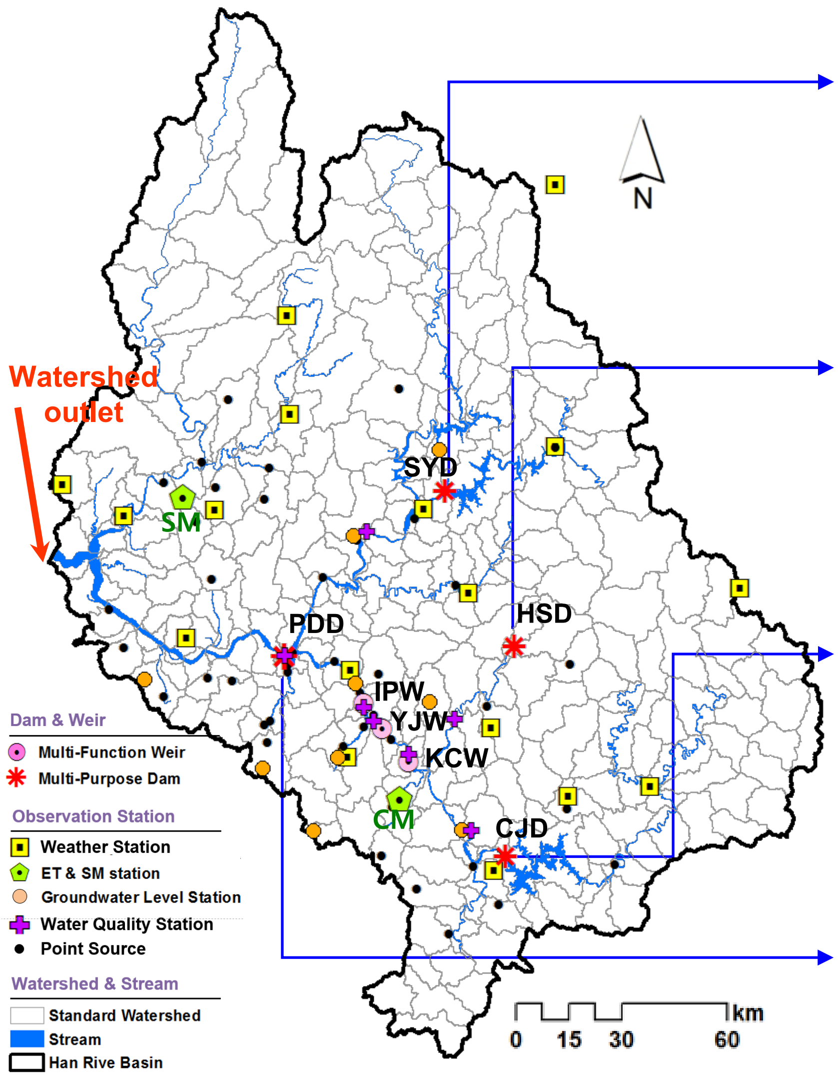


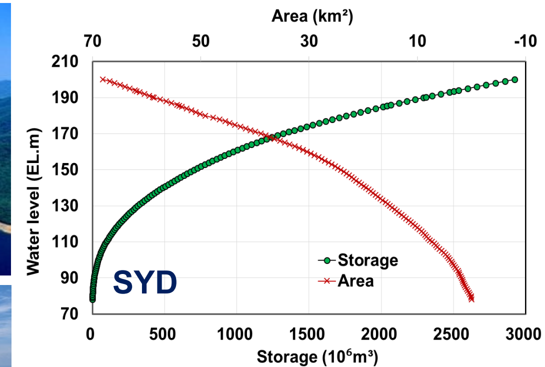


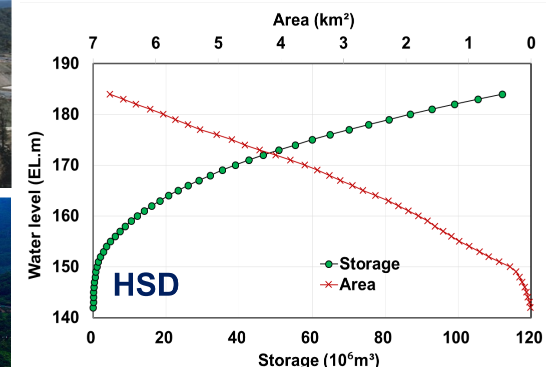


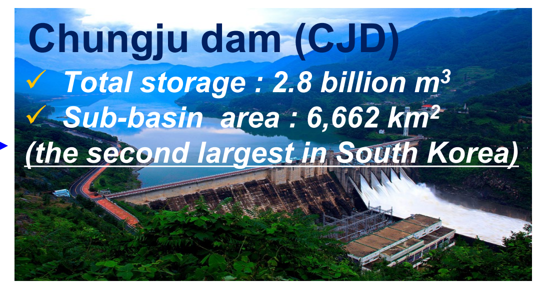
◆ Data for SWAT model evaluation

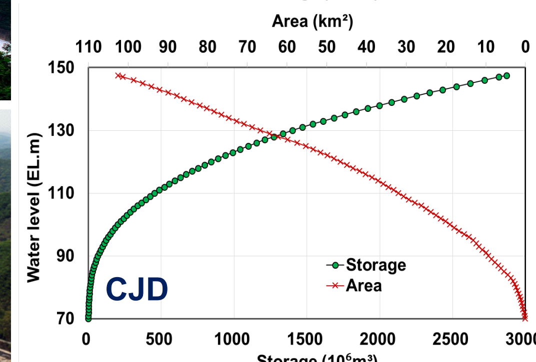
4 Multipurpose dam data (area-level and storage-level relationship curve)

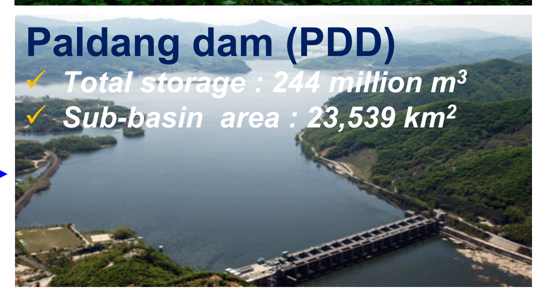






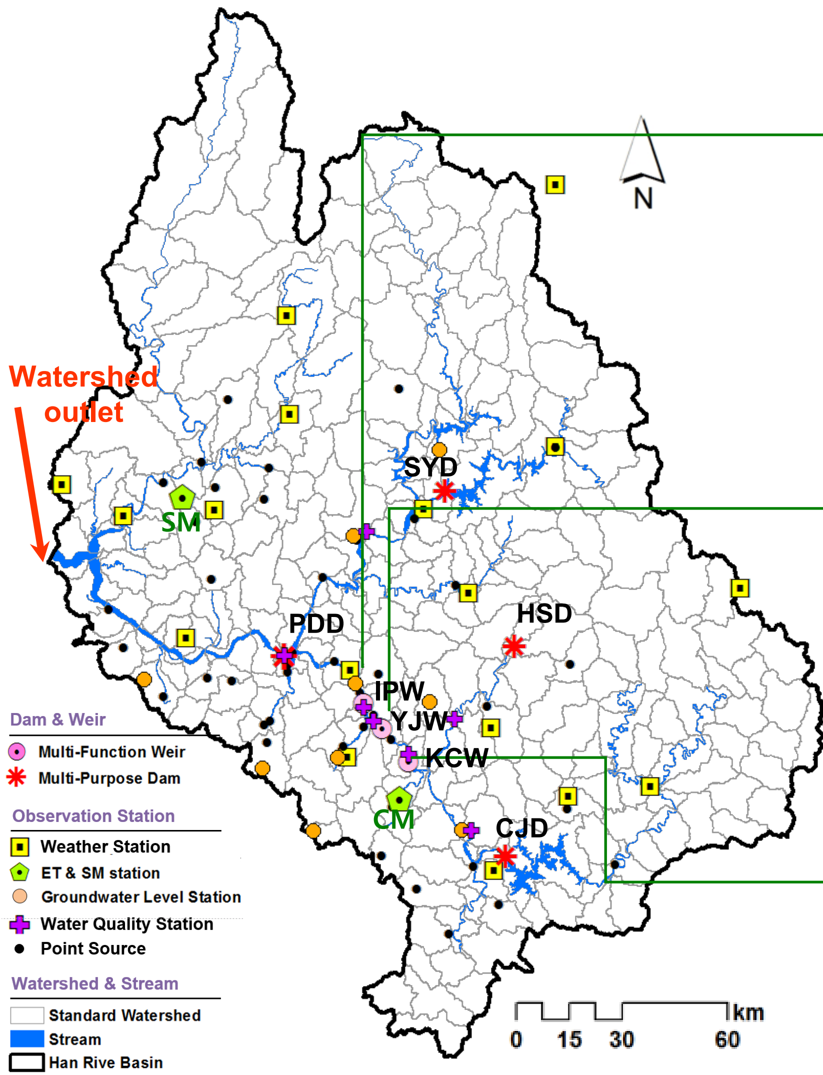


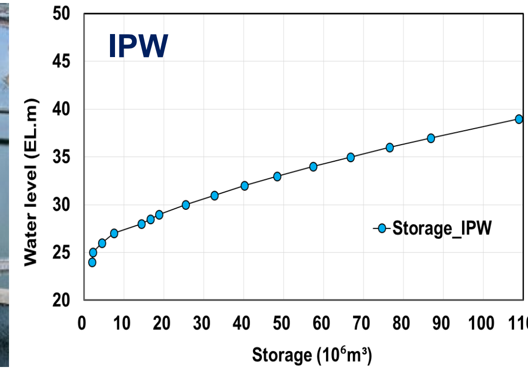


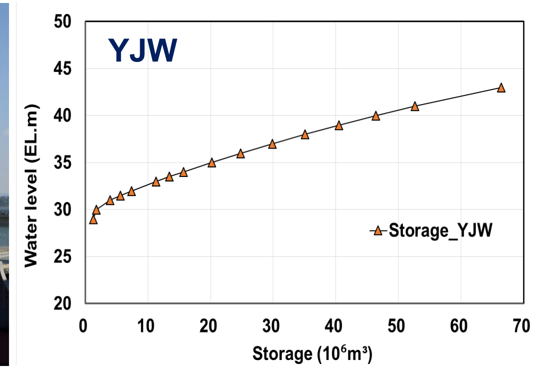


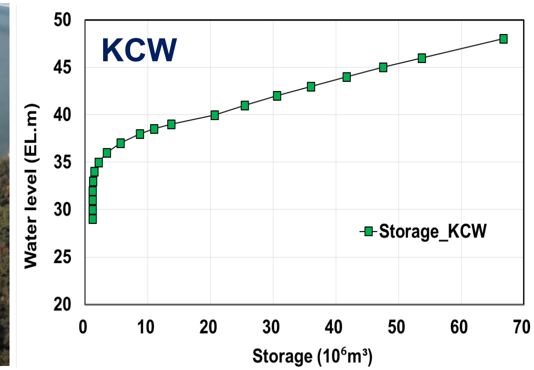
Data for SWAT model evaluation

3 Multifunction weir data (area-level and storage-level relationship curve)





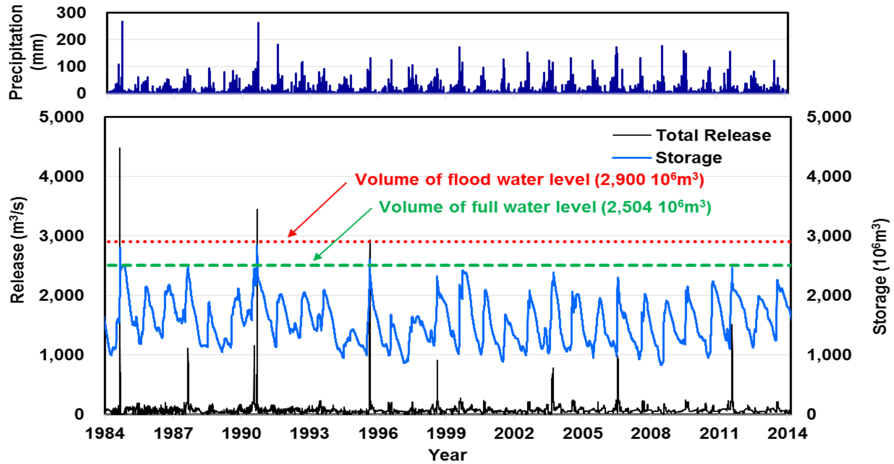




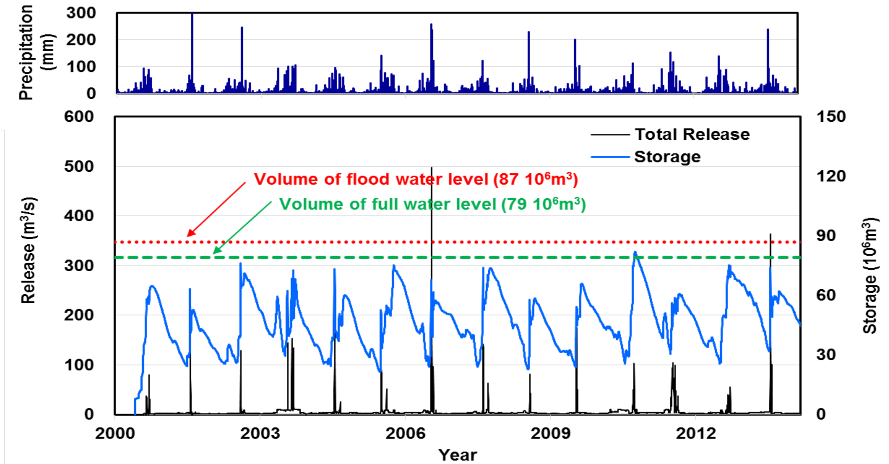
◆ Data for SWAT model evaluation

4 Multipurpose dam data (release and storage : 1984-2014)

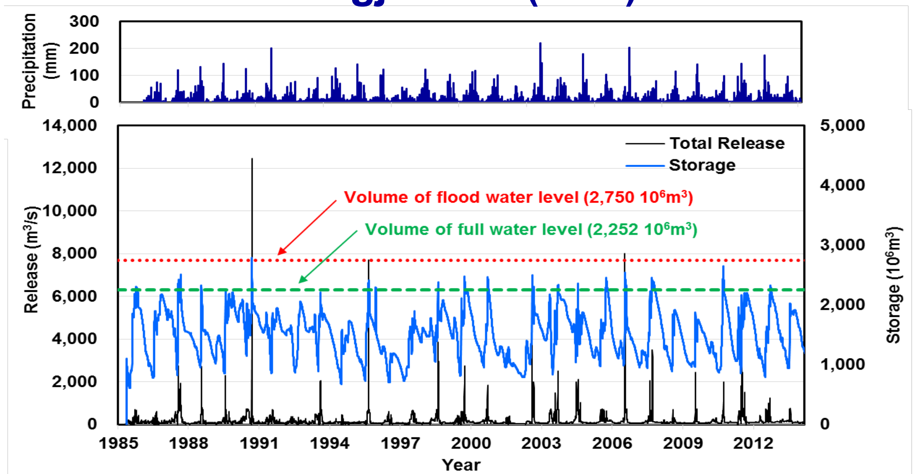
Soyang dam (SYD)



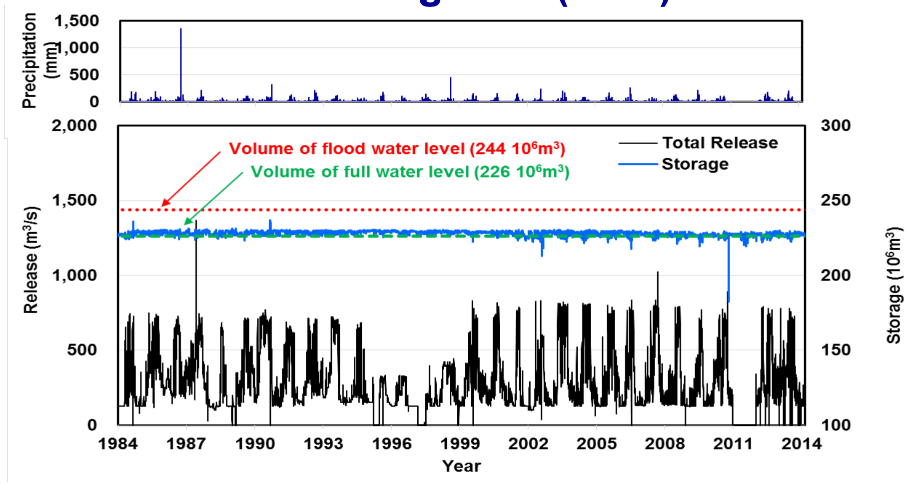
Hoengseong (HSD)



Chungju dam (CJD)



