IMPUTATION FOR MISSING STREAM FLOW USING THE SWAT AND MACHINE LEARNING MODELS IN THE TAEHWA RIVER, KOREA

Minjeong Kim, Sangsoo Baek, Mayzonee Ligaray, Kyunghwa Cho

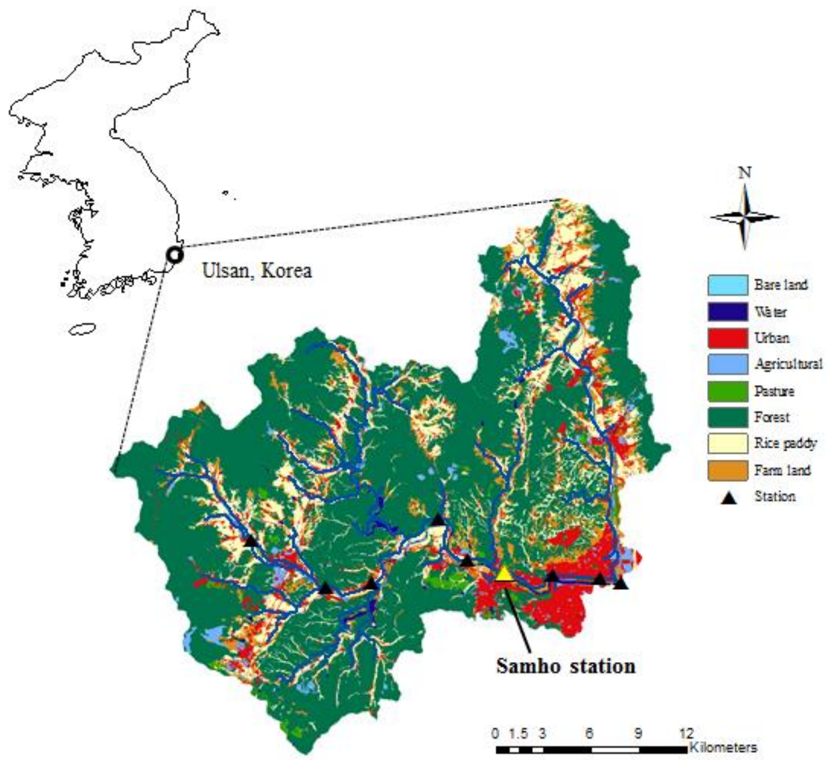
Ulsan National Institute of Science and Technology (UNIST), Korea

CONTENTS

PURPOSE AND BACKGROUND

- **Hydrologic model always requires observed data input to calibrate parameters and assess the model.**
- **However, field sensors sometimes fail, making the given observed data unreliable with abnormalities or missing values.**
- **It causes substantial uncertainties in estimation of hydrological parameters or/and in prediction of streamflows and hydrological cycle.**
- **To resolve these problems, data imputation can be used, which replaces incorrect and missing values in the dataset with probable ones.**
- **This study recovered missing stream flow using the SWAT and two machine learning models in the Taehwa river, Korea.**

STUDY AREA



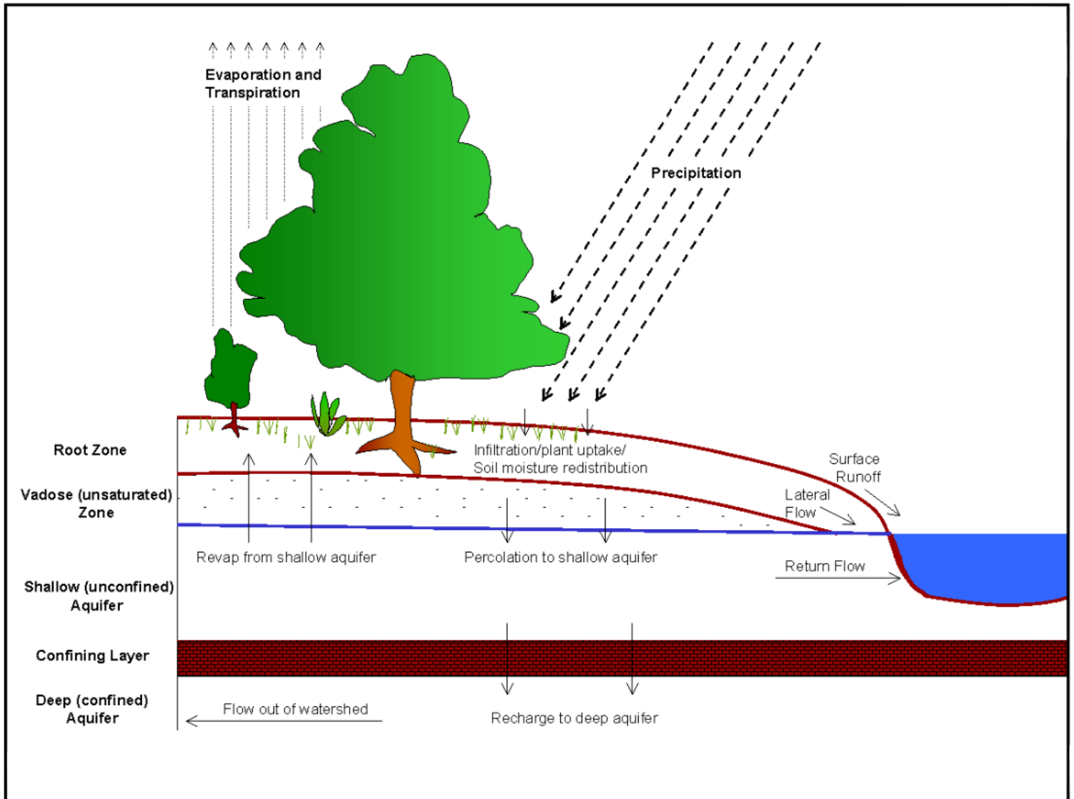
- The area of the watershed is 643.96km^2
- The watershed consists of forest (62%), rice paddy (14%) and urban (10%) areas.
- Samho station had sensor faults from 2004 to 2006 with constant signals for stream flow.(# : 350)
- No substitute for flow data in TR during that period.

