THA 2019 International Conference in Bangkok
Swissôtel Bangkok Ratchada, 2019.01.23.~2019.01.25.

Outlier Detection of Reservoir Water Level Data Using Artificial Neural Network Model

MAGA KIM*, JIN-YONG CHOI**

*Graduate student, Department of Rural Systems Engineering, College of Agriculture and Life Sciences, Seoul National University
**Professor, Department of Rural Systems Engineering, College of Agriculture and Life Sciences, Seoul National University

○Contents

Background

1) Reservoir water level data

- Hydrological data have been utilized as primary data of the analysis for water resource management
- Especially, reservoir water level data have been utilized to estimate reservoir capacity using capacity curve (Jeong and Kim, 2007)
- Moreover, reservoir water level data have been used in the various studies as basic data or reference data

-Development of the reservoir operation model (Shim et al., 1997; Jang et al., 2007)
-Prediction of behavior of fill dam subjected to water level change (Lee et al., 2014)
-Effectiveness evaluation of drainage systems in wetland based on the change in groundwater level near the reservoir (Kim et al., 2016)
-Correlation analysis of deep landslide occurrence and variation of groundwater level(Lim et al., 2017)

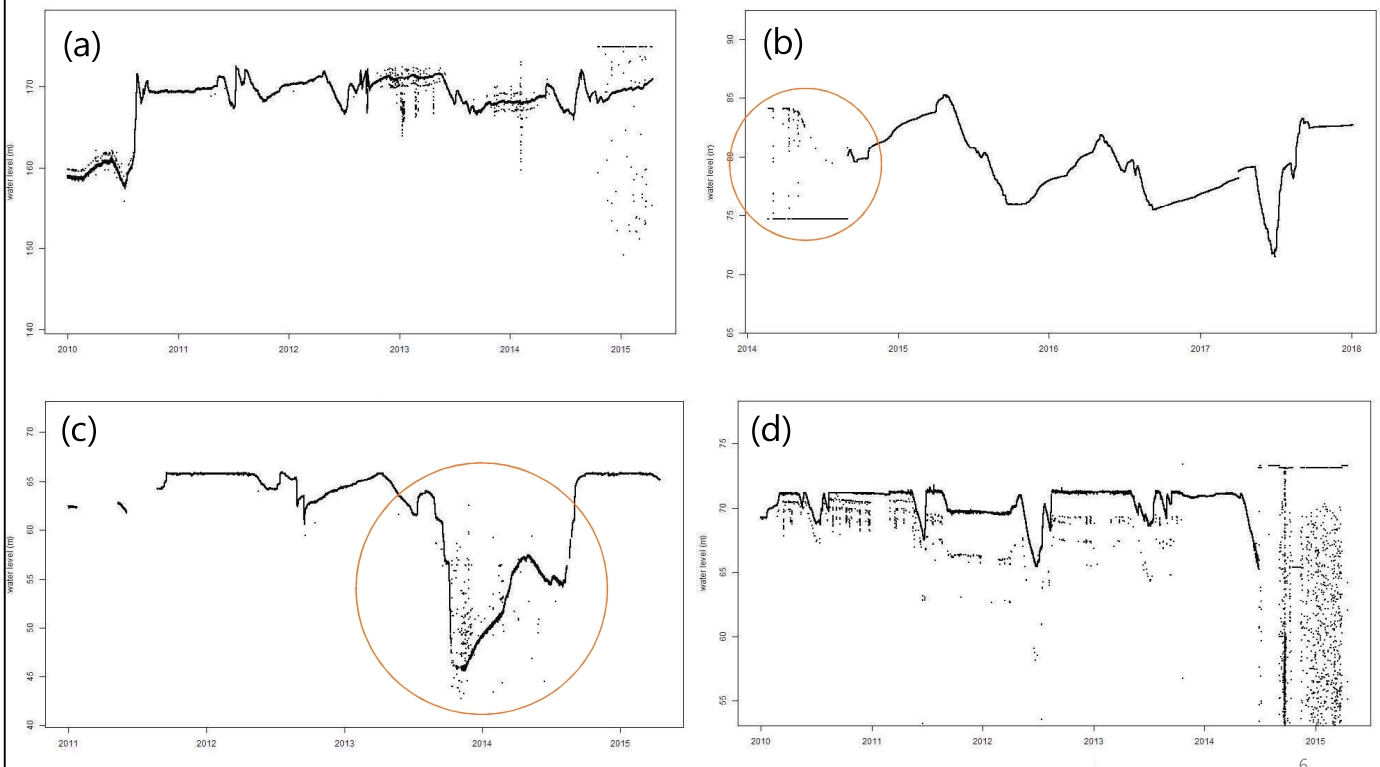
2) Status of reservoir water level data

- Korea Rural Community Corporation (KRC) installed water level measuring devices at around 1,600 agricultural reservoirs
- Reservoir water level data have been collected using the pressure sensor and ultrasonic sensor in every 10 minutes
- But various types of outliers have arisen due to physical and environmental causes

Sensor type	Causes of outlier
pressure sensor	• inflow of sediment in inclined conduit • silted deposit near the sensor
ultrasonic sensor	• hindrance of wave or plants near the sensor • changes in atmospheric temperature and humidity

- Reliability of the reservoir water level data should be ensured to utilize the data effectively
  - ➢ Quality control of the reservoir water level data is essential

2) Outliers of reservoir water level data



