Geophysical Research Letters

Supporting Information for

Generalization of Runoff Risk Prediction at Field Scales to a Continental-scale Region Using Cluster Analysis and Hybrid Modeling

Chanse M. Ford[1], Yao Hu[2,3], Chirantan Ghosh[4], Lauren M. Fry[5], Siamak Malakpour-Estalaki[2], Lacey Mason[5], Lindsay Fitzpatrick[6], Amir Mazrooei[7], Dustin C. Goering[8]

[1]Department of Earth and Environmental Sciences, Michigan State University; [2]Department of Geography and Spatial Sciences, University of Delaware; [3]Department of Civil and Environmental Engineering, University of Delaware; [4]Department of Computer and Information Sciences, University of Delaware; [5]Great Lakes Environmental Research Lab (GLERL), National Oceanic and Atmospheric Administration; [6]Cooperative Institute for Great Lakes Research (CIGLR), University of Michigan; [7]Research Application Laboratory, National Center for Atmospheric Research (NCAR); [8]North Central River Forecast Center, National Weather Service, National Oceanic and Atmospheric Administration

Contents of this file

Introduction

Below contains the information related to the process for the development of the XGBoost models to predict the occurrence probability and magnitudes of daily EOF runoff. It also contains information related to the input data and the tools used to perform cluster analysis.

Development of XGBoost Models

Step1: Selection of causal variables

Causal variables are defined as variables that can have causal influence on the target variable in the sense of Granger Causality, which measures the ability to predict the future values of one time series using prior values of another time series (Granger 1969). For each cluster, different causal variables as the model outputs from NOAA's National Water Model (NWM) are selected to predict the daily EOF runoff (Tables S5 and S6). The method used for the selection of these casual variables is based on the causal inference using Directed Information algorithm as introduced in detail by Hu et al., (2021). Please note that the NWM runoff output (i.e. QQSFC) was considered as one of the potential causal variables but not selected by the Directed Information algorithm due to the large discrepancy between the observed EOF runoff and simulated runoff by the NWM. Please refer to the detailed explanation in Hu et al., (2021).

Step 2: Data preparation

For each EOF site in a cluster, we combined the observations of the daily EOF runoff with the model outputs for the selected causal variables from the NOAA's NWM on the grid (1km x 1km) where the EOF site falls within. Next, we repeated this process for all EOF sites in the cluster, which generates the dataset for training of the XGBoost model to predict the magnitude of daily EOF runoff. For the prediction of occurrence probability, we converted the observation of daily EOF runoff in the dataset to binary values: 1 when the magnitude of EOF runoff is positive and 0 otherwise.

Step 3: Validation of XGBoost models

Hyperparameter selection. XGBoost models contain a set of hyperparameters, whose values are used to control the performance of the XGBoost algorithm. In our case, we considered nine hyperparameters for each XGBoost model (Table S7). However, not all hyperparameters are critical to the model performance. A variance-based global sensitivity analysis approach (i.e., Sobol decomposition; Sobol, 2001) was thus used to identify three influential hyperparameters based on the values of the first order and total order indices, including learning rate, maximum depth of the tree and subsample rate.

Validation strategy. To evaluate the performance of the XGBoost models for ungauged locations, we randomly split EOF sites by 70%/30% within the cluster: EOF measurements from 70% of the EOF sites were used for training and the remaining 30% were for validation. Additionally, we also applied five-fold cross-validation to the training of the XGBoost models to mitigate overfitting. Figures S1 – S6 show the validation results using the XGBoost models to predict the magnitude of daily EOF runoff for each cluster under different split scenarios.

Table S1. List of Variables for Cluster Analysis

Variable	Temporal Resolution	Spatial Resolution	Source
Annual Rainfall	Daily	4-km	PRISM
Max Annual SWE	Daily	1-km	SNODAS
Annual Snowmelt	Daily	1-km	SNODAS
Annual PET	Hourly	$1/8^{th}$ °	NLDAS-2
Max Annual Soil Ice Content	Daily	1-km	NWM
Max Annual Vegetation Extent	Daily	1-km	NWM
Mean Depth to Water Table	Daily	250-m	NWM
Soil Moisture Content (Top Soil Layer)	Daily	250-m	NWM
Depth to Bedrock	Aggregated	HUC-10 Basin Scale	NWM*
Mean Urban LULC	Aggregated	HUC-10 Basin Scale	NWM*
Mean Water LULC	Aggregated	HUC-10 Basin Scale	NWM*
Mean Forested LULC	Aggregated	HUC-10 Basin Scale	NWM*
Mean Grassland LULC	Aggregated	HUC-10 Basin Scale	NWM*
Mean Shrubland LULC	Aggregated	HUC-10 Basin Scale	NWM*
Mean Wetland LULC	Aggregated	HUC-10 Basin Scale	NWM*
Mean Soil Sand Content (4 Layers)	Aggregated	HUC-10 Basin Scale	NWM*
Mean Soil Clay Content (4 Layers)	Aggregated	HUC-10 Basin Scale	NWM*
Mean Elevation	Aggregated	HUC-10 Basin Scale	NWM*
Percent Flatland	Aggregated	HUC-10 Basin Scale	NWM*
Percent Lowland	Aggregated	HUC-10 Basin Scale	NWM*
Percent Upland	Aggregated	HUC-10 Basin Scale	NWM*
Mean Relief	Aggregated	HUC-10 Basin Scale	NWM*
Mean Surface Runoff	Daily	250-m	NWM
Mean Subsurface Runoff	Daily	250-m	NWM