LANDUSE	Bare land	Water	Urban	Agricultural	Pasture	Forest	Rice paddy	Farm land
%	0.024	1.62	8.843	2.97	1.979	62.395	14.705	7.464

< Land use property in TR watershed >

METHODOLOGY
1. SWAT

Soil and Water Assessment Tool



< Schematic representation of the hydrologic cycle >

- Hydrologic cycle
- Sediment
- Nutrients and pesticide
- Bacteria transport

METHODOLOGY
1. SWAT

- DEM and Land use

from Water Management Information System (WAMIS)

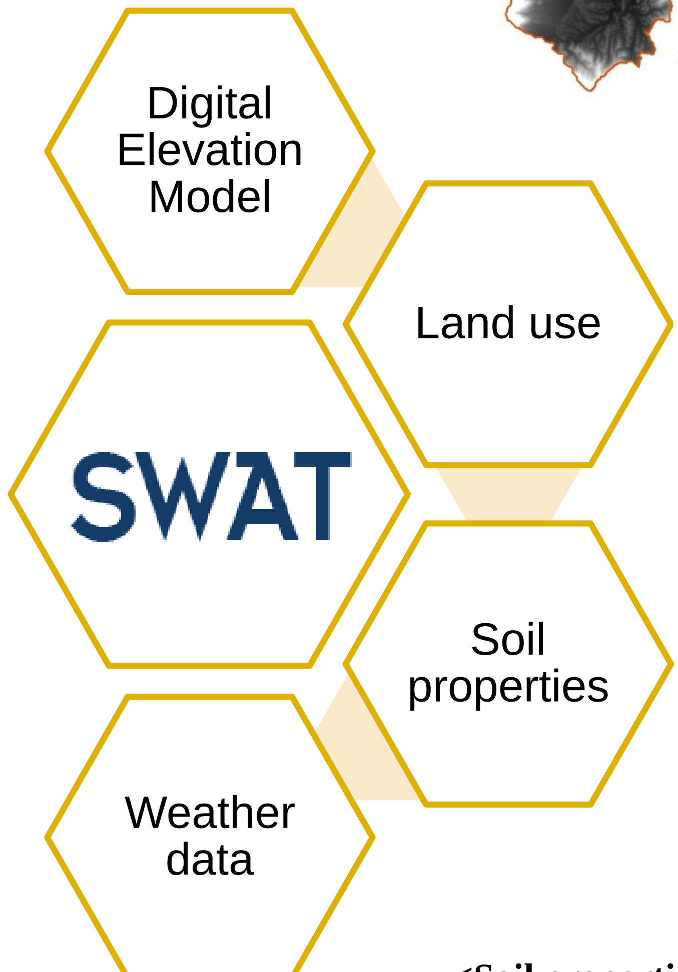
- Soil properties

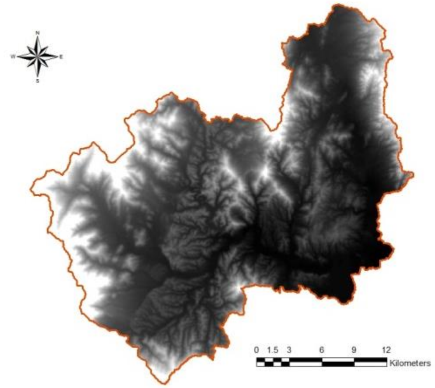
from Korean Soil information System

- Weather data

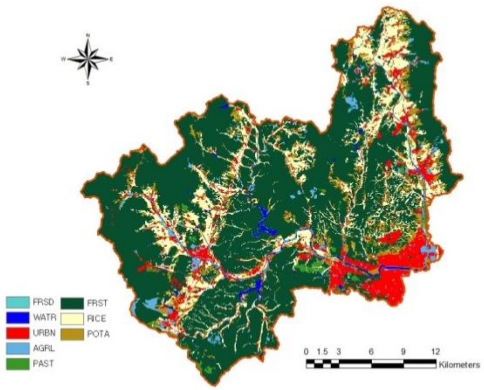
from Meteorological Information Portal Service System-Disaster Prevention

