ESCAP/WMO Typhoon Committee

19^{th} INTEGRATED WORKSHOP / AP-TCRC FORUM

The Deep-Learning Algorithm for Hydrological Data Quality Control and Flood Forecasting in TC member countries

20 Nov. 2024

Ph.D. Chung-Soo Kim

WGH AOP 2 &3 Leader

KICT

KOREA INSTITUTE of CIVIL ENGINEERING and BUILDING TECHNOLOGY

Contents

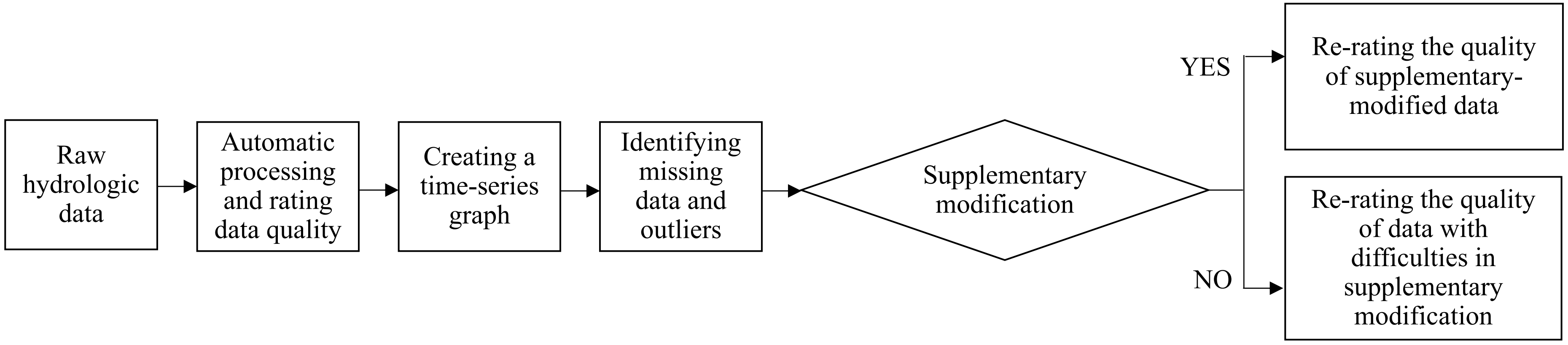
1 Hydrological Data Quality Control
using AI Deep-Learning Techniques

2 Application of AI Hydrological Data Quality Control
in TC member countries

3 Flood Forecasting
using AI Deep-Learning Techniques

Hydrological Data Quality Control using AI Deep-Learning Techniques

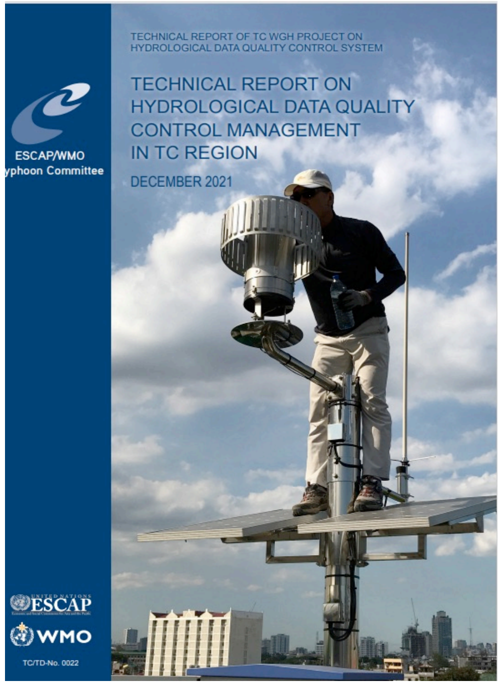
Hydrological Data Quality Control (HDQC)



<Systemic framework of hydrological data quality management >

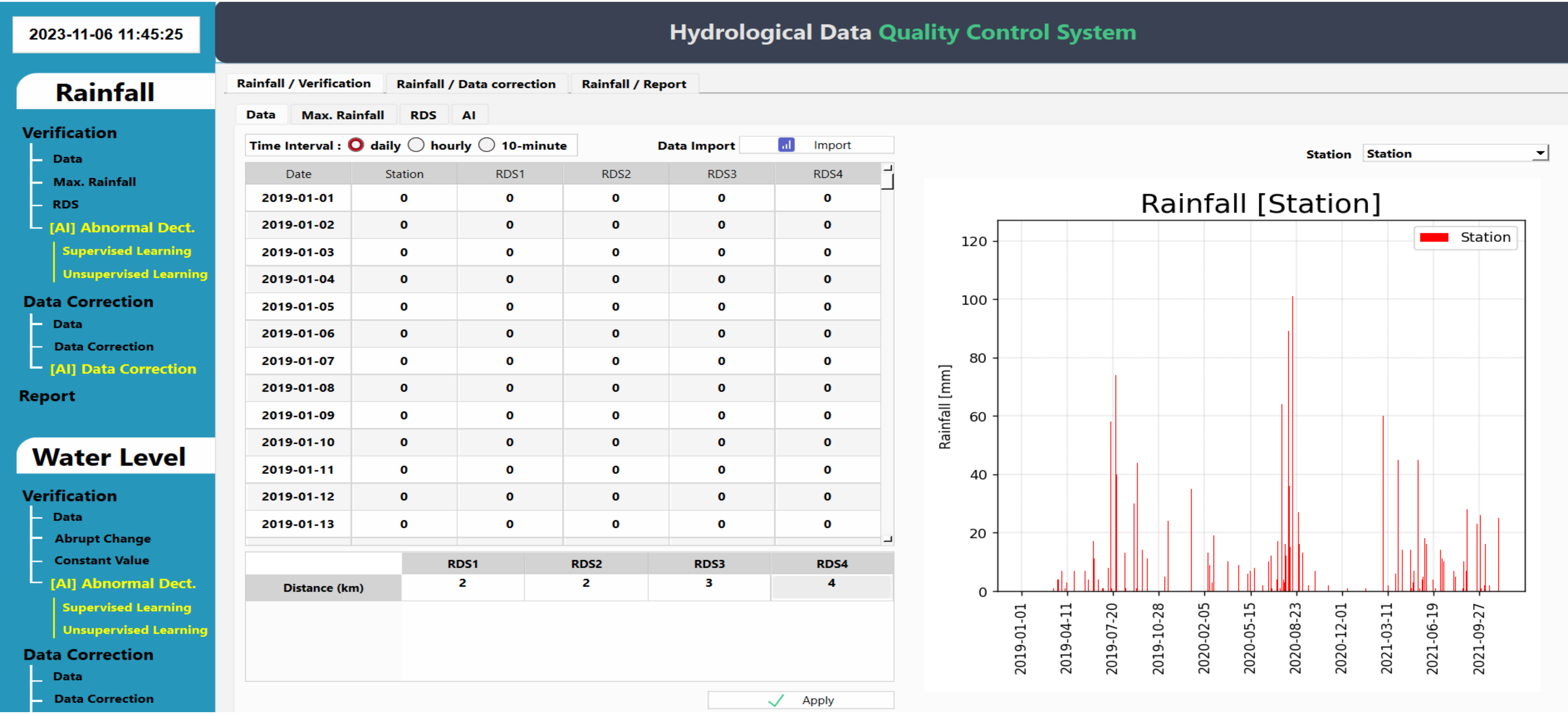
- ✓ Semi-automatic quality control
- ✓ Statistical data analysis, quality ratings & flag
- ✓ Historical data analysis and establish criteria for screening abnormal data
- ✓ Criteria:
  - Rainfall: Max. Rainfall, RDS
  - Water Level: Abrupt change, Constant

Quality Code		Data status and correction method			Remarks
0100		Good quality data of continuous time series			
02 (020)	11	Corrected Outliers (Raw data)	Rainfall	Corrected by RDS (Reciprocal Distance Squared) method	Outlier Handling
	12			Corrected by the judgment of the person in charge (reason of judgment, correcting method should be written)	
	21		Water Stage	Corrected by linear interpolation	
	22			Corrected by cubic spline interpolation	
	23			Corrected by relation equation with upstream & downstream stations' data	
	24			Corrected by the judgment of the person in charge (reason of judgment & correcting method should be written)	
03 (030)	11 - 12	Corrected Missing Data (Raw data)	Rainfall	Methods in the 2?? codes applied to missing data	Missing data Handling
	21 - 24		Water Stage		
04		Outlier that cannot be corrected			No change
05		Missing data that cannot be corrected			No change
09 (090)	10	AQC result (MQC needed with caution)	Rainfall	Too large value	Too large rainfall value
	20			Observed value is more than the ??? % and less than the ??? % of the RDS value	Spatial Inconsistency
	30			Observed value is less than zero	
	91			Missing data	Not recorded
	50		Water Stage	Abrupt zero value	Abrupt change
	60			Slope is more than ?? times or than -?? times of the slope for previous two time steps	
	70			Constant water stage for more than ??? hours	Constant value
	92			Missing data	



Hydrological Data Quality Control System (HDQCS)

Rainfall Data Import



Hydrological Data Quality Control System (HDQCS)

Rainfall Criteria (Max. Rainfall)

2023-11-06 11:45:25

Rainfall

Verification
- Data
- Max. Rainfall
- RDS
- [AI] Abnormal Dect.
  - Supervised Learning
  - Unsupervised Learning

Data Correction
- Data
- Data Correction
- [AI] Data Correction

Report

Water Level

Verification
- Data
- Abrupt Change
- Constant Value
- [AI] Abnormal Dect.
  - Supervised Learning
  - Unsupervised Learning

Data Correction
- Data
- Data Correction

Hydrological Data Quality Control System

Rainfall / Verification | Rainfall / Data correction | Rainfall / Report

Data | Max. Rainfall | RDS | AI

Rainfall distribution by amount

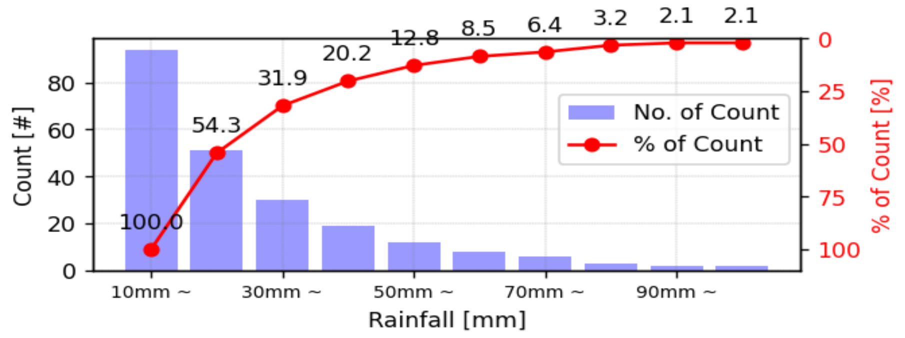
	10mm~	20mm~	30mm~	40mm~	50mm~	60mm~	70mm~	80mm~	90mm~	100mm
No. of Count	94.0	51.0	30.0	19.0	12.0	8.0	6.0	3.0	2.0	2.0
% of Count	100.0	54.3	31.9	20.2	12.8	8.5	6.4	3.2	2.1	2.1

Statistics

Mean	Std.	Mean+2*Std.	Mean-2*Std.	Max (mm)	0.9max (mm)
13.5	18.2	49.8	-22.8	121.0	108.9

Outlier Verification Criteria

% of count	2.1	%	Calculation
Max. rainfall	90.0	mm	Apply



Download

Hydrological Data Quality Control System (HDQCS)

Rainfall Data Criteria (RDS)

2023-11-06 11:45:25

Hydrological Data Quality Control System

Rainfall

Verification
- Data
- Max. Rainfall
- RDS
- [AI] Abnormal Dect.
  - Supervised Learning
  - Unsupervised Learning

Data Correction
- Data
- Data Correction
- [AI] Data Correction

Report

Water Level

Verification
- Data
- Abrupt Change
- Constant Value
- [AI] Abnormal Dect.
  - Supervised Learning
  - Unsupervised Learning

Data Correction
- Data
- Data Correction

Rainfall / Verification | Rainfall / Data correction | Rainfall / Report

Data | Max. Rainfall | RDS | AI

RDS Calculation | RDS Criteria

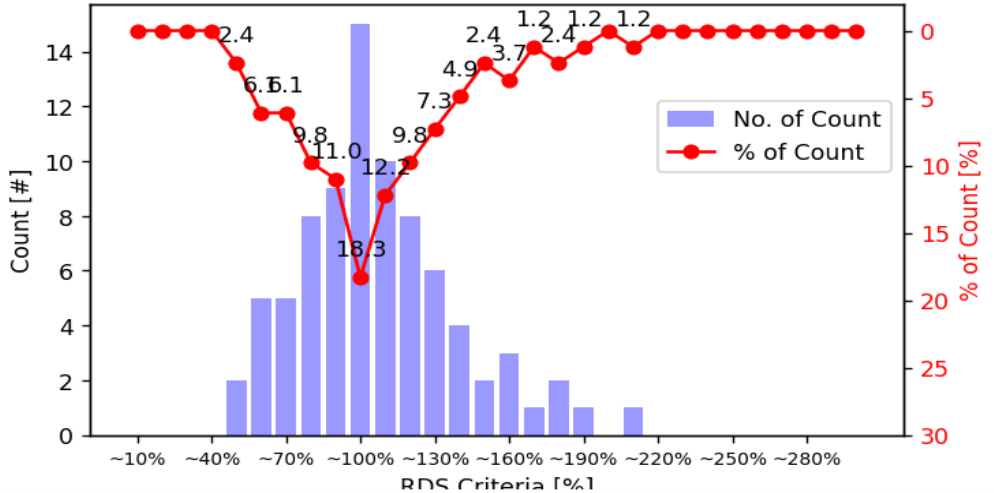
	~10%	~20%	~30%	~40%	~50%	~60%	~70%	~80%	~90%	~100%
No. of Count	0.0	0.0	0.0	0.0	2.0	5.0	5.0	8.0	9.0	15.0
% of Count	0.0	0.0	0.0	0.0	2.4	6.1	6.1	9.8	11.0	18.3

	~110%	~120%	~130%	~140%	~150%	~160%	~170%	~180%	~190%	~200%
No. of Count	10.0	8.0	6.0	4.0	2.0	3.0	1.0	2.0	1.0	0.0
% of Count	12.2	9.8	7.3	4.9	2.4	3.7	1.2	2.4	1.2	0.0

	~210%	~220%	~230%	~240%	~250%	~260%	~270%	~280%	~290%	~300%
No. of Count	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
% of Count	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Statistics

Mean	Std.	Mean+2*Std.	Mean-2*Std.	Max (mm)	0.9Max (mm)
102.2	32.1	166.3	38.1	208.6	187.8



Download

Outlier Verification Criteria

Upper limit	5	%	Calculation	Upper difference	139.6	%	Apply
Lower limit	5	%	Calculation	Lower difference	57.0	%	Apply

Hydrological Data Quality Control System (HDQCS)

Water Level Data Import

2023-11-06 11:45:25

Hydrological Data Quality Control System

Rainfall

Verification
- Data
- Max. Rainfall
- RDS
- [AI] Abnormal Dect.
  - Supervised Learning
  - Unsupervised Learning

Data Correction
- Data
- Data Correction
- [AI] Data Correction

Report

Water Level

Verification
- Data
- Abrupt Change
- Constant Value
- [AI] Abnormal Dect.
  - Supervised Learning
  - Unsupervised Learning

Data Correction
- Data
- Data Correction

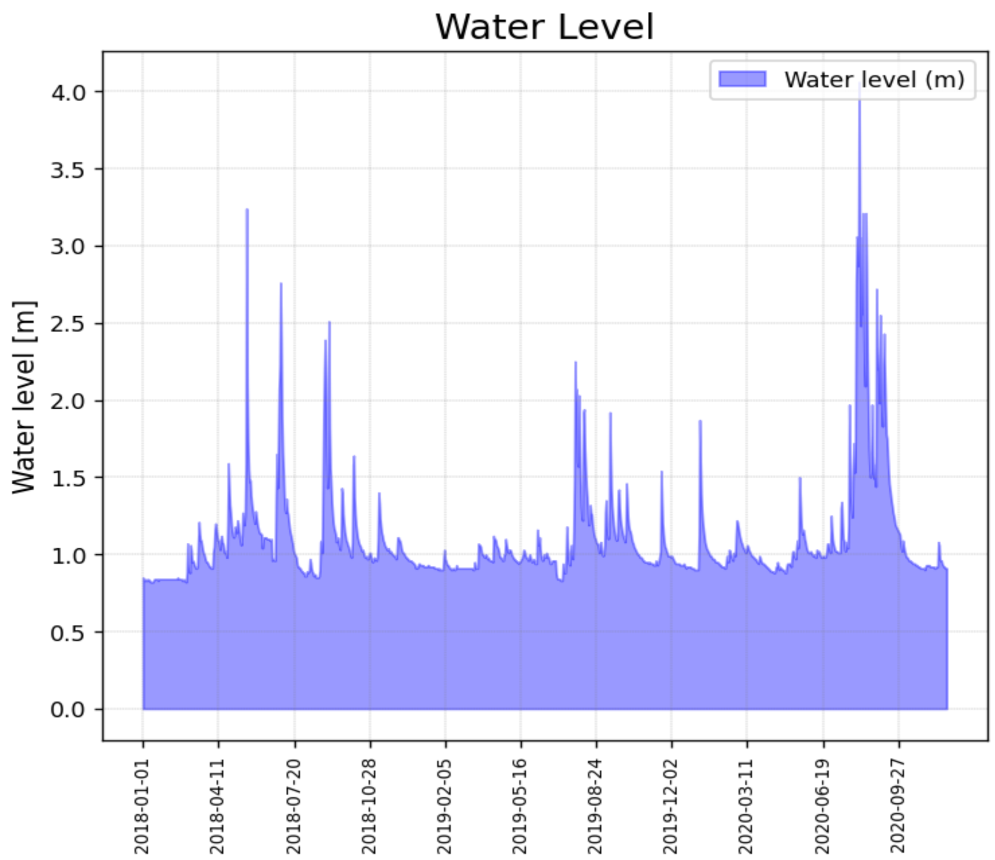
Water level / Verification Criteria | Water level / Data Correction | Water level / Report

Data | Abrupt Change | Constant Value | AI

Time Interval : daily hourly 10-minute

Data Import Import

	Date	Water level (m)
1	2018-01-01	0.85
2	2018-01-02	0.84
3	2018-01-03	0.83
4	2018-01-04	0.83
5	2018-01-05	0.84
6	2018-01-06	0.83
7	2018-01-07	0.83
8	2018-01-08	0.84
9	2018-01-09	0.83
10	2018-01-10	0.83
11	2018-01-11	0.82
12	2018-01-12	0.82
13	2018-01-13	0.82
14	2018-01-14	0.82
15	2018-01-15	0.83
16	2018-01-16	0.84
17	2018-01-17	0.84
18	2018-01-18	0.84
19	2018-01-19	0.84



Hydrological Data Quality Control System (HDQCS)

Water Level Criteria (Abrupt change)

2023-11-06 11:45:25

Hydrological Data Quality Control System

Rainfall

- **Verification**
  - Data
  - Max. Rainfall
  - RDS
  - [AI] Abnormal Dect.
    - Supervised Learning
    - Unsupervised Learning
- **Data Correction**
  - Data
  - Data Correction
  - [AI] Data Correction
- **Report**

Water Level

- **Verification**
  - Data
  - Abrupt Change
  - Constant Value
  - [AI] Abnormal Dect.
    - Supervised Learning
    - Unsupervised Learning
- **Data Correction**
  - Data
  - Data Correction

Water level / Verification Criteria | Water level / Data Correction | Water level / Report

Data | Abrupt Change | Constant Value | AI

Water level change ratio | Abrupt change criteria

	~(-X5)	~(-X4)	~(-X3)	~(-X2)	~(-X1)	~(X0)
No. of Count	6.0	2.0	15.0	22.0	94.0	339.0
% of Count	0.82	0.27	2.05	3.0	12.82	46.25

	(+X1)~	(+X2)~	(+X3)~	(+X4)~	(+X5)~	(+X6)~	(+X7)~	(+X8)~	(+X9)~	(+X10)~
No. of Count	203.0	16.0	16.0	3.0	4.0	0.0	0.0	0.0	0.0	1.0
% of Count	27.69	2.18	2.18	0.41	0.55	0.0	0.0	0.0	0.0	0.14

Statistics

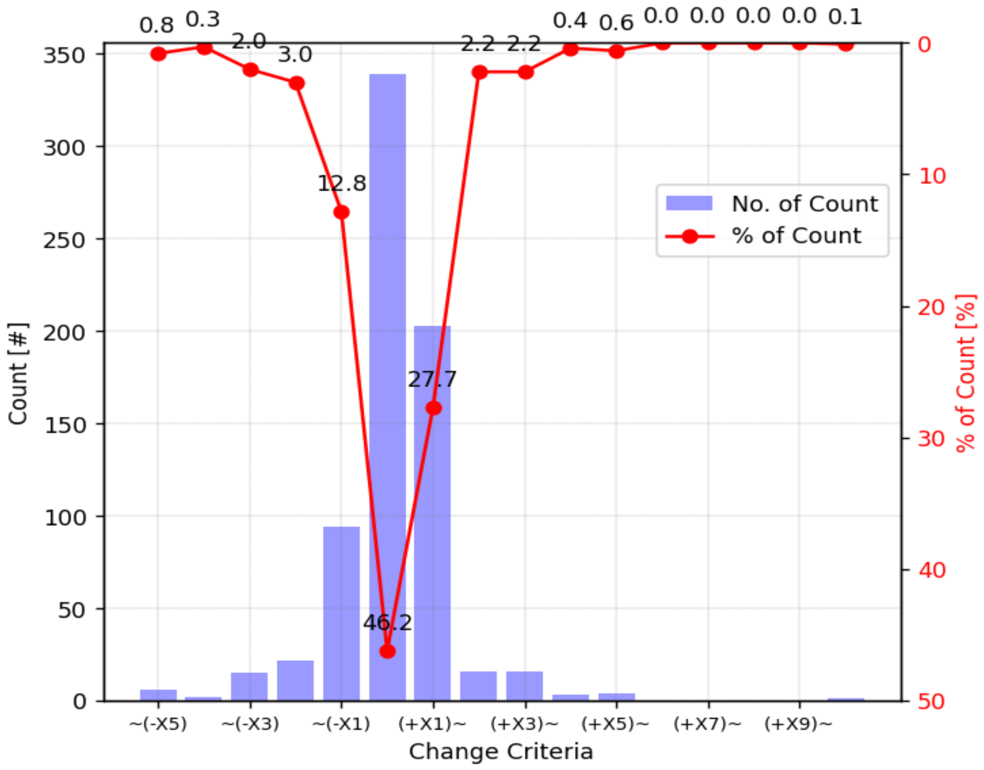
	Mean	Std.	Mean+2*Std.	Mean-2*Std.	Max
1	-0.26	3.84	7.43	-7.95	47

Outlier Verification Criteria

Upper limit	1	%
Lower limit	1	%
Upper change	3.7	
Lower change	-3.6	

Calculation

Apply



Download

Hydrological Data Quality Control System (HDQCS)

Water Level Criteria (Constant value)

2023-11-06 11:45:25

Hydrological Data Quality Control System

Rainfall

Verification
- Data
- Max. Rainfall
- RDS
- [AI] Abnormal Dect.
  - Supervised Learning
  - Unsupervised Learning