Table S2. List of Packages for Cluster Analysis

Package Name	Version	Author
curl	4.3	Ooms, J.
devtools	2.3.1	Wickham, H., Hester, J., and Chang, W.
doMC	1.3.6	Revolution Analytics and Weston, S.
dplyr	1.0.0	Wickham, H. et al.
factoextra	1.0.7	Kassambara, A. and Fabian, M.
foreach	1.5.0	Microsoft and Weston, S.
ggplot2	2016	Wickham, H.
latticeExtra	0.6-29	Sarkar, D. and Andrews, F.
maptools	1.0-1	Bivand, R. and Lewin-Koh, N.
ncdf4	1.17	Pierce, D.
prism	0.0.6	Hart, E.M. and Bell, K.
raster	3.4-5	Hijmans, R.J.
rgdal	1.5-12	Bivand, R., Keitt, T., and Rowlingson, B.
rgeos	0.5-3	Bivand, R. and Rundel, C.
rwrfhydro	1.0.0.9100	McCreight, J. et al.
sp	NA	Bivand, R. et al.
tidyr	1.1.0	Wickham, H. and Henry, L.

Table S3. Definition of Level of Severity (LS)

LS Interval	Definition[1]
[0, 1)	$P_{EOF} < M^{2}_{EOF,\ 20\%}$ (0.4mm)
[1, 2)	$M_{EOF,\ 20\%}$ (0.4mm) <= $P_{EOF} < M_{EOF,\ 50\%}$ (2.3mm)
[2, 3)	$M_{EOF,\ 50\%}$ (2.3mm) <= $P_{EOF} < M_{EOF,\ 80\%}$ (11.5mm)
[3, 4]	$P_{EOF} >= M_{EOF,\ 80\%}$ (11.5mm)

[1]Intervals are defined based on historical EOF measurements (M_{EOF}) and predicted magnitude of EOF runoffs (P_{EOF}).
[2]Magnitude of daily EOF runoff equals to 20% of all measured runoff events, i.e., 0.4mm.

Table S4. Relationship between Clusters, HUC-10 Watersheds and EOF Sites.

Cluster	HUC-10 Watershed	EOF Site
1	757	10
2	1085	5
3	399	6
4	595	4
5	1716	54

Table S5. List of selected causal variables from the National Water Model

No.	Variable Name	Definition	Units*
1.	RAINRATE	Precipitation	mm s^{-1}
2.	SFHD	Depth of ponded water on the surface	mm
3.	ACSNOM	Accumulated melting water out of snow bottom	mm
4.	SOILSAT	Fraction of soil saturation, column integrated	fraction
5.	SHG	Ground sensible heat	W m^{-2}
6.	TGV	Ground temperature with vegetated ground	K
7.	SOIL_W1	Liquid volumetric soil temperature at the top layer	m^3 m^{-3}
8.	SOIL_M3	Volumetric soil moisture in layer 3	m^3 m^{-3}
9.	FIRA	Total net LW radiation (+ to atmosphere)	W m^{-2}
10.	SOIL_M4	Volumetric soil moisture in the bottom layer	m^3 m^{-3}
11.	T2MV	2m temperature with vegetated ground	K
12.	QQSUB	Subsurface runoff	mm h^{-1}

* Values were calculated for each day.