- **Precipitation**
- **Temperature**
- **Relative humidity**
- **Solar radiation**
- **Wind speed**

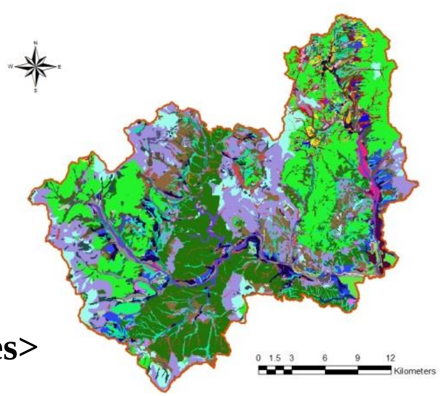




< DEM>



< Land use>



<Soil properties>

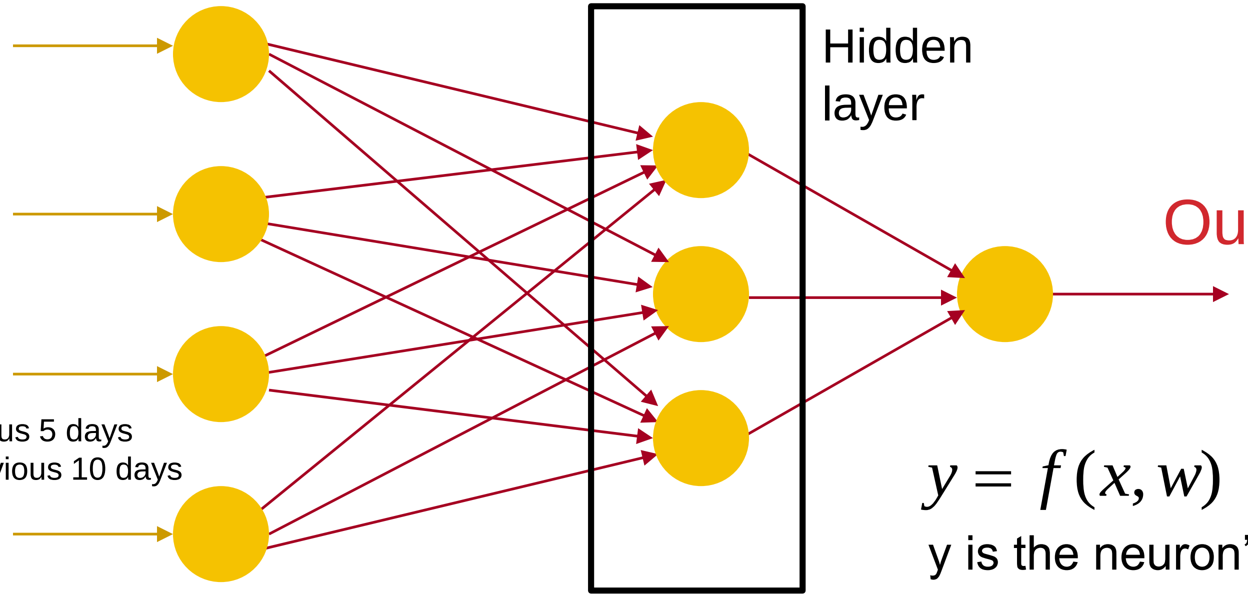
METHODOLOGY
2. ANN

Artificial Neural Network

Animals are able to react adaptively to changes in their external and internal environment, and they use their nervous system to perform these behavior.

METHODOLOGY
2. ANN

1. Constructing ANN frame
2. Initializing weights
3. Feedforward network
4. Back-propagation algorithm



-daily precipitation
-daily mean temp
-total rainfall of previous 5 days
-mean temp over previous 10 days

$$y = f(x, w)$$

y is the neuron's output

$$y_1^1 = f(x_1, w_1^1)$$
$$y_2^1 = f(x_2, w_2^1)$$
$$y_3^1 = f(x_3, w_3^1)$$
$$y_4^1 = f(x_4, w_4^1)$$