Data Correction
- Data
- Data Correction
- [AI] Data Correction

Report

Water Level

Verification
- Data
- Abrupt Change
- Constant Value
- [AI] Abnormal Dect.
  - Supervised Learning
  - Unsupervised Learning

Data Correction
- Data
- Data Correction

Water level / Verification Criteria | Water level / Data Correction | Water level / Report

Data | Abrupt Change | Constant Value | AI

Water level distribution by constant time

	48hr	72hr	96hr	120hr	144hr	168hr	240hr	360hr	480hr	600hr	720hr
No. of Count	92.0	17.0	15.0	11.0	5.0	1.0	0.0	0.0	1.0	1.0	0.0
% of Count	100.0	18.5	16.3	12.0	5.4	1.1	0.0	0.0	1.1	1.1	0.0

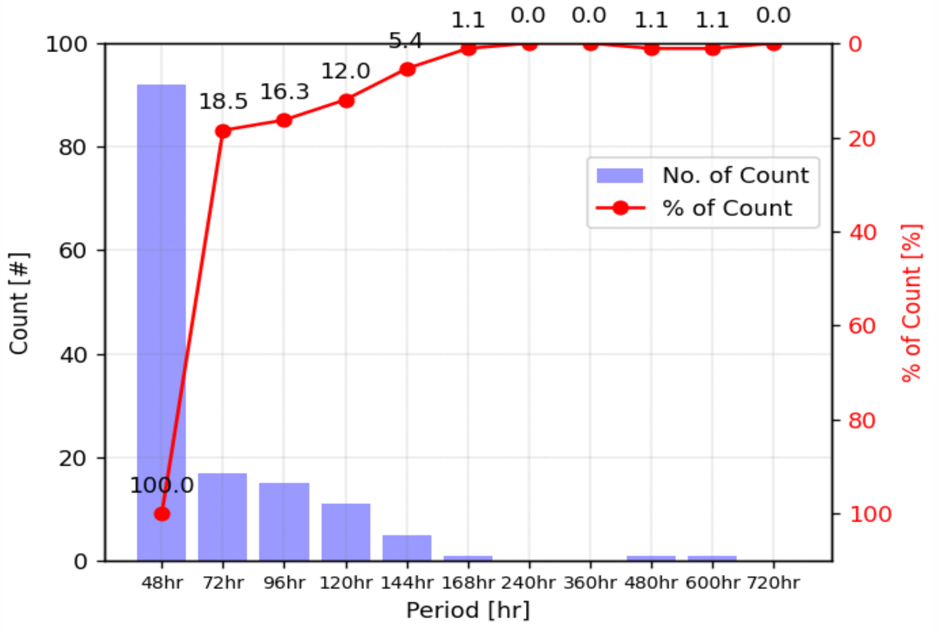
Statistics

	Mean	Std.	Mean+2*Std.	Mean-2*Std.	Max
1	3.03	2.67	8.37	-2.31	25

Outlier Verification Criteria

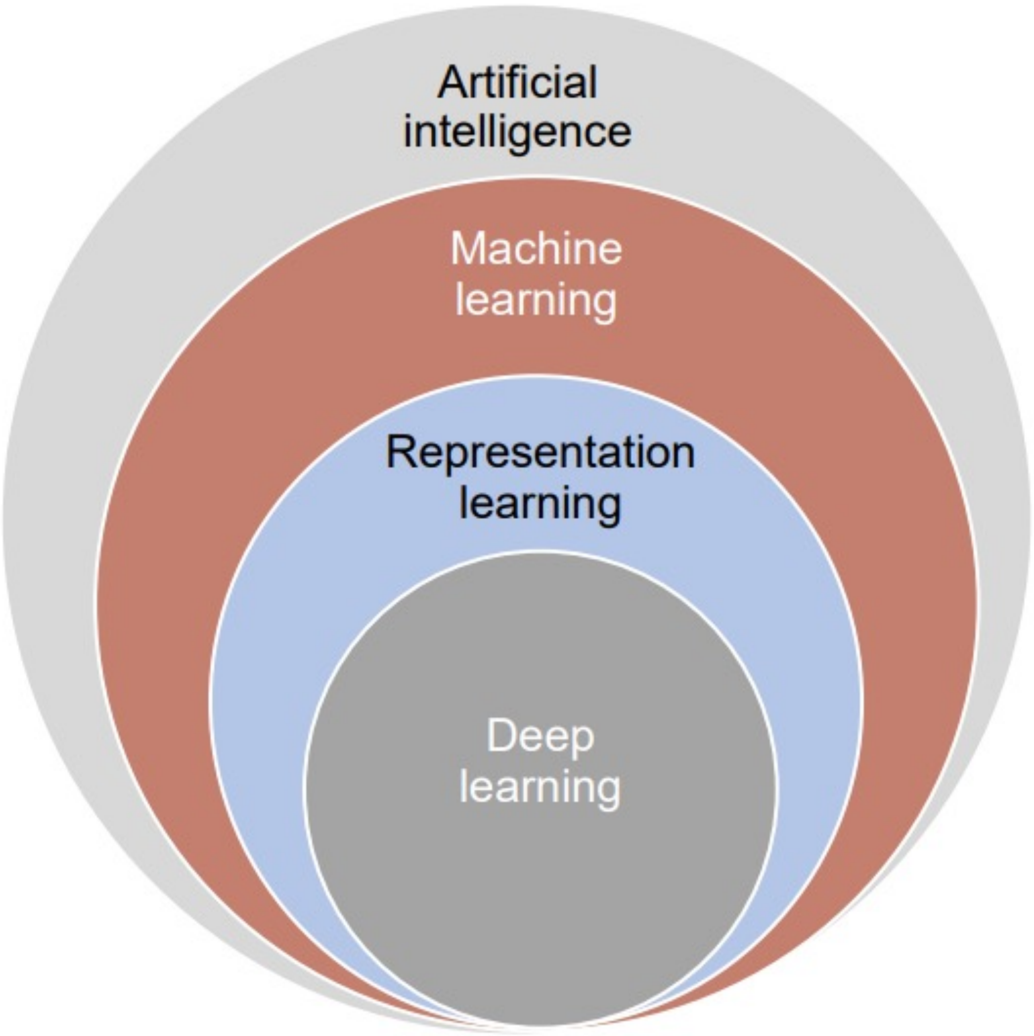
% of count: 1 % — Calculation

Max. period: 7.3 day — Apply



Download

Description of
AI (Artificial Intelligent), Machine Learning, Deep Learning



Artificial Intelligence (AI):
The study of intelligence demonstrated by a machine manifested by its capability to perceive the environment and take actions to achieve its goals and tasks through flexible adaptation.

Machine Learning (ML)
A sub-set of AI, which are learning methods and algorithms that enable computers to automatically improve performance through experience.

Representation Learning (RL)
Techniques that automatically discover representation (or featur es) that are useful for subsequent learning tasks. Also known as feature learning.

Deep Learning (DL)
A class of machine learning algorithms based on artificial neural networks (ANNs) and using hierarchical architectures to extract higher level features from input data via representation learning
.

History of AI Application in Hydrology Sector

***1980s**: Artificial neural networks (ANNs) were first applied to hydrology, primarily for rainfall-runoff modeling.*

***1990s**: Support vector machines (SVMs) were introduced and applied to hydrology, showing promising results for streamflow forecasting.*

***Early 2000s**: Other machine learning techniques, such as decision trees and random forests, were applied to hydrology, mainly for flood risk mapping and water quality modeling.*

***Mid-2000s**: Ensemble methods, such as bagging and boosting, were applied to hydrology modeling, showing improved accuracy compared to individual models.*

***Mid-2000s**: Ensemble methods, such as bagging and boosting, were applied to hydrology modeling, showing improved accuracy compared to individual models.*

***Late 2000s**: Deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), were introduced and applied to hydrology, showing improved accuracy for various hydrologic tasks, including streamflow prediction, rainfall-runoff modeling, and flood forecasting.*

***2010s**: Transfer learning, a technique where knowledge learned in one task is transferred to another task, was applied to hydrology, showing improved accuracy for streamflow prediction.*

***2020s**: With the increasing availability of large datasets, there has been a growing interest in applying ML to hydrology modeling, including applications of generative adversarial networks (GANs) for rainfall simulation and unsupervised learning for anomaly detection in hydrologic data.*

Selection of AI Deep-Learning Technique for HDQC

1) Supervised Learning for Anomaly Detection Using XGBoost

- XGBoost, an extension of gradient boosting, can be used as a supervised anomaly detection algorithm by treating ***one class as anomalies and the other as normal data***.

- ***Labelled data is required*** to enable the model to learn to distinguish between the abnormal and normal data. XGBoost cannot be carried out without labelled data.

2) Unsupervised Learning for Anomaly Detection Using Isolation Forest

- An ensemble method that isolates anomalies by constructing random forests and isolating data points that require fewer splits in the tree to be isolated.

- It is a simple yet effective approach for detecting anomalies. ***No labelled data is required***, the anomaly can be recognized by just specifying the contamination level.

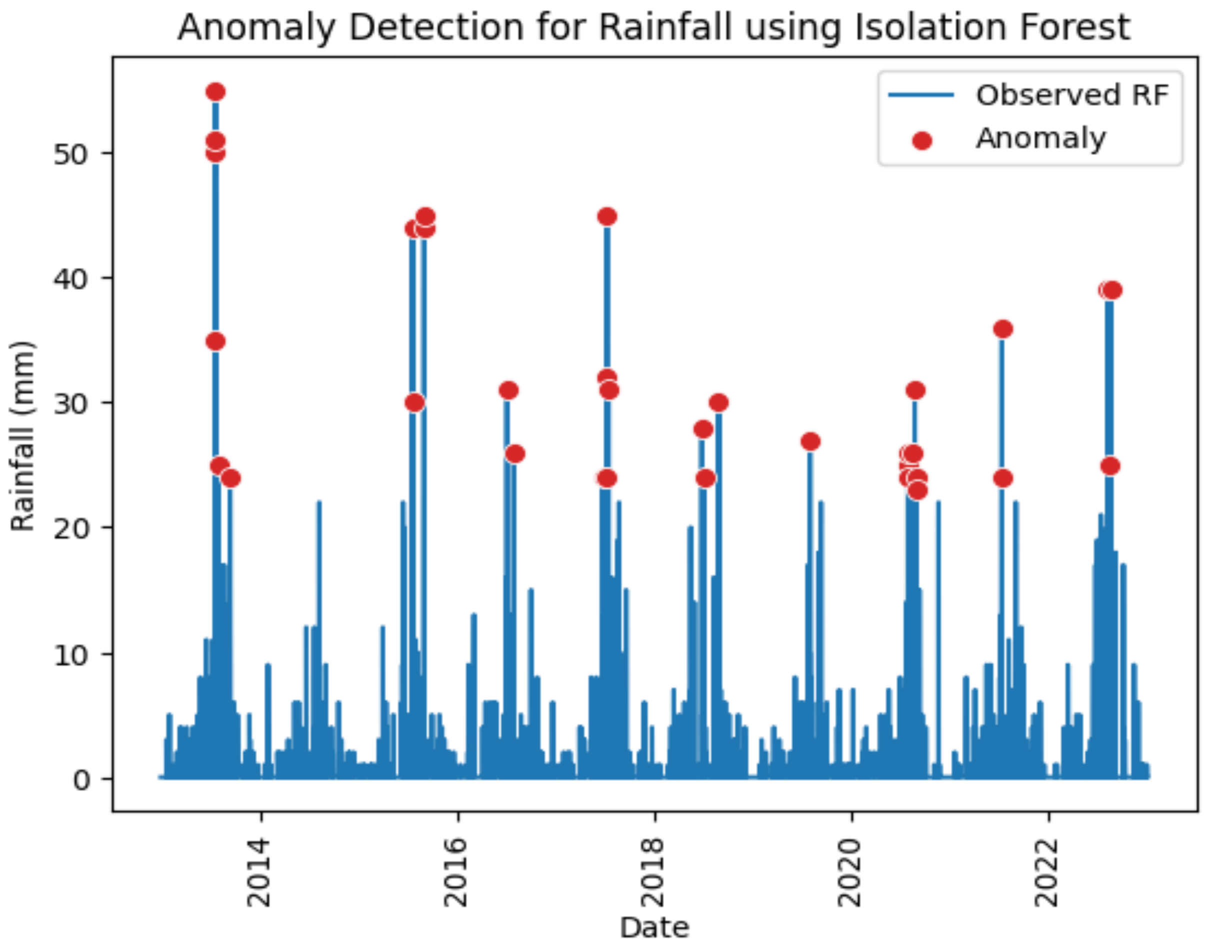
 # All the codes are written using open-source Python language

Application of AI HDQCS in TC member countries

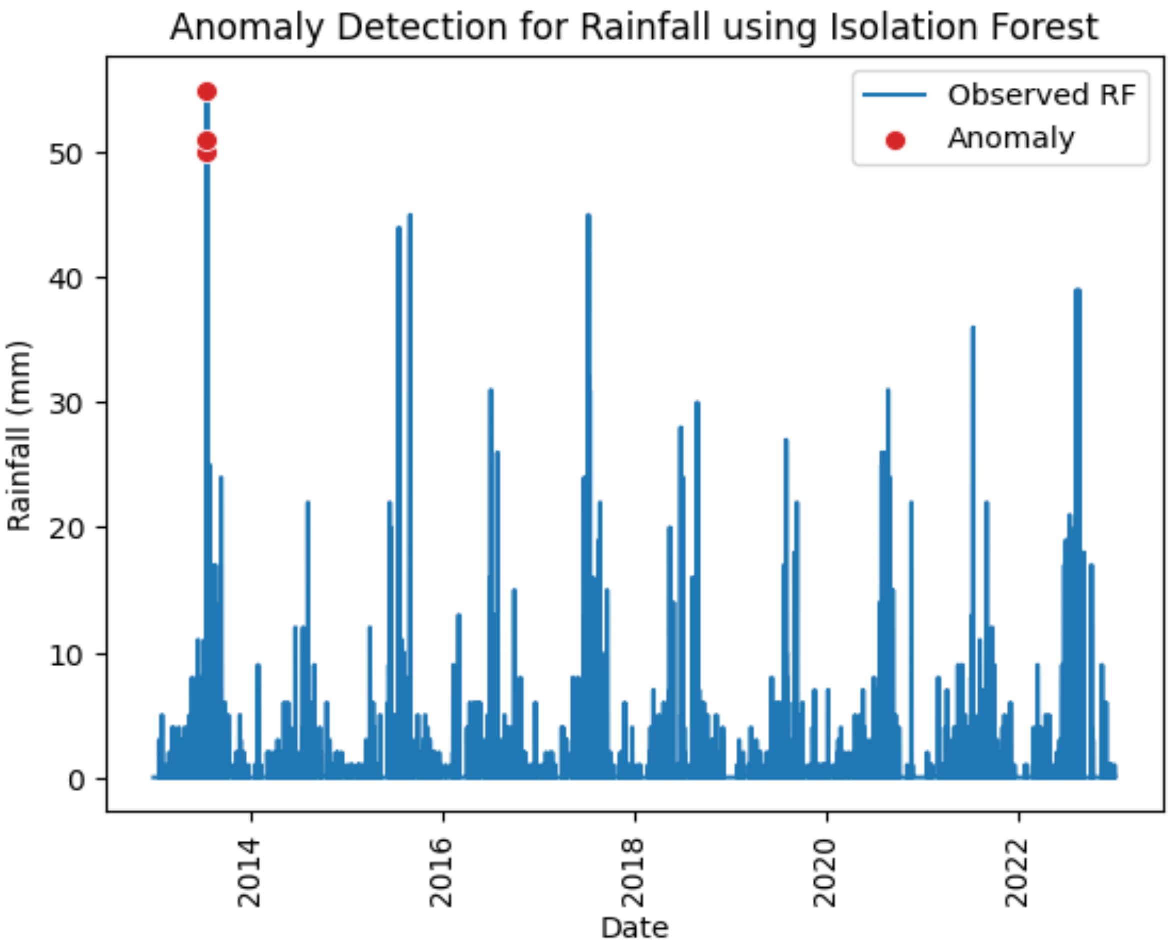
Using Isolation Forest for Identifying Abnormal Rainfall Data in Republic of Korea (Hourly Data) & Lao PDR (Daily Data)

Max. Rainfall for hourly data in ROK (Yeoju Bridge)
Zero data is excluded

Contamination Level=0.01, Threshold=23mm



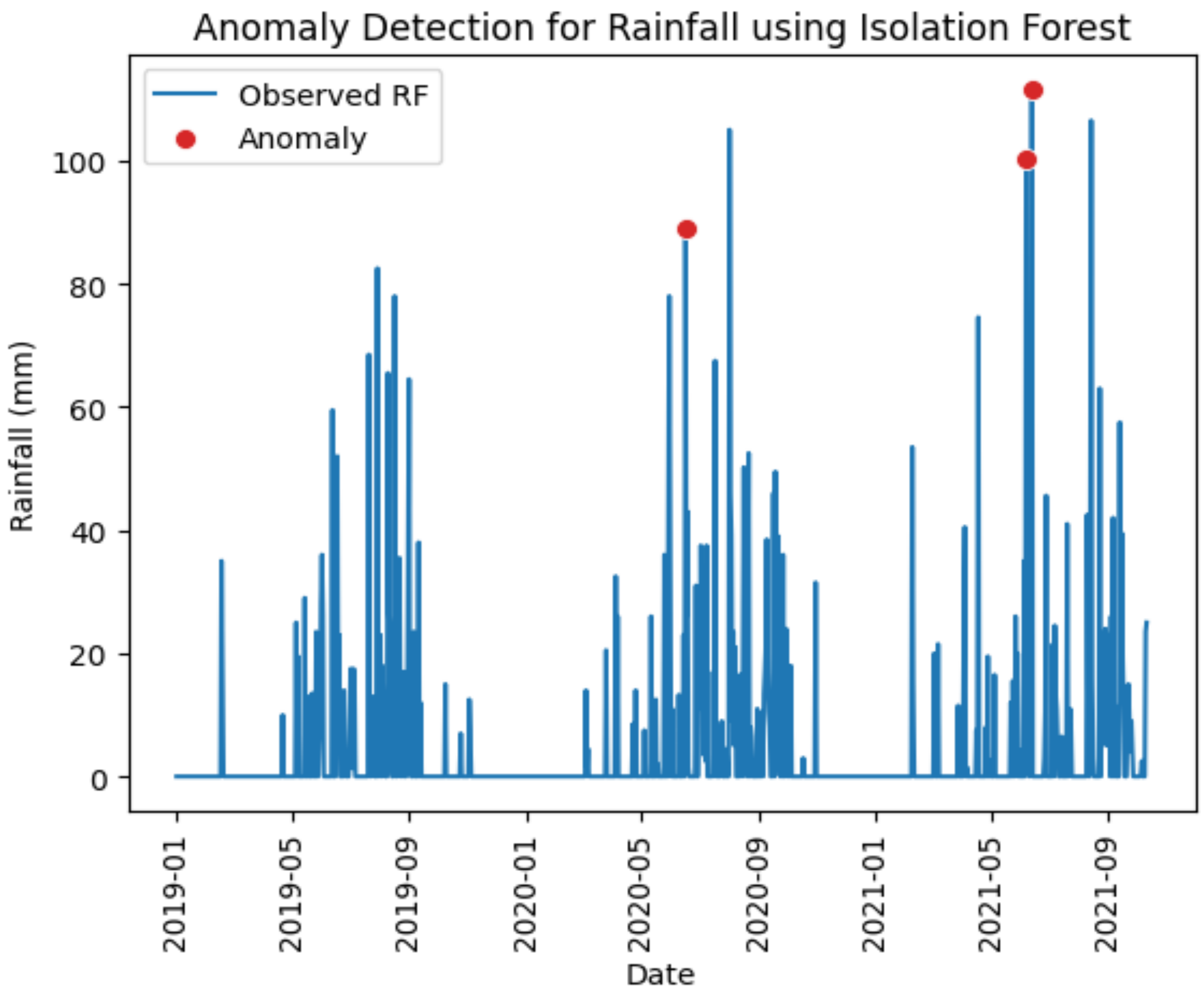
Contamination Level=0.001, Threshold=50mm



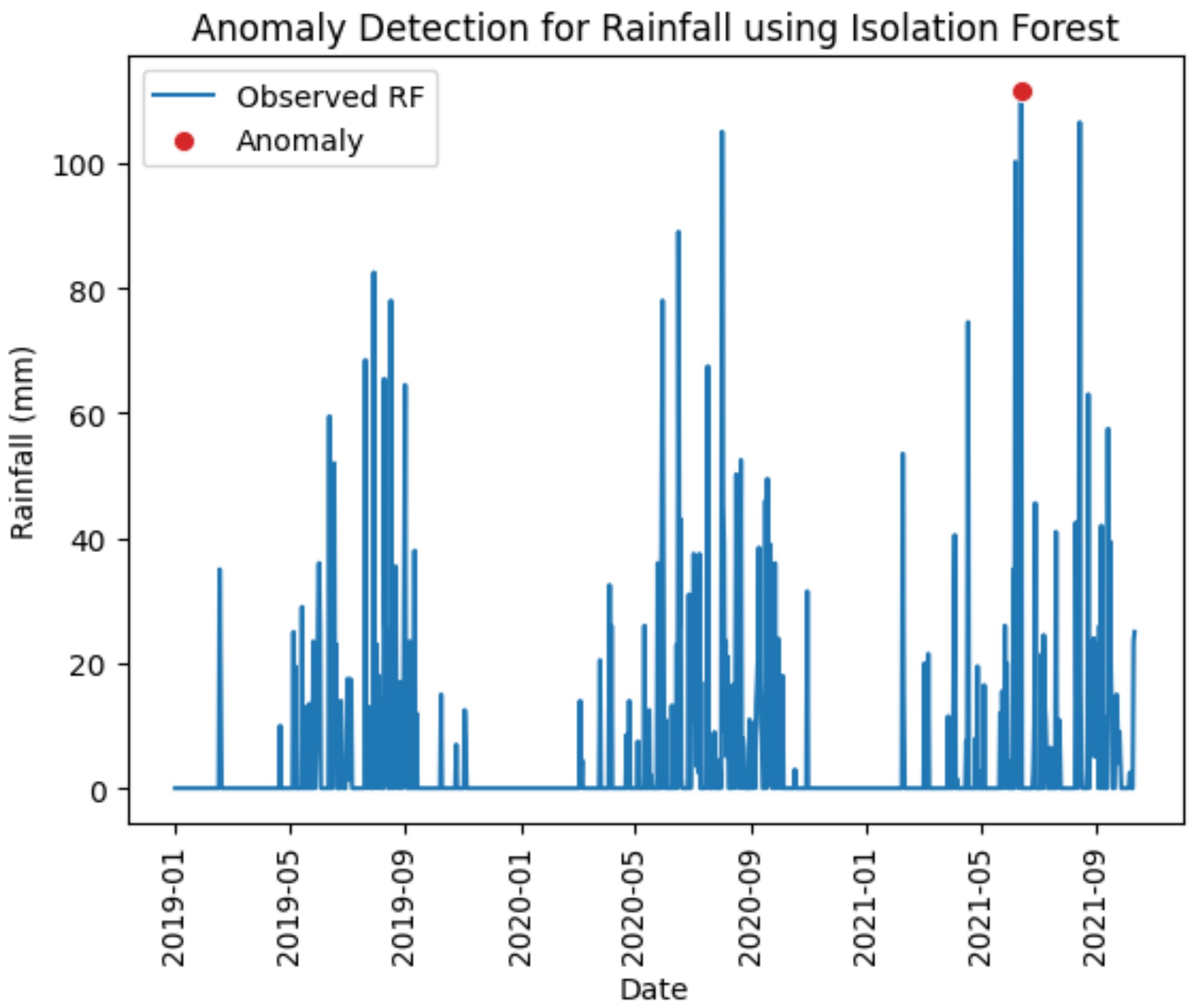
- The concept of contamination level is similar with percentile where contamination level of 0.01 is equivalent to 99 percentile and contamination level of 0.001 is equivalent to 99.9 percentile.
- Low threshold value will be obtained for outlier identification when zero data is included due to increasing amount of irrelevant data sample.