3) History of research – quality control & outlier detection of hydrological data

- Research on quality control and outlier detection for hydrological data

Quality control & Calibration	-Quality control of radar precipitation estimation data using actual rain gage data (Steiner et al., 1999; Harrison et al., 2000) -Quality control of long times series of data such as rainfall, flow rate, concentration of the pollutant in urban hydrology (Mourad and Bertrand-Krahewski, 2002) -Development of observation data model (ODM) and classification of quality control level (Horsburgh et al., 2008)
Outlier Detection	-Outlier detection and homogeneity test of monthly precipitation data (González-Rouco et al., 1999; Feng et al., 2004) -Outlier detection for time series data in water distribution systems (Mounce et al., 2011) -Outlier detection and semi-automatic quality control of monthly precipitation data (Schneider et al., 2014)

➢ Most quality control method uses neighboring station or measured value of different types of sensor

3) History of research – artificial neural network (ANN) model

- Artificial neural network model (ANN model) is a computational model created by simplifying brain structure of human (Yeo et al., 2010)
- ANN model can solve problems without direct knowledge or algorithm but only with data given and it used to solve regression or classification problems (Yeo et al., 2010)
- It can model the natural phenomenon which has nonlinear characteristics such as whether of hydrological data (Choi and Kang, 2000)

Hydrological data	-Prediction of runoff (Kim, 2000), flood elevation (Kim and Salas, 2000), quality of river water (Oh et al., 2002), inflow of dam in the flood season (Yoon et al., 2004), and probability of occurrence of precipitation by the water system (Kang and Lee, 2008) using the ANN model -Estimation of surface runoff in the paddy field using the ANN model (Ahn et al., 2012)
Hydrological data Calibration	-Complement missing data of precipitation using the ANN model (Ahn et al., 2000; Oh et al., 2008) -Quality control of radar precipitation estimation data using the ANN model (Kim et al., 2010)