$$y^1 = \begin{pmatrix} y_1^1 \\ y_2^1 \\ y_3^1 \\ y_4^1 \end{pmatrix}$$

$$y_1^2 = f(y^1, w_1^2)$$
$$y_2^2 = f(y^1, w_2^2)$$
$$y_3^2 = f(y^1, w_3^2)$$

$$y^2 = \begin{pmatrix} y_3^2 \\ y_3^2 \\ y_3^2 \end{pmatrix}$$

$$y_{Out} = f(y^2, w_1^3)$$

METHODOLOGY
2. ANN

1. Constructing ANN frame
2. Initializing weights
3. Feedforward network
4. Back-propagation algorithm

$$E = \frac{1}{N}\sum_{i=1}^{N}\left(f(x_i; W_i) - yobs_i\right)^2$$

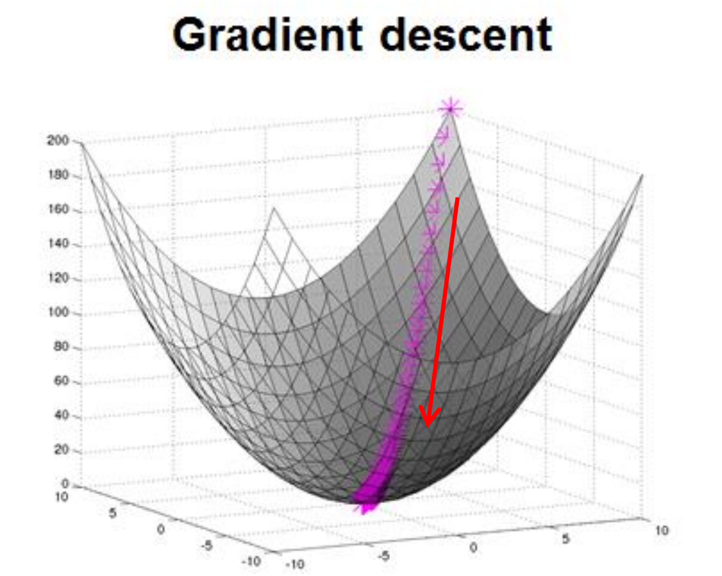
By adjusting every weights, minimize E

Gradient descent with momentum algorithm

$$\Delta w_i^{j+1} = -c \cdot \frac{\partial E}{\partial w_i^j}(w_i^j) + \underline{a \cdot \Delta w_i^j}$$

$$w_i^{j+1} = w_i^j + \Delta w_i^{j+1}$$

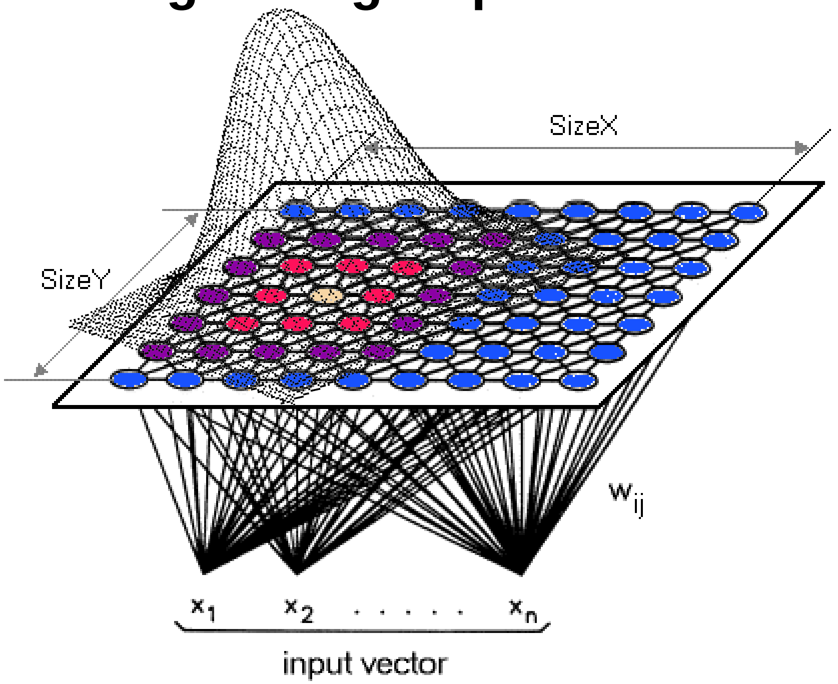
j = iteration number
c = learning parameter
a = momentum constant