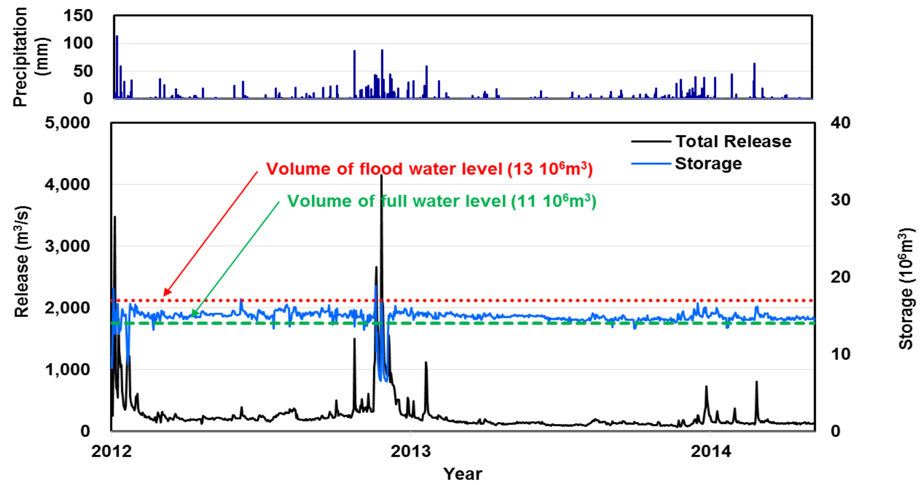
Paldang dam (PDD)



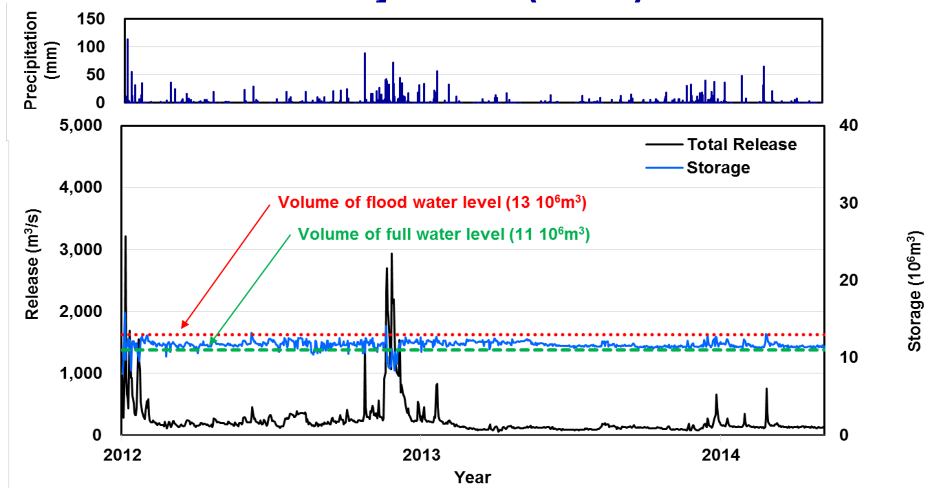
◆ Data for SWAT model evaluation

3 <u>Multifunction weir</u> data (release and storage : 2012-2014)

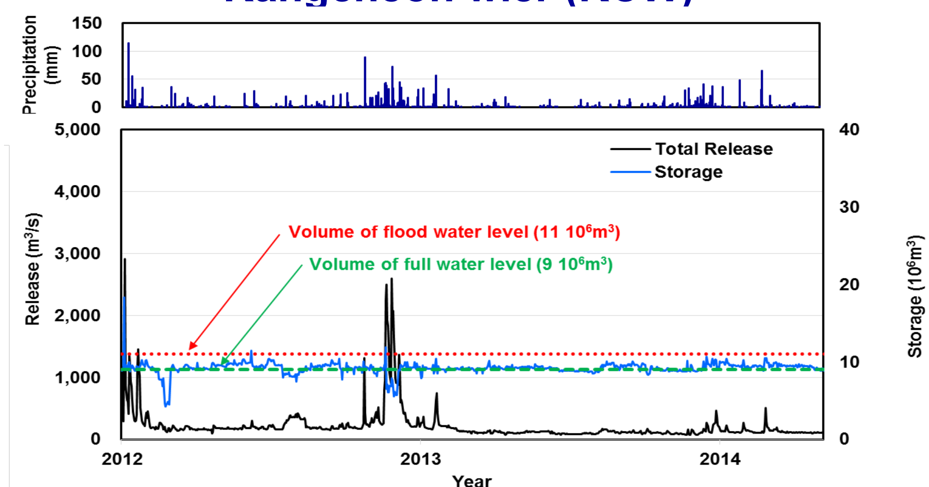
Ipo weir (IPW)



Yeoju weir (YJW)



Kangcheon wier (KCW)



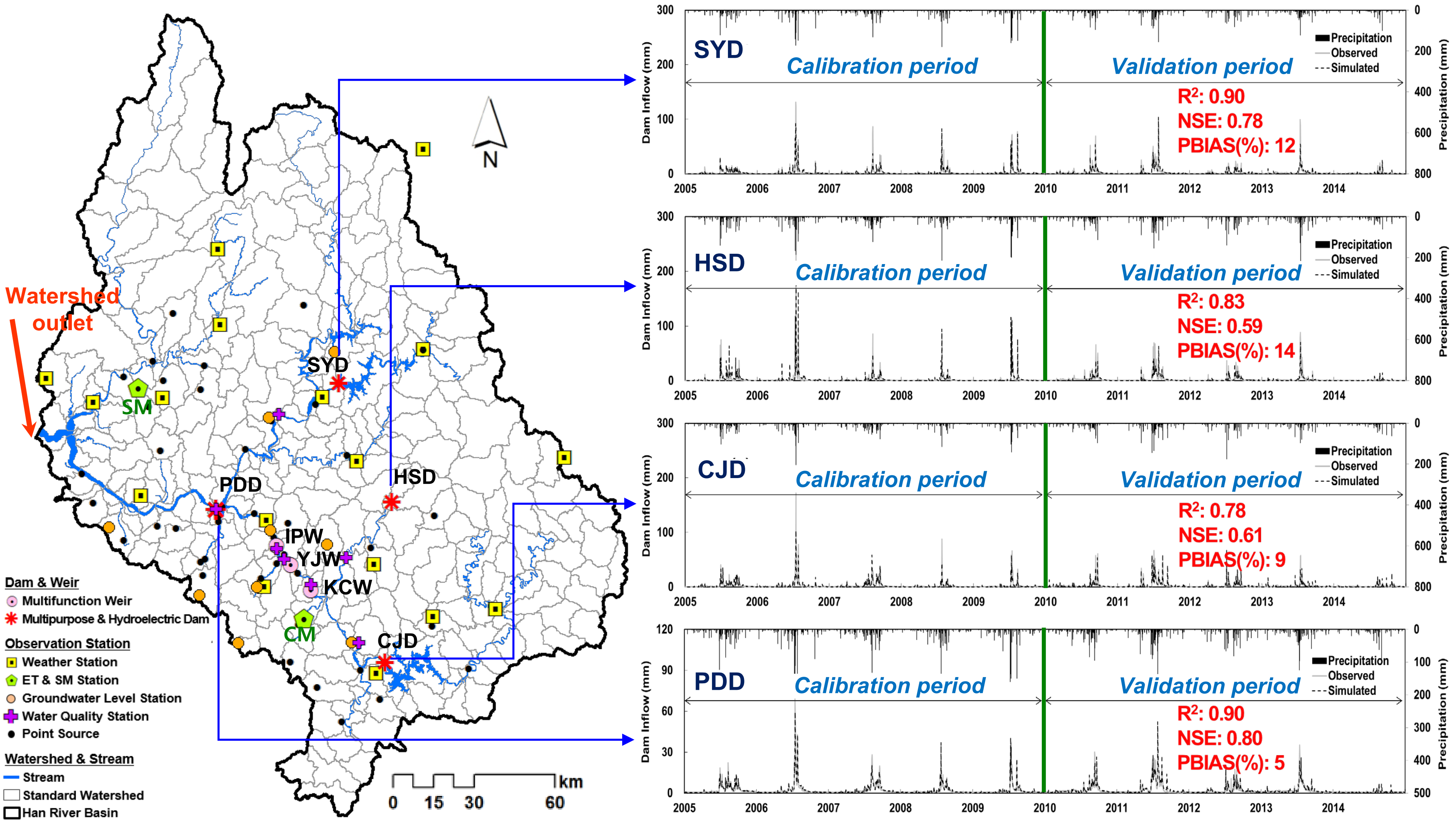
SWAT Model calibration and validation

- Hydrology Results

Ahn et al. (2015)

Observed vs. simulated streamflow results of model calibration and validation

✓ Calibration : 5 years (2005-2009) / Validation : 5 years (2010-2014)



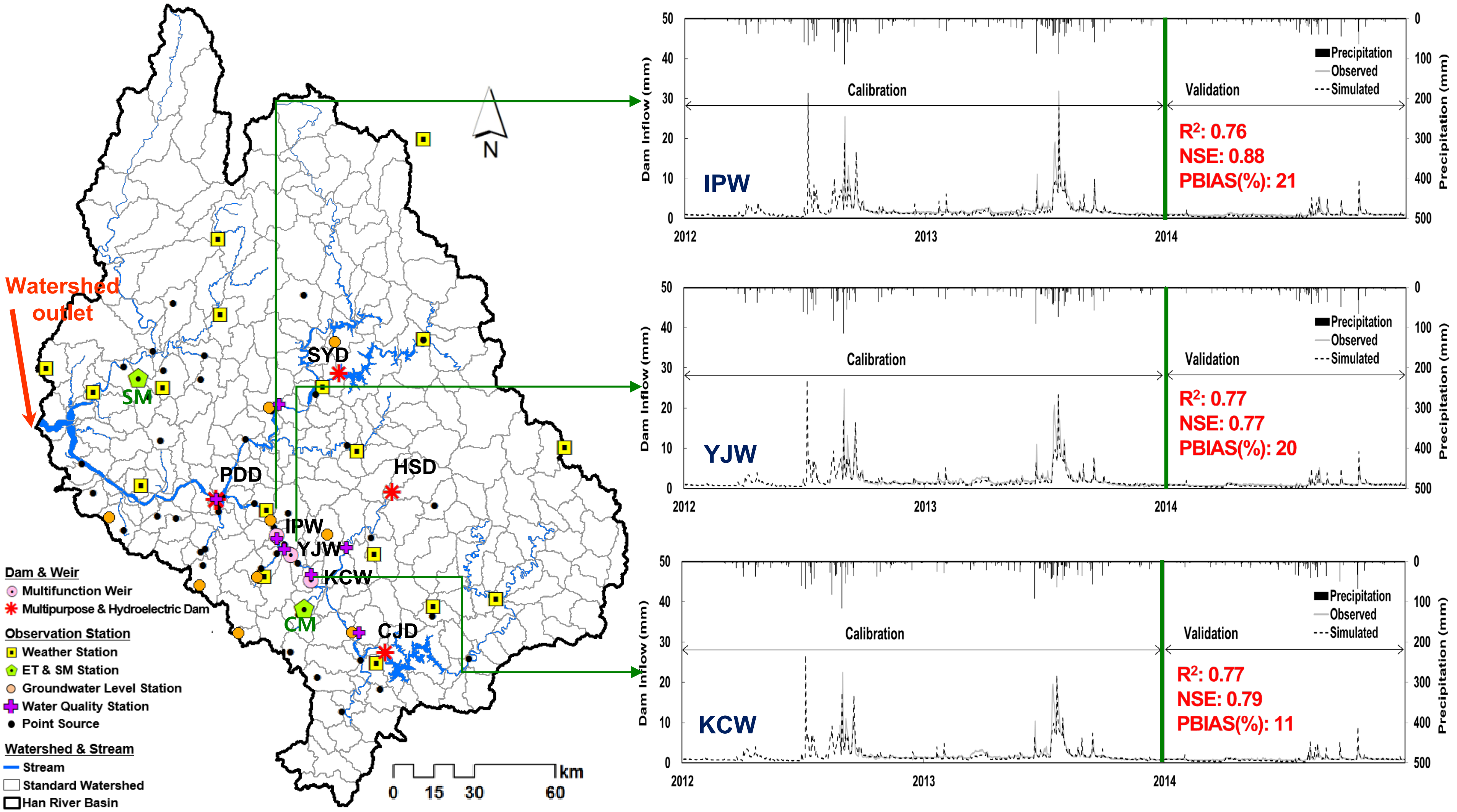
◆ SWAT Model calibration and validation

- **Hydrology Results**

Ahn et al. (2015)

Observed vs. simulated <u>streamflow</u> results of model calibration and validation

✓ **Calibration : 2 years (2012-2013) / Validation : 1 year (2014)**



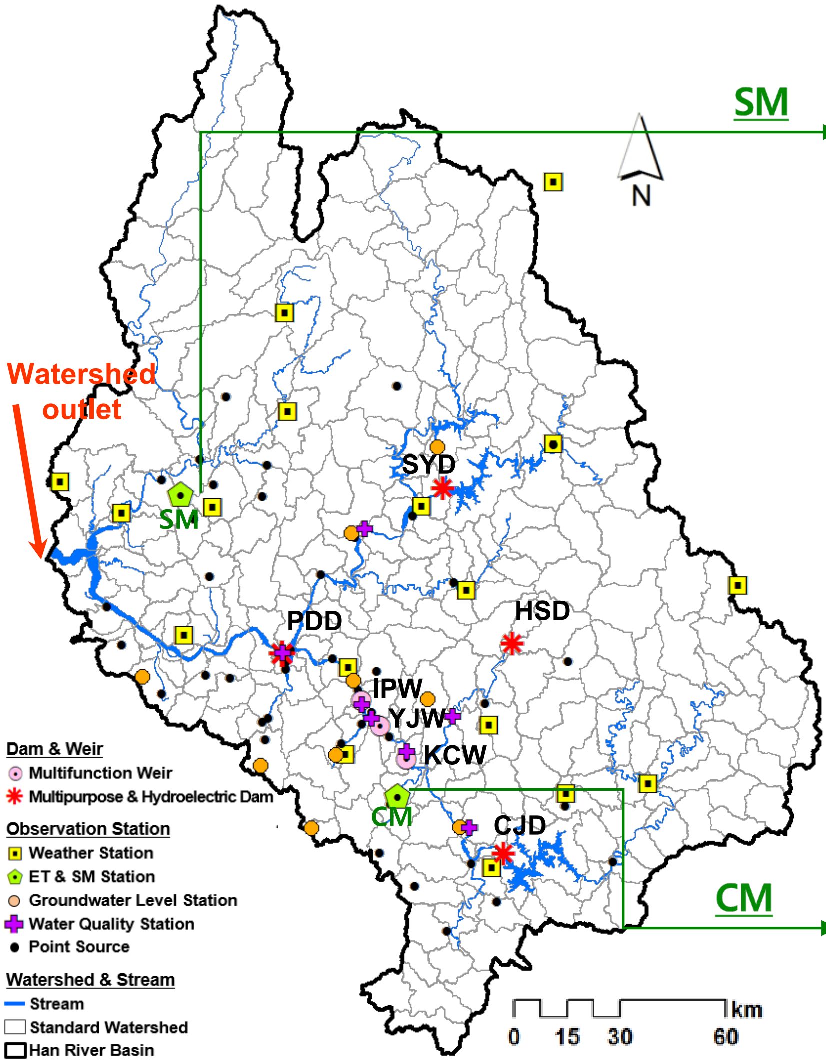
SWAT Model calibration and validation

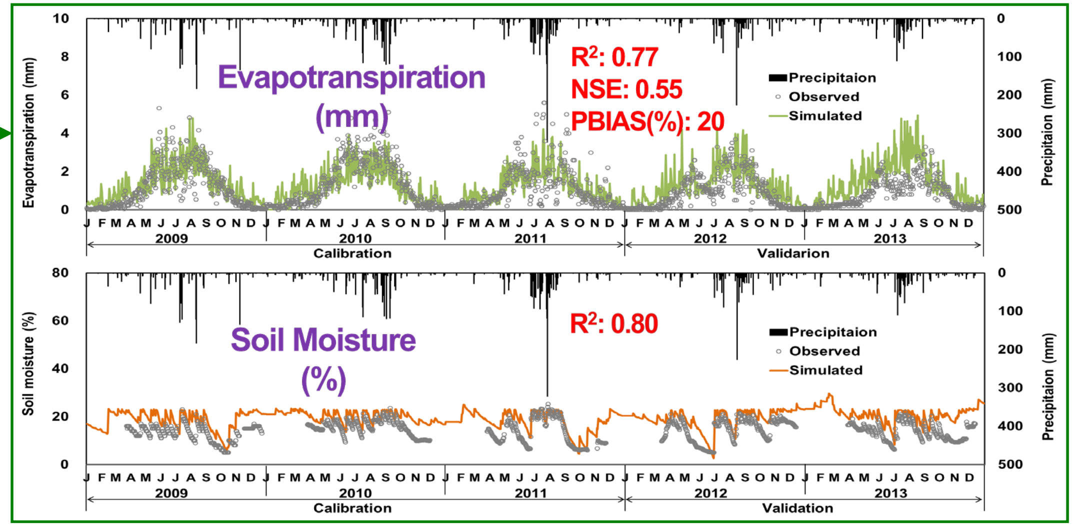
- Hydrology Results

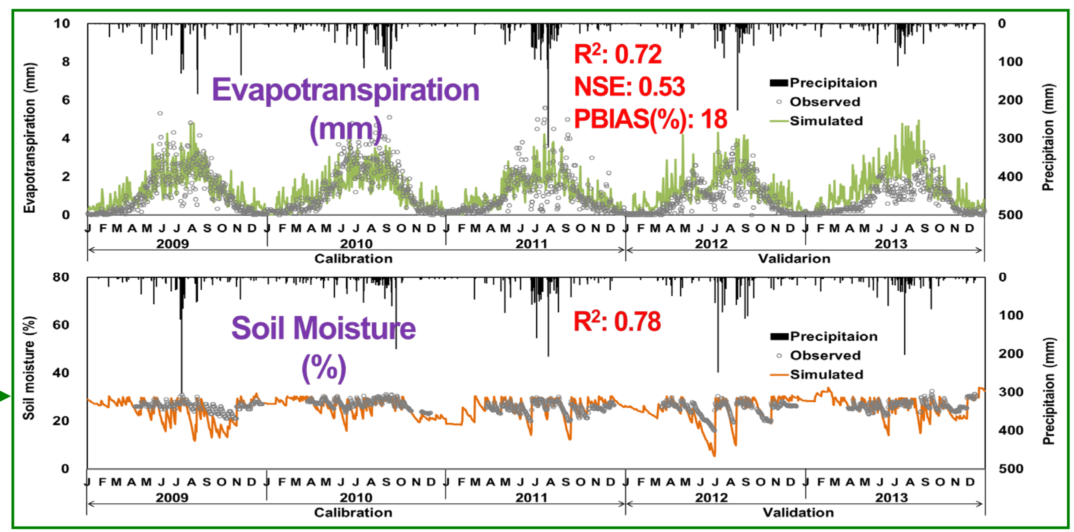
Ahn et al. (2015)