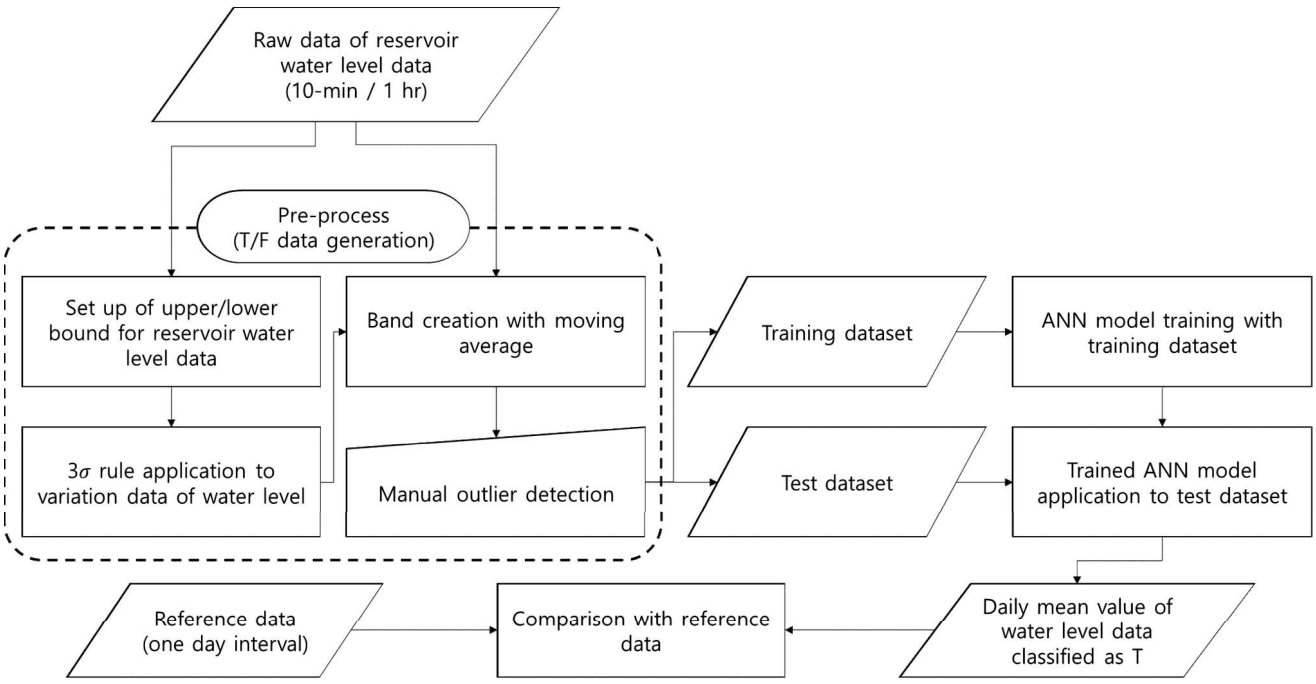
4) Object of Study

- Set up of an outlier detection model with artificial neural network model
- Application of outlier detection model to the reservoir water level data
- Evaluation of the artificial neural network model using reference data

Outlier detection of reservoir water level data
using artificial neural network model