Error space

METHODOLOGY
3. SOM

Self Organizing Map



-Input
-Output

$$w_i \ni \{wi_1, w_{i2}, w_{i3}, w_{i4}, w_{i5}\}$$

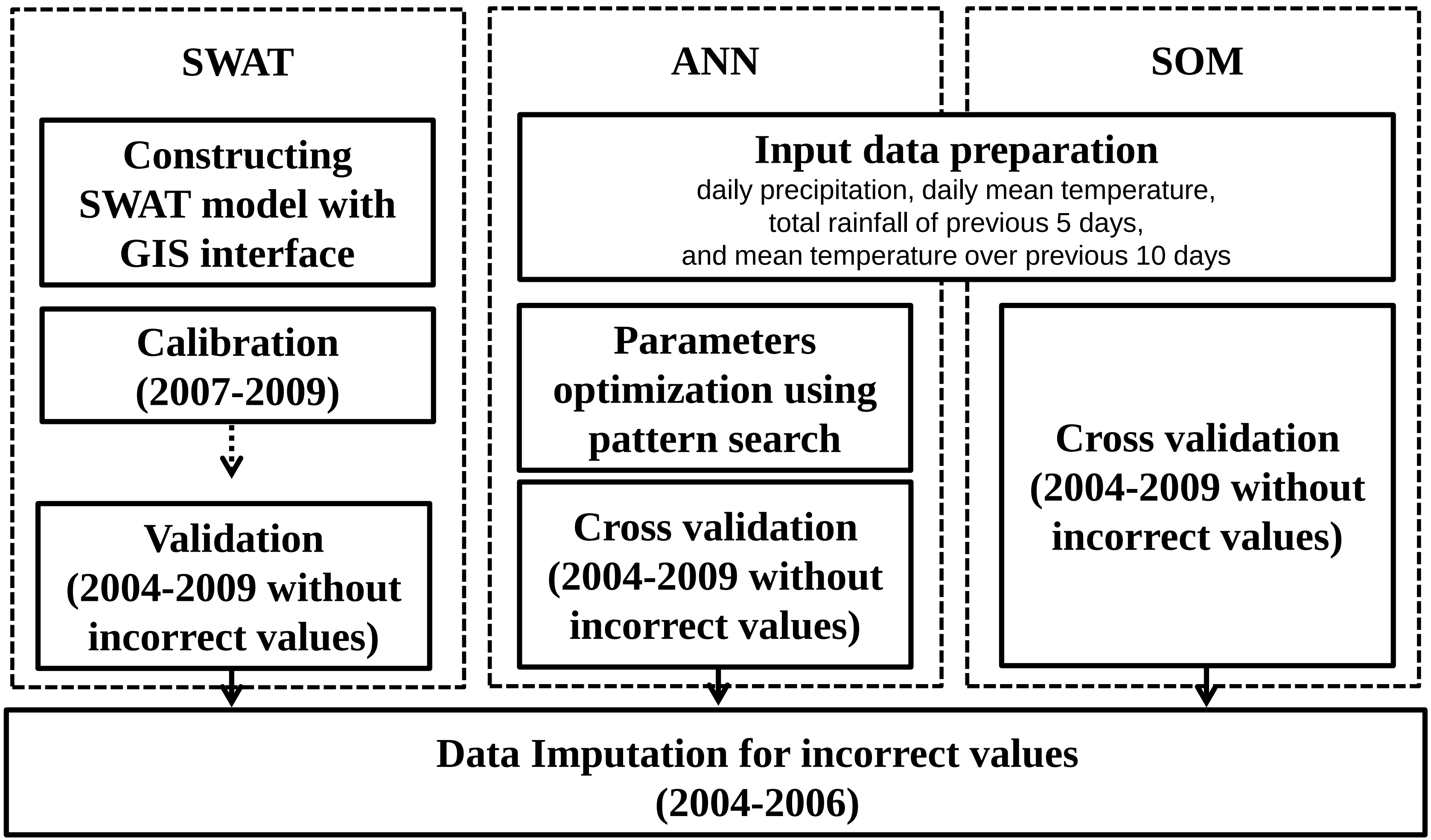
$$c_j = \operatorname{argmin}_i \{\|w_i - x_j\|\}$$

$$h_{c_j,i} = \exp\left(-\frac{\|r_i - r_{c_j}\|^2}{2\sigma^2}\right)$$

: Gaussian distribution
r = coordinate

$$w_i^{new} = \frac{\sum_{j=1}^{N} h_{c_{j,i}} * x_j}{\sum_{j=1}^{N} h_{c_{j,i}}}$$