Table S6. Test results for the XGBoost Models used to predict daily EOF runoff

Clusters	Causal Variables	Test R^2
1	RAINRATE, SFHD, ACSNOM, SOILSAT	0.16
2	RAINRATE, ACSNOM, SOILSAT, SHG, TGV	0.38
3	RAINRATE, SOIL_W1, SHG, SOIL_M3, ACSNOM, SFHD	0.12
4	RAINRATE, SOIL_W1, FIRA, SFHD, SOIL_M4, ACSNOM	0.55
5	RAINRATE, ACSNOM, SFHD, SHG, T2MV, SOILSAT, QQSUB, FIRA	0.40

Table S7. List of hyperparameters of the XGBoost model

No.	Hyperparameter Name	Definition	Range
1.	LR*	Learning Rate	[0, 1]
2.	MTD*	Maximum Tree Depth	[0, ∞)
3.	MCW	Minimum Child Weight	[0, ∞)
4.	SR*	Subsample Rate	(0, 1]
5.	ES	Number of the estimators	[1, ∞)
6.	CB	Subsampling of the columns	(0, 1]
7.	Gamma	Minimum loss reduction parameter	[0, ∞)
8.	SD	Random Seed	[0, ∞)
9.	Lamda	L2 Regularization parameter	[0, ∞)

*Selected influential hyperparameters

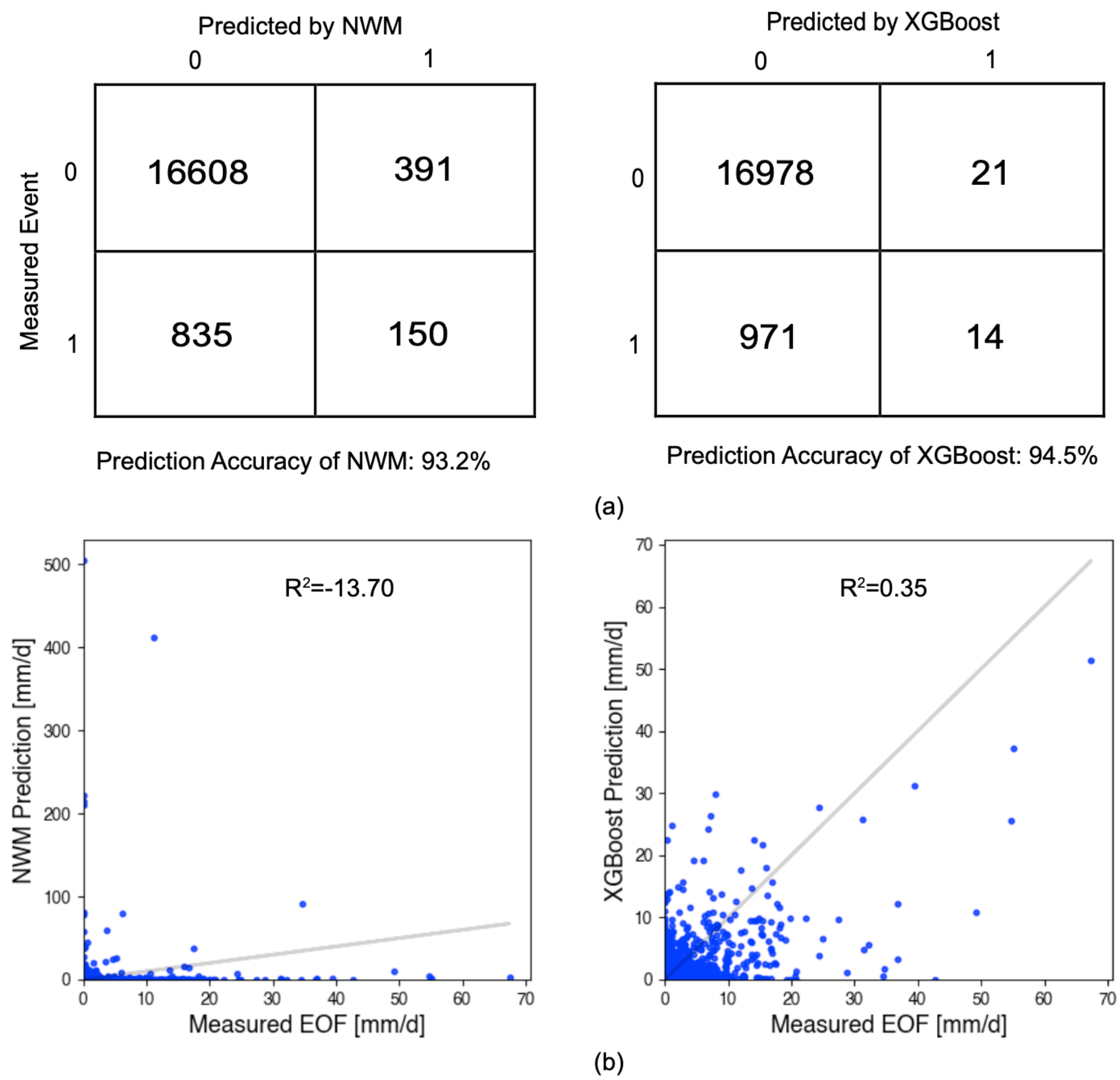


Figure S1: Comparison of predictions by the National Water Model (i.e., Predicted by NWM) and XGBoost Model (i.e., Predicted by XGBoost) for Cluster5 under the split scenario with the medium R^2 value (R^2 = 0.35): (a) Confusion matrices of the occurrence predictions of daily runoff events: 0/1: no/yes for a runoff event. (b) Scatter plots of the comparisons between the observed runoff events and the predictions by the NWM and XGBoost model measured by R^2.