Materials & Methodology

1) Flow chart



2) Subject reservoir & Building reservoir water level data

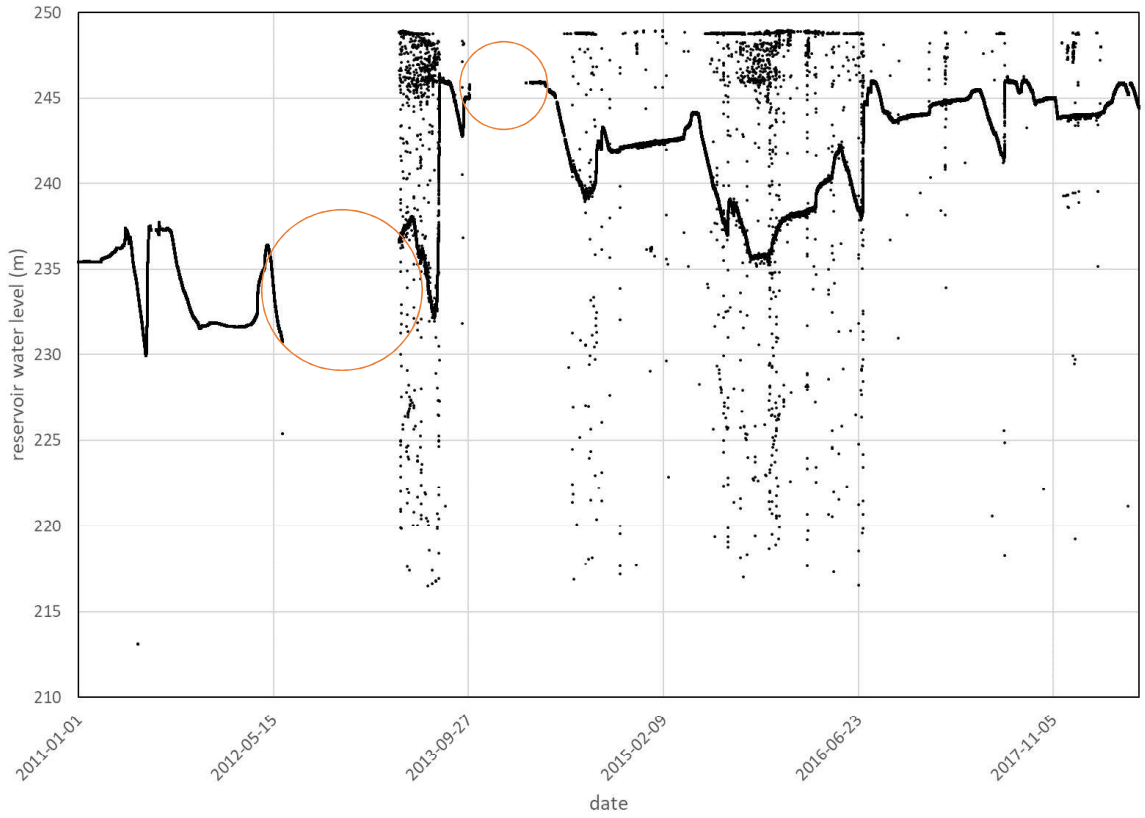
- Properties of the Gaeun reservoir

Dam property	Contents	Dam property	Contents
Dam length (m)	224	Water-intake type	intake tower
Dam height (m)	39.7	Total storage (㎥)	1,649,375
Dead storage water level (EL.m)	222.50	Effective storage (㎥)	1,636,475
Full water level (EL.m)	245.80	Benefitted area (ha)	103
Flood water level (EL.m)	246.80	Basin area (ha)	474

- Properties of reservoir water level data

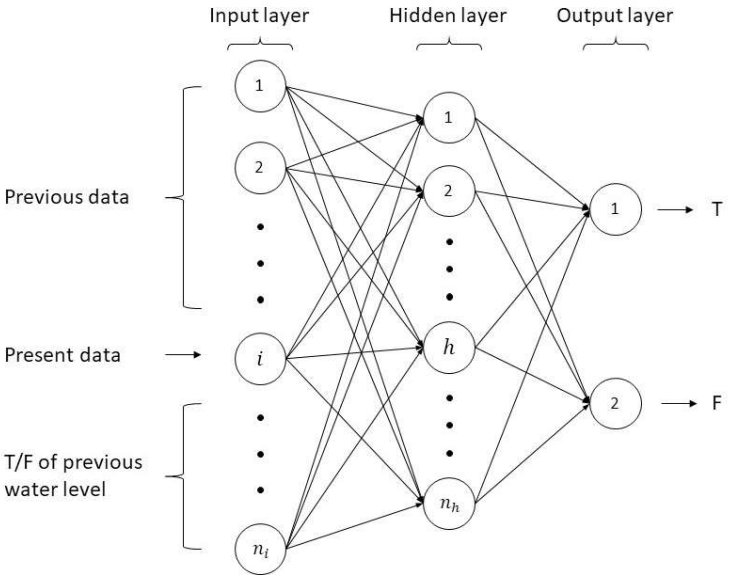
Data	Measurement method	Measurement cycle	Measurement period
Sensor data	ultrasonic sensor	10 minutes	2011. 01. 01. 0:00~ 2018. 06. 12. 11:40
Reference data	sensor/eye measurement	1 day	2009. 08. 19.~ 2018. 05. 23.