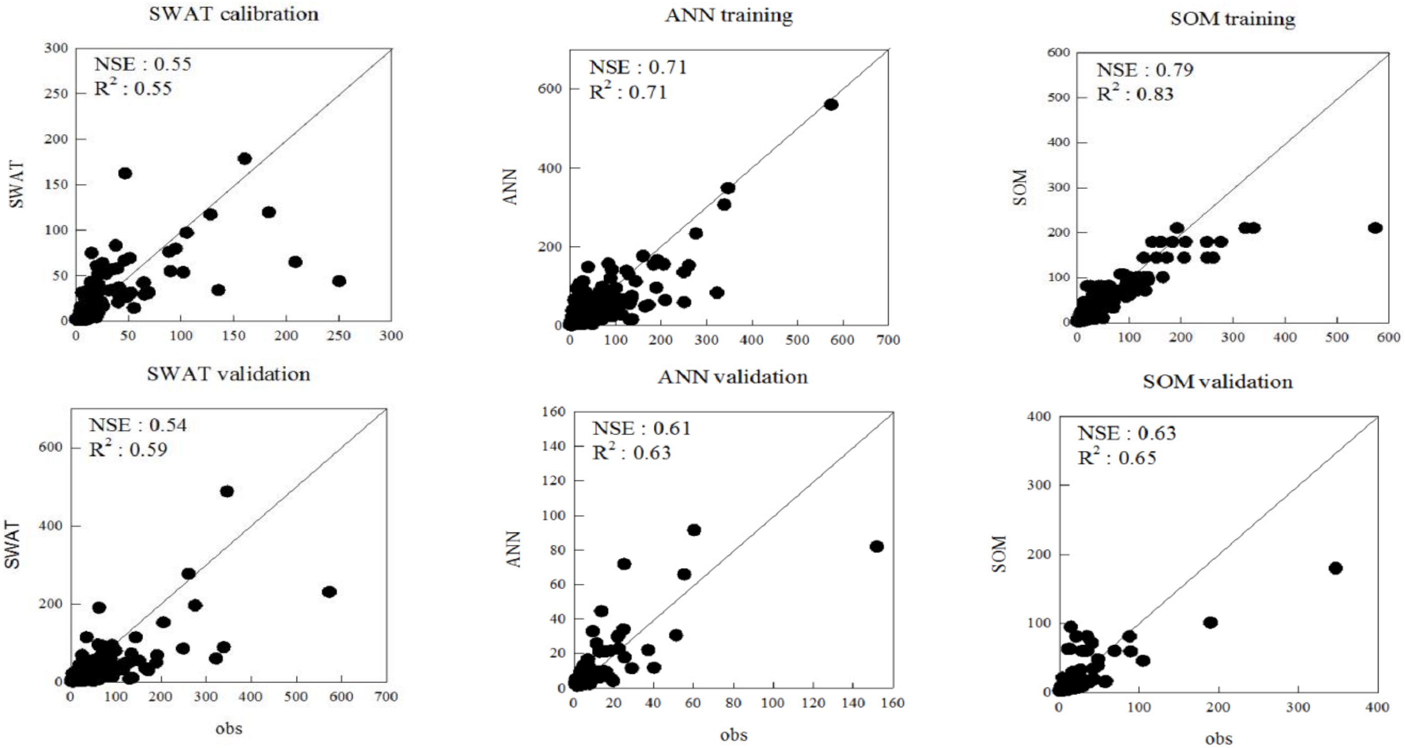
METHODOLOGY
4. FLOW CHART



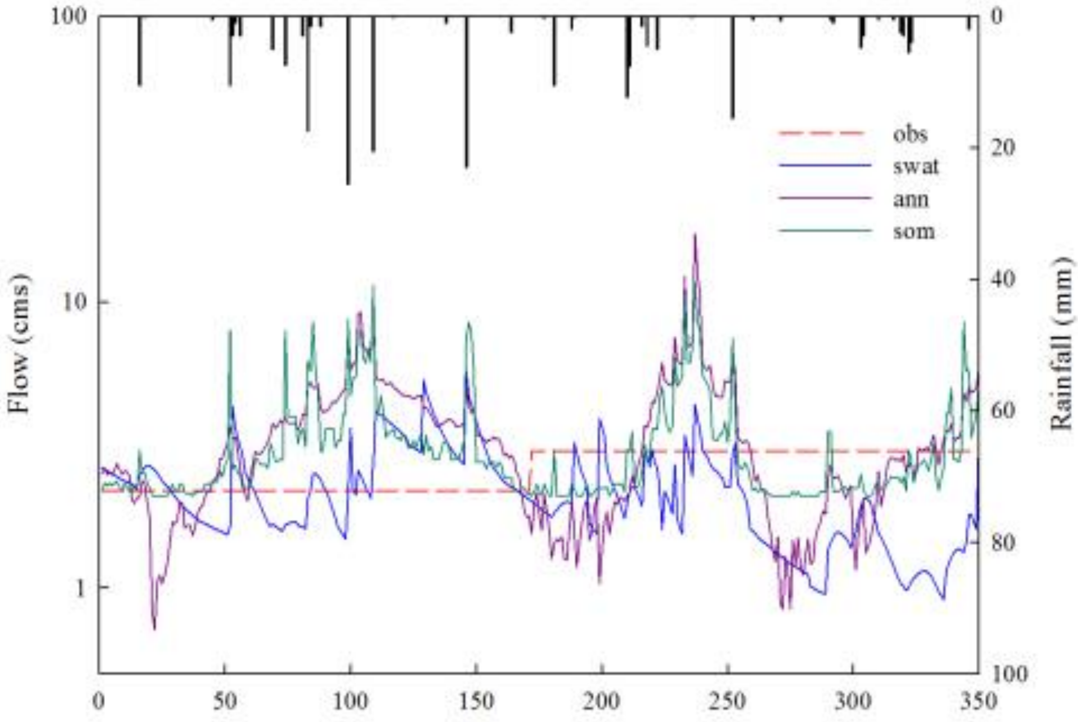
RESULT

Method	NSE		R^2		RMSE	
	calibration	validation	calibration	validation	calibration	validation
SWAT	0.55	0.54	0.55	0.59	10.82	27.77
ANN	0.71	0.61	0.71	0.63	16.92	6.86
SOM	0.79	0.63	0.83	0.65	13.64	15.44