Observed vs. simulated ET & SM results of model calibration and validation

✓ Calibration : 3 years (2009-2011) / Validation : 2 years (2012-2013)







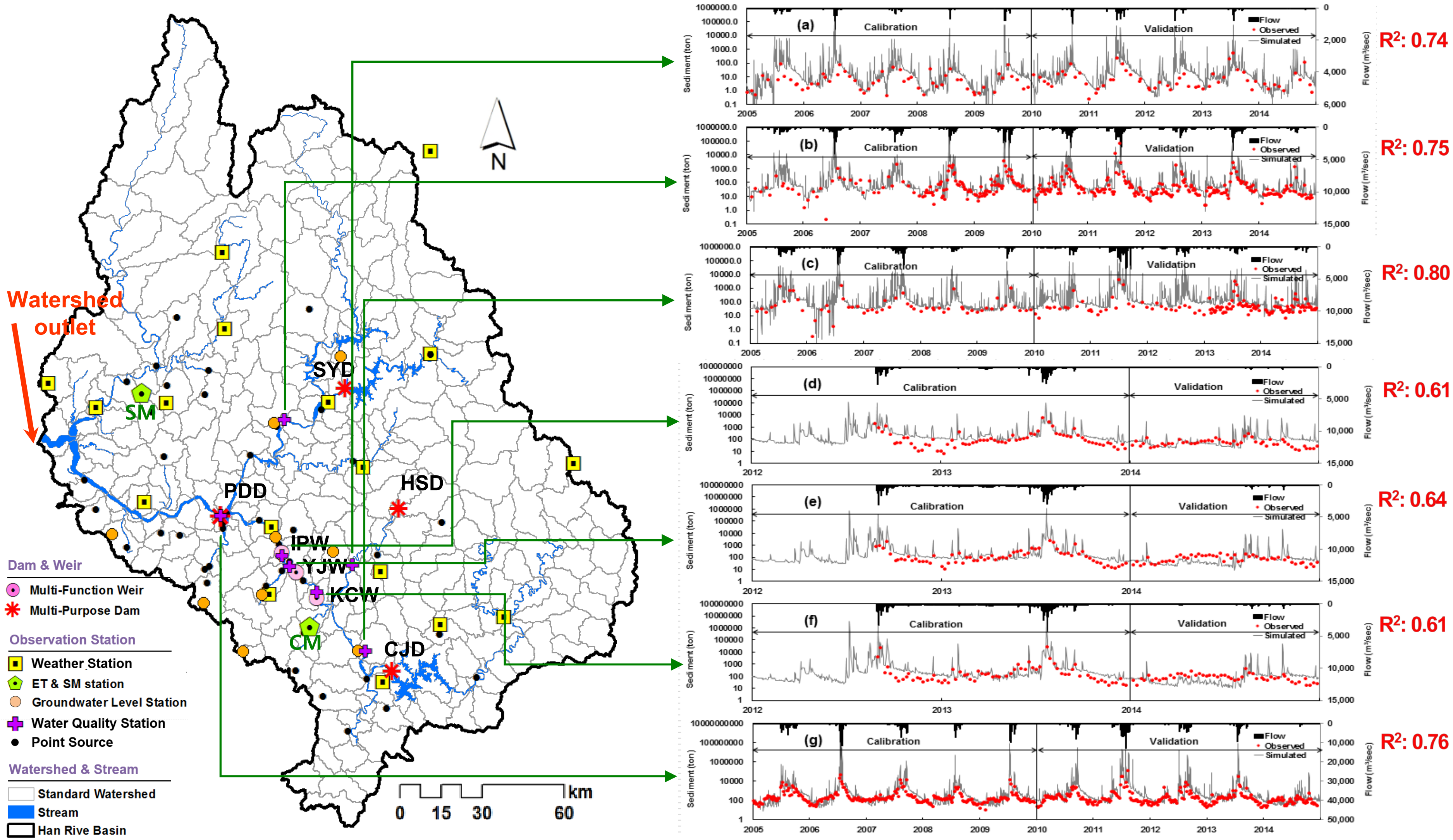
◆ SWAT Model calibration and validation

- Water quality

Ahn et al. (2015)

Observed vs. simulated sediment results of SWAT model calibration and validation

✓ Calibration : 5 years (2005-2009) / Validation : 5 years (2010-2014) at 7 stations



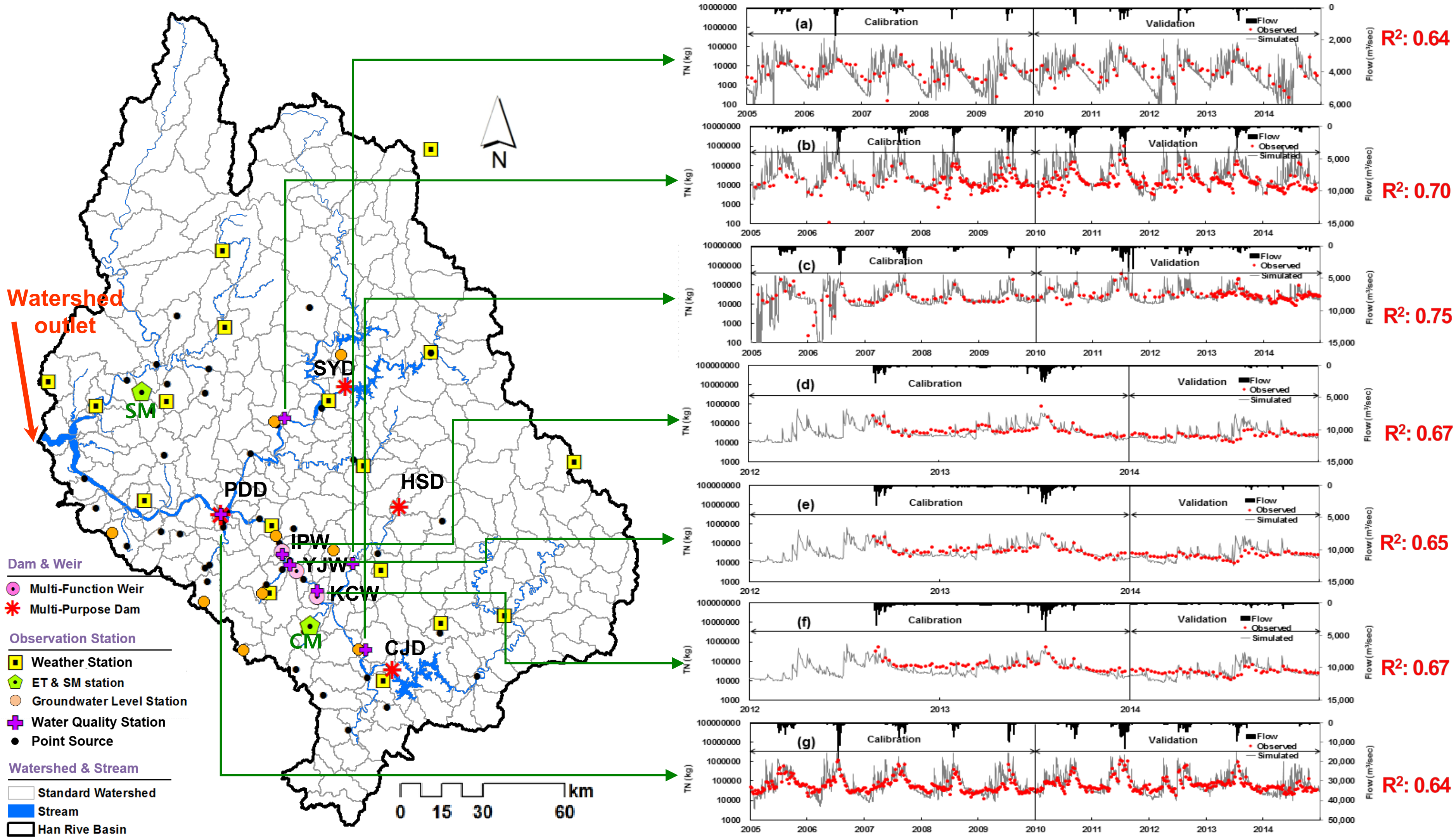
◆ SWAT Model calibration and validation

- **Waterquality**

Ahn et al. (2015)

Observed vs. simulated <u>T-N</u> results of SWAT model calibration and validation

✓ **Calibration : 5 years (2005-2009) / Validation : 5 years (2010-2014) at 7 stations**



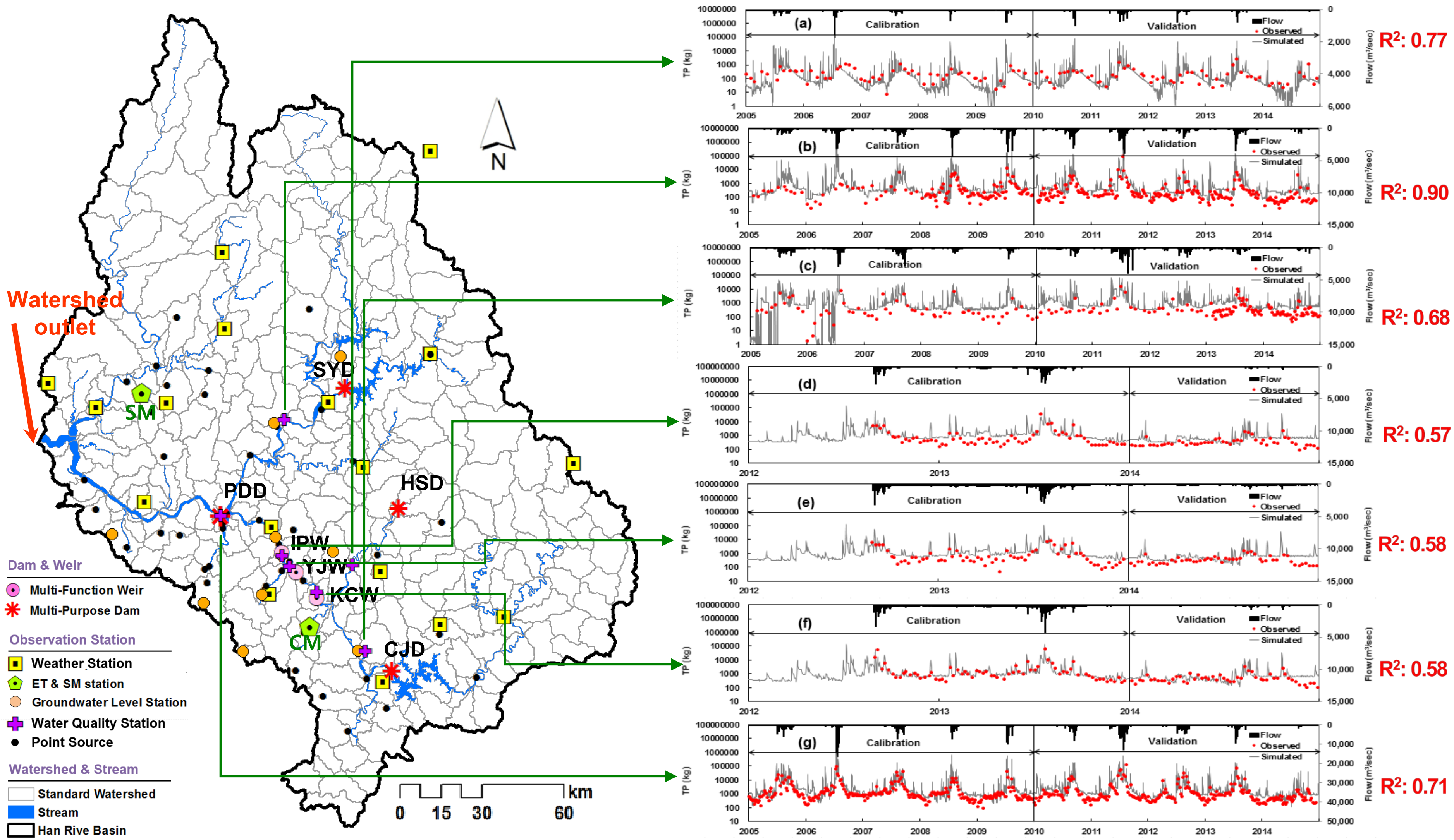
SWAT Model calibration and validation

- Waterquality

Ahn et al. (2015)

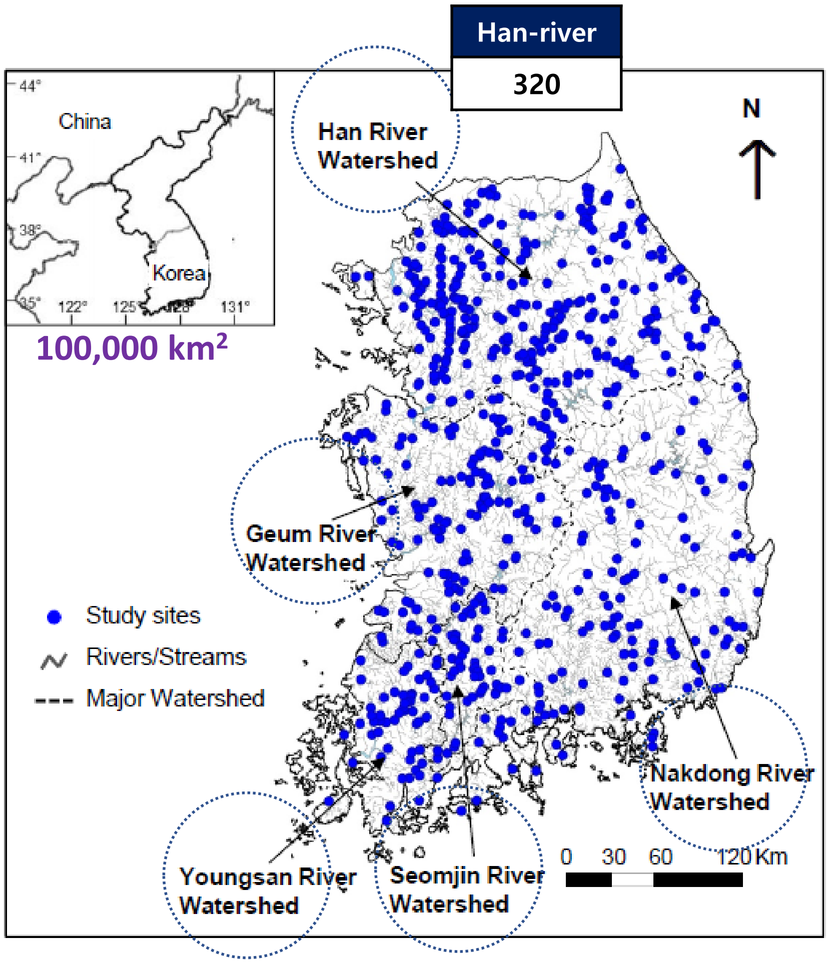
Observed vs. simulated T-P results of SWAT model calibration and validation

✓ Calibration : 5 years (2005-2009) / Validation : 5 years (2010-2014) at 7 stations



◆ Data for Aquatic Ecology Health Index

- South Korea has the **National Aquatic Ecological Monitoring Program (NAEMP)** operated by the Ministry of Environment and the National Institute of Environmental Research, Korea.