2) Building reservoir water level data – raw data (10-minute interval)



3) Structure of the ANN model

- ANN model consists of three input, hidden, and output layers
- It learned from training data, and detect outlier



- input data:
  previous data and T/F label data
- output data:
  T/F label at that point
- T/F label data:
  water level data divided as normal data and outlier
- number of input nodes(n_i) and hidden nodes(n_h)
  → trial and error method

3) Structure of the ANN model

- Back-propagation is general training method in ANN model
- Back-propagation spreads error of output layer into the reverse direction in a differential form and adjusts weights and biases (Kim and Koo, 2000)
- In this study, Adam method (Adaptive Moment Estimation) is applied
- Adam method is similar to combine Momentum and RMSProp method

- Momentum method

-considering inertia
-remember the way model moved before and made model to move further

+

- RMSProp

-apply different step size to each variable
-apply exponential averaging to the Adagrad method and reduce influence of past variables

3) Component of the ANN model

- Sigmoid function

$$f(x) = \frac{1}{1 + e^{-x}}$$

- Softmax function

$$f(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

- Cross entropy error

$$E = -\sum_{k} t_k \log y_k$$

- Hyper-parameter

-number of input nodes n_i
-number of hidden nodes n_h
-learning rate η

4) Generation pre-processed T/F data

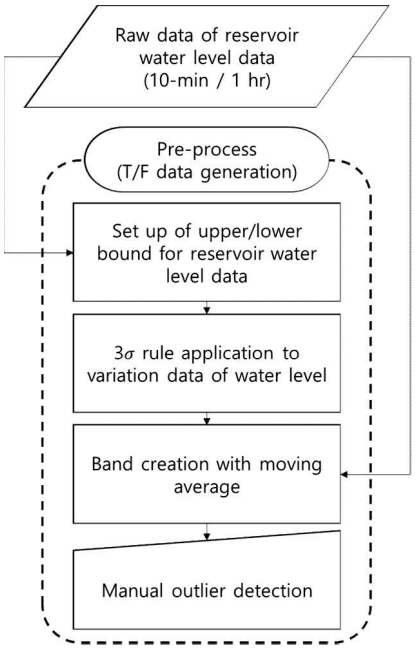
- Pre-processed T/F data is the reservoir water level data classified as normal data (T) and outlier/missing data (F)
- Upper/lower bound set up for reservoir water level

  -upper bound: flood water level
  -lower bound: dead water level
- Application of 3σ rule to water level variation

  -upper/lower bound: $m \pm 3\sigma$ (m: mean, σ: standard deviation)
- Band creation with moving average

  -$n = 10$ (n: the number of data for calculating moving average)
  -$\alpha = 0.05$ (α: the width of band, unit: m)
- Manual outlier detection

  -Elimination of remaining spike noise
  -Classification of outlier as normal in rapidly changing period



5) Input data of ANN model

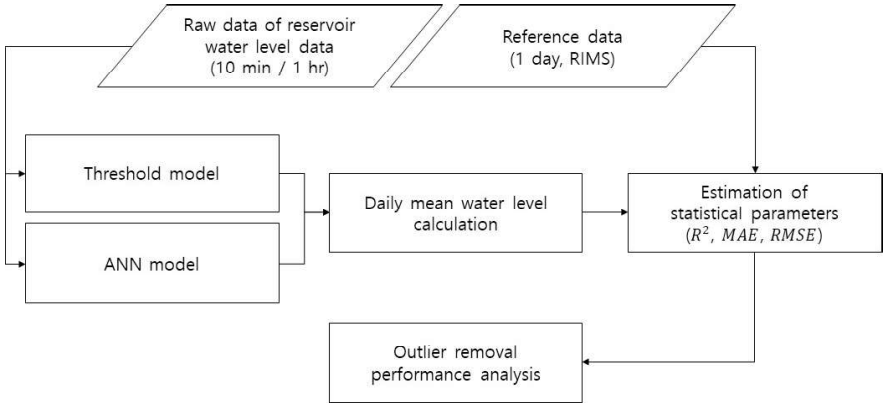
- Application of 6 types of input data with identical T/F data
- Each types of input data is pre-processed by different method
- Input data type and pre-process method

Input data type	Pre-process method
water level (wl_t)	wl_t
variation (v_t)	$v_t = wl_t - wl_{t-1}$
reservoir regularization (rr_t)	$rr_t = \frac{w_{;t} - dl}{fl - dl}$
limited regularization (lr_t)	$lr_t = \begin{cases} 1 & (\text{if } rr_t < 1) \\ rr_t & (\text{if } 0 < rr_t < 1) \\ -1 & (\text{if } rr_t < 0) \end{cases}$
regularization (r_t)	$r_t = \frac{wl_t - \min(wl)}{\max(wl) - \min(wl)}$
variation regularization (vr_t)	$vr_t = \frac{v_t - \min(v)}{\max(v) - \min(v)}$