Max. Rainfall for daily data in Lao PDR (Pakkayoung Station)
Zero data is excluded

Contamination Level=0.01, Threshold=104mm



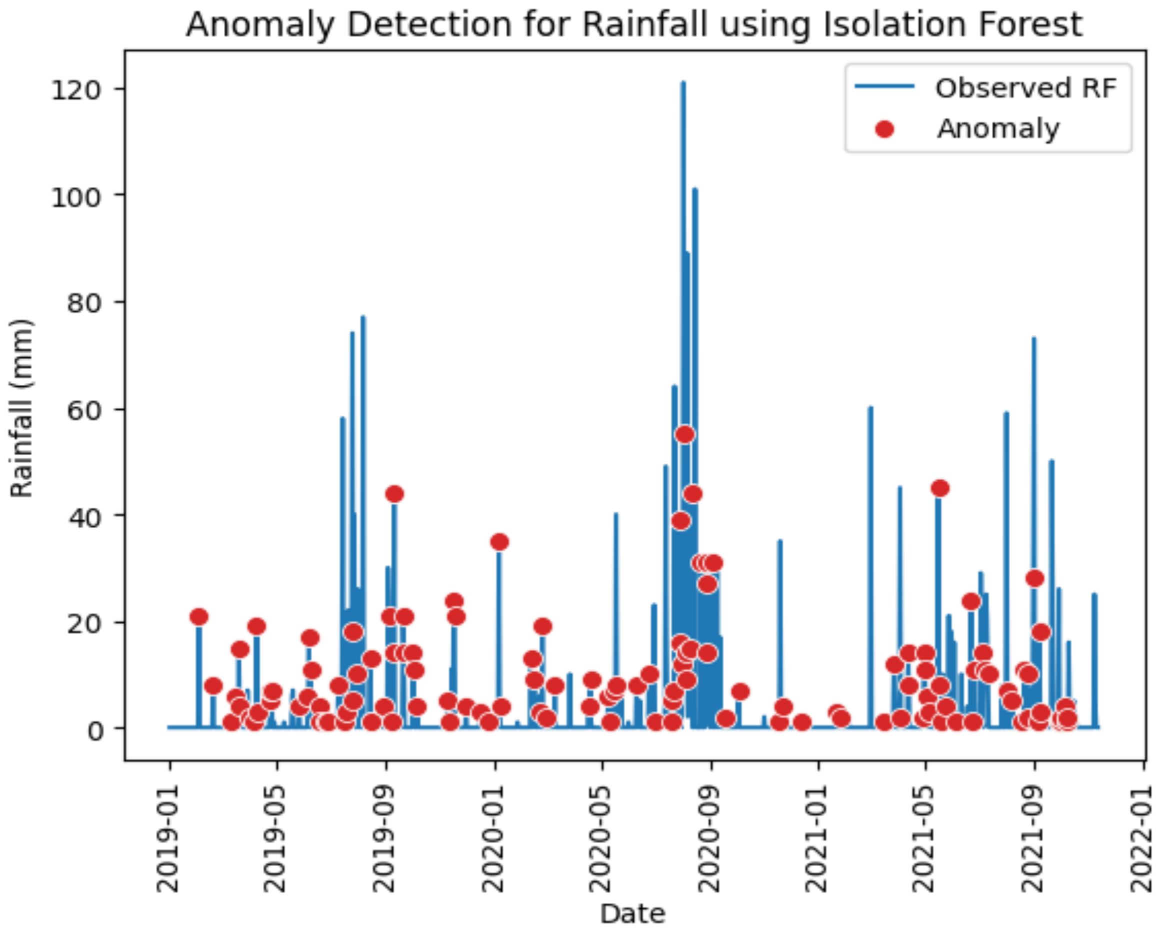
Contamination Level=0.001, Threshold=110mm



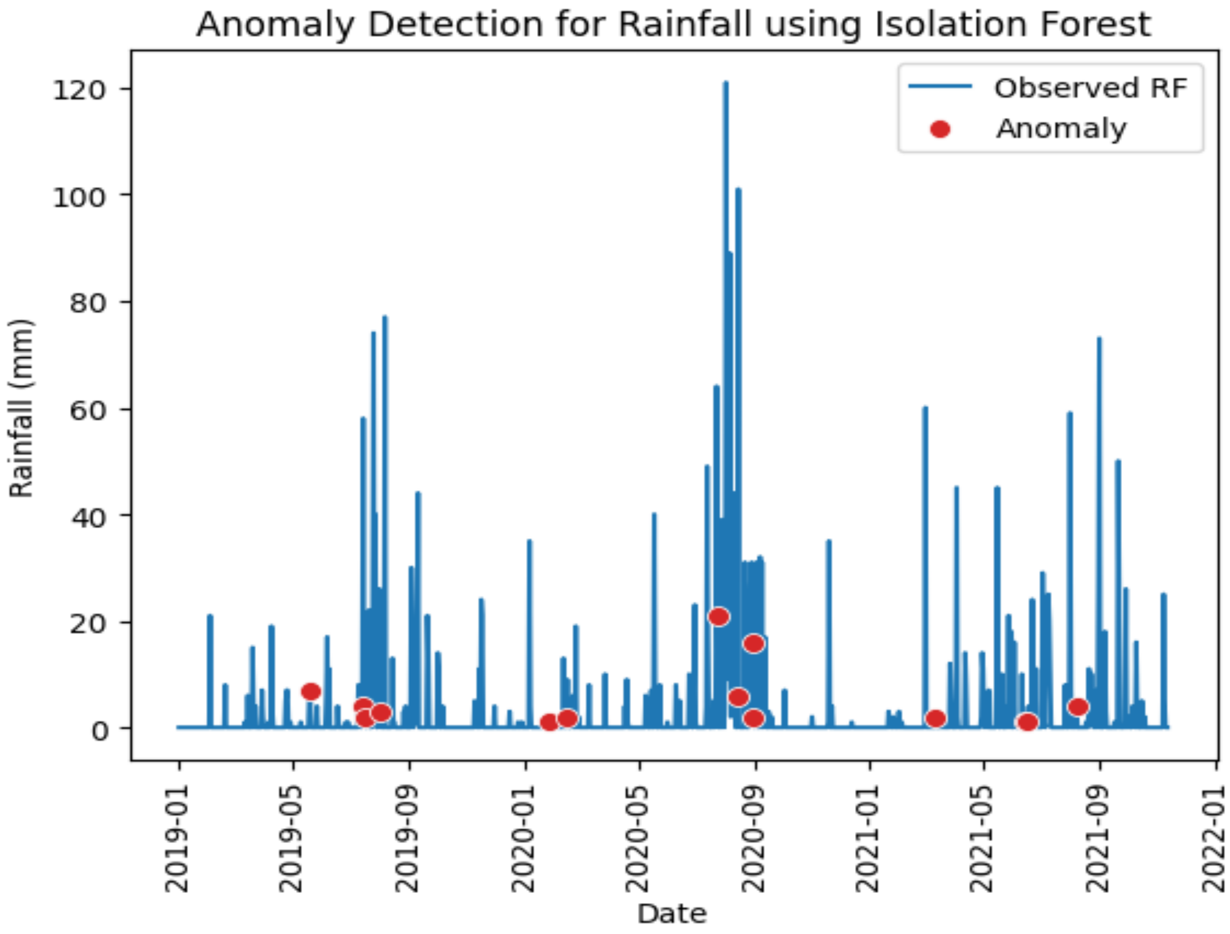
- The lesser the contamination level specified, the higher the threshold value is, and the lesser data will be identified as abnormal data.

RDS for daily data in **Lao PDR** (Thalad Station)

Contamination Level=0.15, Upper Threshold=+55.5%



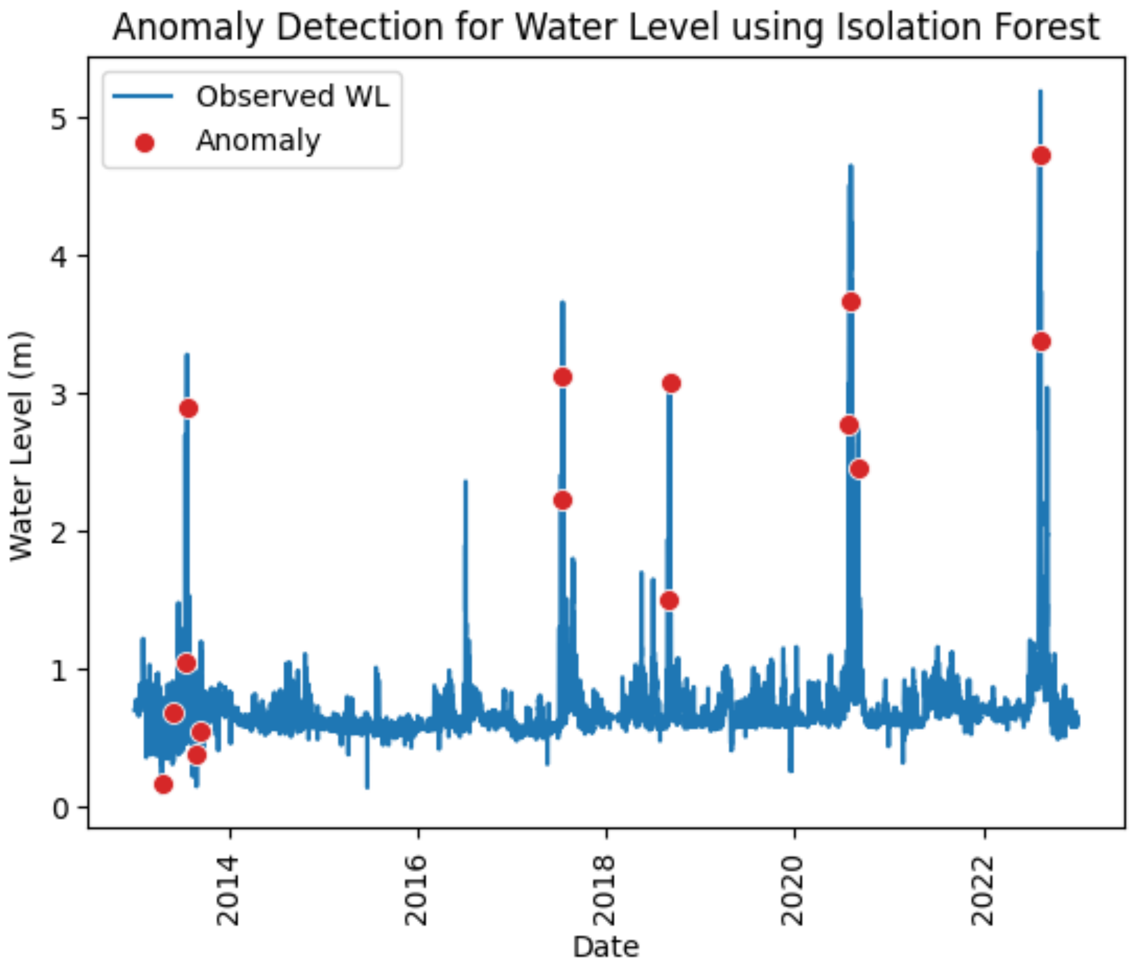
Contamination Level=0.15, Lower Threshold=-8.2%



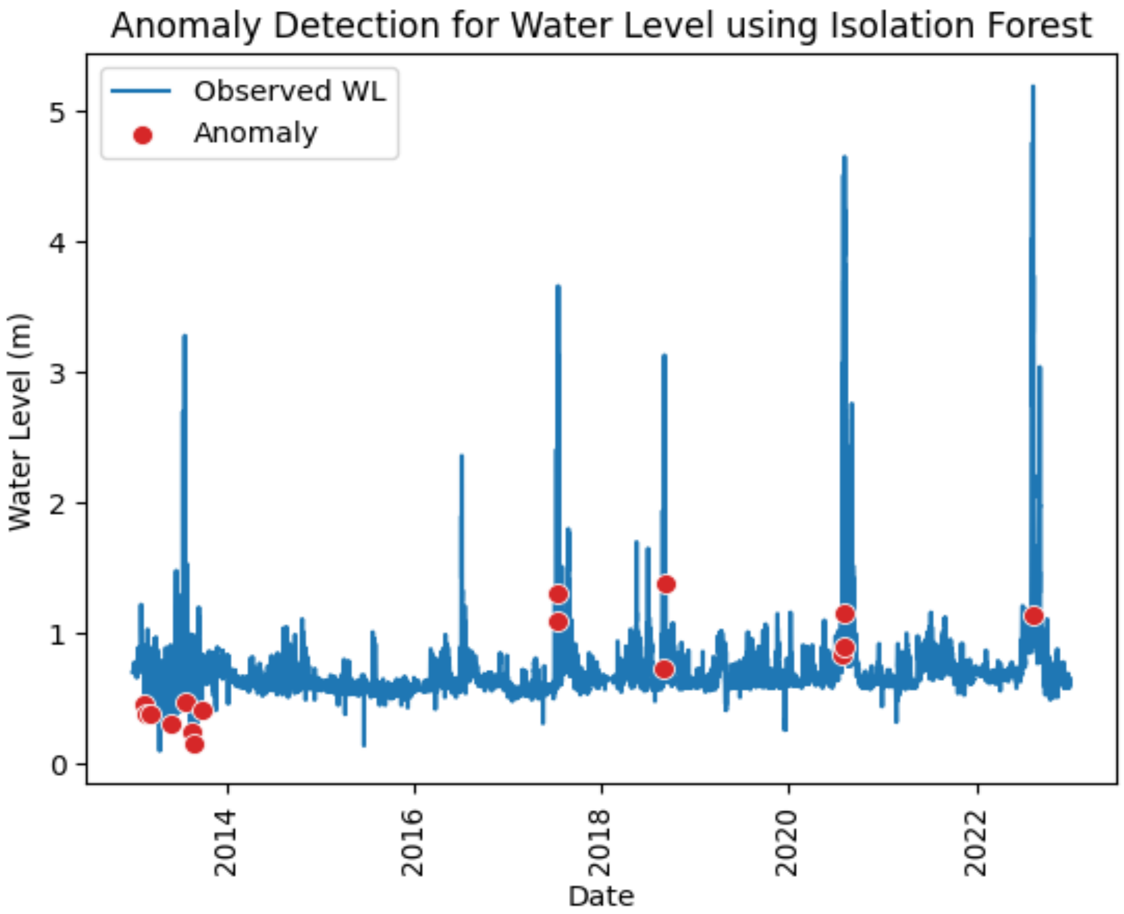
- The RDS abnormal is identified from the differences between the rainfall value at the target station and the RDS rainfall value calculated from nearby rainfall stations.

Using Isolation Forest
for
Identifying Abnormal Water Level Data
in
ROK(Hourly Data)
and
Lao PDR (Daily Data)

Water Level Gradient (*between current and previous one hour data*) **for hourly data in ROK (Yeoju Bridge)**
WL Gradient for Upper Limit, Contamination level = 0.001
WL Gradient for Lower Limit, Contamination level = 0.001

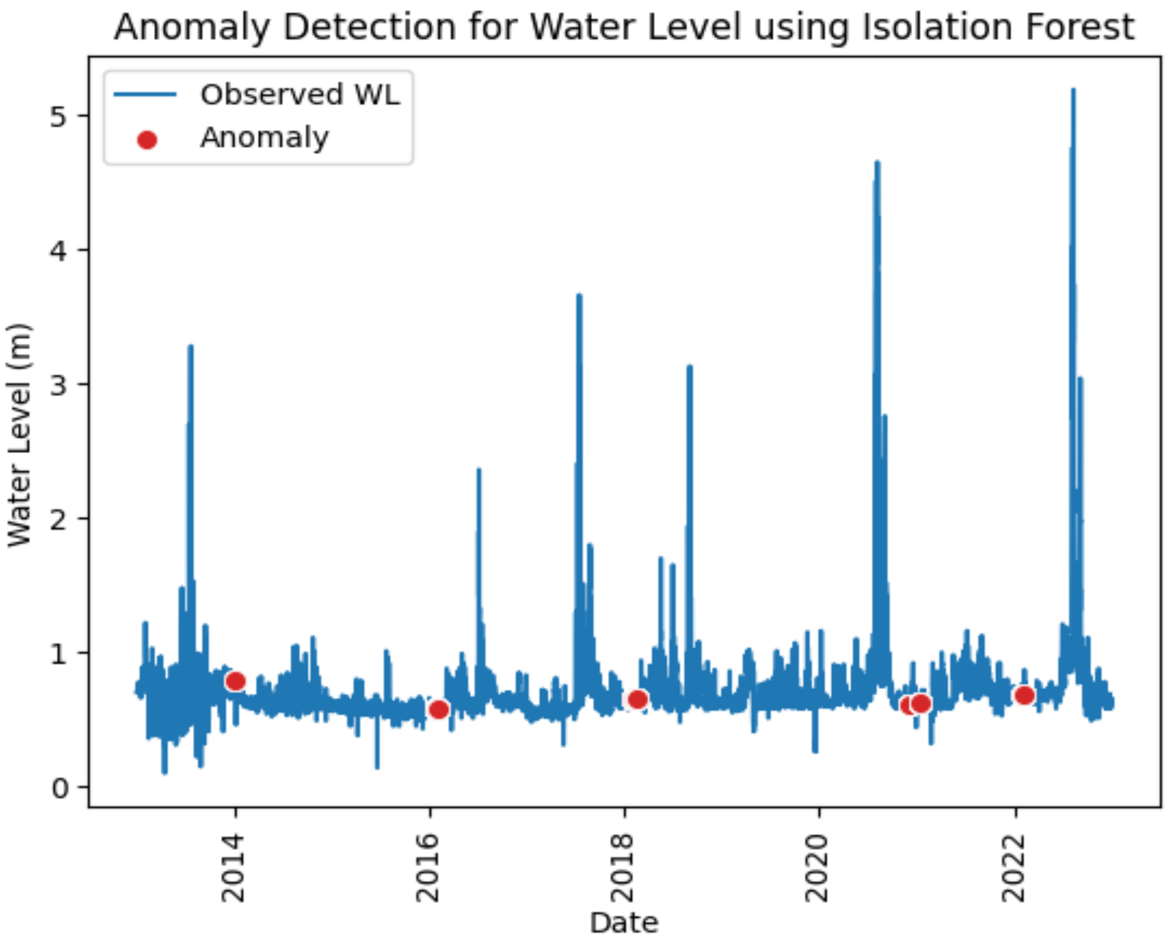


Upper Threshold=+68%

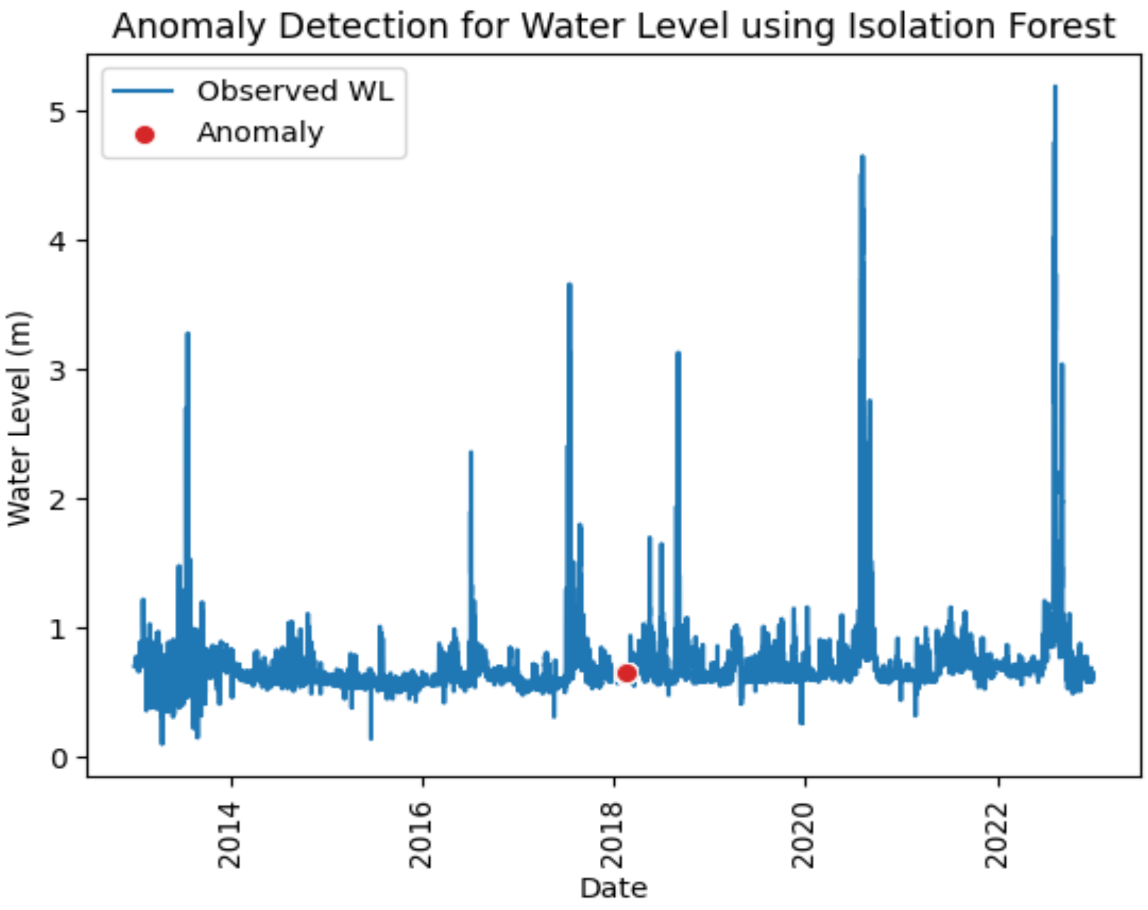


Lower Threshold=-41.3%

Constant water level for hourly data in ROK (Yeoju Bridge)