- since 2007 for entire country
- spring (April to May) and autumn (September to October) in twice a year
- Measurement components : water temperature, pH, DO, BOD, NH_4, NO_3, T-N, T-P, PO_4, Chlorophyll-a, and so on.
- from that components, TDI (Trophic Diatom Index), FAI (Fish Assessment Index), and BMI (Benthic Macroinvertebrate Index) have be estimated

Locations of the sampling sites in five major watersheds in South Korea

Int. J. Environ. Res. Public Health 2012, 9, 3629-3653

◆ Aquatic Ecology Health Index

- **Fish Assessment Index, FAI (U. S. EPA, 1993)**
  : Calculate the score for 4 metrics (*M1, M2, M3 and M7*) that depends on the stream order of Korea and the other 4 metrics (*M4, M5, M6 and M8*), using class division "0", "6.25", "12.5"

$$\text{FAI} = M1 + M2 + M3 + M4 + M5 + M6 + M7 + M8$$

M1 : Total number of domestic species
M2 : Number of riffle benthic species
M3 : Number of sensitive species
M7 : Total population of sampled domestic species
M4 : Population ratio of Tolerant species
M5 : Population ratio of Omnivores
M6 : Population ratio of domestic Insectivores
M8 : Population ratio of Anormalities


Mandarin fish

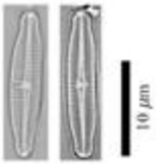

Blue gill

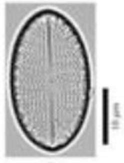

Lon-nosed barbel

- **Trophic Diatom Index, TDI (Kelly and Withon, 1995)**
  : Calculate the score using relative density, contamination sensitivity and indicative value of emerging species for about 400 species of Trophic Diatom

$$\text{TDI} = 100\text{-}(25(\sum ai * si * vi / \sum ai * vi) - 25)$$

ai: Relative abundance of species in specimens, %
si: Contamination sensitivity of species (1 - 5)
vi: Indicative value of species (1 - 3)


Bacillariales


Genus Cocconeis

- **Benthic Macroinvertebrate Index, BMI (Merritt and Cummins 1984)**
  : Calculate the score using benthic macroinvertebrate organisms with contaminate and indicative weighted value

$$\text{BMI} = (4\text{-}\sum si * hi * gi / \sum hi * gi) * 25$$

i: the number of individuals in the sample
si: pollution unit index of species i
hi: appearance rank of species i
gi: weight index of species i


ephemera orientalis mclachlan


Chironomidae **spp.**

Index	A (Very Good)	B (Good)	C (Fair)	D (Poor)	E(Very Poor)
FAI	$100 \geq \sim \geq 80$	$80 > \sim \geq 60$	$60 > \sim \geq 40$	$40 > \sim \geq 20$	$20 > \sim \geq 0$
TDI	$100 \geq \sim \geq 90$	$90 > \sim \geq 70$	$70 > \sim \geq 50$	$50 > \sim \geq 30$	$30 > \sim \geq 0$
BMI	$100 \geq \sim \geq 80$	$80 > \sim \geq 65$	$65 > \sim \geq 50$	$50 > \sim \geq 35$	$35 > \sim \geq 0$

◆ Aquatic Ecology Health Index

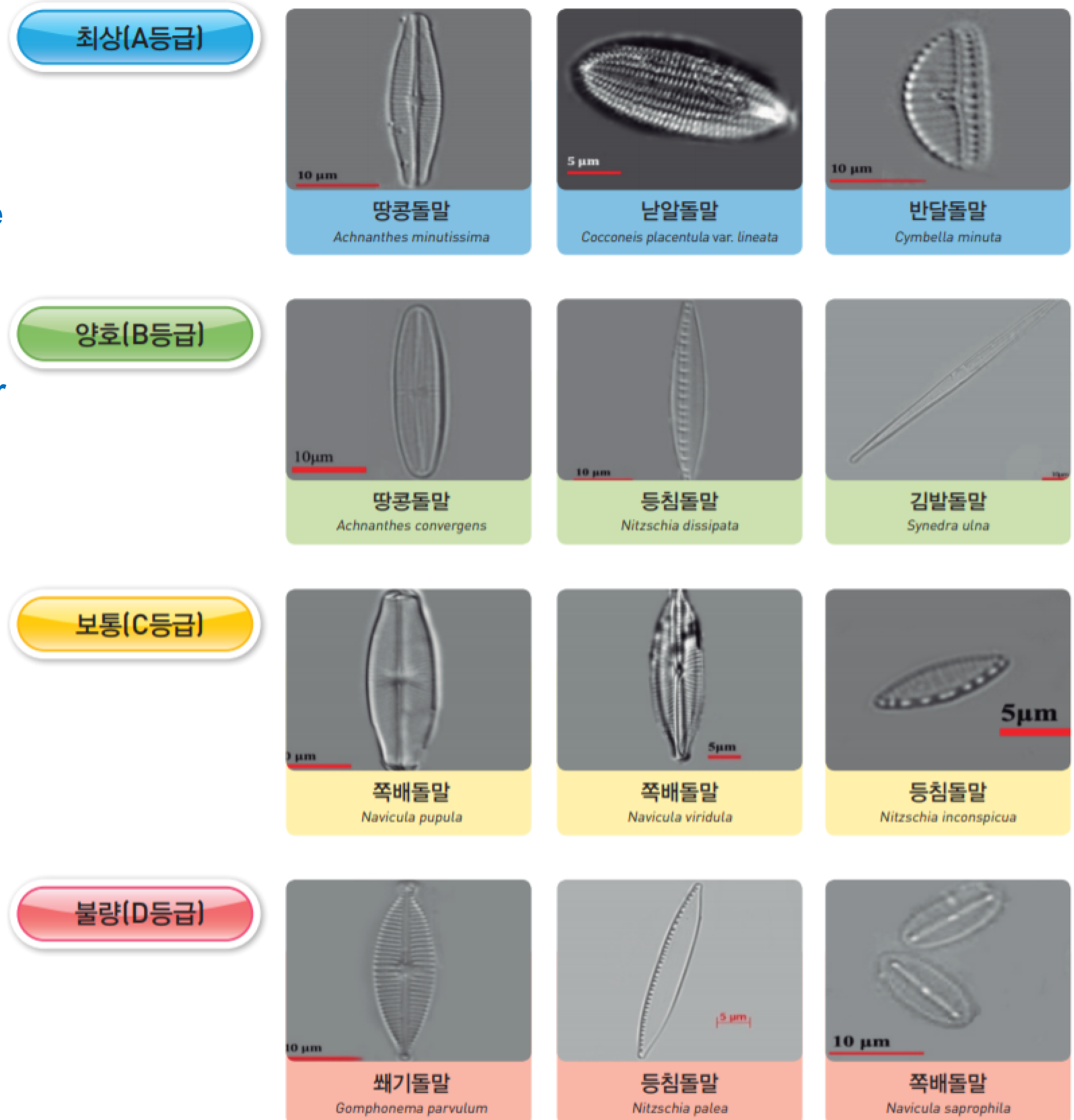
Fish Assessment Index, FAI :

- It refers to the organism at the top level of the food chain in the water body that represents omnivorous, herbivorous, inflorescence, and carnivorous at various nutritional stages.

Aquatic Ecology Health Index

Trophic Diatom Index, TDI :

- As the primary producer of the river ecosystem food chain, it refers to the diatom attached to the stone (substrate) such as gravel or cobble stone in the bottom.
- It is responsible for energy transfer in the ecosystem.
- Also, It is sensitive to changes in water quality (TN, TP) and environment.



◆ Aquatic Ecology Health Index

Benthic Macroinvertebrate Index, BMI :

- It is a biologic indicator that reflects local characteristics as a primary or secondary consumer of river ecosystem.

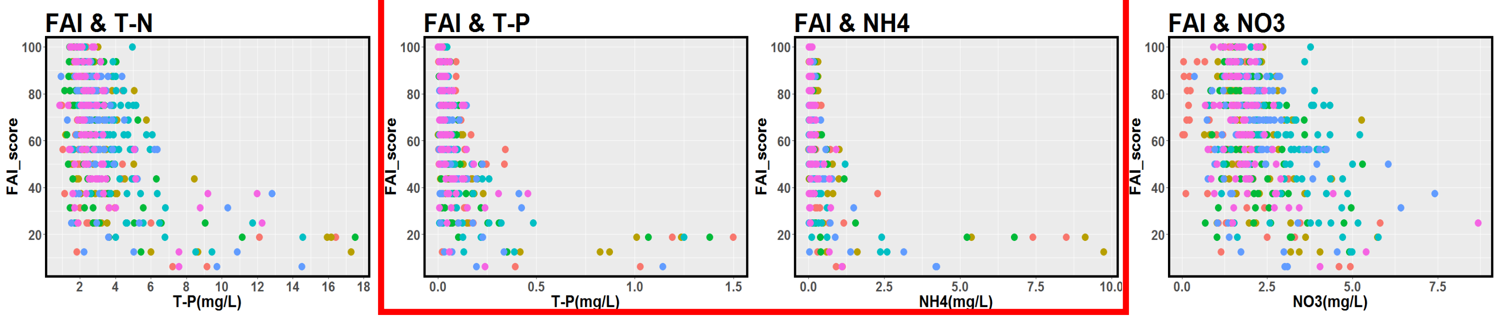
- It refers to a group of aquatic insects.

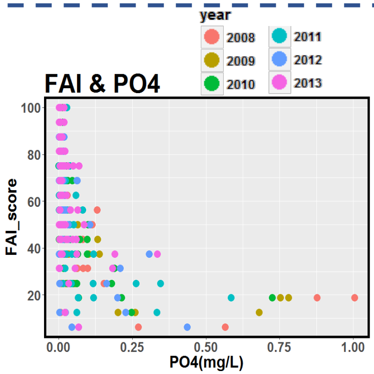
- As a sub-consumer linking producers and upper consumers in aquatic ecosystem food chain, they are sensitive to environmental changes and have excellent indicators and are used as indicators of water quality evaluation.

◆ Correlation Analysis

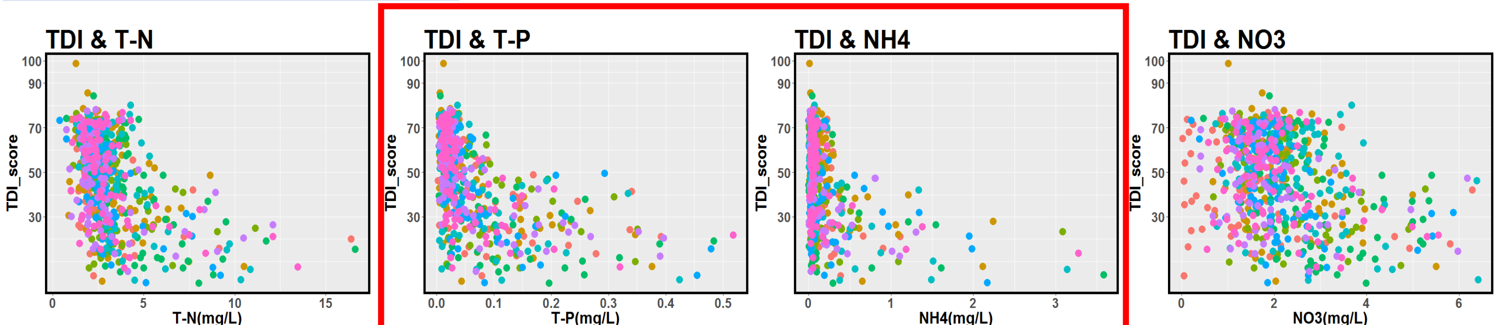
- The relationship between the score and the water quality is not clear

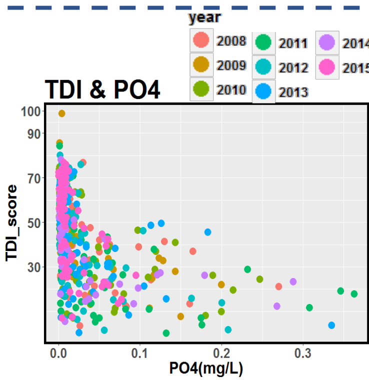
FAI & WQ Relationship



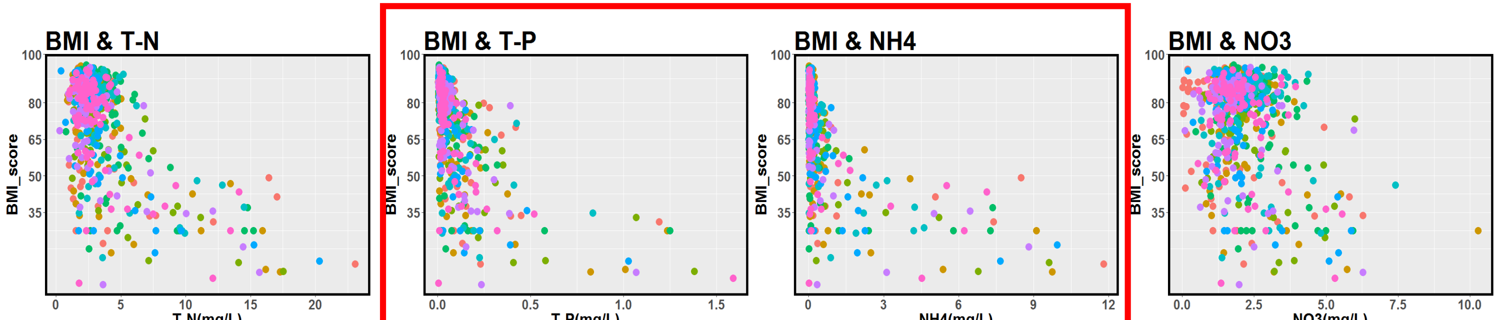


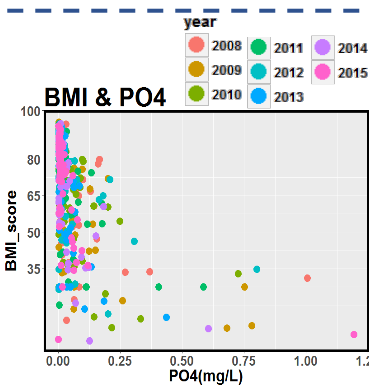
TDI & WQ Relationship





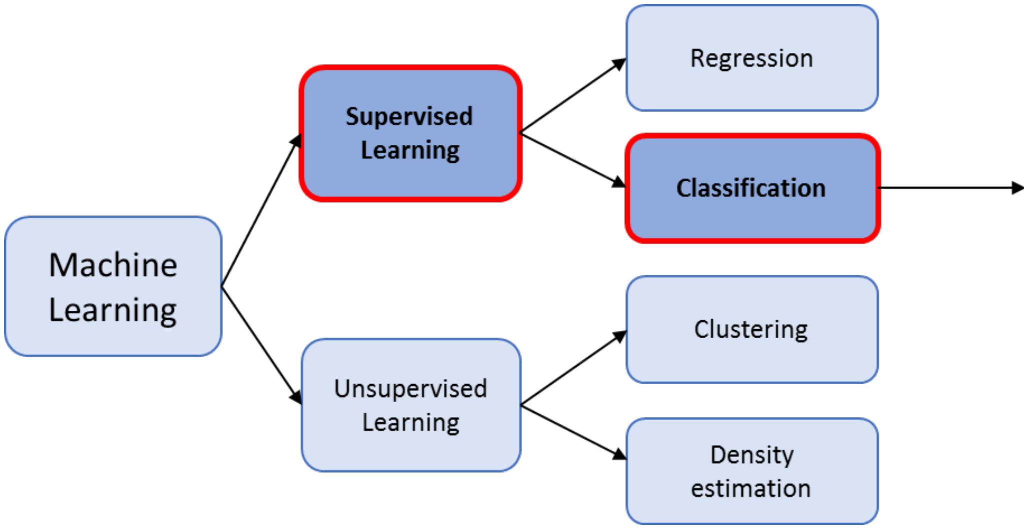
BMI & WQ Relationship





◆ Machine Learning

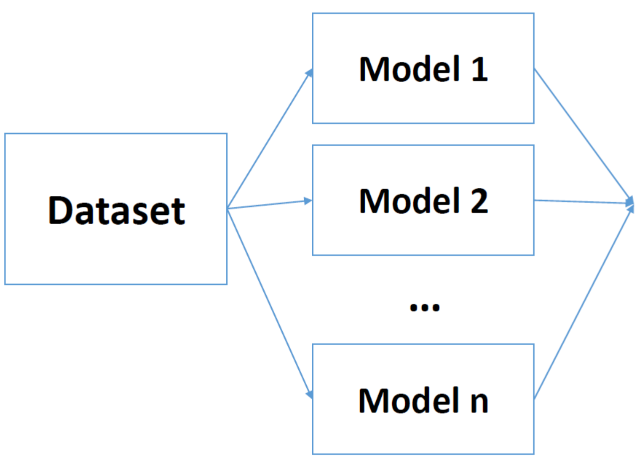
General Machine Learning

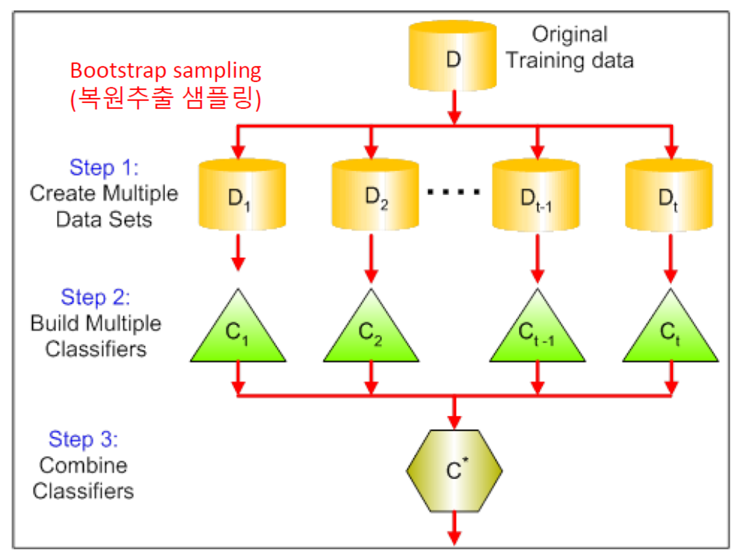


Classification

- Logistic Regression (LR)
- Linear Discriminant Analysis (LDA)
- K-nearest Neighbors Classifier (KNN)
- Classification And Regression Tree, Decision Tree (CART)
- Gaussian Naive Bayes (NB)
- Random Forest Classifier (RF)