< Results in statistics indicators>



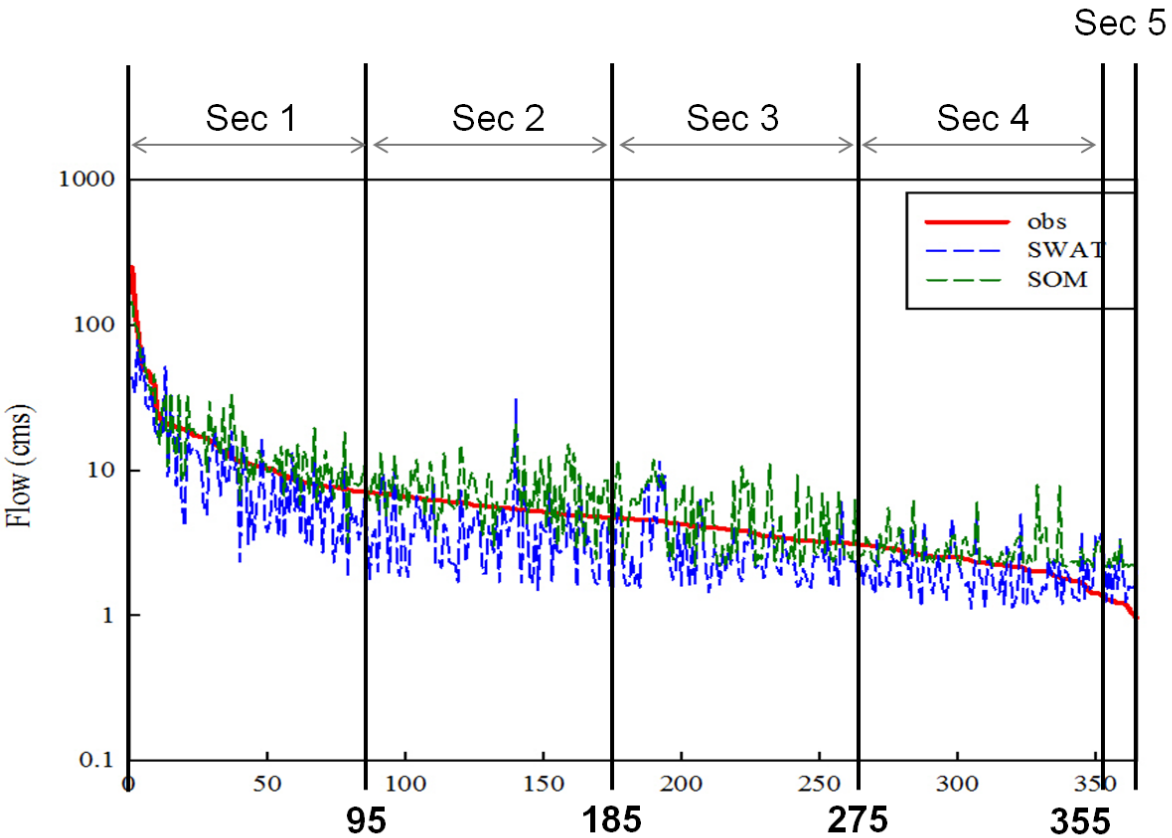
RESULT



< Constant flow comparison>

- **SOM and ANN** had **similar trend** and they were **sensitive** to rainfall event.
- **SWAT underestimated** flow compared to ANN and SOM

RESULT



- To understand model efficiency according to the flow characteristics, we adopted Flow Duration Curve (FDC).

- By separating FDC into five sections, we compared SWAT and SOM in terms of RMSE.