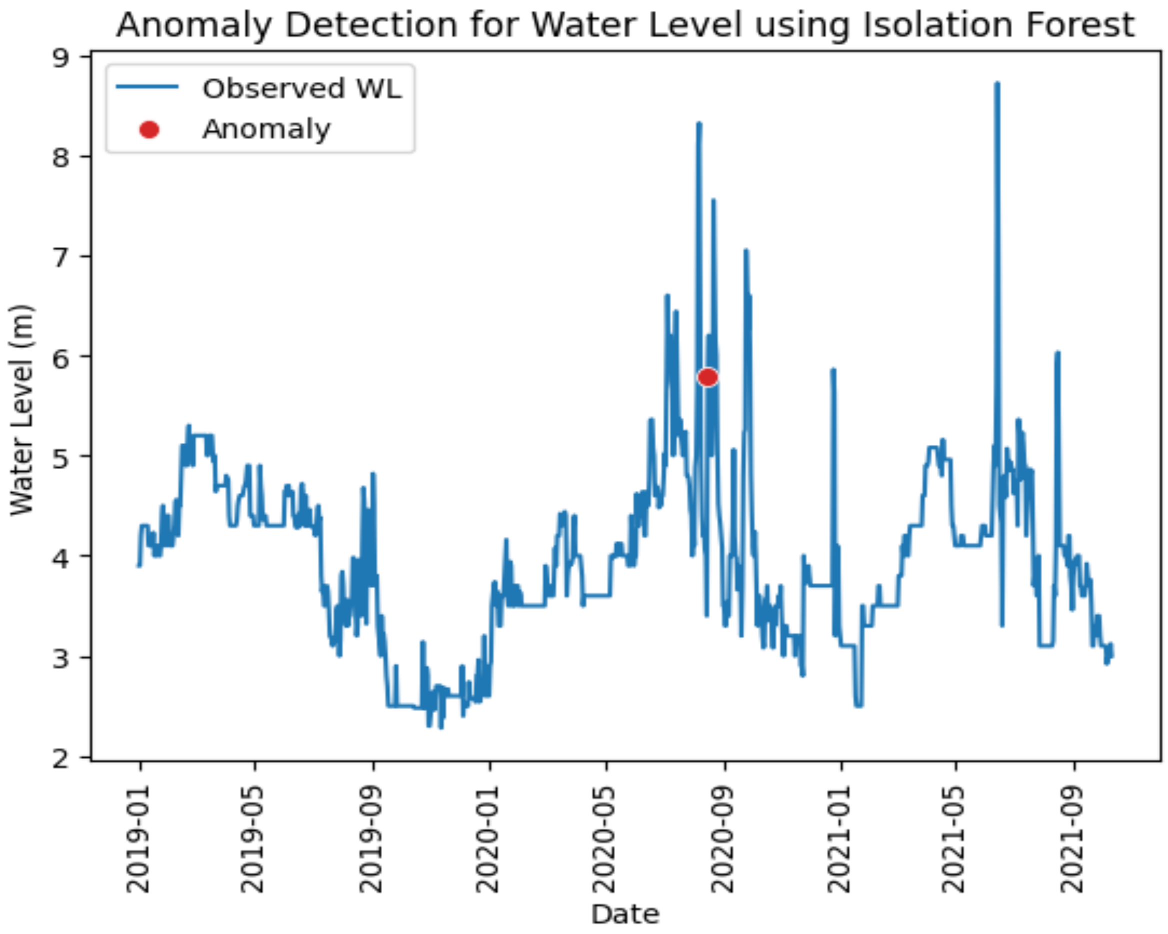


Contamination Level=0.01, Threshold=102 hours

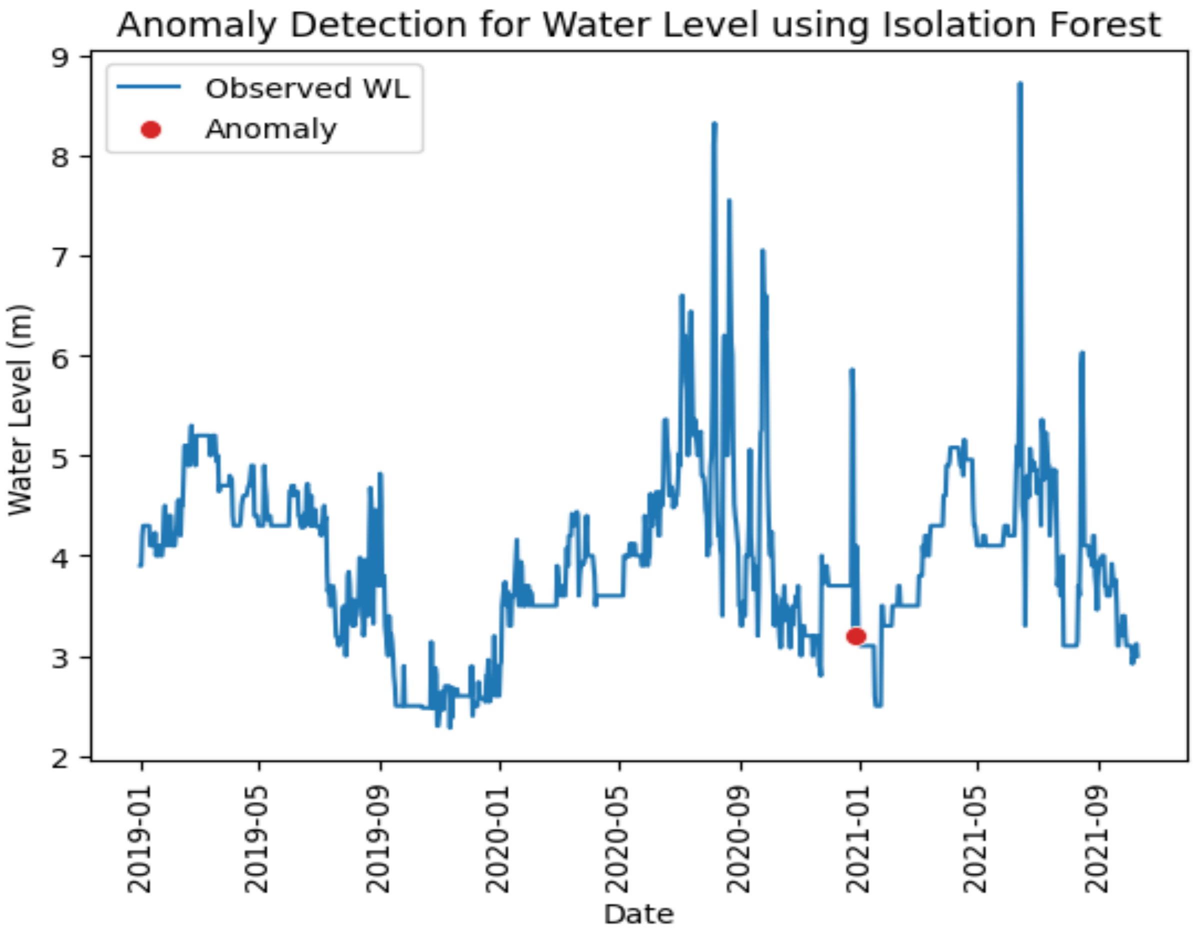


Contamination Level=0.001, Threshold=407 hours

Water Level Gradient (*between current and previous one day data*) **for daily data in Lao PDR (Pakkayoung)**
Dummy WL Gradient for Upper Limit, Contamination level = 0.001
Dummy WL Gradient for Lower Limit, Contamination level = 0.001

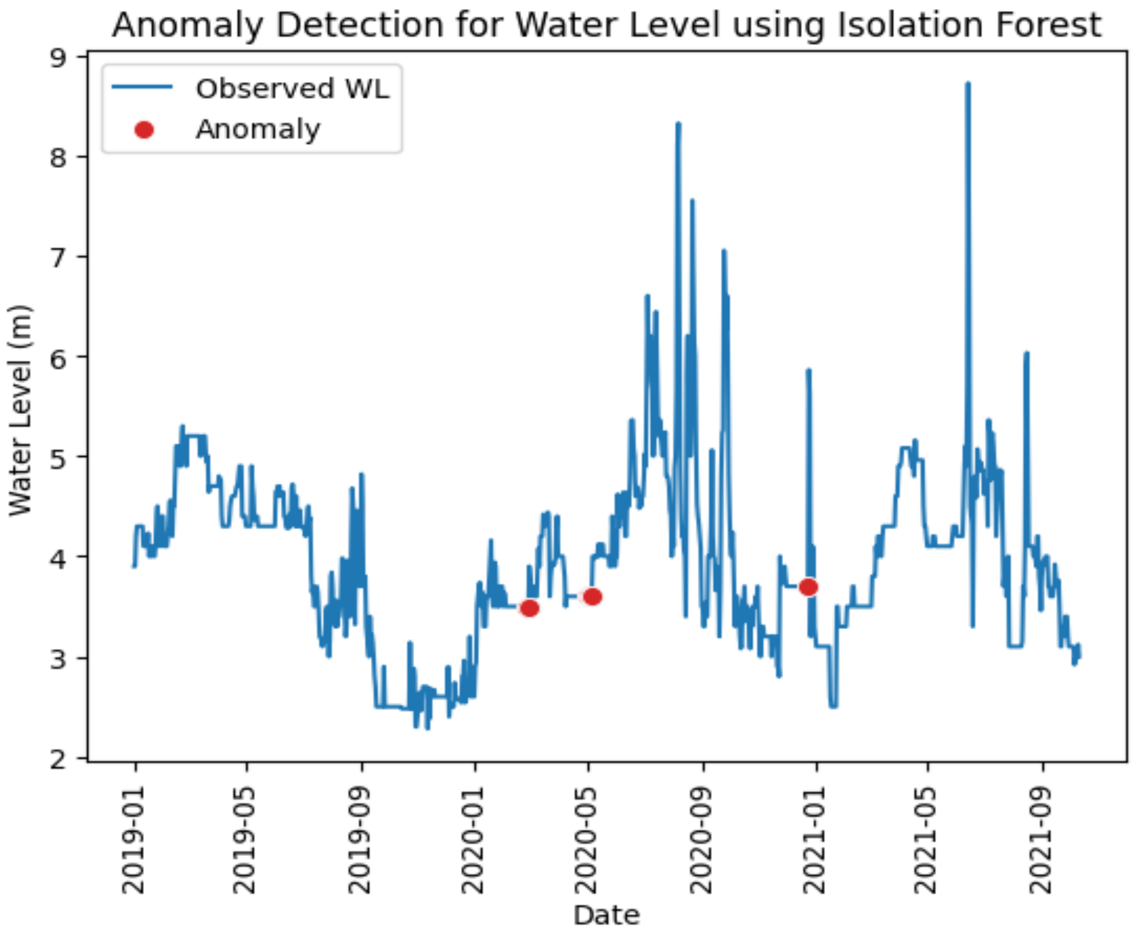


Upper Threshold=+67.5%

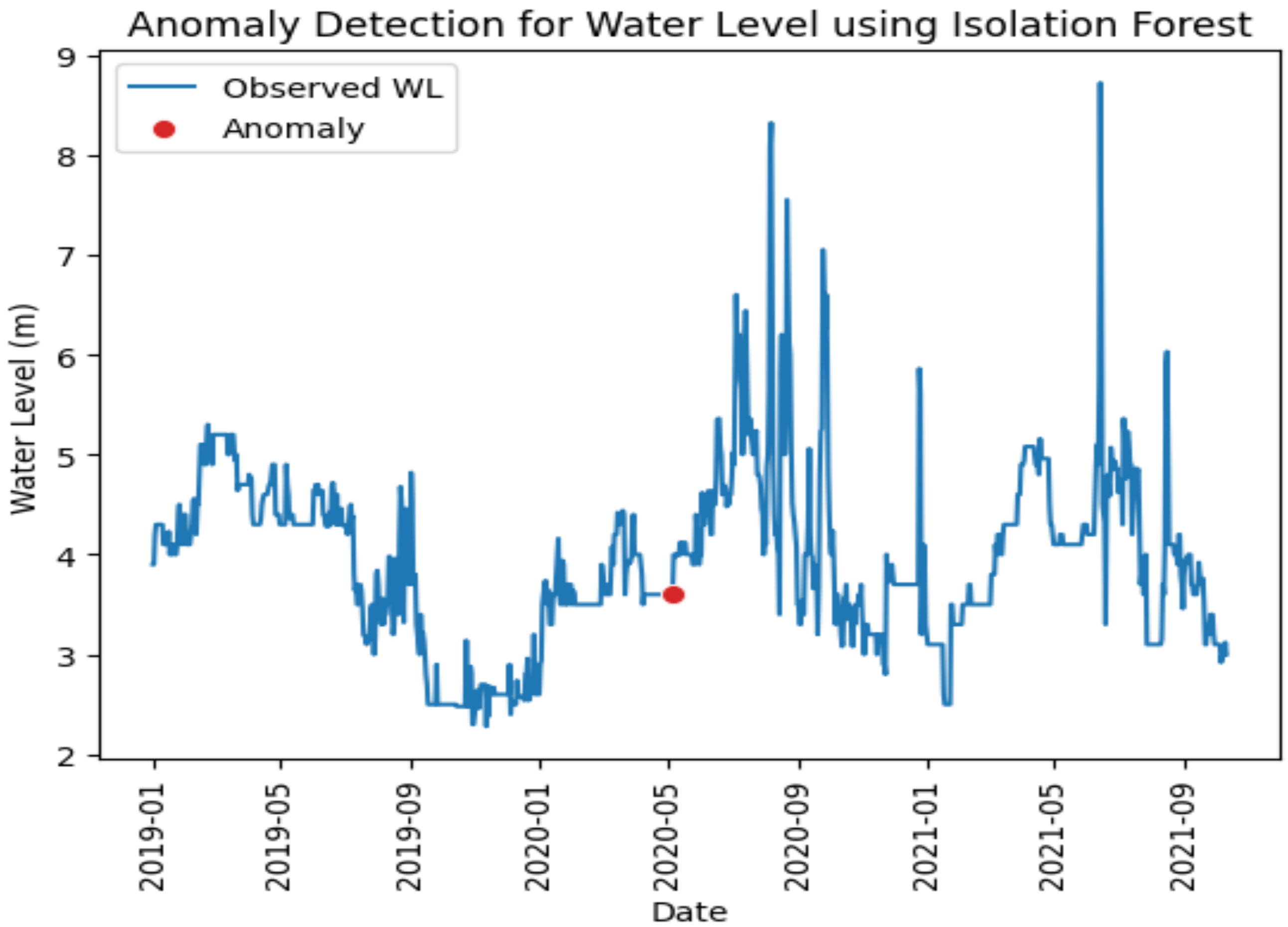


Lower Threshold=-40.9%

Constant water level for daily data in Lao PDR (Pakkayoung)



Contamination Level=0.01, Threshold=22 days

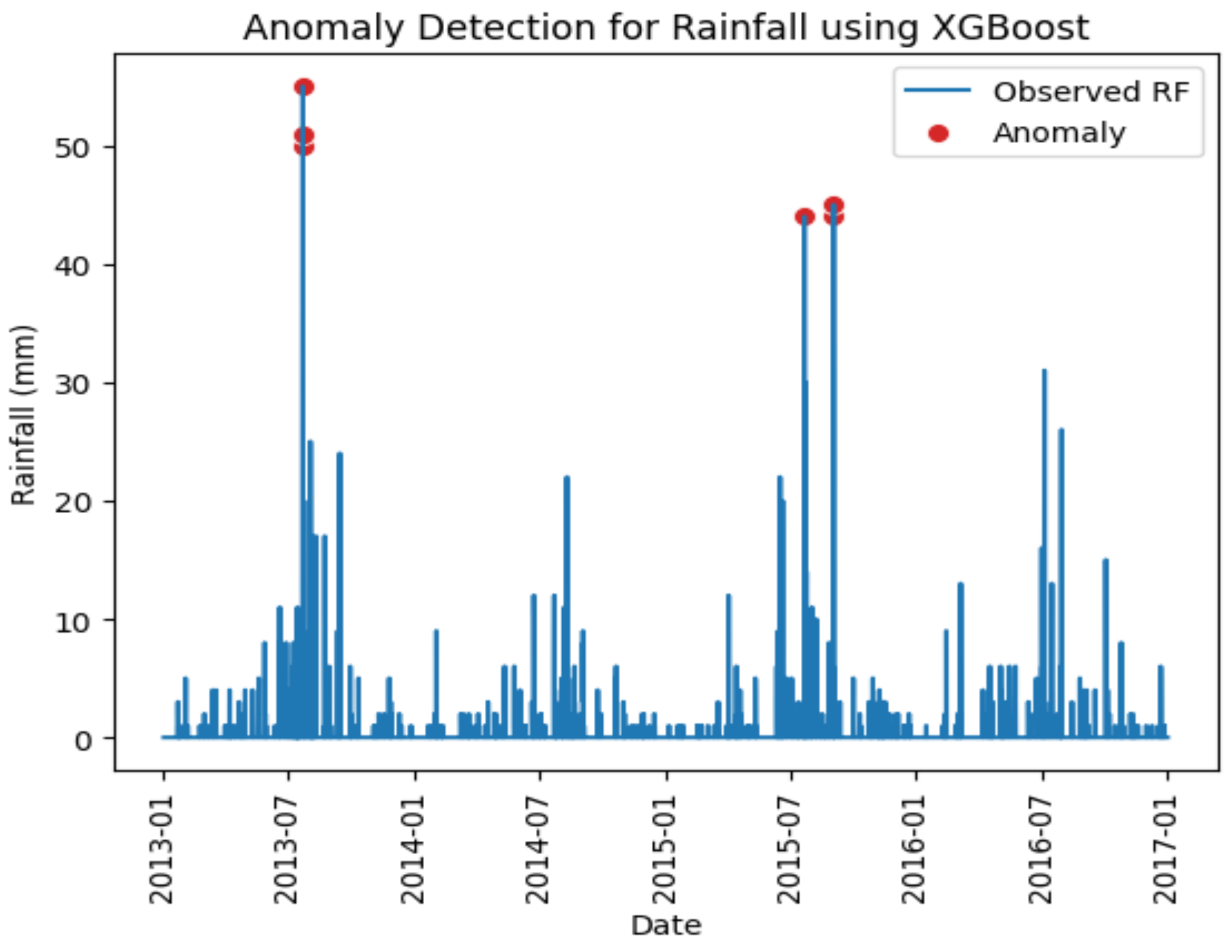


Contamination Level=0.001, Threshold=26 days

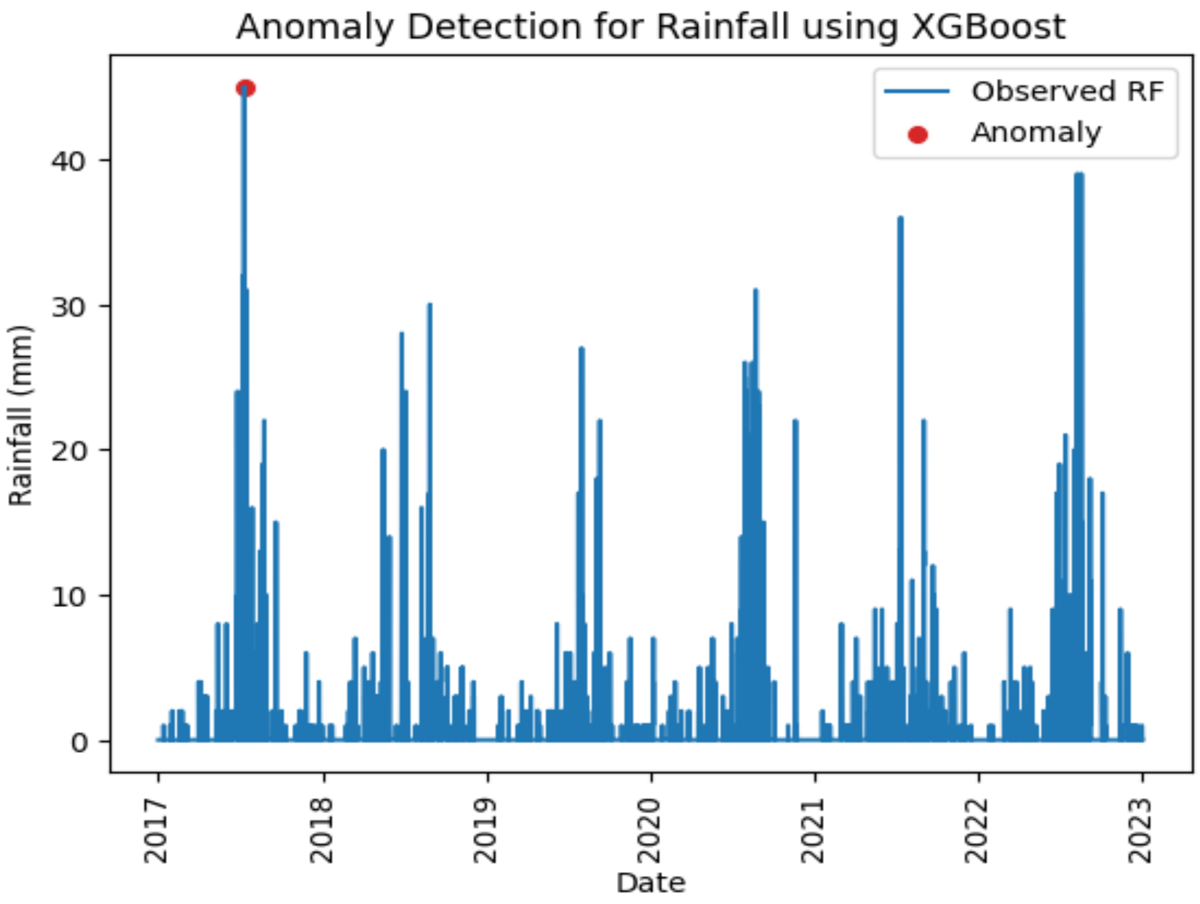
Using XGBoost for Identifying Abnormal Rainfall Data in ROK(Hourly Data) and Lao PDR (Daily Data)

Max. Rainfall for hourly data in ROK (Yeoju Bridge)
Different Percentage of Training Data (40%) and Testing Data (60%)
Dummy RF threshold >= 40mm (labelled data as abnormal)

Anomalies in Training (2013-2016)

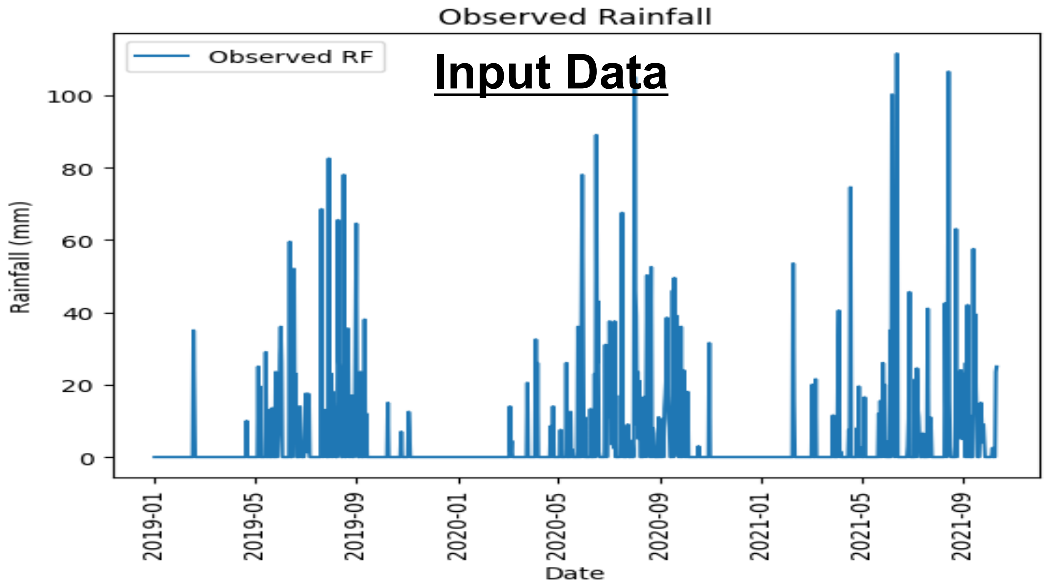


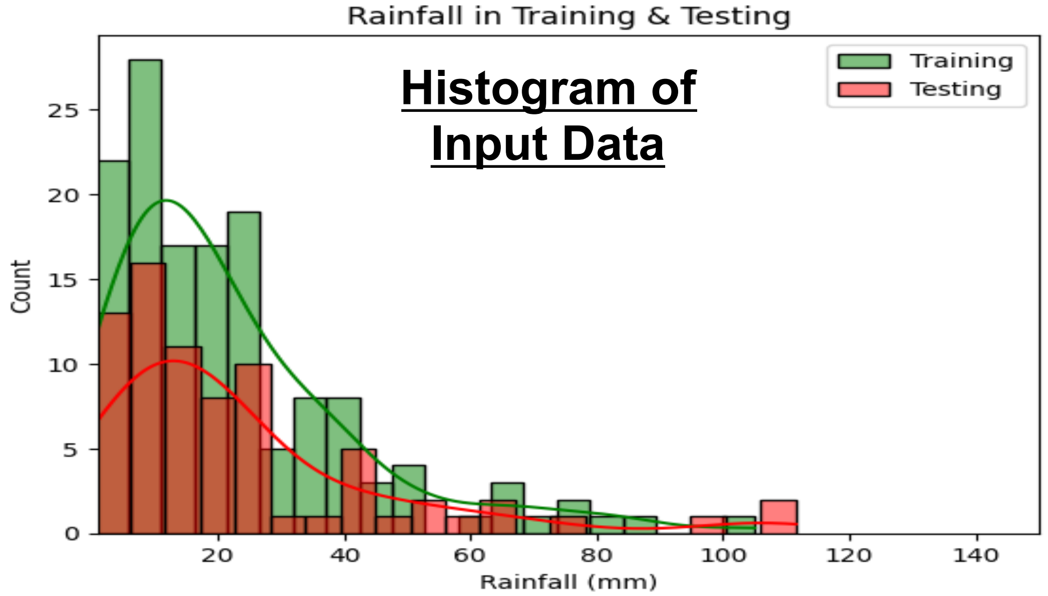
Anomalies in Testing (2017-2022)