Ensemble Machine Learning

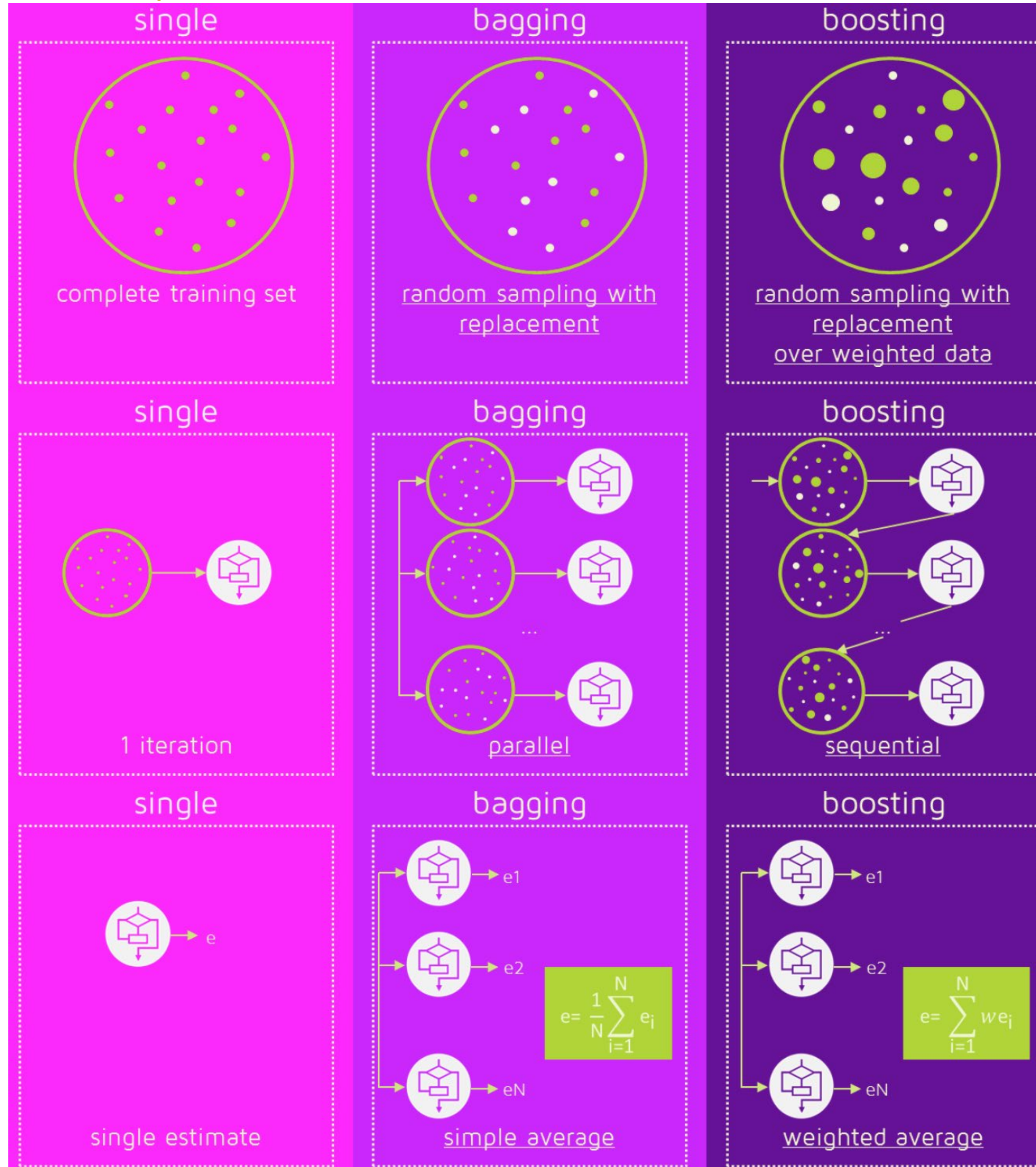




Machine Learning

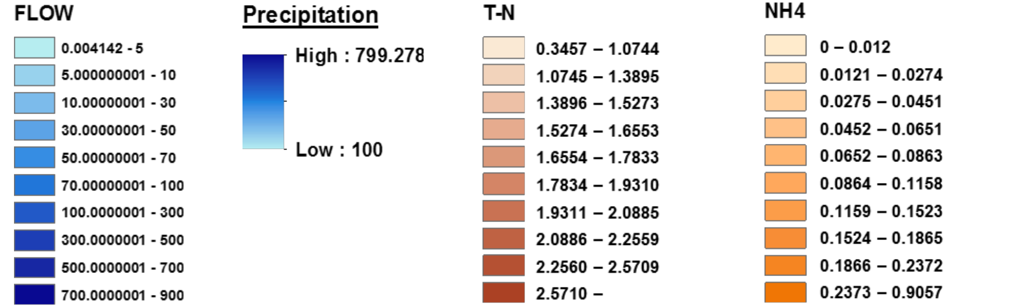
eXtreme Gradient Boosting tree (XGBoost)

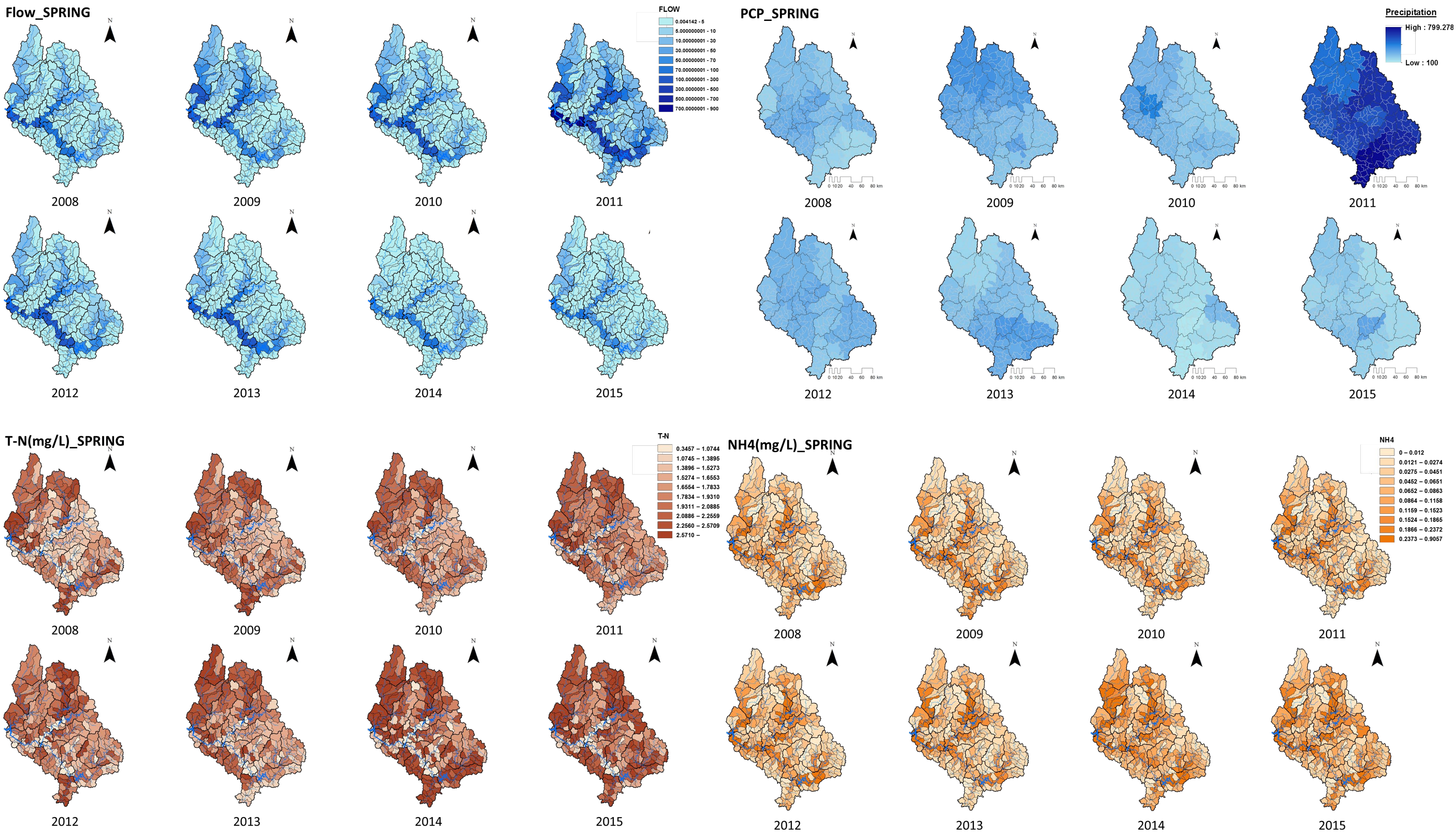
- Algorithm to make decision **by mixing several models** (tree boosting model)
- The algorithm that **learn**s results **sequentially**, with previous result and previous result affect the next model result in the current stage (**additive training**)
- This is a system in which the predicted value **(Y) gradually approaches** the **target value ($\hat{Y}$) as the stage progresses for next stage**
- So, this learn **weakly learning**, and gradually get closer to actual value unlike random forest model
- The method that weakly fit current data has a **high bias but low variance**.
- The **high bias** can be improved sufficiently **by sequentially learning data weakly.**



◆ Machine Learning

Input variables





Machine Learning

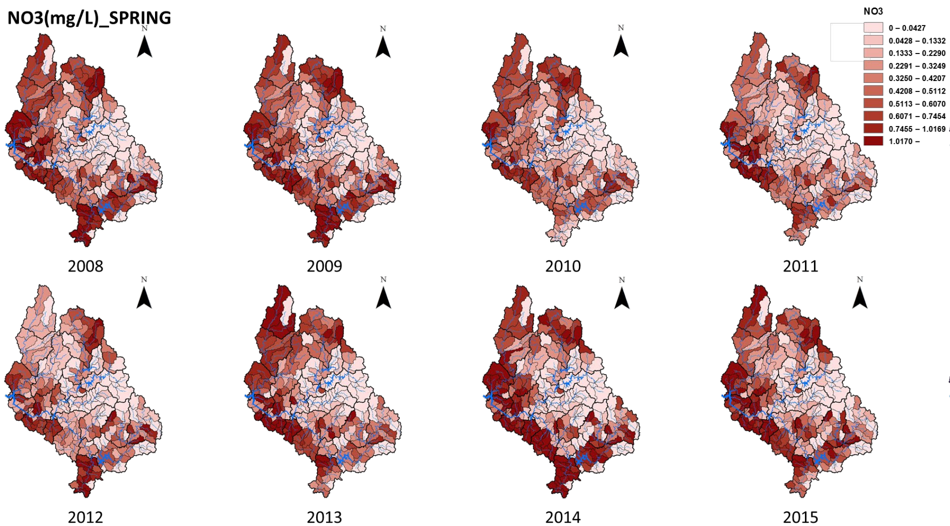
Input variables

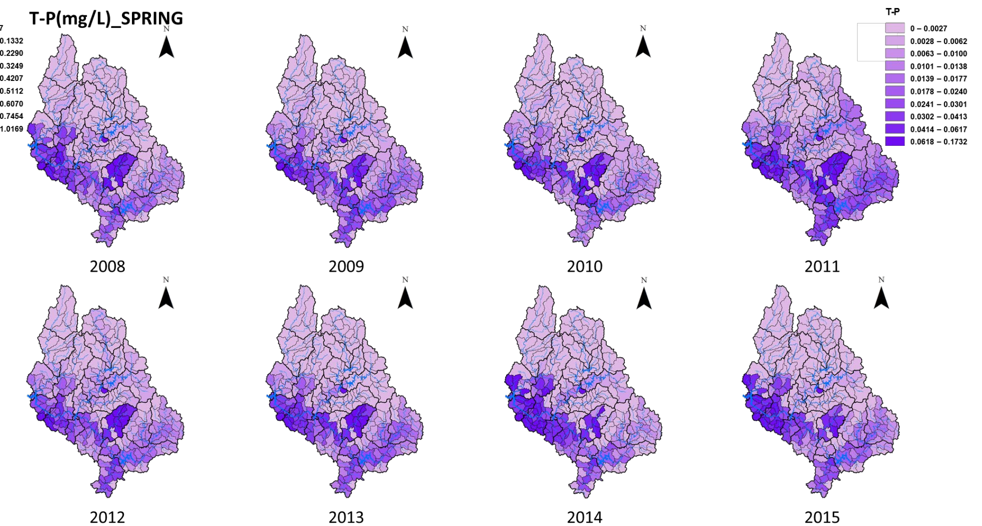
NO3
0 – 0.0427
0.0428 – 0.1332
0.1333 – 0.2290
0.2291 – 0.3249
0.3250 – 0.4207
0.4208 – 0.5112
0.5113 – 0.6070
0.6071 – 0.7454
0.7455 – 1.0169
1.0170 –

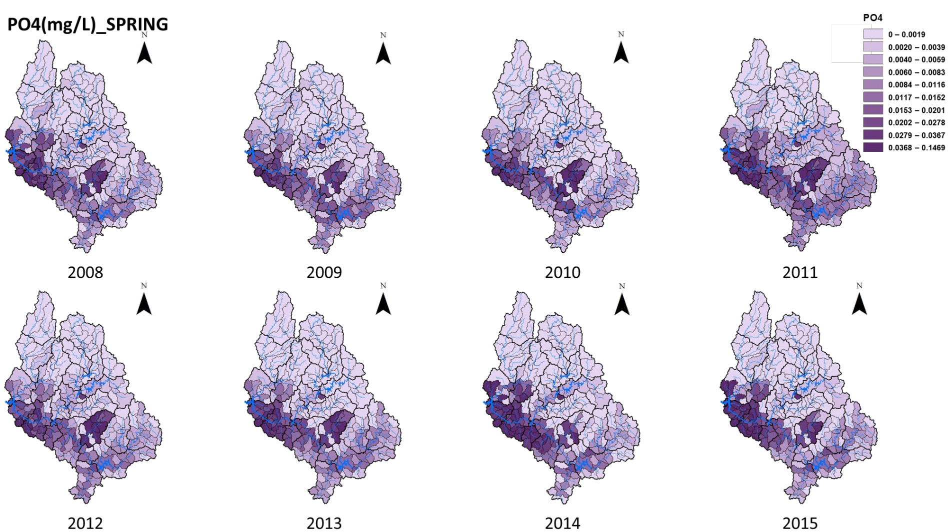
T-P
0 – 0.0027
0.0028 – 0.0062
0.0063 – 0.0100
0.0101 – 0.0138
0.0139 – 0.0177
0.0178 – 0.0240
0.0241 – 0.0301
0.0302 – 0.0413
0.0414 – 0.0617
0.0618 – 0.1732