-fl: flood water level
-dl: dead water level

6) Validation and evaluation of model

- Application of daily water level data provided from KRC as reference data
- Manager checks the measured data at fixed time, and if there is abnormality, the measured data is adjusted by eye measurement value
- Water level data before and after the model application are averaged daily to equalize the time interval
- Estimation and evaluation R^2, MAE, $RMSE$ by comparing water level data before and after the model application with reference data



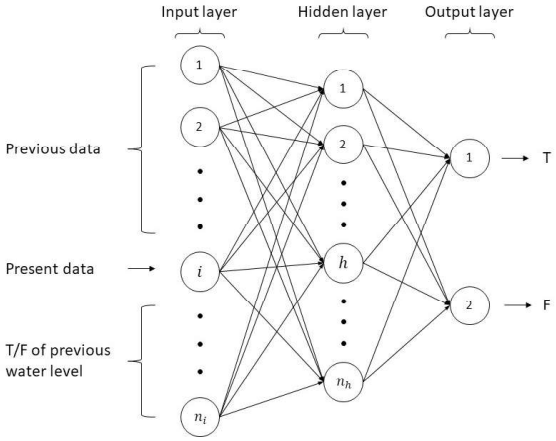
19

Results & Discussion

1) Structure of the ANN model

- Number of input nodes n_i, number of hidden nodes n_h, and learning rate η were decided by trial & error method
-

Reservoir	Data type	Input nodes	Hidden nodes	Learning rate
Gaeun	Raw data	27	15	0.005
	Hourly mean data	9	18	0.001



- input data:
  previous data and T/F label data
- output data:
  T/F label at that point

2) Training data of the ANN model

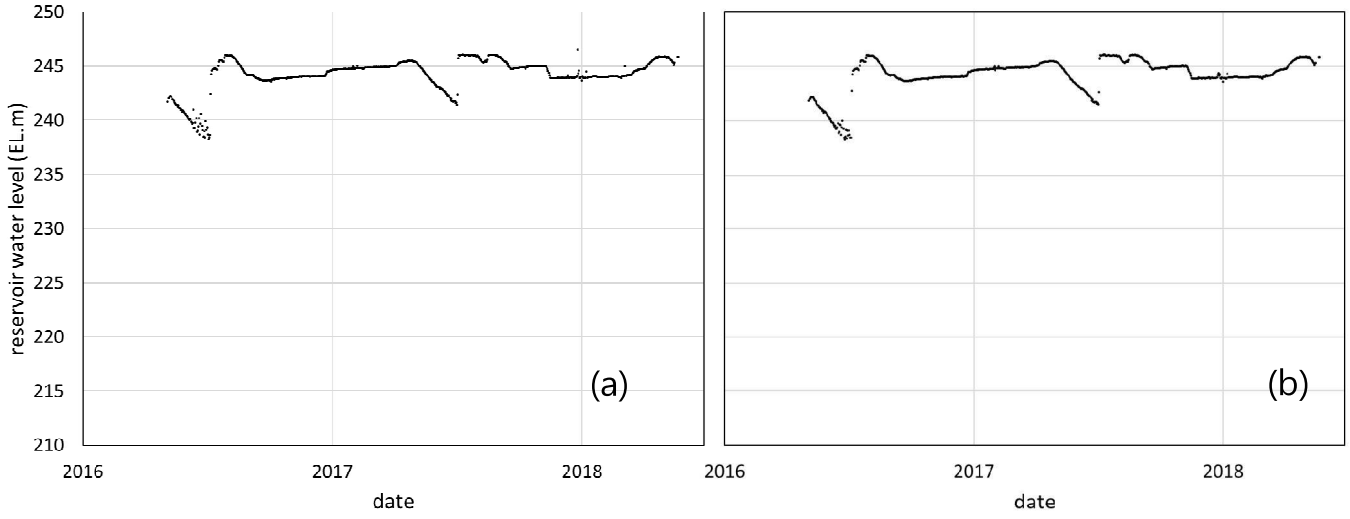
- Reservoir water level data classified as normal data (T) and outlier/missing data (F)
- Long-term missing periods were excluded from training data when the ANN model trained and applied to detect outlier
- Long-term missing period is defined more than 4 hours missing for raw water level data (10-minute interval) and 6 hours missing for hourly mean water level data (1 hour interval)