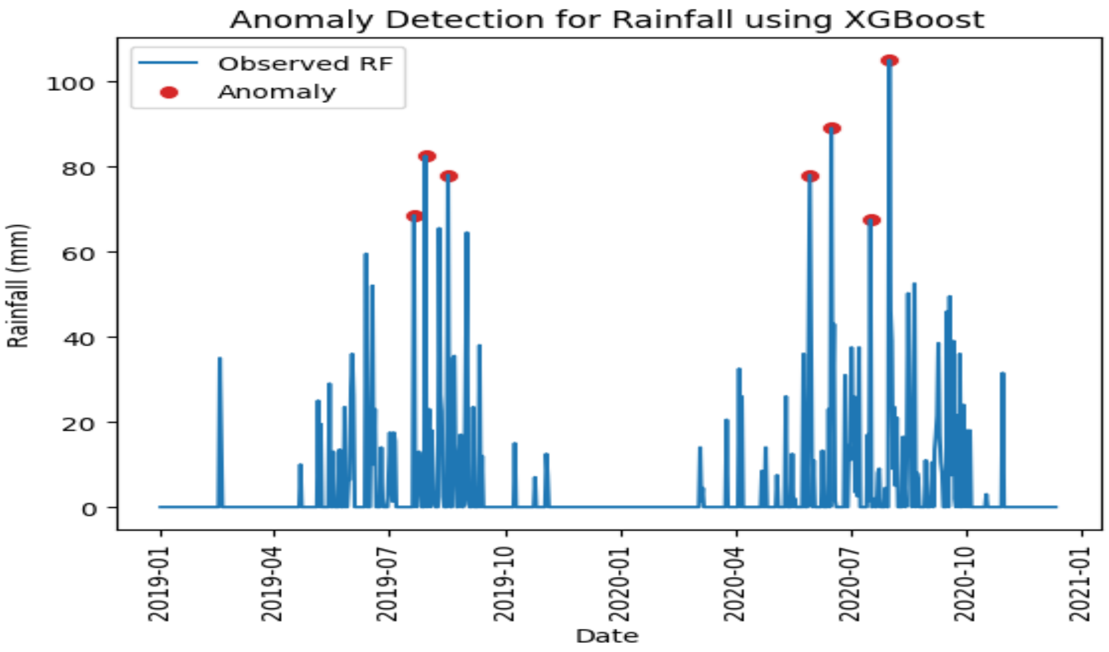
It is advised to specify longer period for training to allow the model to study and recognize the labelled data as much as possible for identifying the abnormal data in the testing period.

Max. Rainfall for daily data in Lao PDR (Pakkayoung Station)
(70% Training, 30% Testing)
Dummy Threshold >= 66mm (labelled data as abnormal)

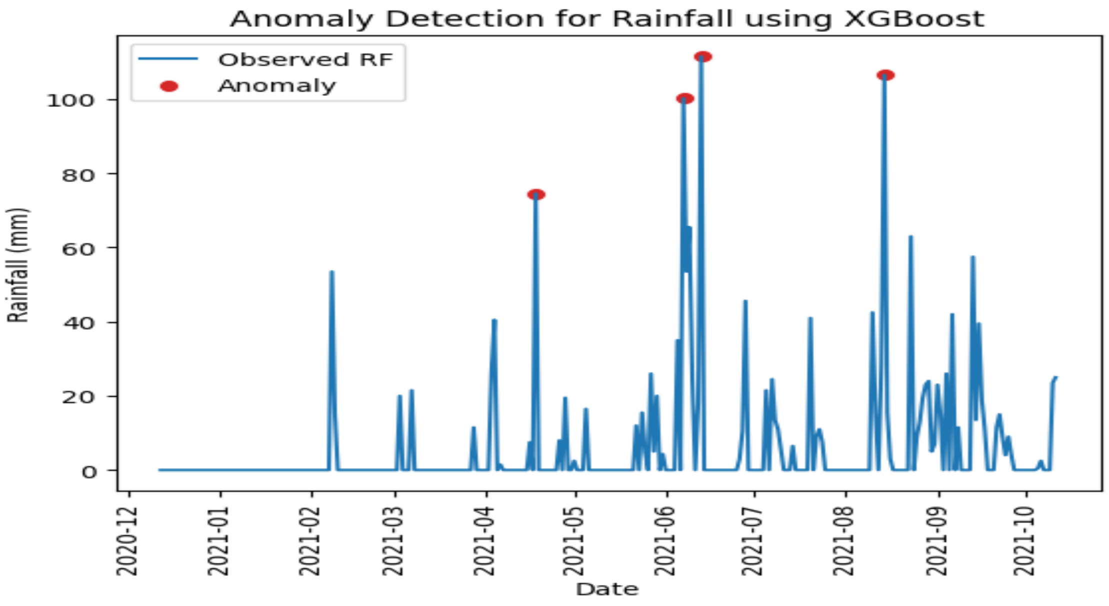




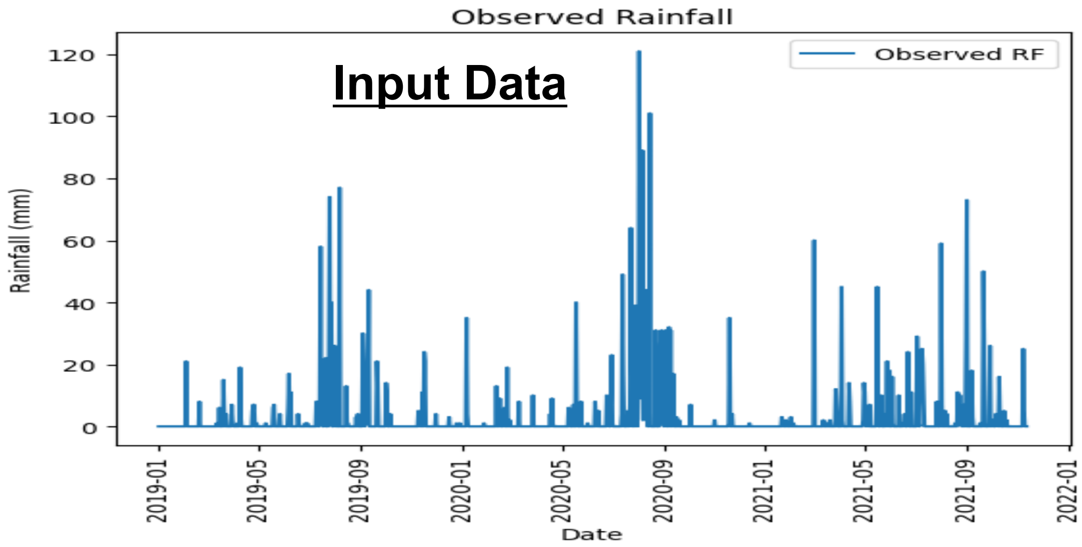
Anomalies in Training

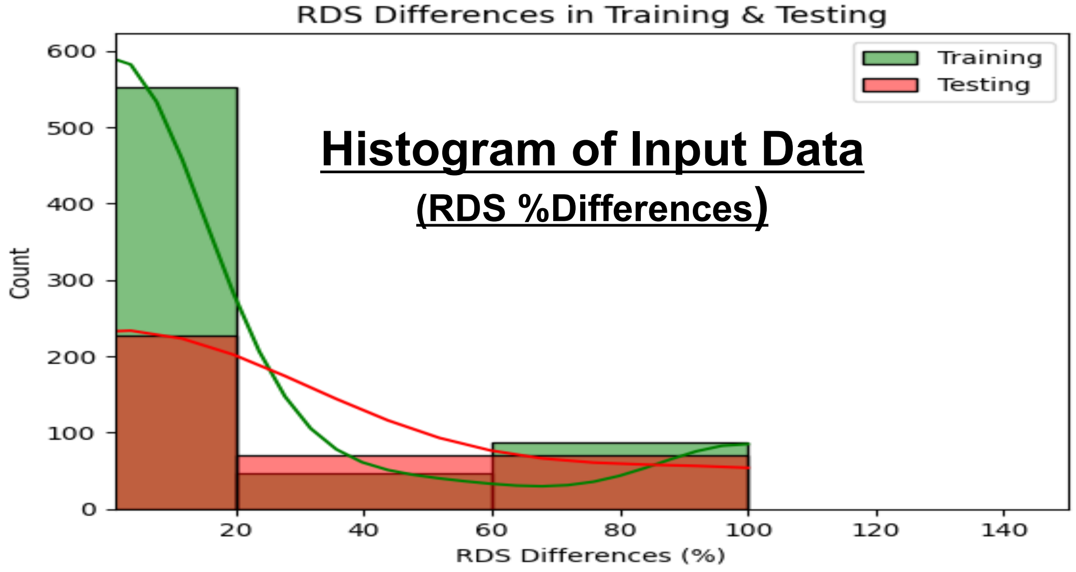


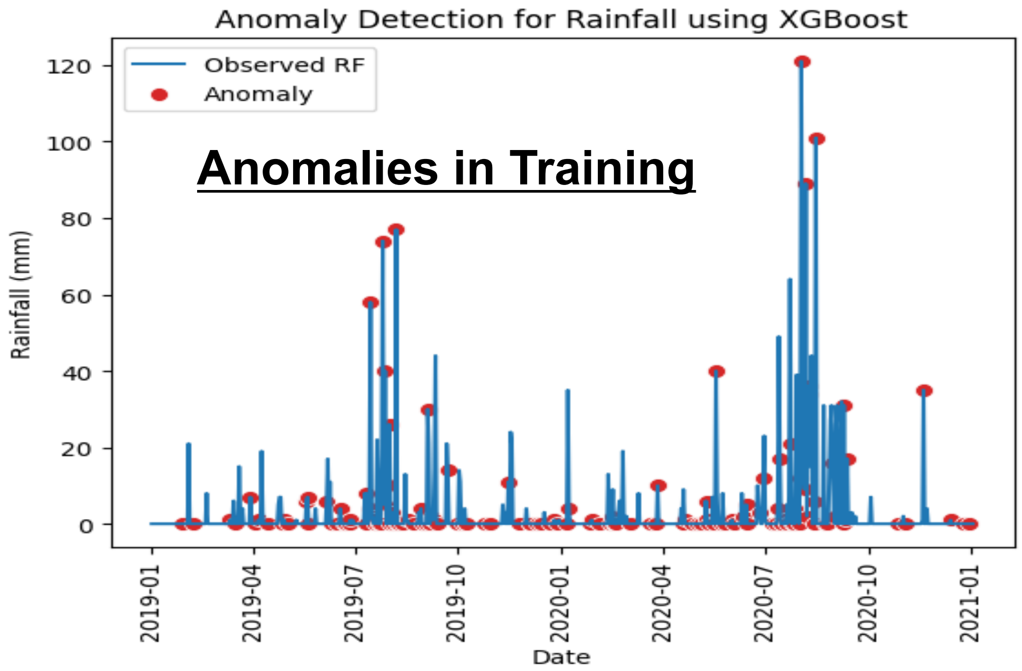
Anomalies in Testing

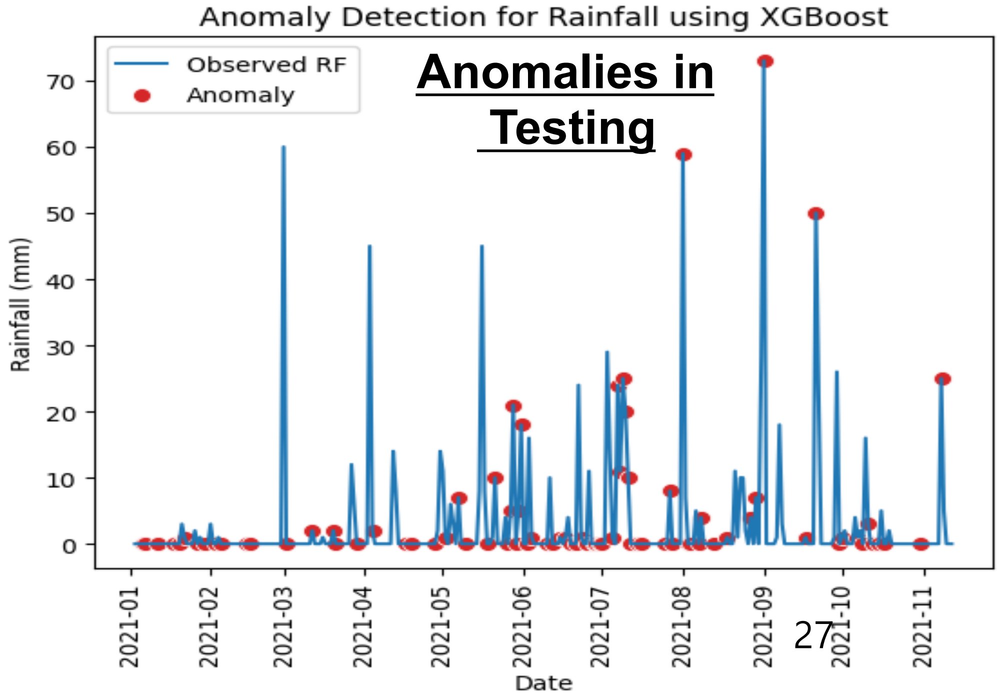


RDS for daily data in Lao PDR (Thalad Station)
(70% Training, 30% Testing)
Dummy Upper Limit >= 50% Differences (labelled data as abnormal)
Dummy Lower Limit <= -10% Differences (labelled data as abnormal)



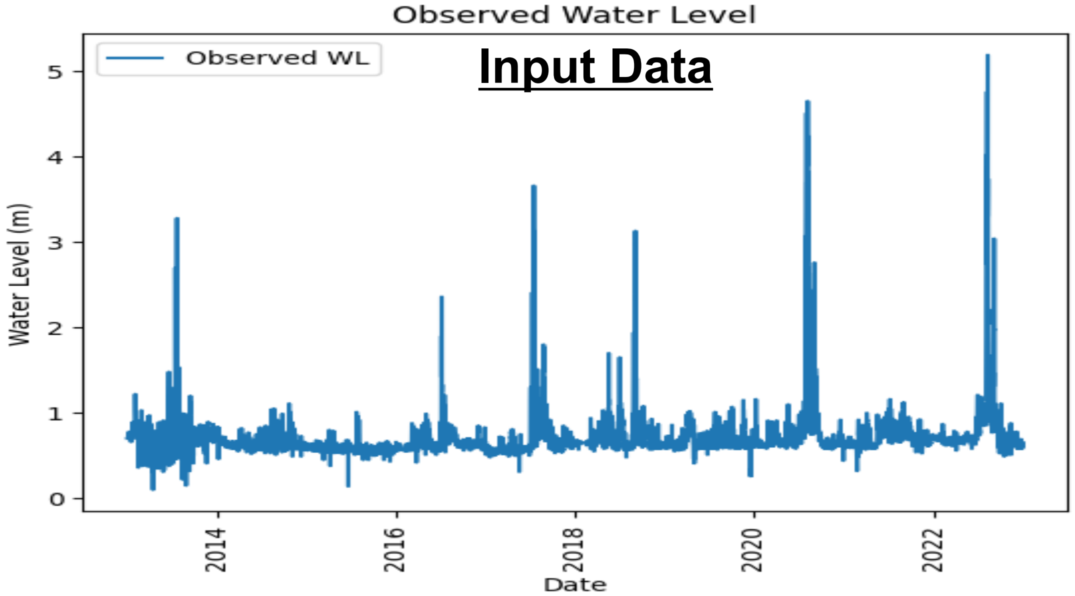


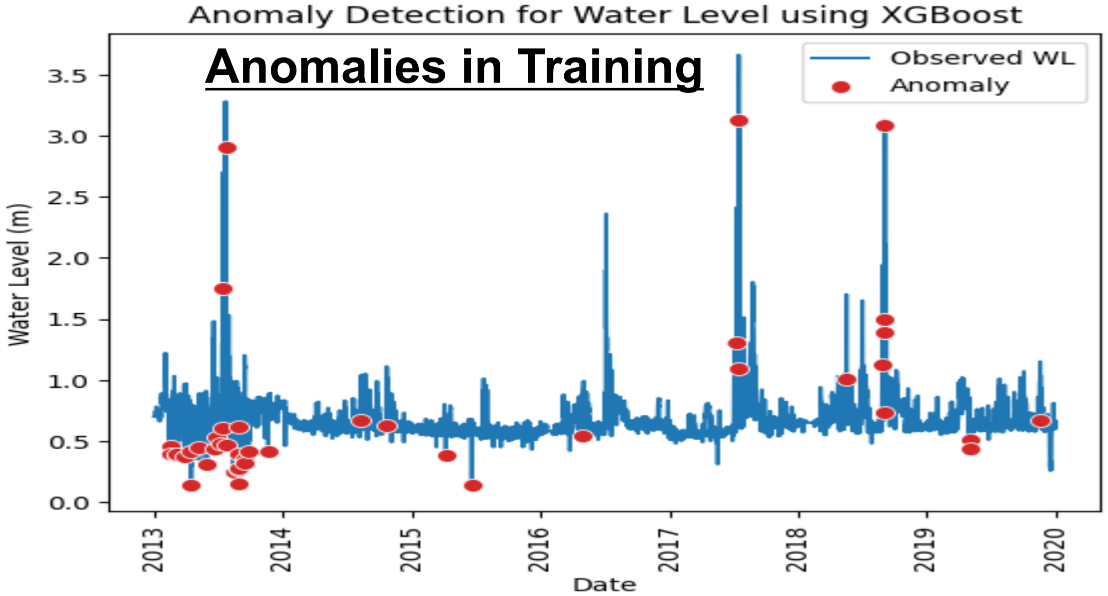


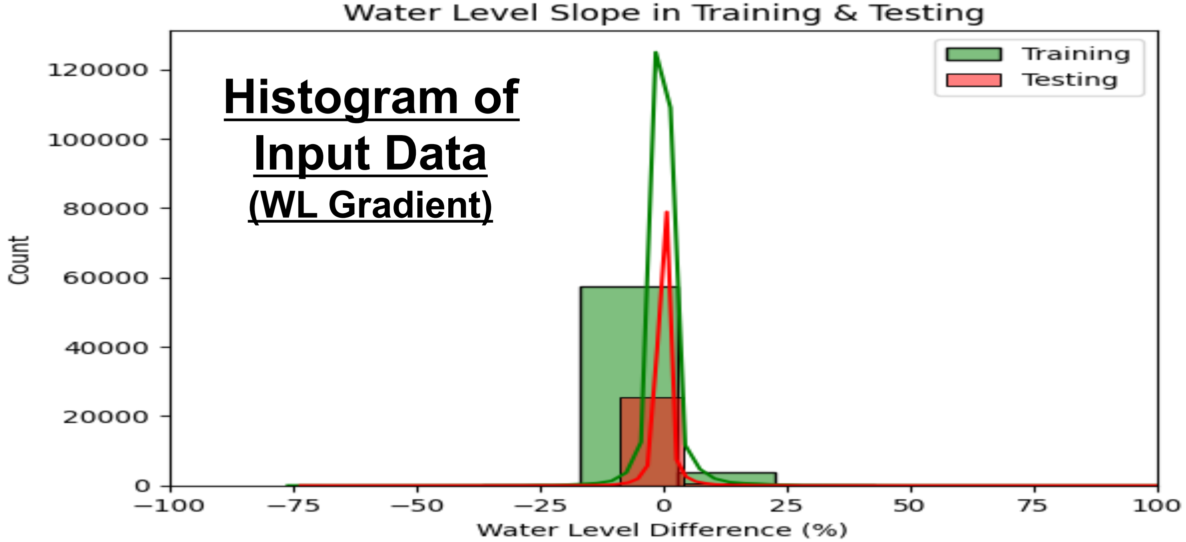


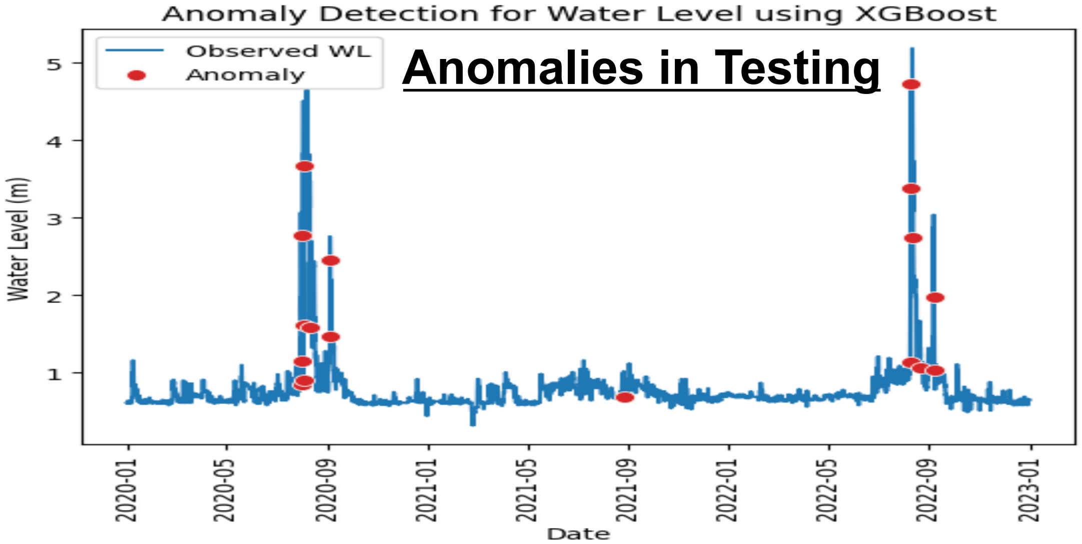
Using XGBoost for Identifying Abnormal Water Level Data in ROK(Hourly Data) and Lao PDR (Daily Data)