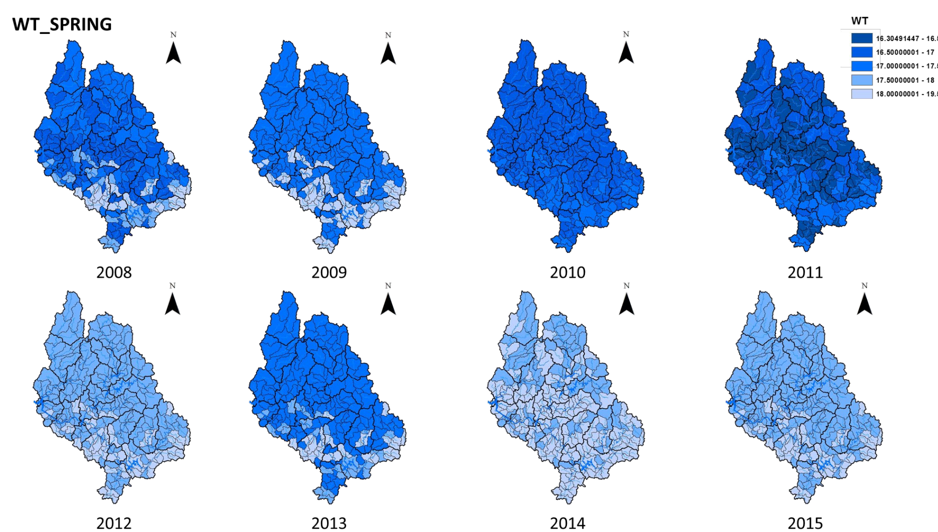
PO4
0 – 0.0019
0.0020 – 0.0039
0.0040 – 0.0059
0.0060 – 0.0083
0.0084 – 0.0116
0.0117 – 0.0152
0.0153 – 0.0201
0.0202 – 0.0278
0.0279 – 0.0367
0.0368 – 0.1469

WT (Water temperature, °C)
16.30491447 - 16.5
16.50000001 - 17
17.00000001 - 17.5
17.50000001 - 18
18.00000001 - 19.5









Training technique (Random forest)

K-fold Cross validation

- Generally, the data divided into **training (70%) and test (30%)**.
- To reduce variability, cross-validation are performed using different partitions and results of the **different partitions** are combined (e.g. **averaged**) to estimate a final predictive model.
- The k-fold cross validation separate training data into **k folds without overlap**.
- **One set of the k folds is separated by training and validation.**

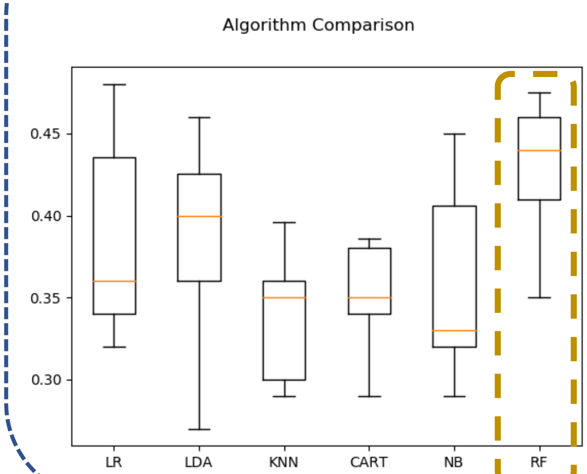
- Logistic Regression (LR)
- Linear Discriminant Analysis (LDA)
- K-nearest Neighbors Classifier (KNN)
- Classification And Regression Tree, Decision Tree (CART)
- Gaussian Naive Bayes (NB)
- **Random Forest Classifier (RF)**

Average accuracy

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_\text{samples}} \sum_{i=0}^{n_\text{samples}-1} 1(\hat{y}_i = y_i)$$

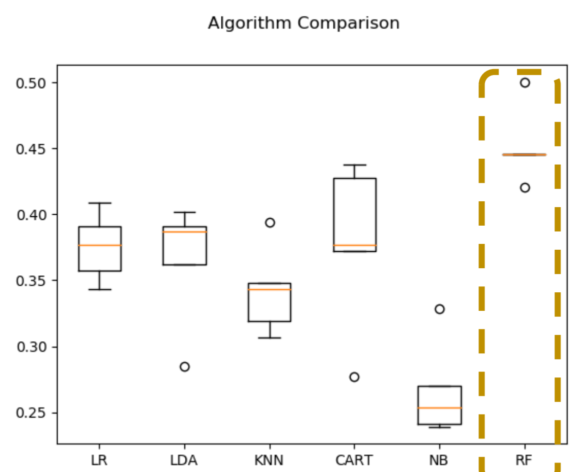
Average accuracy of verification : 0.43

FAI



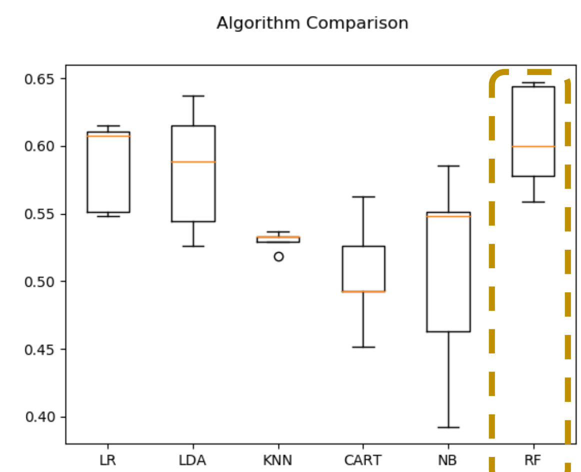
Average accuracy of verification : 0.45

TDI



Average accuracy of verification : 0.58

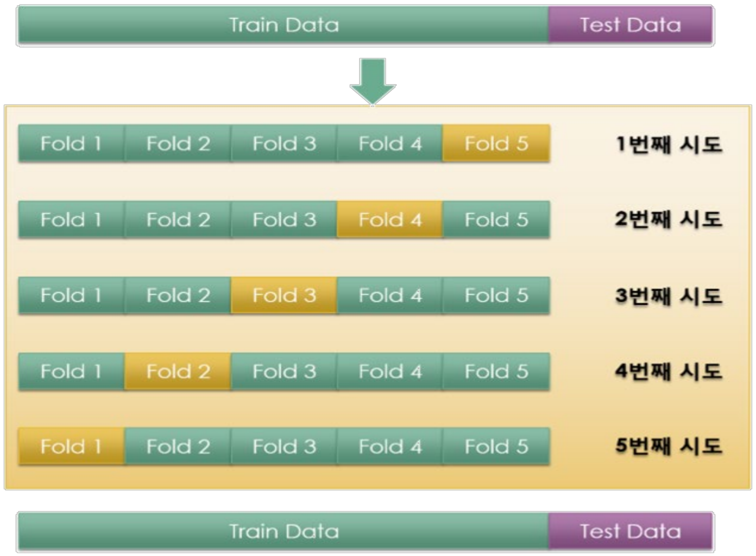
BMI



◆ Training technique (XGBoost)

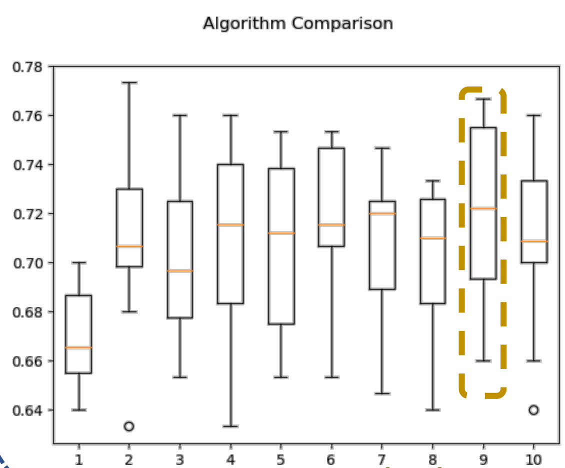
Training & Verification

- The **k-fold cross validation & Grid search**
- Feature importance (**parameter tuning**)
  - **Max-depth** : Maximum tree depth for base learners
  - **Gamma** : Minimum loss reduction required to make a further partition on a leaf node of the tree
- For **optimization** of parameters, max_depth applied from 1 to 10 with 1 intervals and Gamma applied from 0 to 4.5 with 0.5 intervals



FAI

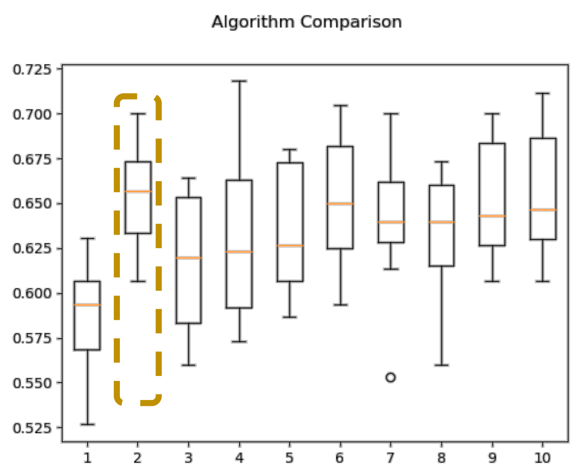
Average accuracy of verification : 0.72



Gamma : 1.5, max_depth : 7

TDI

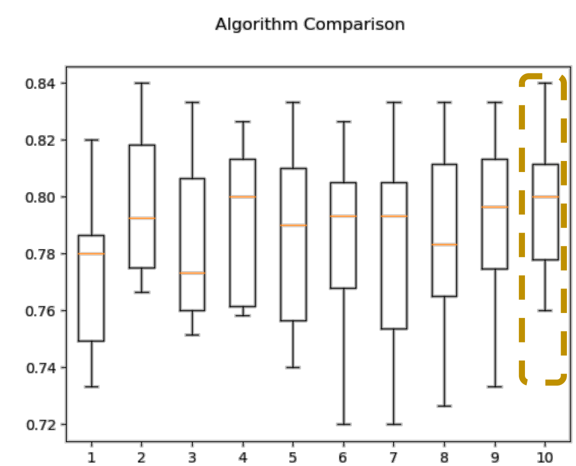
Average accuracy of verification : 0.66



Gamma : 0.15, max_depth : 7

BMI

Average accuracy of verification : 0.80

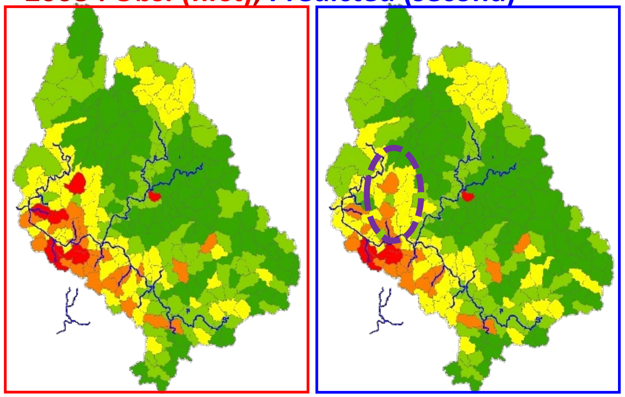


Gamma : 0.05, max_depth : 8

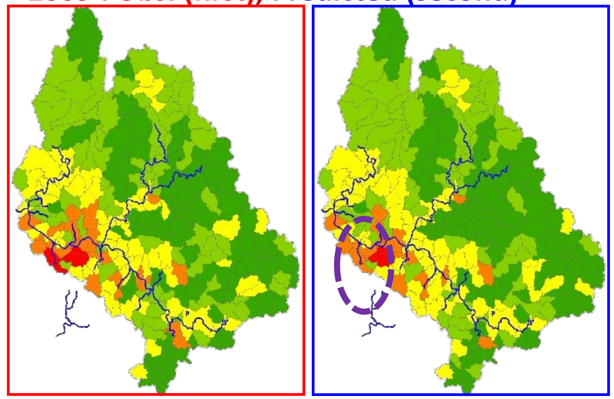
◆ Watershed Health(FAI_spring)

A Very good | B Good | C Normal | D Bad | E Very bad

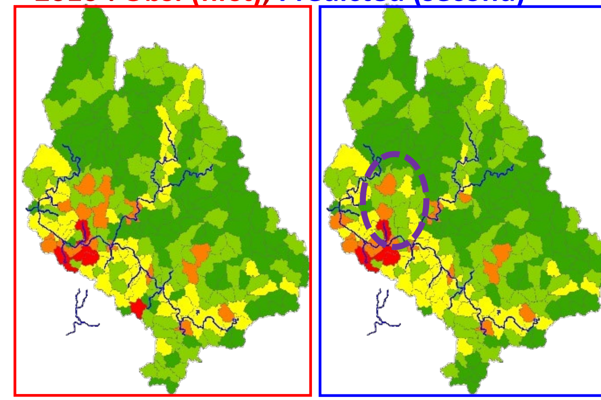
2008 : Obs. (first), Predicted (second)



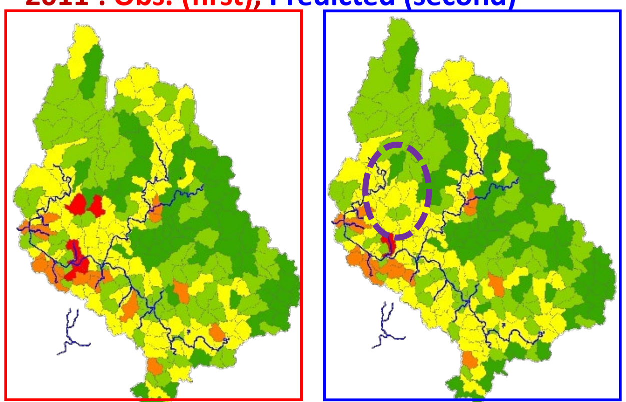
2009 : Obs. (first), Predicted (second)



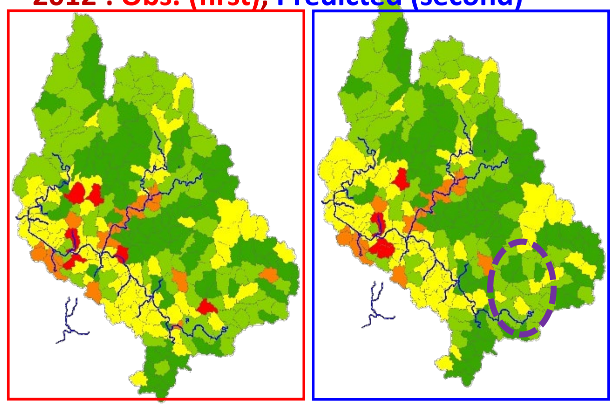
2010 : Obs. (first), Predicted (second)



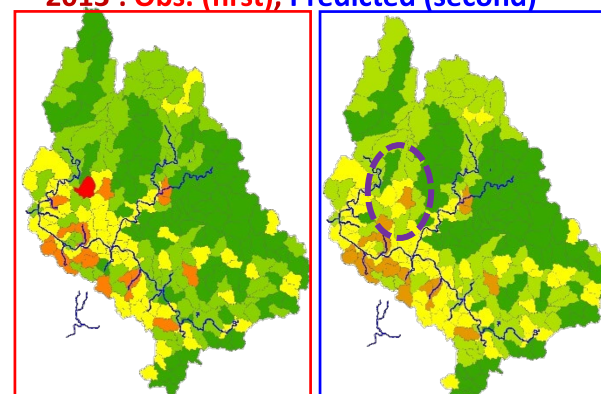
2011 : Obs. (first), Predicted (second)



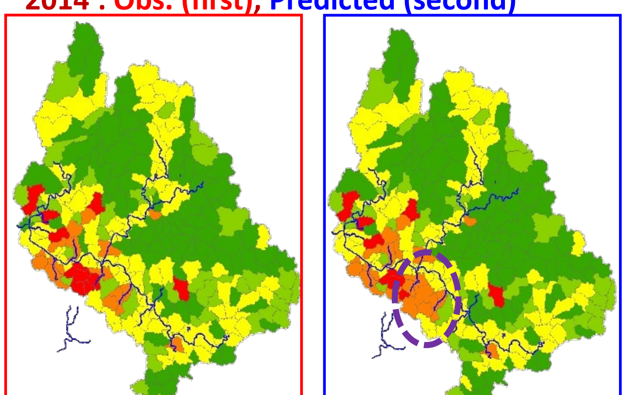
2012 : Obs. (first), Predicted (second)



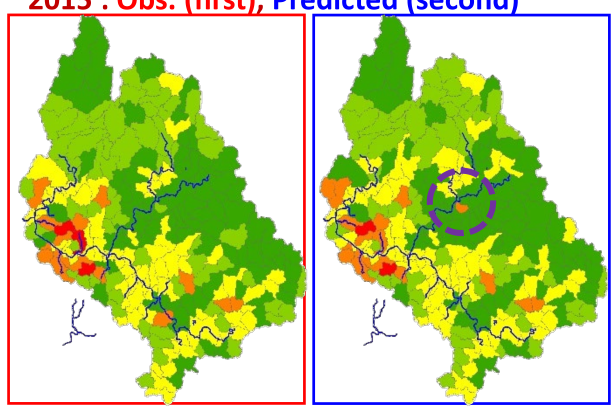
2013 : Obs. (first), Predicted (second)



2014 : Obs. (first), Predicted (second)