RESULT

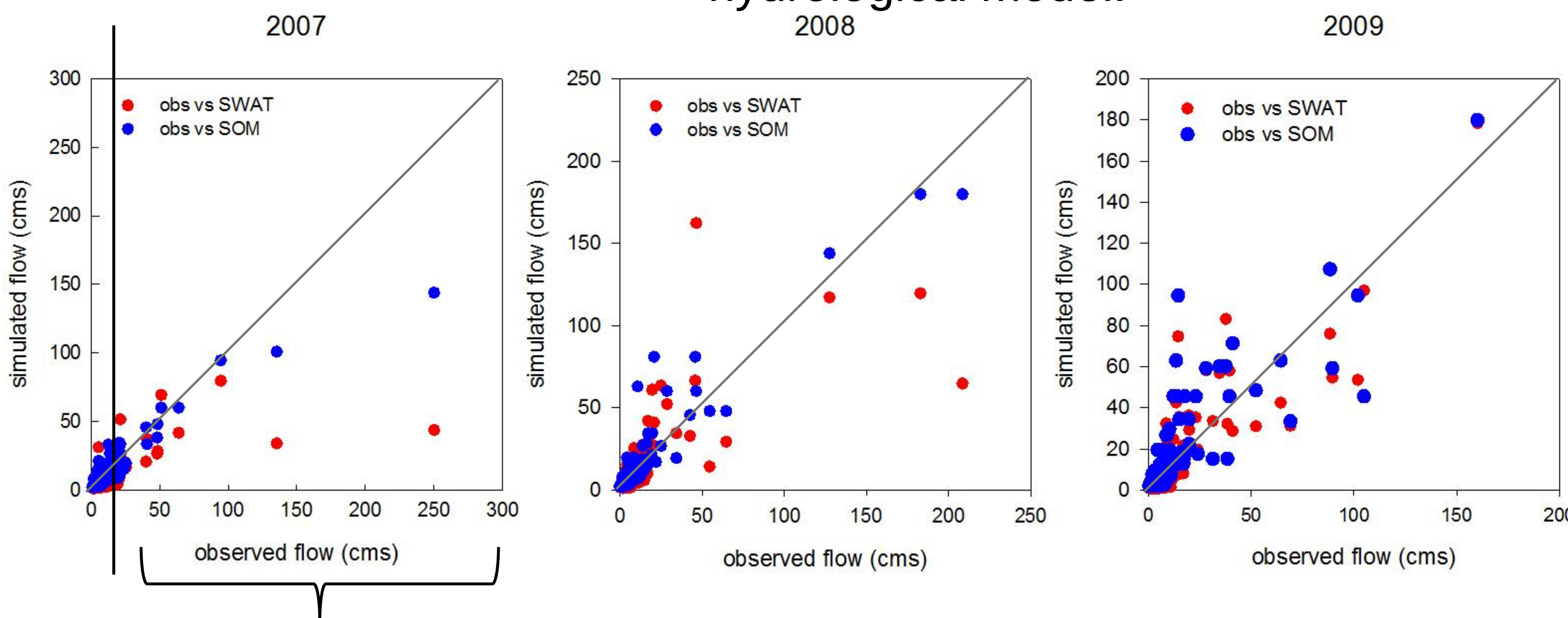
	2007			2008			2009	
section	SWAT	SOM	section	SWAT	SOM	section	SWAT	SOM
1	24.85	**12.86**	1	22.49	**11.48**	1	**3.73**	4.03
2	**3.76**	3.81	2	**2.33**	3.38	2	**1.66**	1.91
3	**1.73**	2.56	3	**1.49**	2.04	3	**1.19**	1.31
4	**1.00**	1.32	4	**1.04**	1.74	4	**0.86**	1.25
5	**0.90**	1.26	5	**1.41**	1.95	5	**0.79**	1.32

- **SWAT** had **always lower RMSE** than SOM for **section 2 - 5**, indicating relatively **low discharge**.
- However, **Section 1** showed **different** result
  - SOM performed better in 2007 and 2008
  - SWAT performed better in 2009

RESULT

	R-square	
year	obs-SWAT	obs-SOM
2007	0.46	**0.86**
2008	0.55	**0.89**
2009	**0.78**	0.73

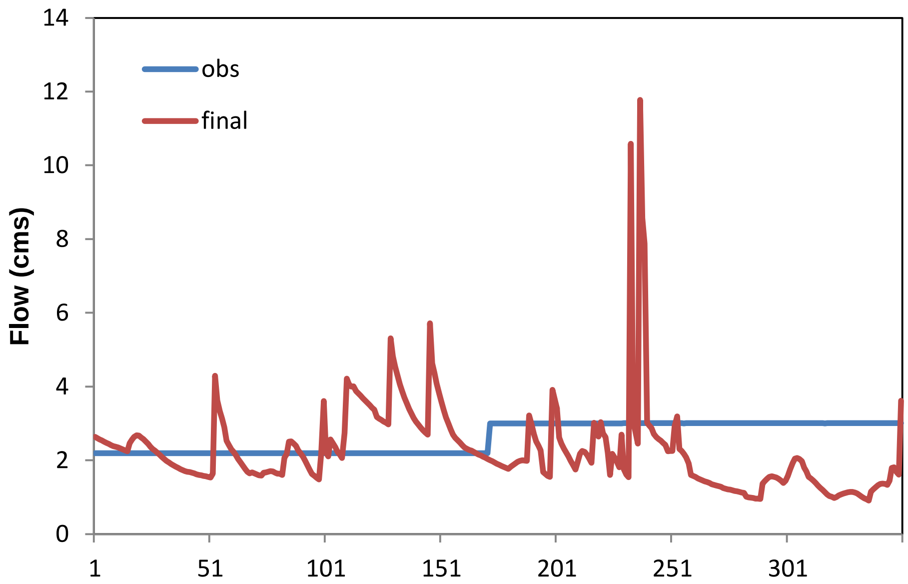
- Concentrated low flow
- **High flow** is more **dependent to model performance**
- Generally, **machine learning is better in capturing high flow** than hydrological model.



Sec 1

RESULT

year	95-day
2004	16.88
2005	6.99
2006	10.38



2004 : 171 missing flow data
171– under 16.88 → **SWAT**

2005 : 111 missing flow data
4 – over 6.99 → **SOM**
107– under 6.99 → **SWAT**

2006 : 68 missing flow data
68 – under 10.38 → **SWAT**

CONCLUSION

- By recovering incorrect flow data, this study expected to increase data efficiency and prevent uncertainty increase.
- Based on FDC, section 1 was well estimated by model which has better performance, and usually, SOM was good at estimating high flow.
- SWAT performed well in low flow estimation as the rest of sections and ANN showed moderate result.
- This study recommends the use of different data imputation techniques according to data characteristics.
- It can be explored further with different methodologies and data sets.

THANK YOU
FOR YOUR ATTENTION