Water Level Gradient (*between current and previous one hour data*) **for hourly data in ROK (Yeoju Bridge)**
(70% Training, 30% Testing)
Dummy WL Gradient Threshold, Upper Limit >= 100% Differences (labelled data as abnormal)
Dummy WL Gradient Threshold, Lower Limit <= -30% Differences (labelled data as abnormal)



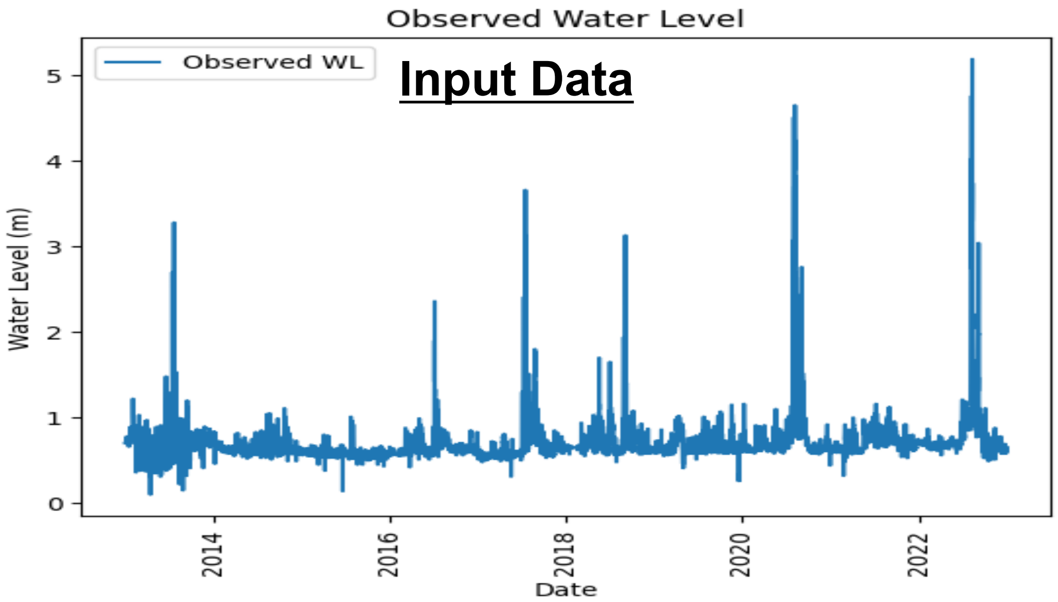




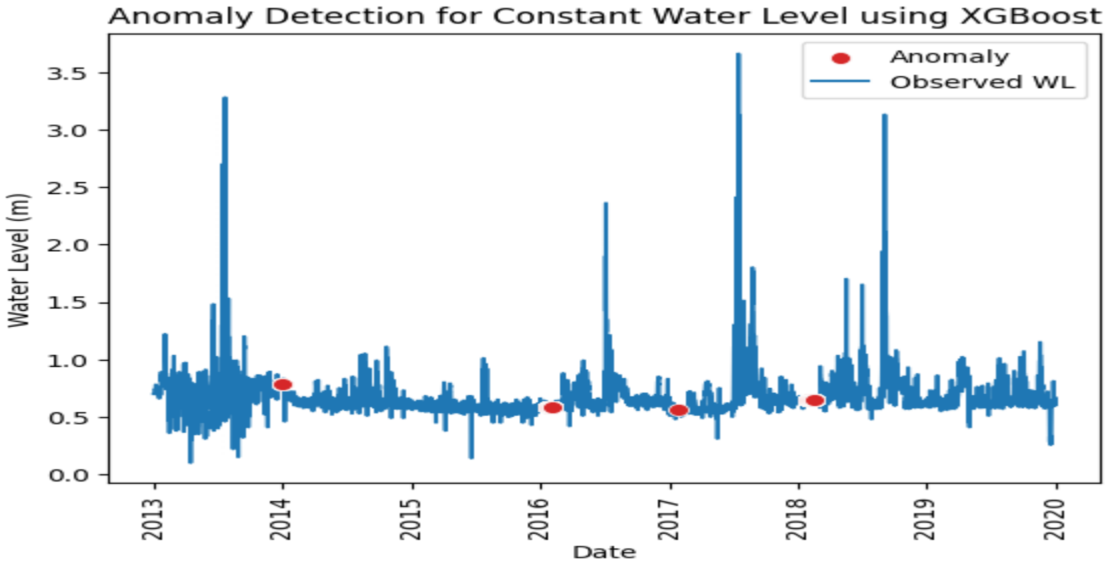


Constant water level for hourly data in ROK (Yeoju Bridge)
(70% Training, 30% Testing)
Dummy constant WL Threshold >= 102 hours (labelled data as abnormal)

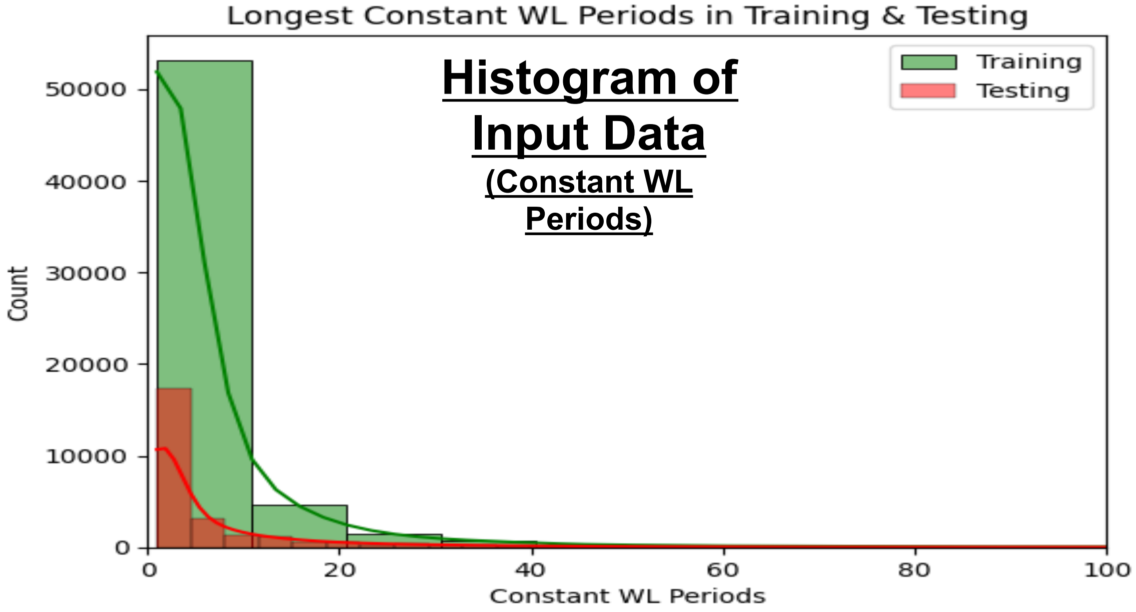
Input Data



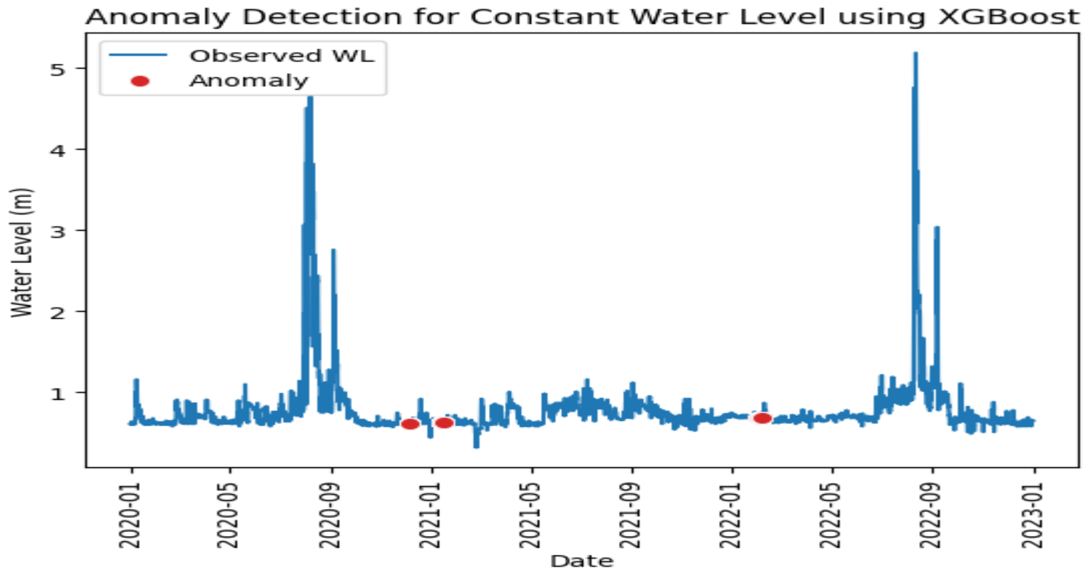
Anomalies in Training



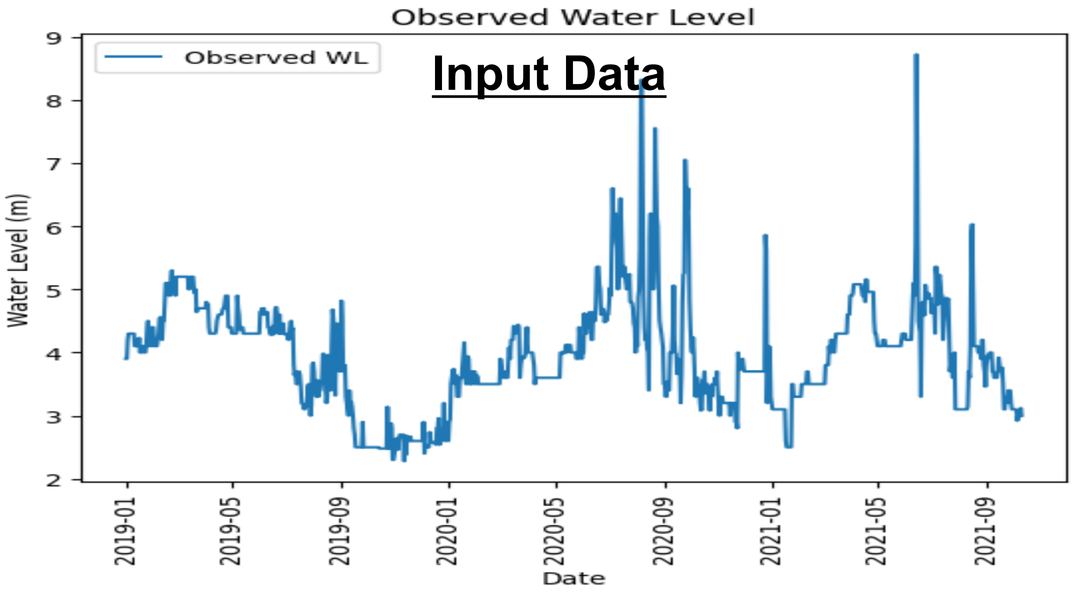
Histogram of Input Data (Constant WL Periods)

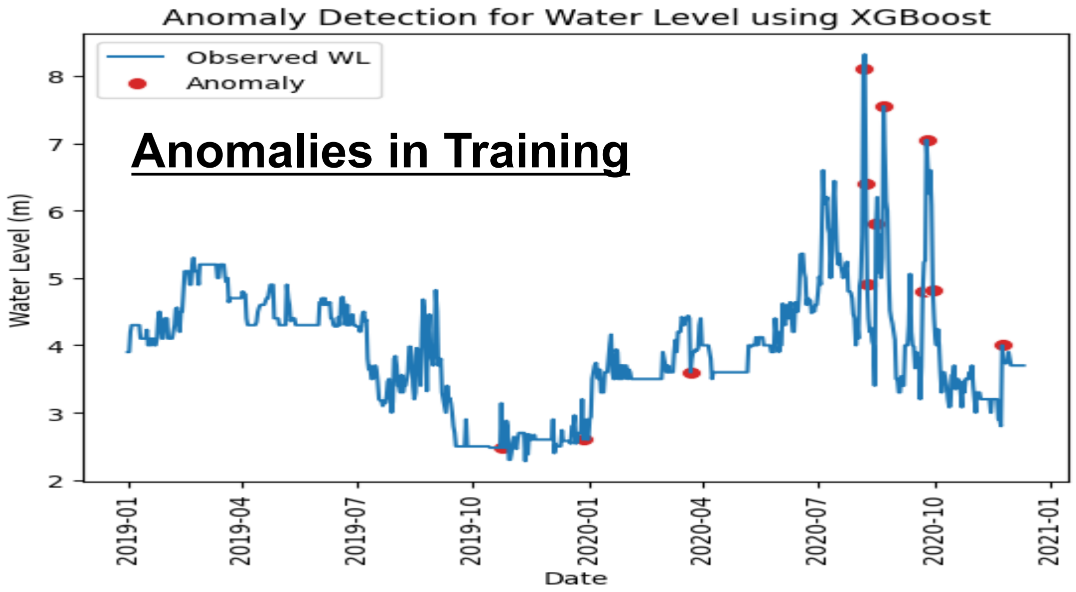


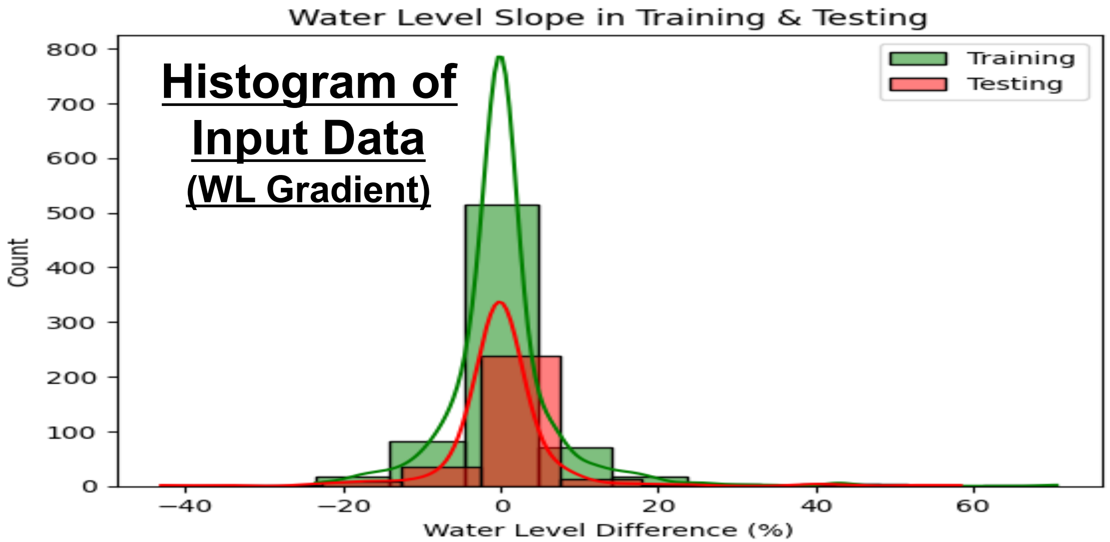
Anomalies in Testing

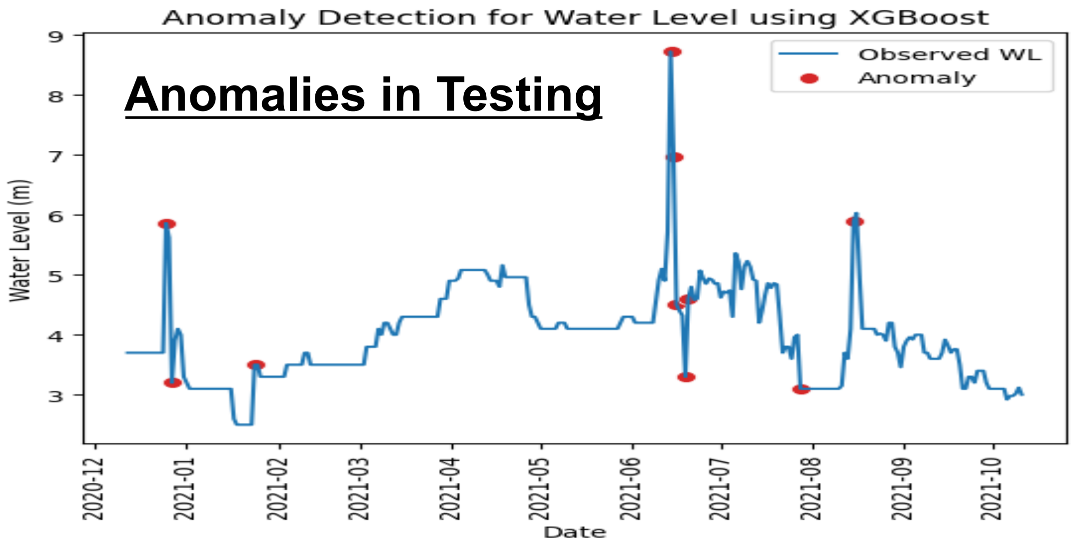


Water Level Gradient (*between current and previous one day data*) **for daily data in Lao PDR (Pakkayoung Station)**
(70% Training, 30% Testing)
Dummy WL Gradient Threshold >= 99 percentile (labelled data as abnormal)

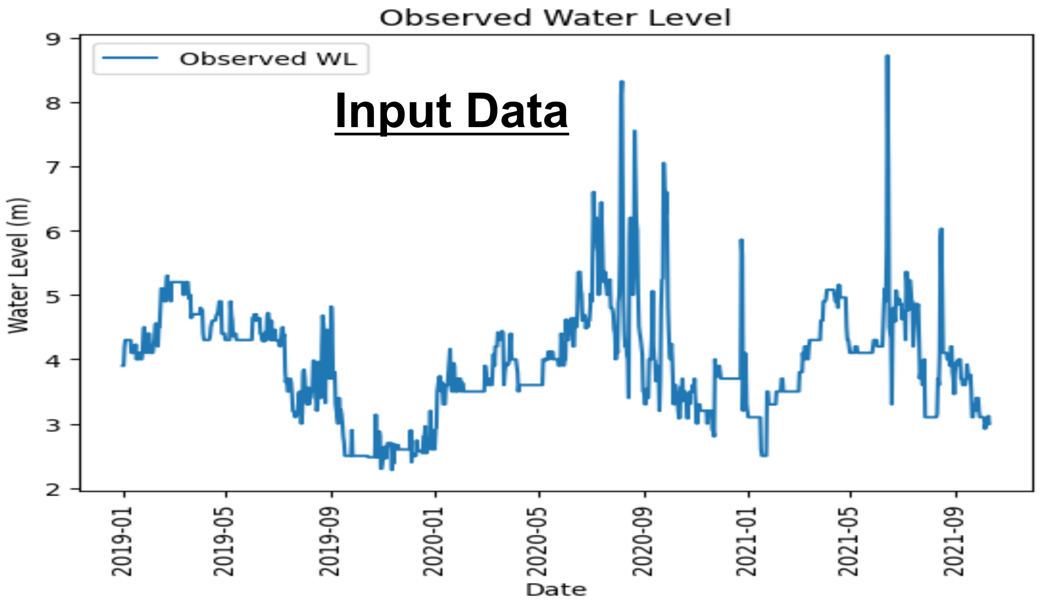


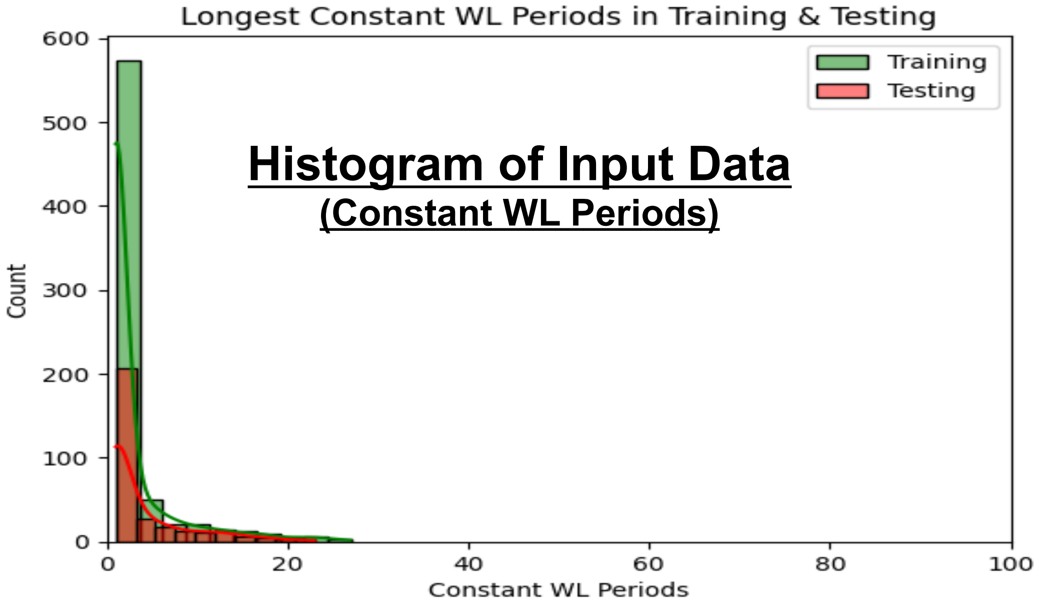




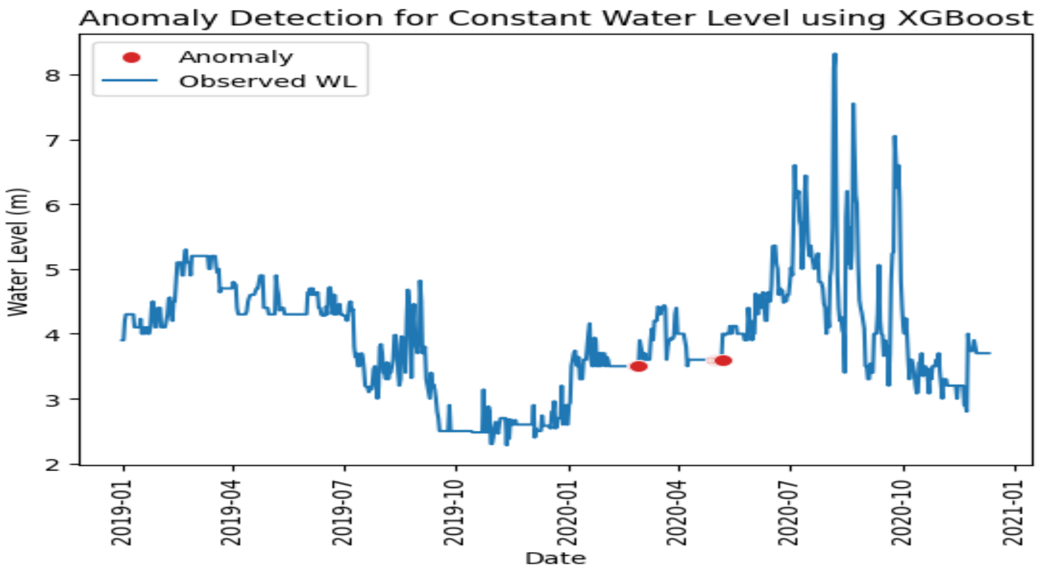


Constant water level for daily data in Lao PDR (Pakkayoung Station)
(70% Training, 30% Testing)
Dummy constant WL Threshold >= 22 days (labelled data as abnormal)