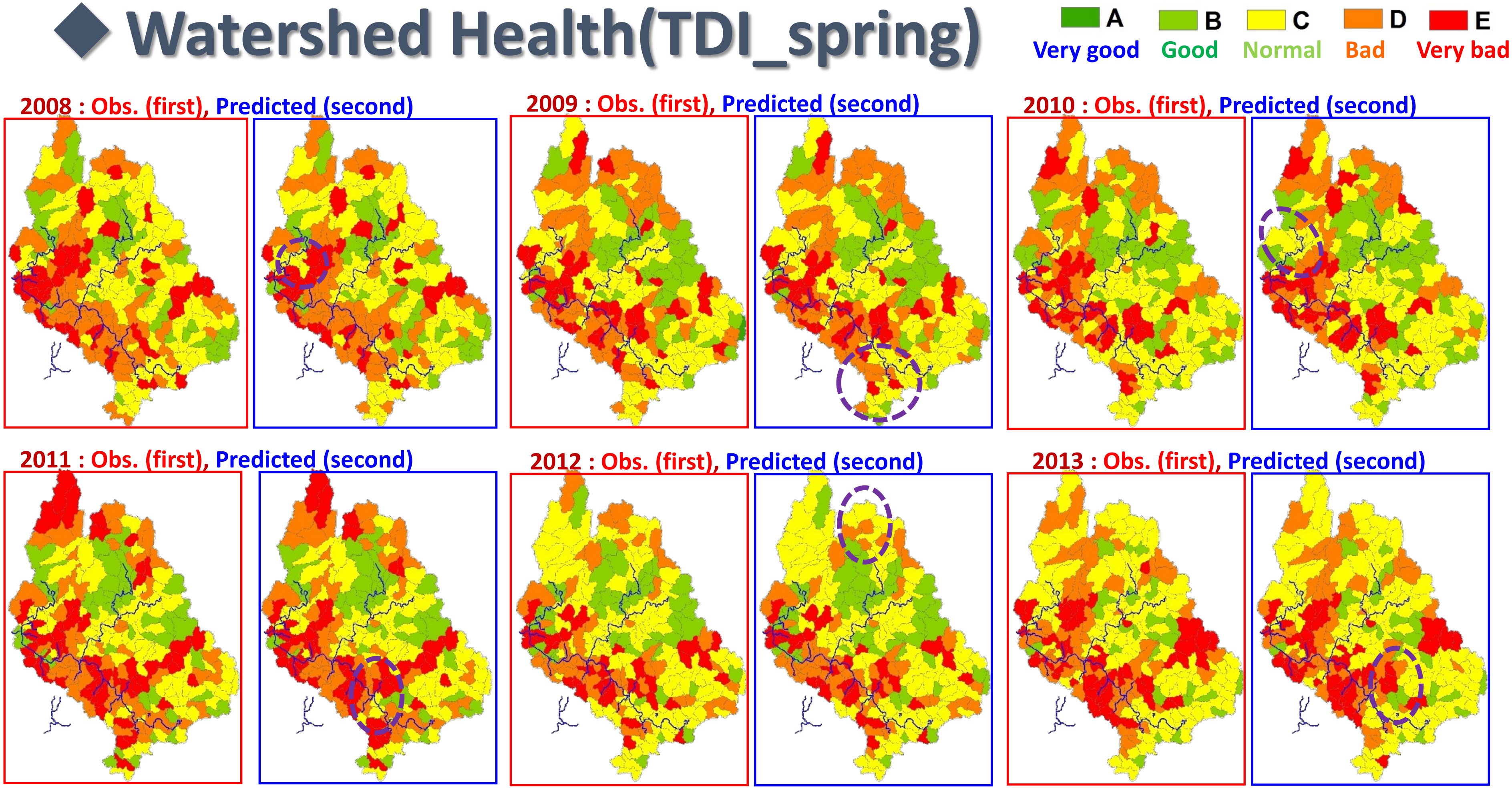


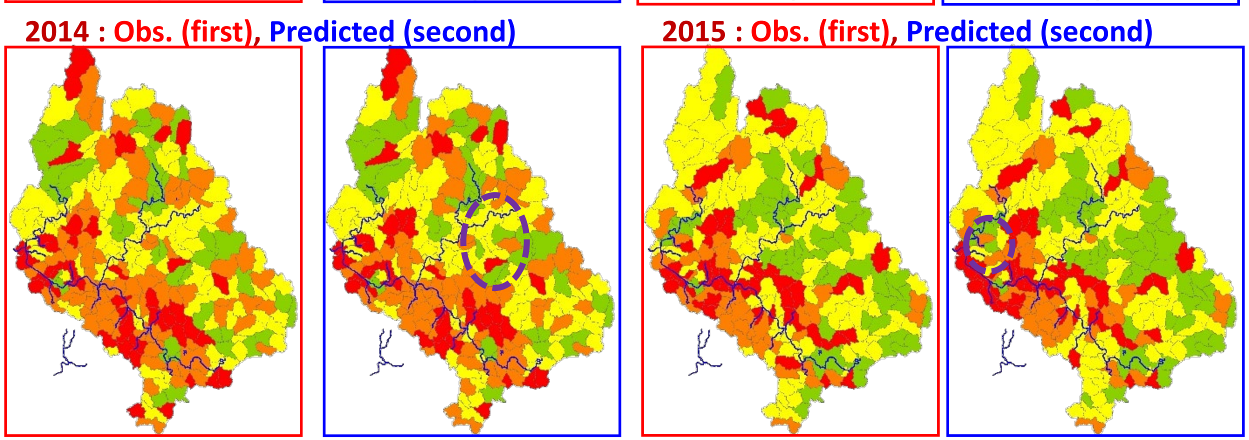
2015 : Obs. (first), Predicted (second)



- ✓ 36 watersheds that have failed to predict.
- ✓ TP and NO_3 in failed watersheds were relatively greater than the overall average TP and NO_3 values.
- ✓ TP was 19.3% greater than whole.
- ✓ PO_4 was 14.4% greater than whole.

◆ Watershed Health(TDI_spring)



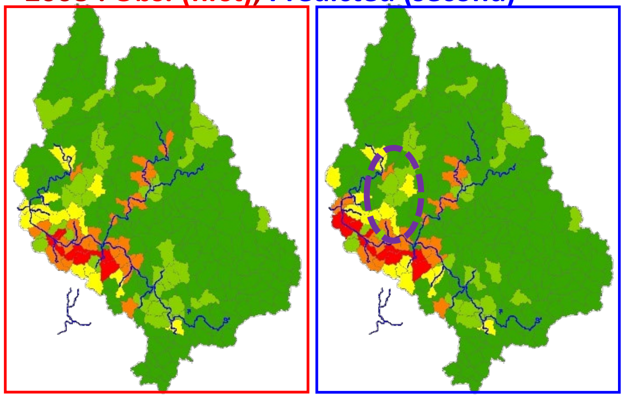


- ✓ 41 watersheds that have failed to predict.
- ✓ TP and PO_4 in failed watersheds were relatively smaller than the whole average TP and PO_4 values.
- ✓ TP was -30.1% smaller than whole.
- ✓ PO_4 was -29.6% smaller than whole.

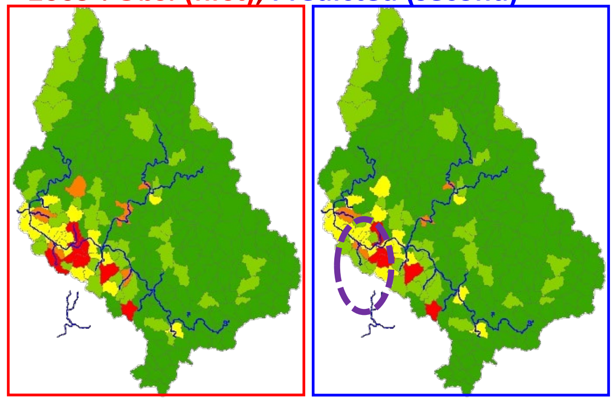
◆ Watershed Health(BMI_spring)

A Very good | B Good | C Normal | D Bad | E Very bad

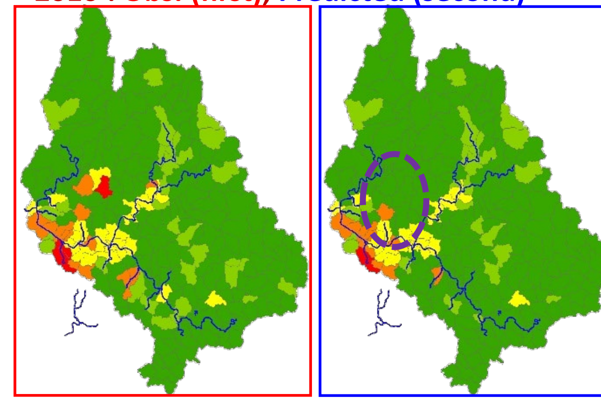
2008 : Obs. (first), Predicted (second)



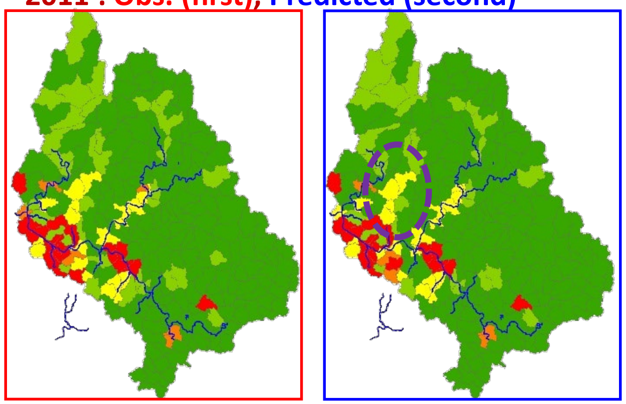
2009 : Obs. (first), Predicted (second)



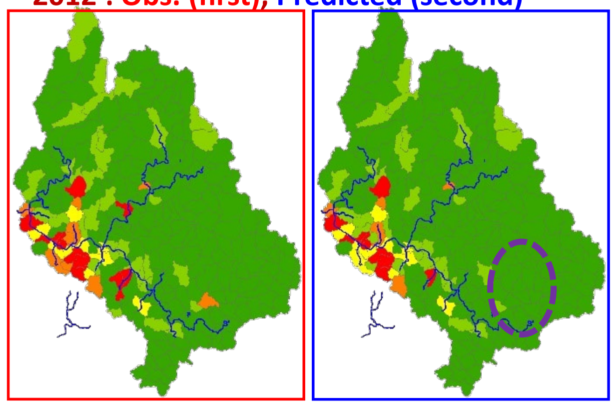
2010 : Obs. (first), Predicted (second)



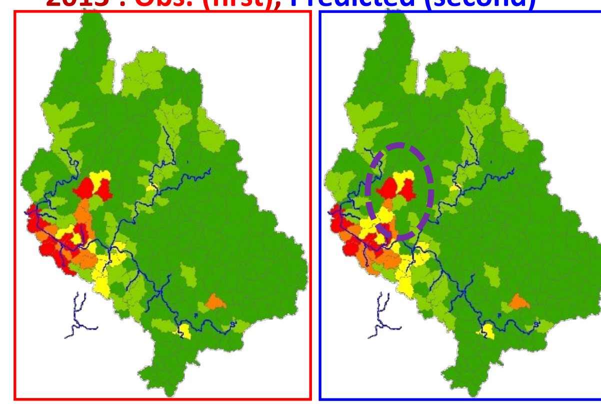
2011 : Obs. (first), Predicted (second)



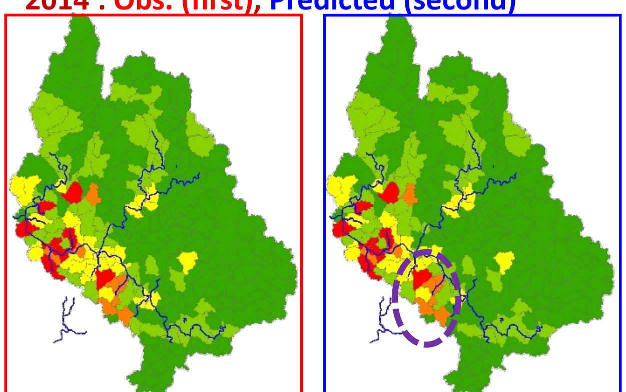
2012 : Obs. (first), Predicted (second)



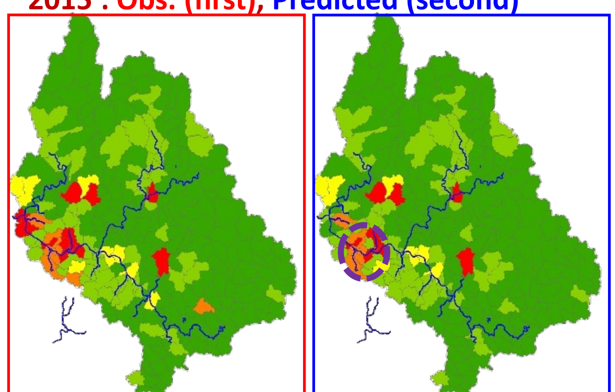
2013 : Obs. (first), Predicted (second)



2014 : Obs. (first), Predicted (second)



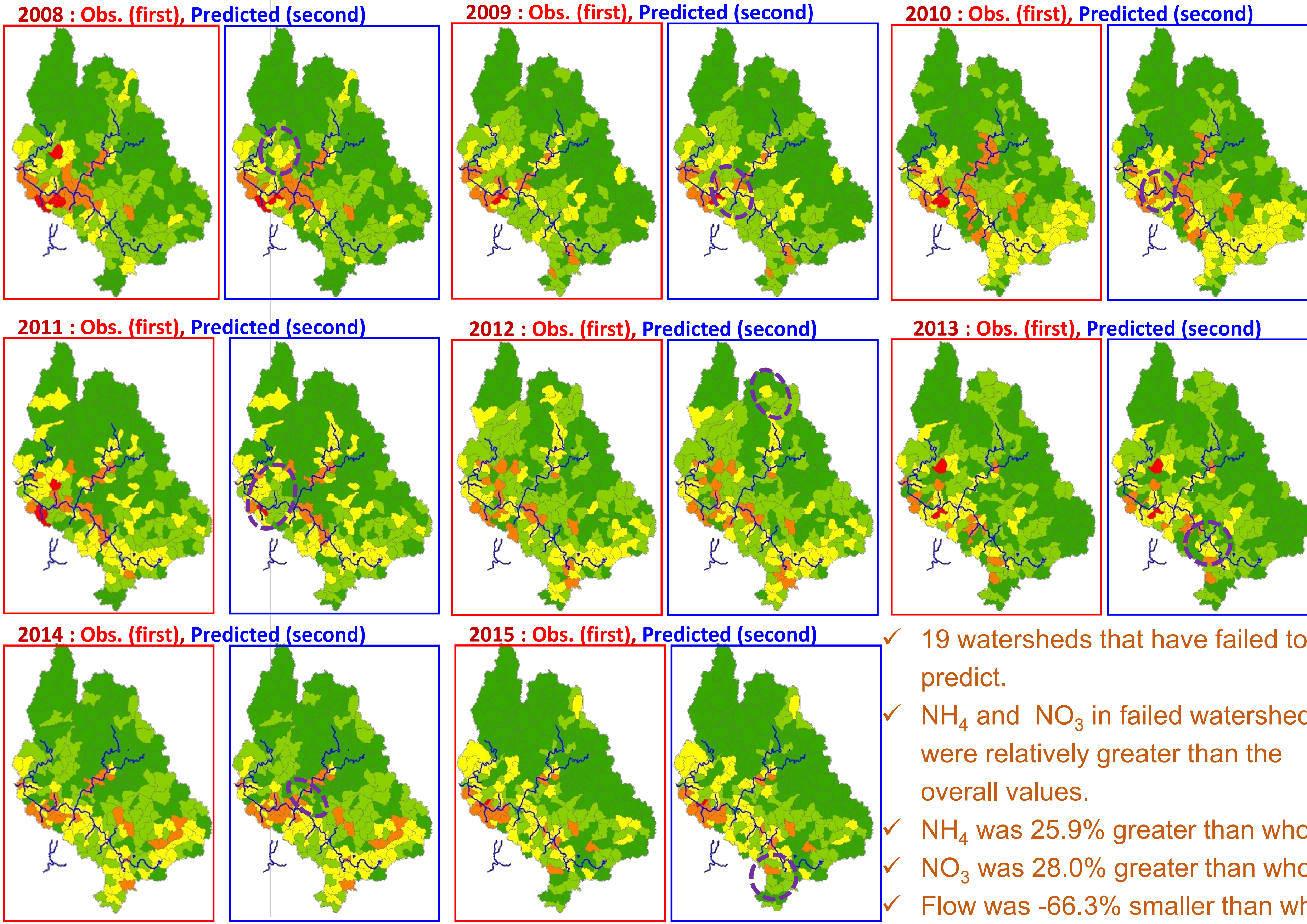
2015 : Obs. (first), Predicted (second)



- ✓ 25 watersheds that have failed to predict.
- ✓ TP and PO_4 in failed watersheds were relatively greater than the overall flow and PO_4 values.
- ✓ TP was 66.1% greater than whole.
- ✓ PO_4 was 81.8% greater than whole.