- Percentage of outlier and missing data (F) to normal data (T) without long-term missing data
-

Reservoir	Data type	Percentage of F to T (%)
Gaeun	Raw data	16.67
	Hourly mean data	7.67

3) Result of ANN model – raw water level data (10-minute interval)

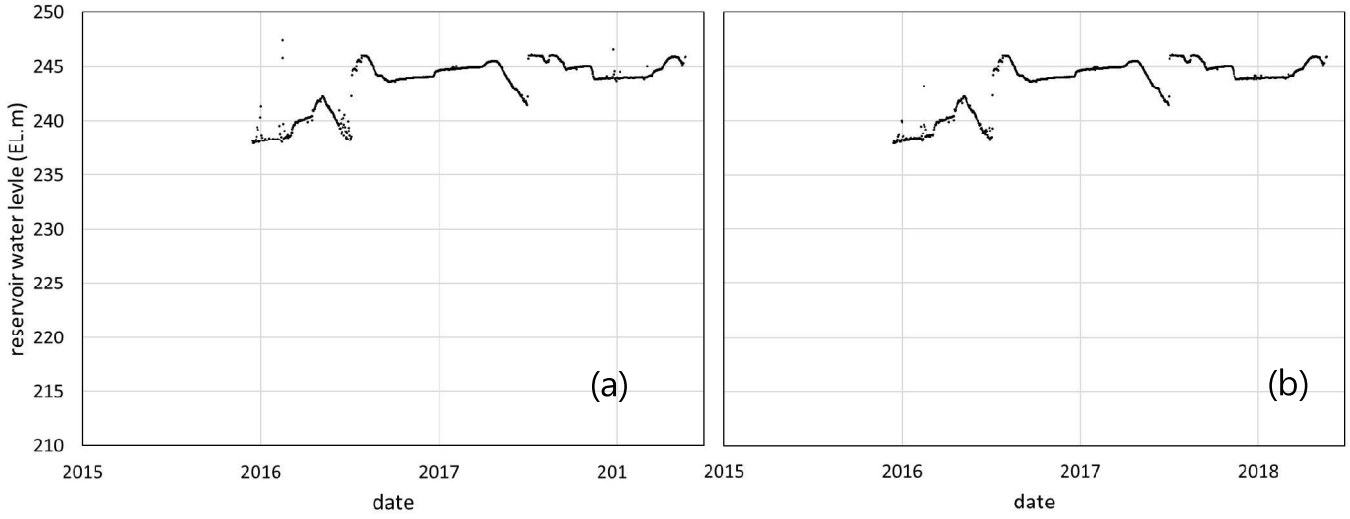
- (a) daily mean of the raw data, (b) daily mean of result data of the ANN data



23

3) Result of ANN model – hourly mean water level data (1 hour interval)

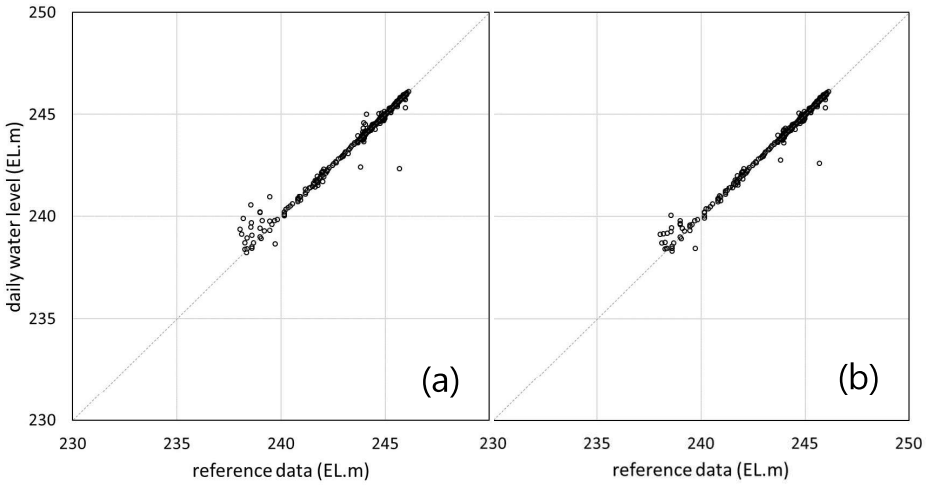
- (a) daily mean of the raw data, (b) daily mean of result data of the ANN data