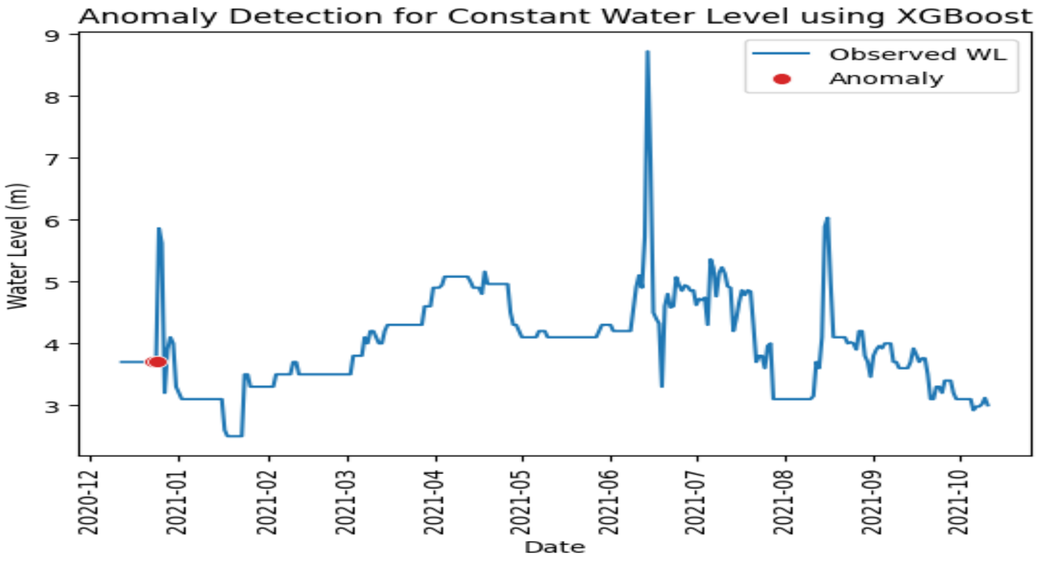




Anomalies in Training

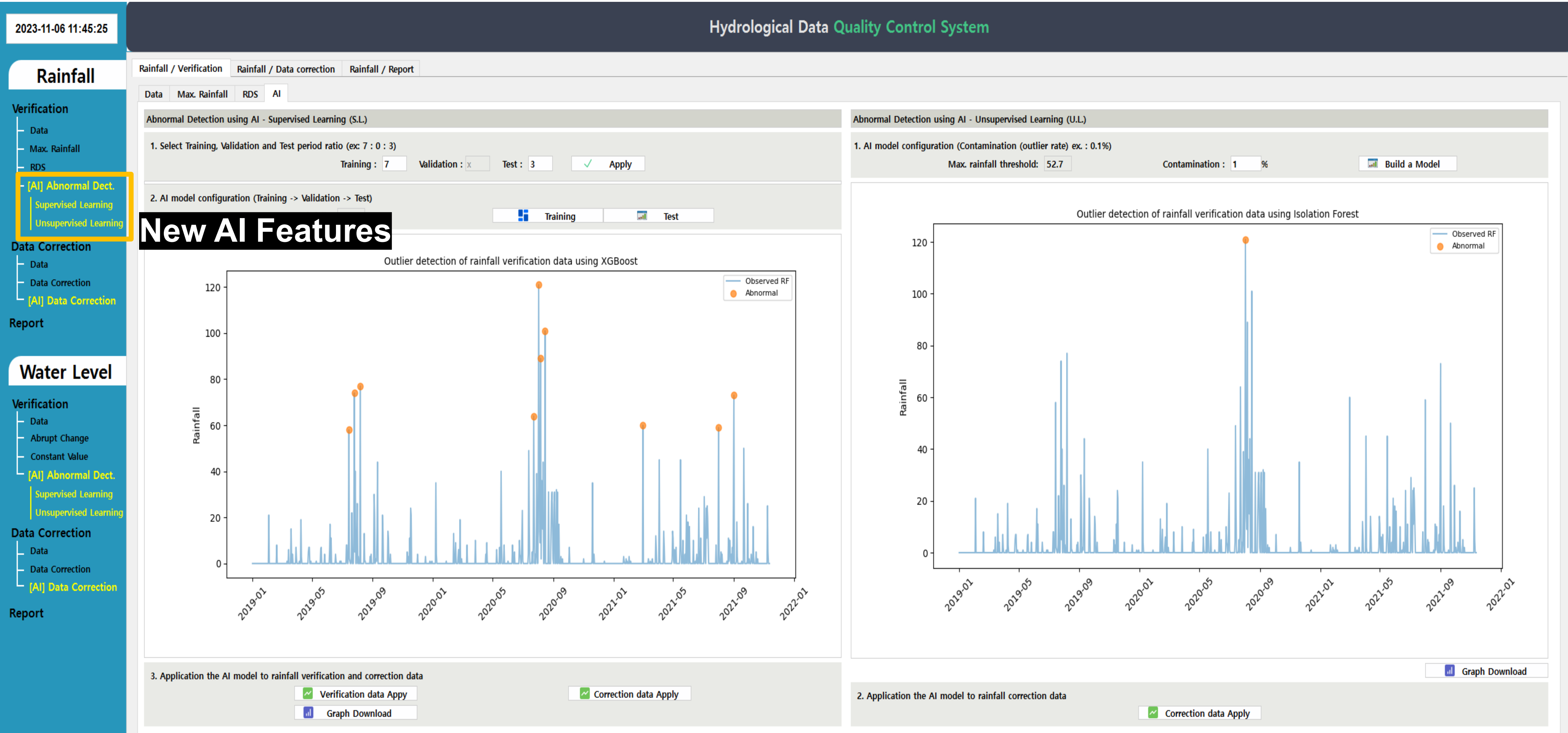


Anomalies in Testing



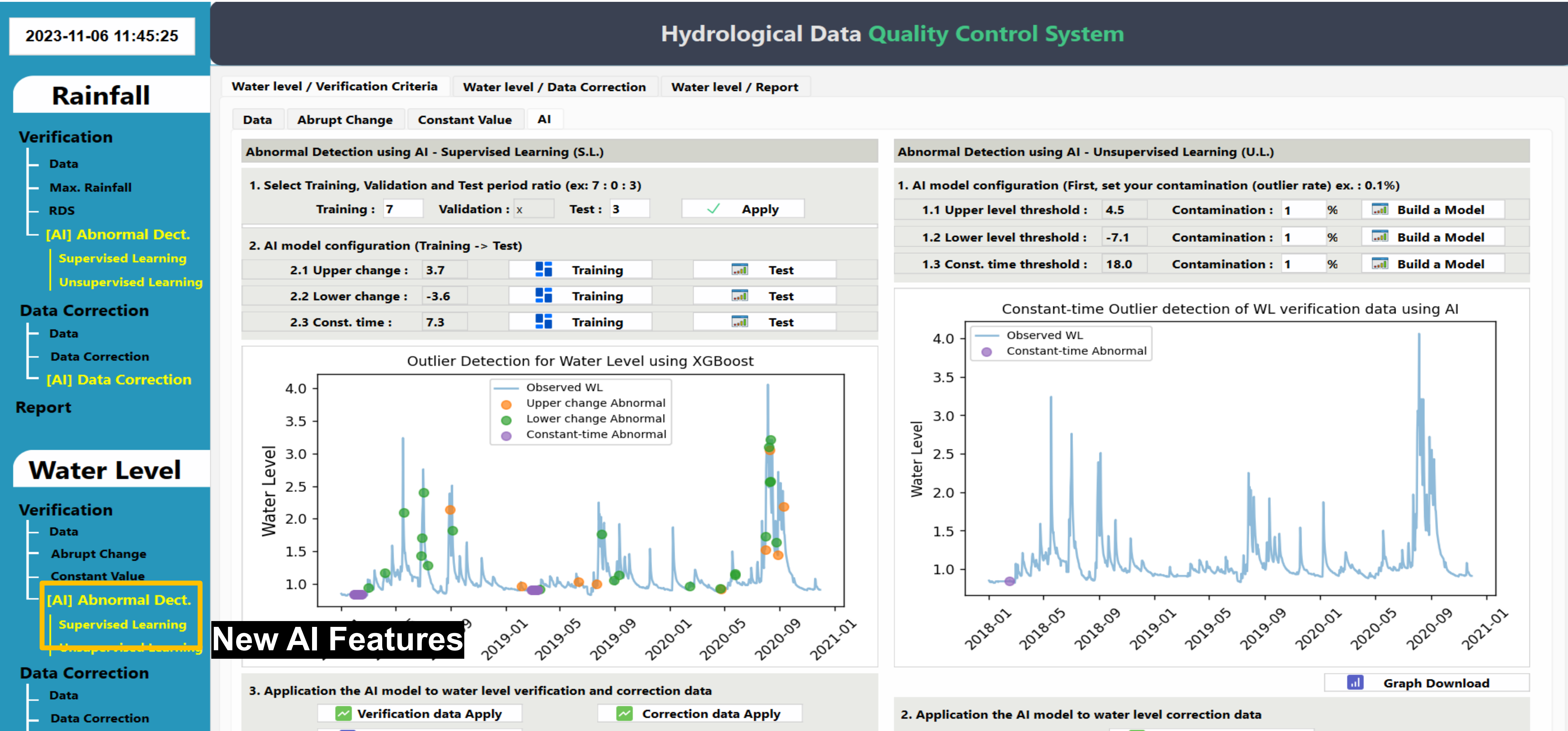
Hydrological Data Quality Control System (HDQCS)

Rainfall Data Criteria (AI)



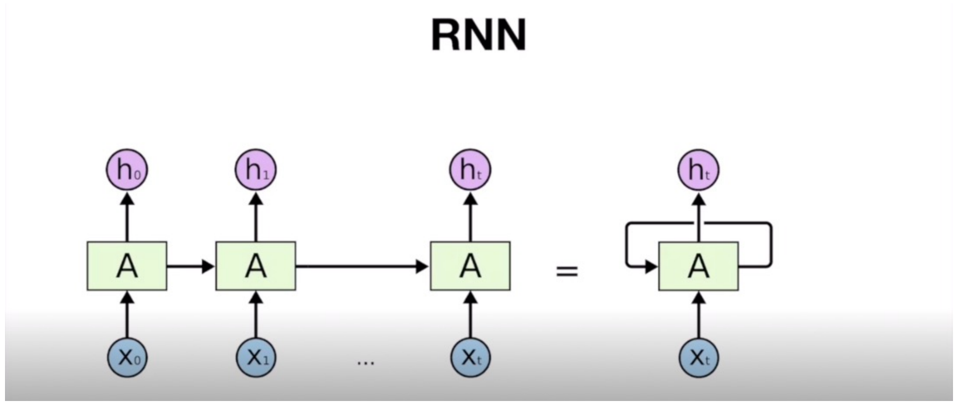
Hydrological Data Quality Control System (HDQCS)

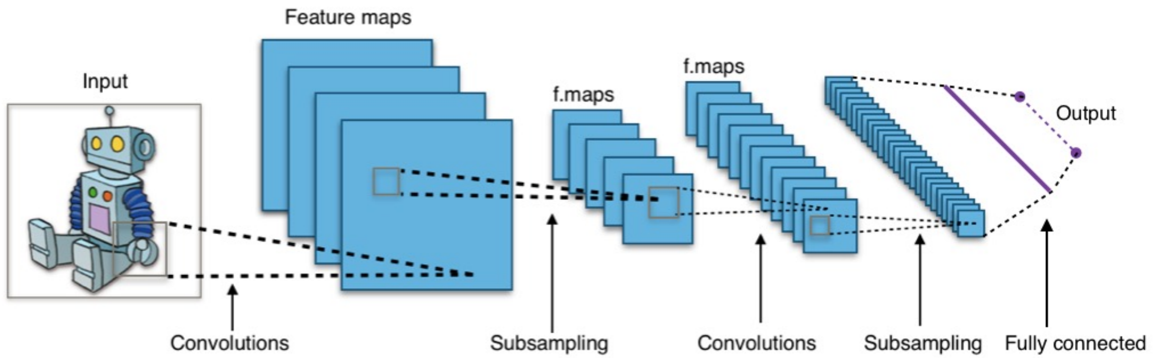
Water Level Data Criteria(AI)

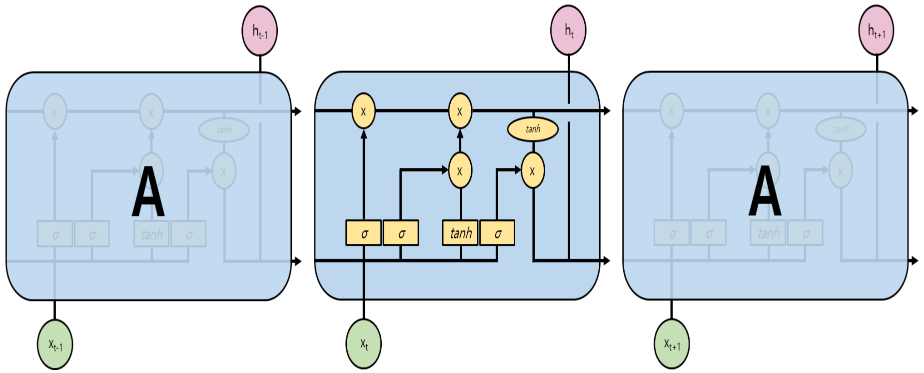


Flood Forecasting
using AI Deep-Learning Techniques

AI Deep-Learning Techniques for Water level prediction







– Recurrent Neural Networks (RNN)

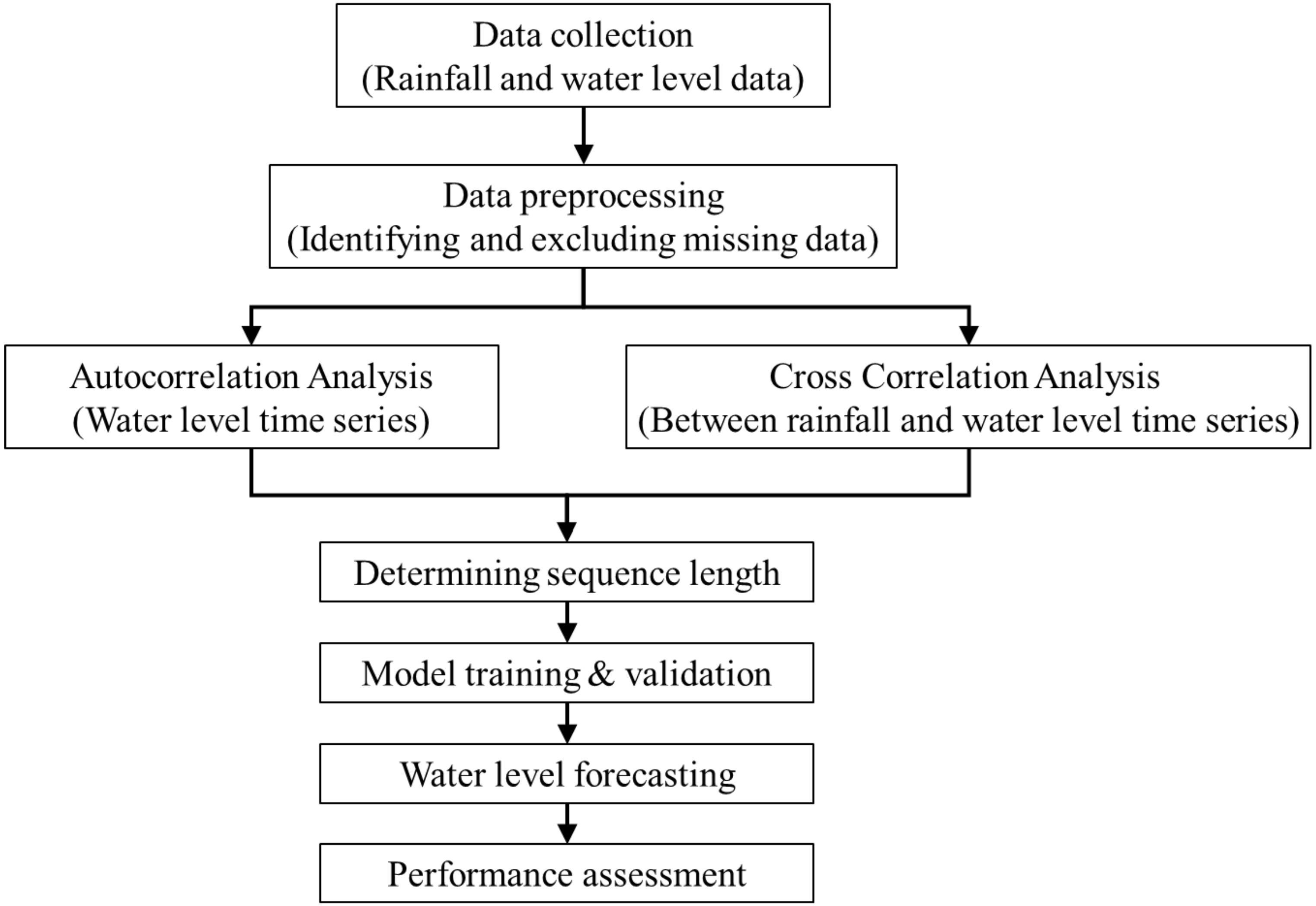
– Convolutional Neural Networks (CNN)

– Long Short-Term Memory (LSTM)

❖ Advantage of using Deep-Learning Techniques

- Adaptability
- Patten recognition capability
- Automatic learning

Water level prediction Procedure using AI Deep-Learning Techniques



Hyperparameter tuning

1) Number of Nodes

(rule of thumb: 1 HL for simple problem and 2HL for complex problem; high accuracy with many nodes in a layer)

2) Number of hidden layers

(rule of thumb: 1-3 units or nodes per layer is a good base)

3) Dropout

(rule of thumb: The 20% (0.2) is widely accepted as the best compromise between overfitting and retaining model accuracy)

4) Input sequence length

5) Learning rate

(rule of thumb: Usually a decaying learning rate is preferred and this hyper-parameter is used in the training phase and has a small positive value, mostly between 0.0 and 0.1.)

6) Batch Size

(rule of thumb: Widely accepted, a good default value for batch size is 32. For experimentation, you can try multiples of 32, such as 64, 128 and 256.)

7) Input dimensionality

8) Number of Epochs

(rule of thumb: Widely accepted, a good default value for batch size is 32. For experimentation, you can try multiples of 32, such as 64, 128 and 256.)

Training evaluation

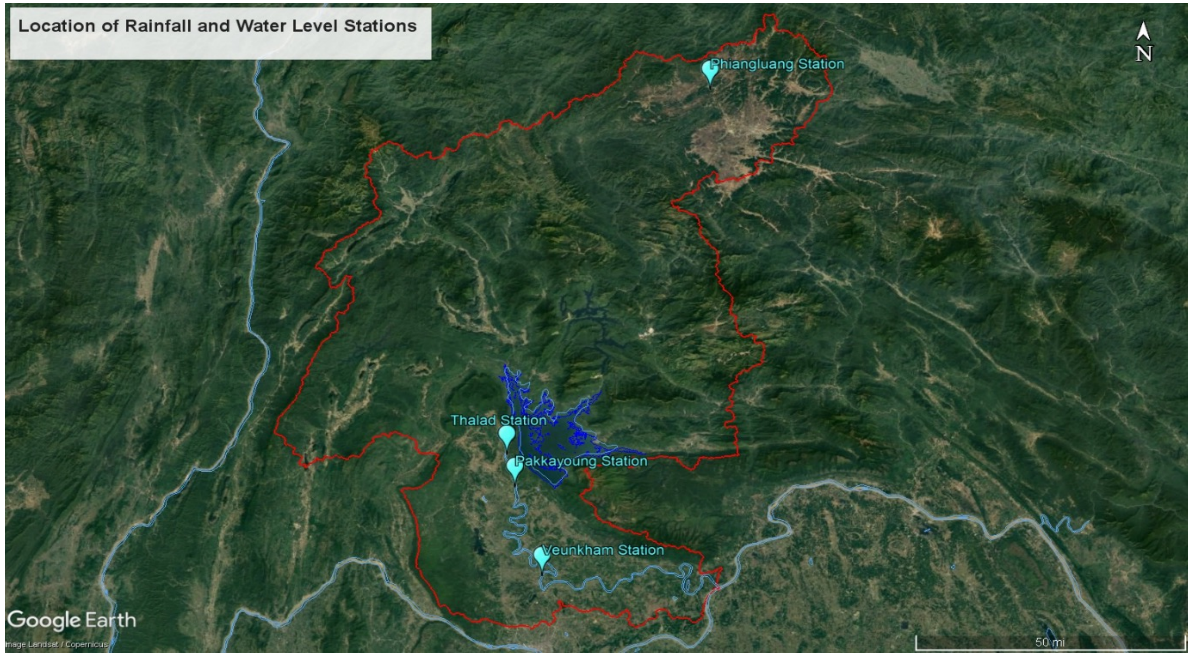
- Correlation coefficient → $r_{OP} = \frac{\sum_{i=1}^{n}(O_i - O^{'})(P_i - P^{'})}{\sqrt{\sum_{i=1}^{n}(O_i - O^{'})^2}\sqrt{\sum_{i=1}^{n}(P_i - P^{'})^2}}$

- RMSE → $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{n}}$

- R^2 → $R^2 = \frac{(\sum_{i=1}^{n}(O_i - O^{'})(P_i - P^{'}))^2}{\sum_{i=1}^{n}(O_i - O^{'})^2 \sum_{i=1}^{n}(P_i - P^{'})^2}$

- MAE → $MAE = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - O^{'})^2}$

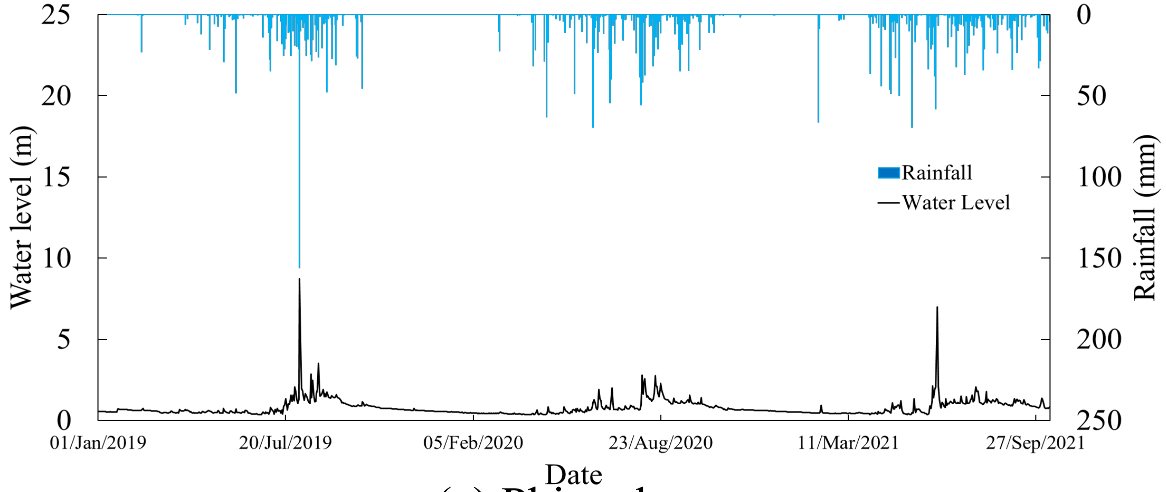
Application Case



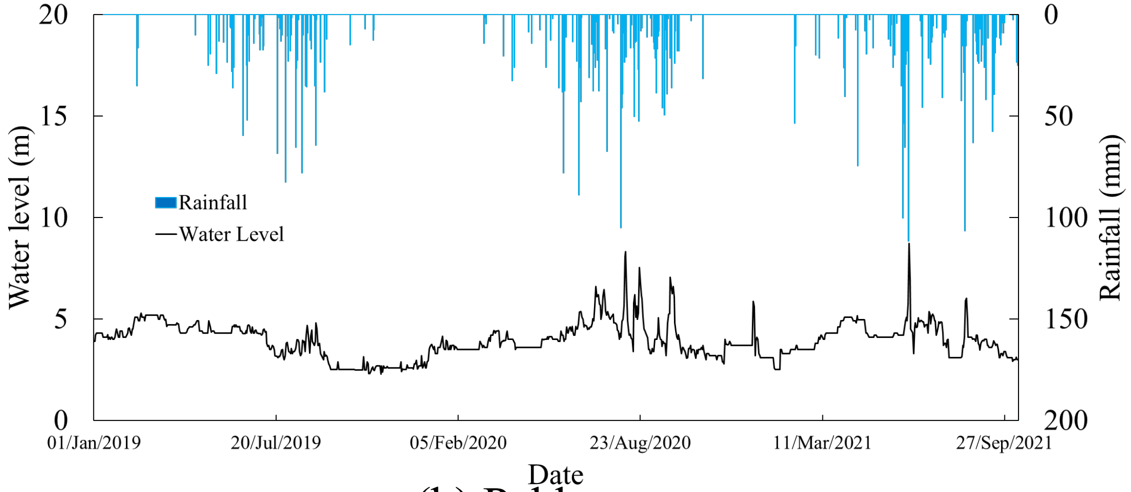
No.	Station	Latitude	Longitude
1	Phiangluang	19°34′06″ N	103°04′17″ E
2	Thalad	18°31′26″ N	102°30′54″ E
3	Pakkayoung	18°25′53″ N	102°32′16″ E
4	Veunkham	18°10′37″ N	102°36′53″ E