◆ Watershed Health(FAI_autumn)

A Very good · B Good · C Normal · D Bad · E Very bad

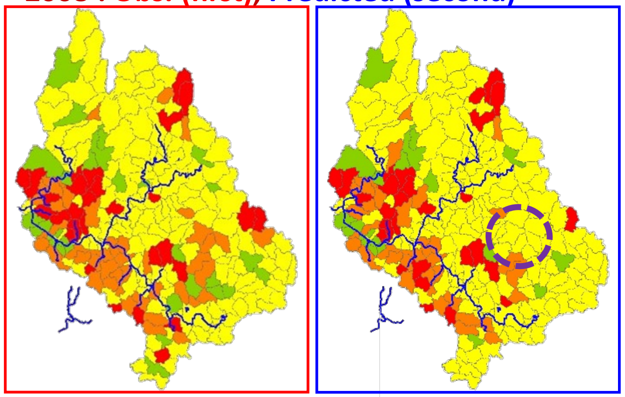


- ✓ 19 watersheds that have failed to predict.
- ✓ NH_4 and NO_3 in failed watersheds were relatively greater than the overall values.
- ✓ NH_4 was 25.9% greater than whole.
- ✓ NO_3 was 28.0% greater than whole.
- ✓ Flow was -66.3% smaller than whole.

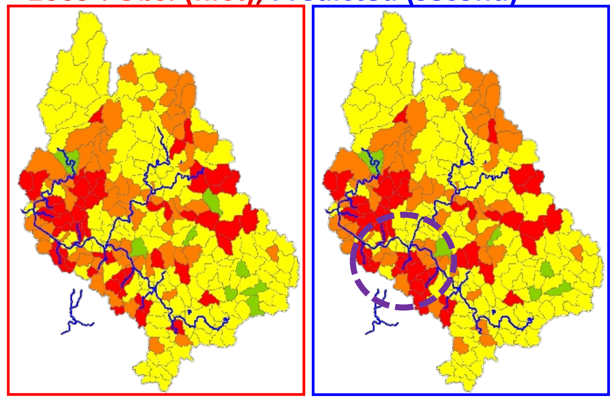
◆ Watershed Health(TDI_autumn)

A Very good | B Good | C Normal | D Bad | E Very bad

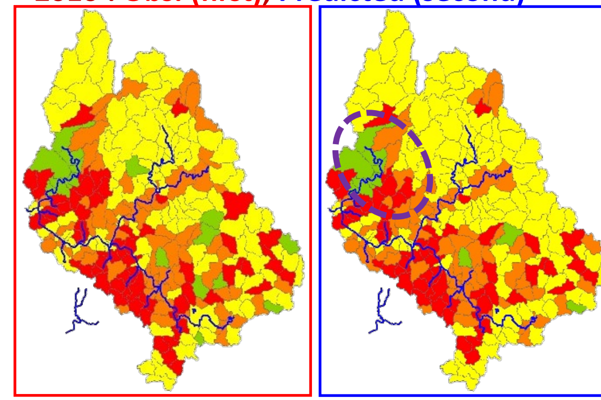
2008 : Obs. (first), Predicted (second)



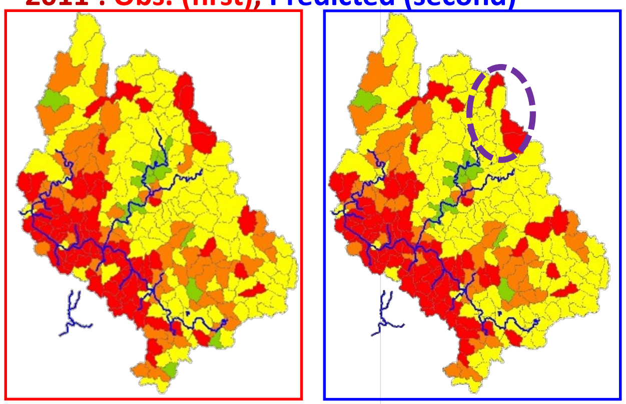
2009 : Obs. (first), Predicted (second)



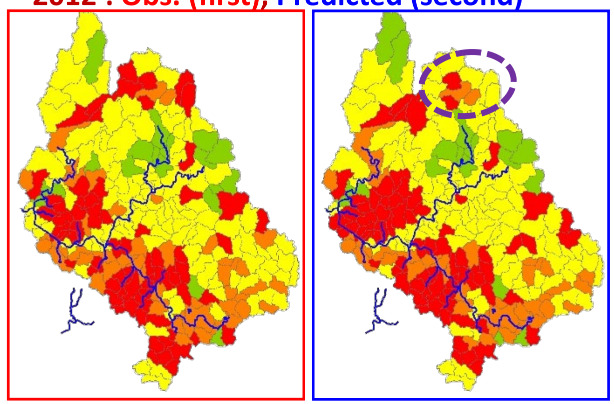
2010 : Obs. (first), Predicted (second)



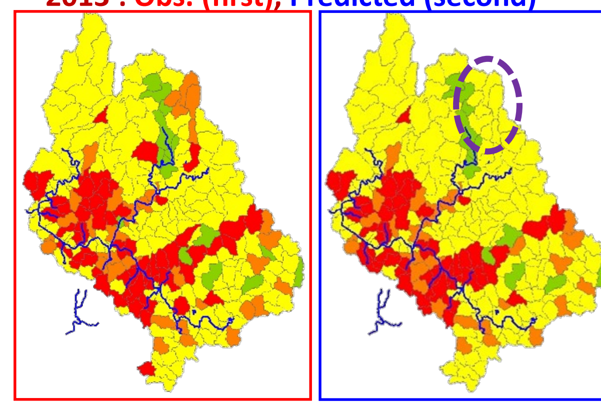
2011 : Obs. (first), Predicted (second)



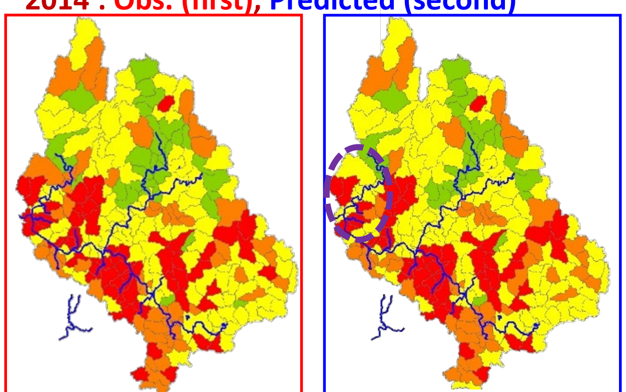
2012 : Obs. (first), Predicted (second)



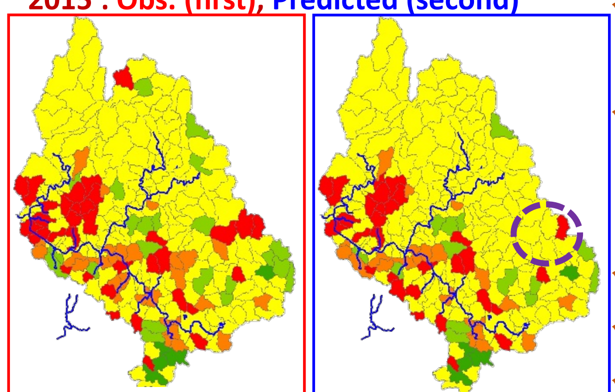
2013 : Obs. (first), Predicted (second)



2014 : Obs. (first), Predicted (second)

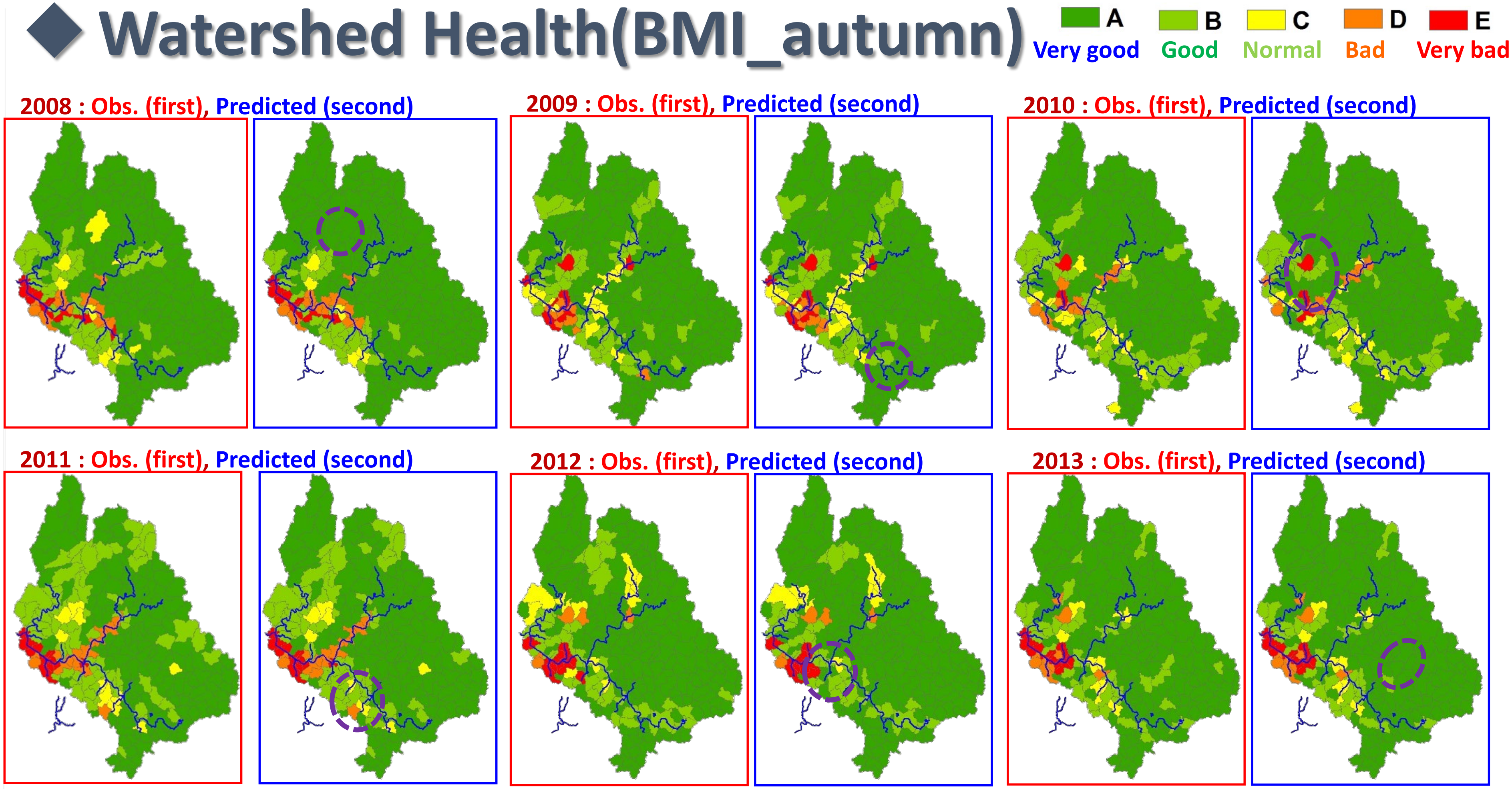


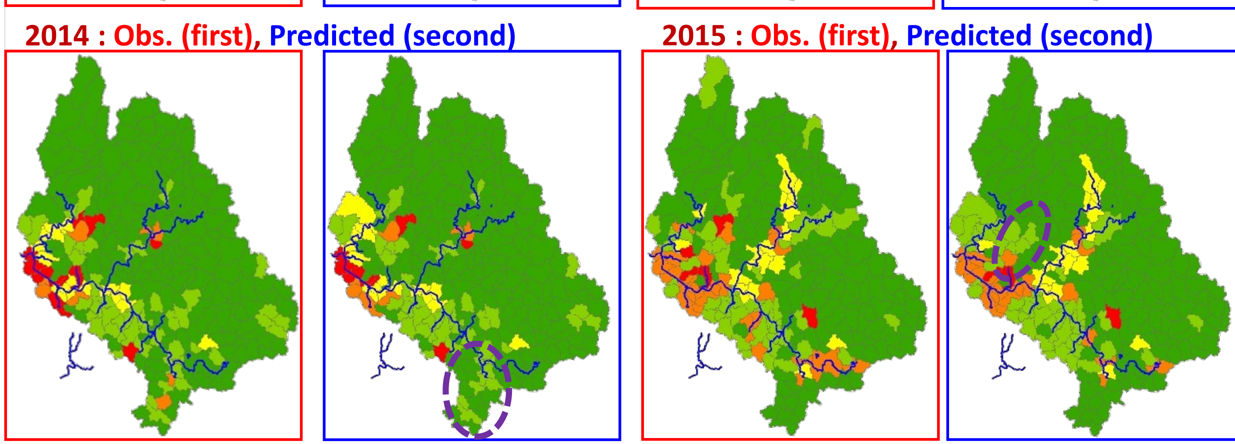
2015 : Obs. (first), Predicted (second)



- ✓ 45 watersheds that have failed to predict.
- ✓ TP and NH_4 in failed watersheds were relatively smaller than the overall values.
- ✓ TP was -23.4% smaller than whole.
- ✓ NH_4 was -31.6% smaller than whole.
- ✓ Flow was -38.6% smaller than whole.

◆ Watershed Health(BMI_autumn)





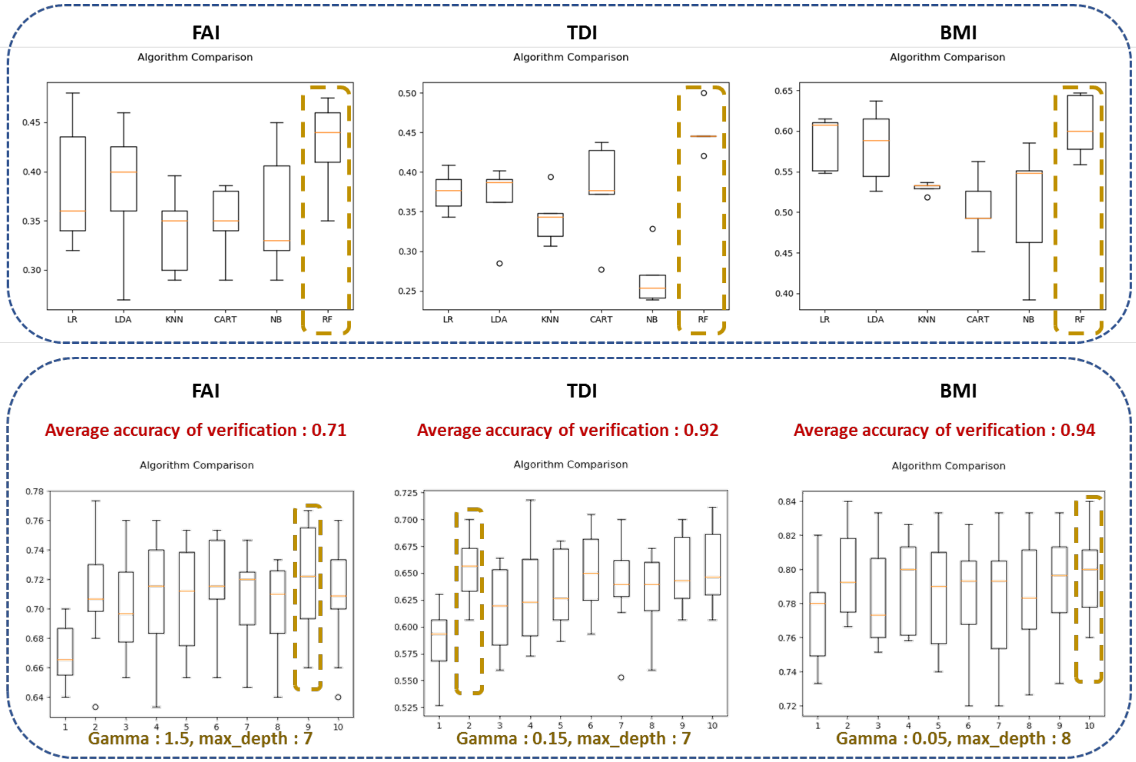
- ✓ 23 watersheds that have failed to predict.
- ✓ NO_3 and NH_4 in failed watersheds were relatively greater and smaller than the overall values.
- ✓ NO_3 was 50.6% greater than whole.
- ✓ NH_4 was -29.8% smaller than whole.
- ✓ Flow was -35.3% smaller than whole.

Findings and Future Researches

- This study was to develop XGBoost which is one of ensemble machine learning algorithms (Random forest vs. XGBoost) for AEH indices prediction using SWAT results.

Random forest

XGBoost



- We could predict AEH indices of ungauged streams via XGBoost with SWAT water quality results at a standard watershed scale.
- For further research, we will include environmental components such as river width, bankside cover treatment with porous material or vegetation cover etc..