24

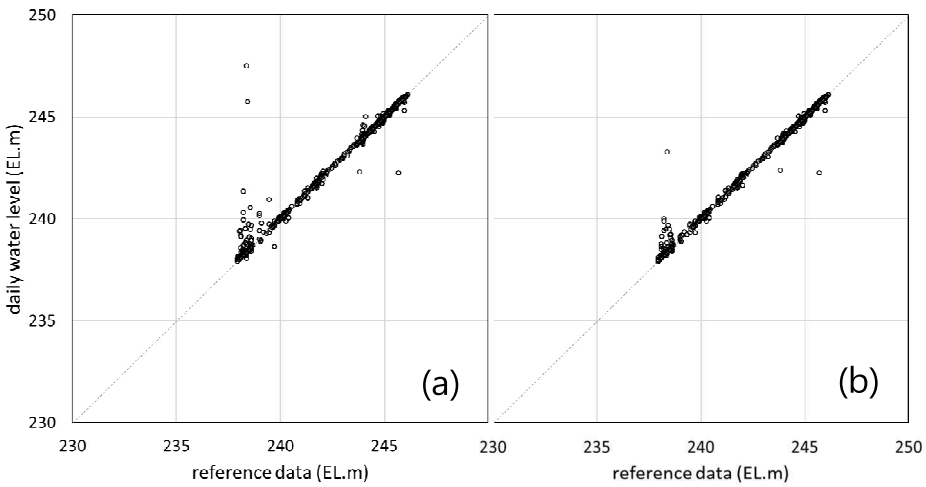
4) Scatter plot – raw water level data (10-minute interval)

- Scatter plot compared with reference data
- (a) daily mean of the raw water level data (b) daily mean of result of the ANN model



4) Scatter plot – hourly mean water level data (1 hour interval)

- Scatter plot compared with reference data
- (a) daily mean of the hourly mean water level data, (b) daily mean of result of the ANN model



5) Result of ANN model – statistical parameters

- Estimation R^2, MAE, $RMSE$ for testing period
- Gaeun reservoir – raw water level data (10-minute interval)

Statistical parameter	Raw	ANN	Target
R^2	0.9804	0.9808	0.9920
MAE (m)	0.0648	0.0491	0.0362
$RMSE$ (m)	0.2545	0.2216	0.1902

- Gaeun reservoir – hourly mean water level data (1 hour interval)

Statistical parameter	Raw	ANN	Target
R^2	0.9768	0.9905	0.9986
MAE (m)	0.1311	0.0465	0.0365
$RMSE$ (m)	0.3620	0.2156	0.1910

27

Conclusion

Summary & Conclusion

- Outlier detection model was applied to the Gaeun reservoir water level data
- MAE is 0.0648 m for raw data, 0.0491 m for result of ANN model in case of raw water level data (10-minute interval)
- MAE is 0.1311 m for raw data, 0.0465 m for result of ANN model in case of hourly mean water level data (1 hour interval)

  → Application of the models has improved quality of the data compared to the raw data
- R^2 and MAE of target data are 0.9920 and 0.0362 m for raw water level data, 0.9988 and 0.0365 m for hourly mean water level data

➢ The ANN model have performed outlier detection properly

➢ Performance of ANN model is better in hourly mean water level data than in raw water level data

→ Further research is needed on more cases

29

Thank You