	Rainfall (mm)			**Water Level (m)**			**Rainfall (RF)**	**Water Level (WL)**
Station	Annual Average	Max.	Average	Max.	Average	Min		
Phiangluang	1283	155.8	3.8	8.72	0.79	0.35	**2019.01.01 ~2021.10.14**	2019.01.01~2021.10.14
Thalad	1515	118.5	4.5	10.95	6.09	0.64	**2019.01.01~2021.10.11**	2019.01.01~2021.10.11
Pakkayoung	1629	111.5	4.8	8.72	3.95	2.28	**2019.01.01~2021.10.11**	2019.01.01~2021.10.11
Veunkham	1651	142.8	4.9	9.10	2.71	0	**2019.01.01~2021.10.10**	2019.01.01~2021.10.10

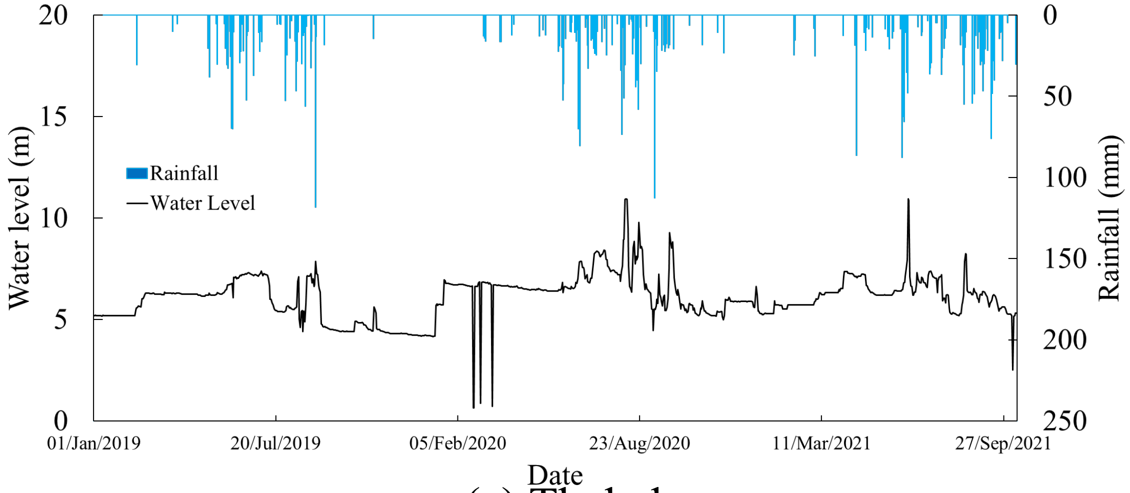
Application Case



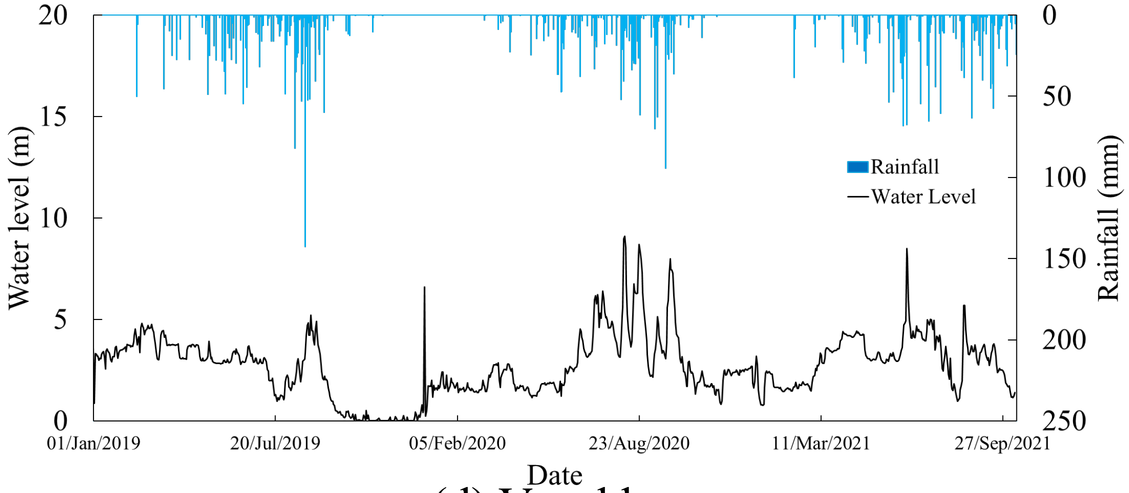
(a) Phiangluang



(b) Pakkayoung

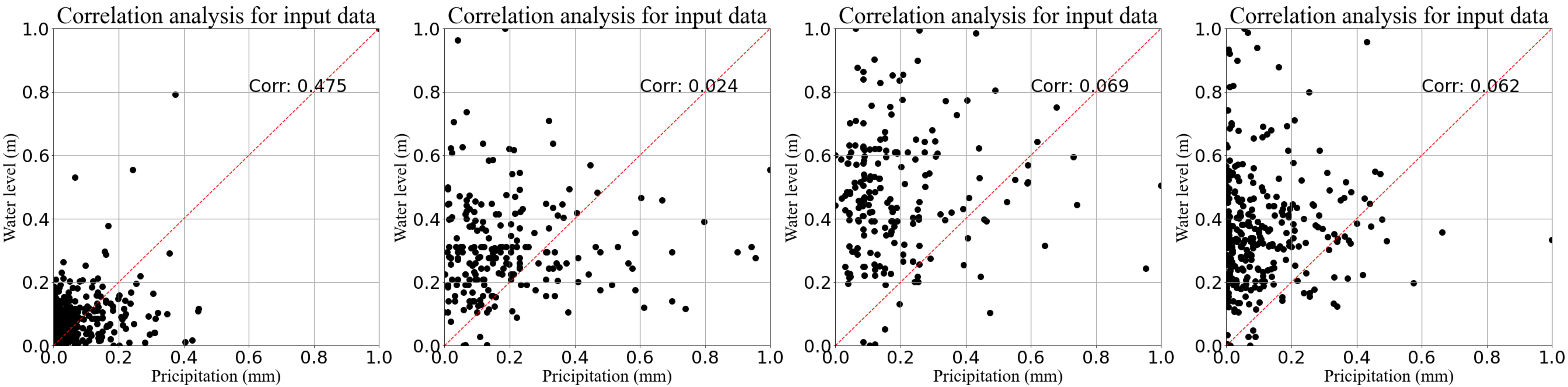


(c) Thalad

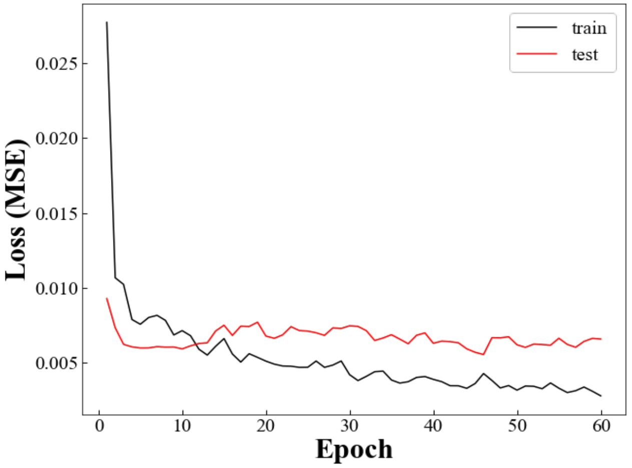


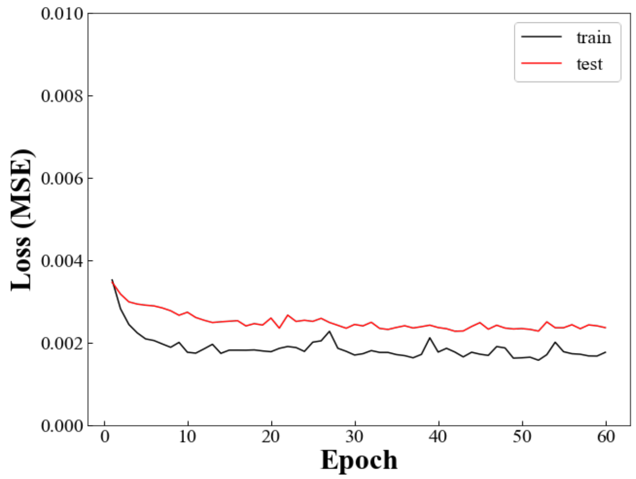
(d) Veunkham

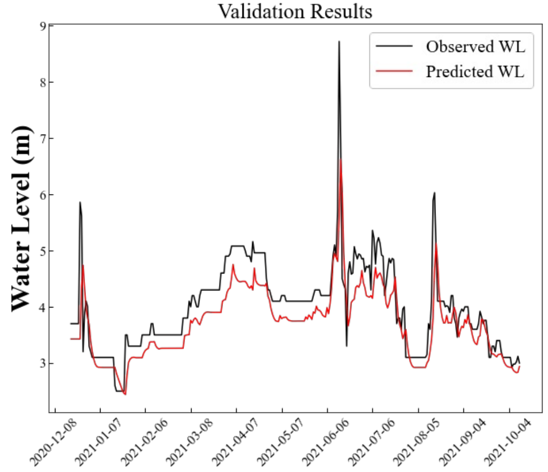
Correlation analysis

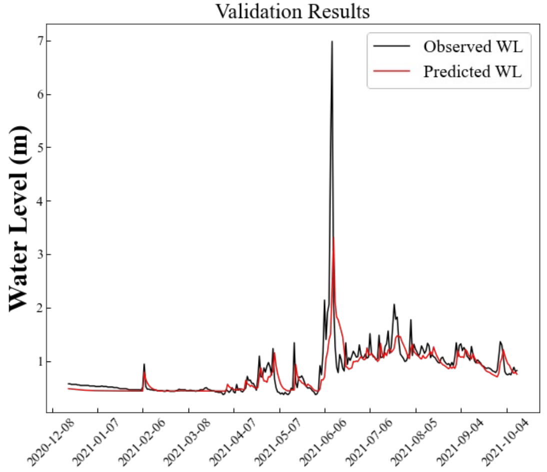


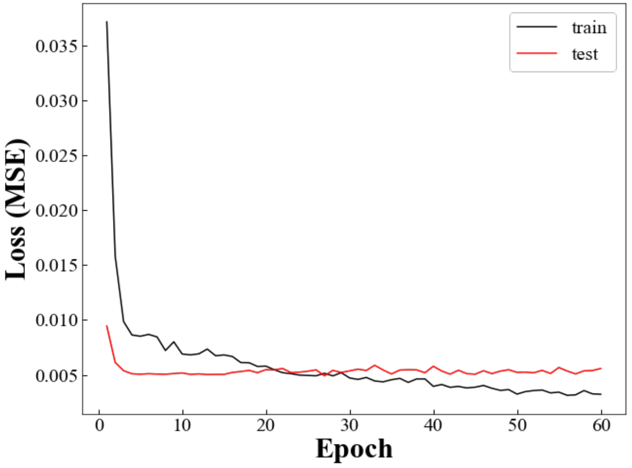
Model Training

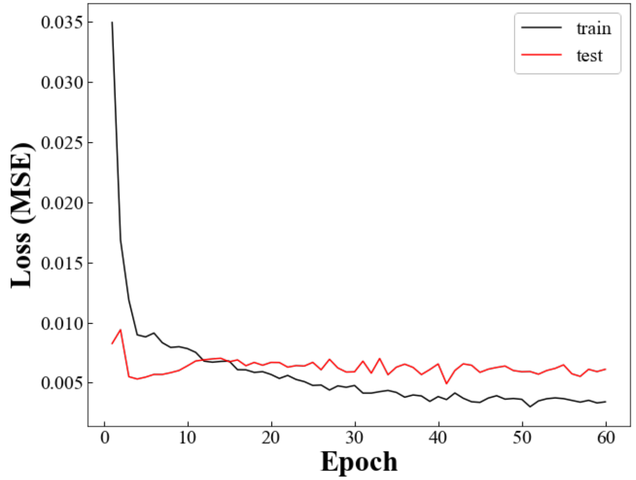


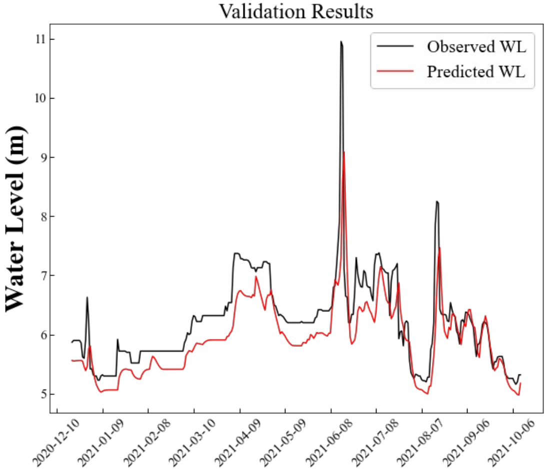


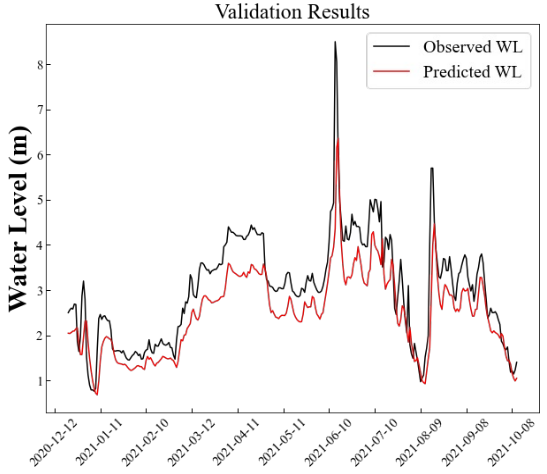




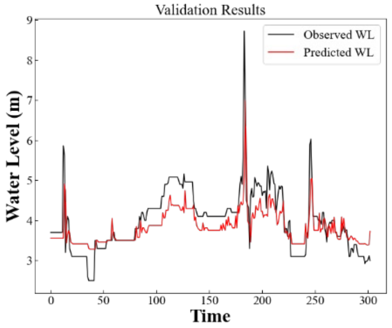


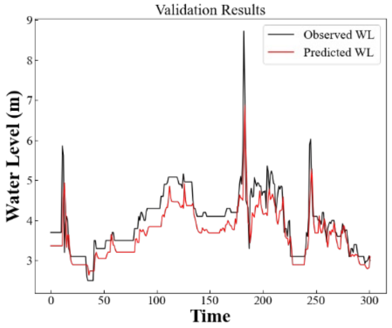


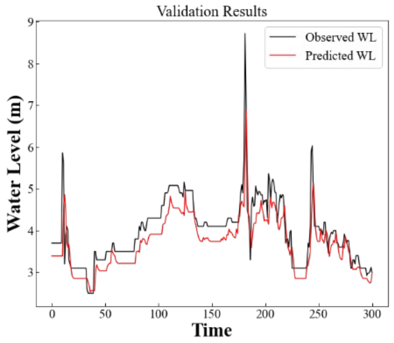


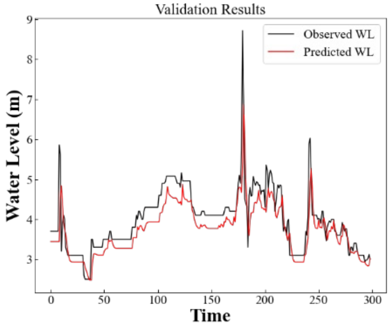


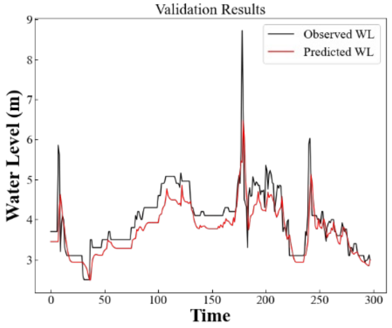
Model Testing

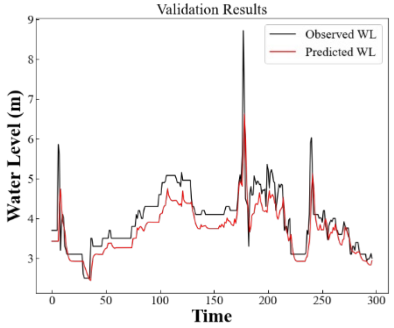


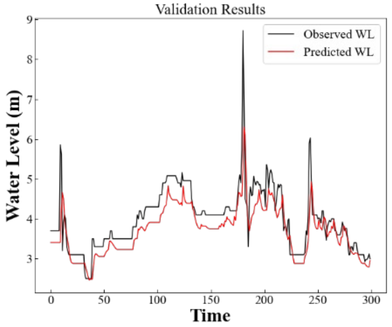


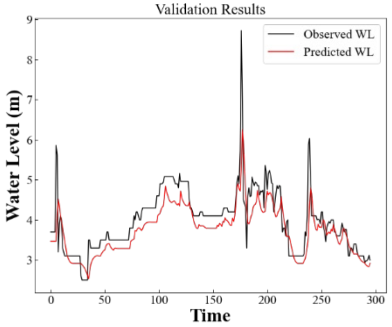


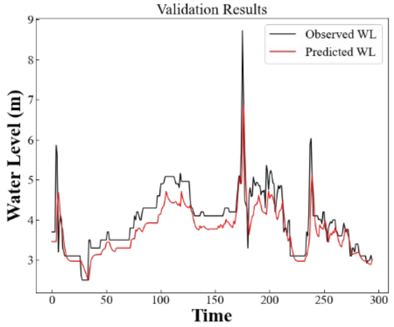




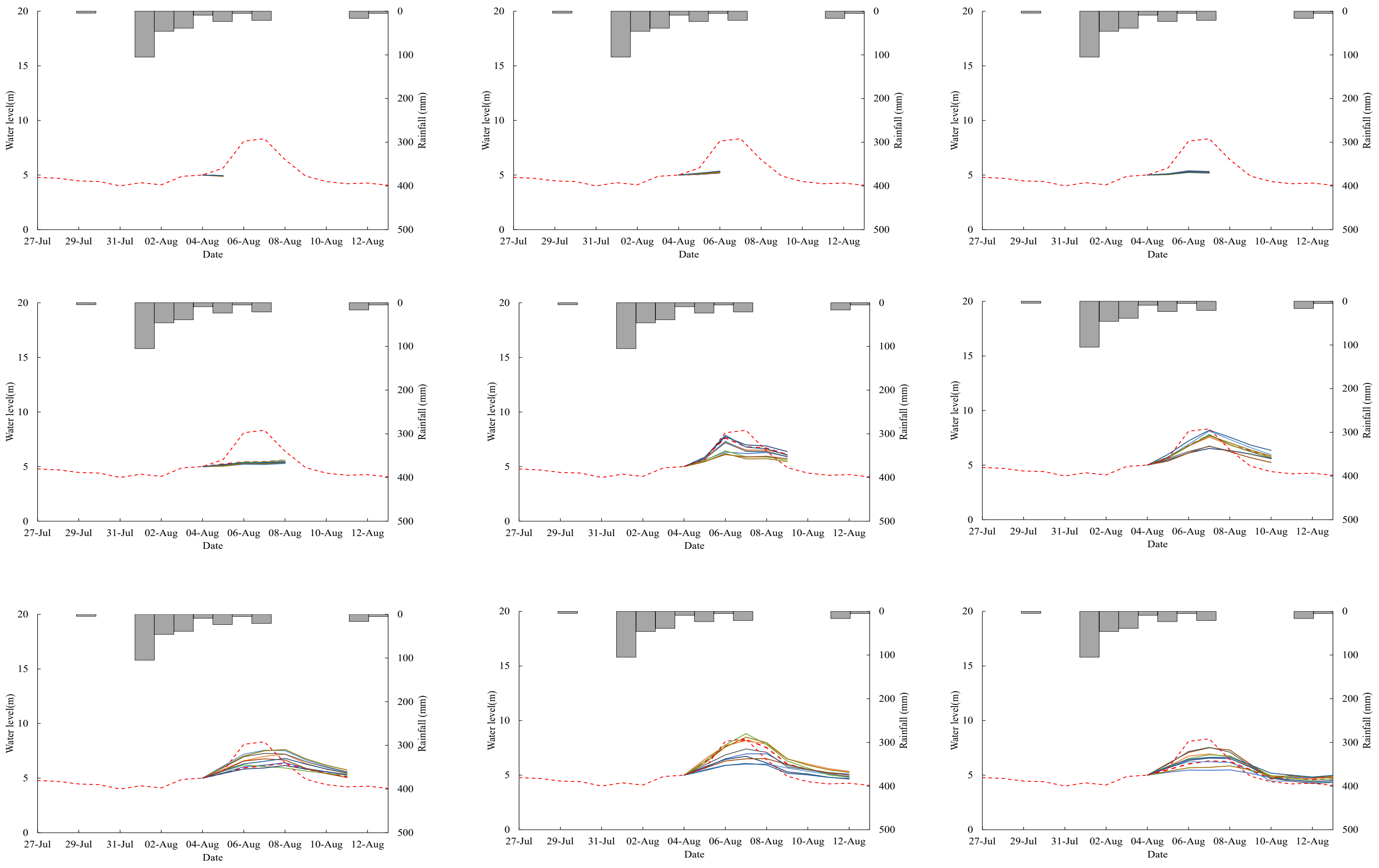








Prediction according to sequence length



Conclusion

❖ The factors that should be considered

- Sufficient correlation of training data
- Selection of appropriate features
- Initial model
- Fine tuning
- Appropriate parameters

❖ Needs

- Non-time series data (Static variables) management
- Deterministic or Stochastic information?

Thanks for your attention.