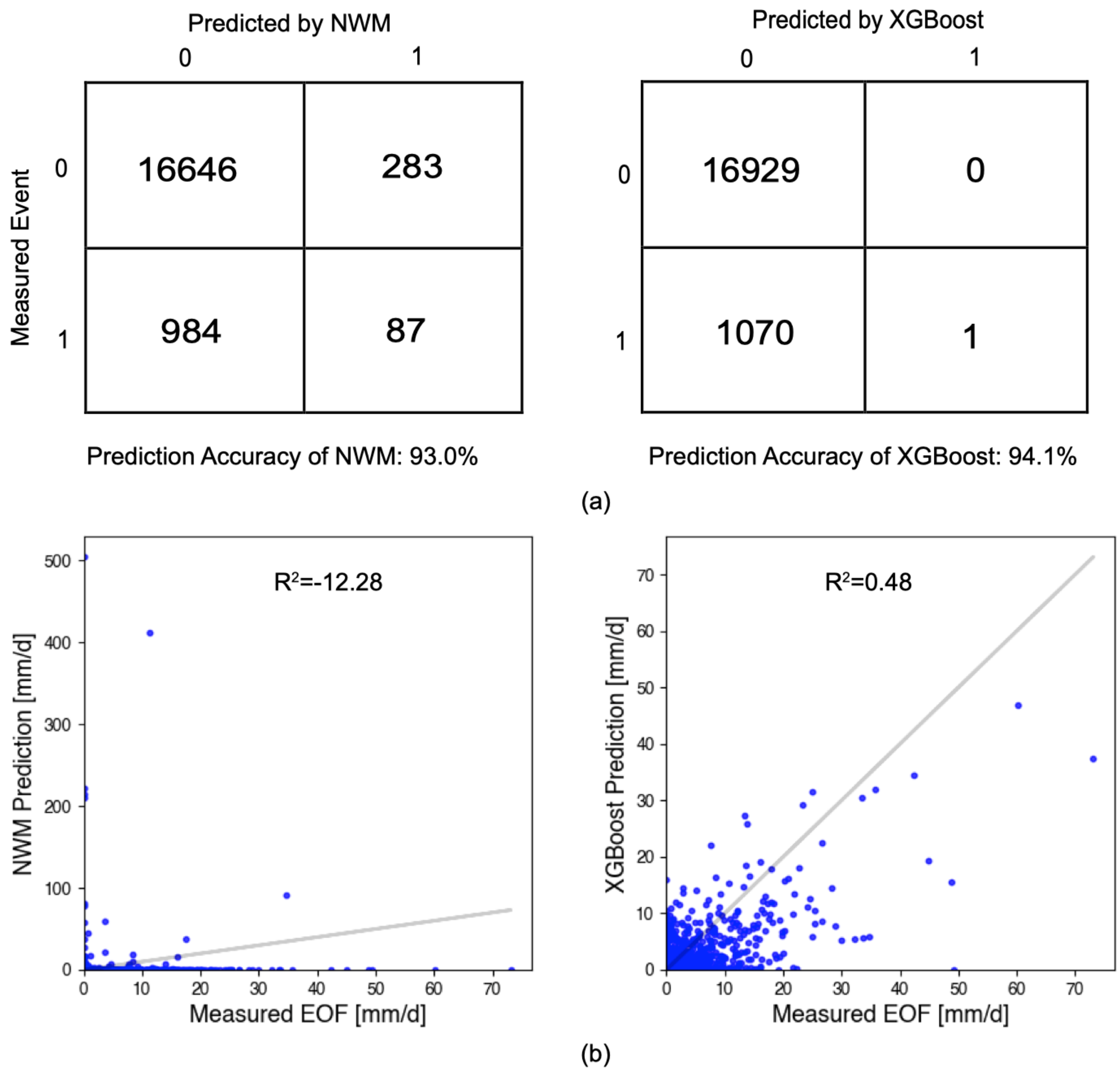


Figure S2: Comparison of predictions by the National Water Model (i.e., Predicted by NWM) and XGBoost Model (i.e., Predicted by XGBoost) for Cluster5 under the split scenario with the maximum R^2 value ($R^2 = 0.48$): (a) Confusion matrices of the occurrence predictions of daily runoff events: 0/1: no/yes for a runoff event. (b) Scatter plots of the comparisons between the observed runoff events and the predictions by the NWM and XGBoost model measured by R^2.

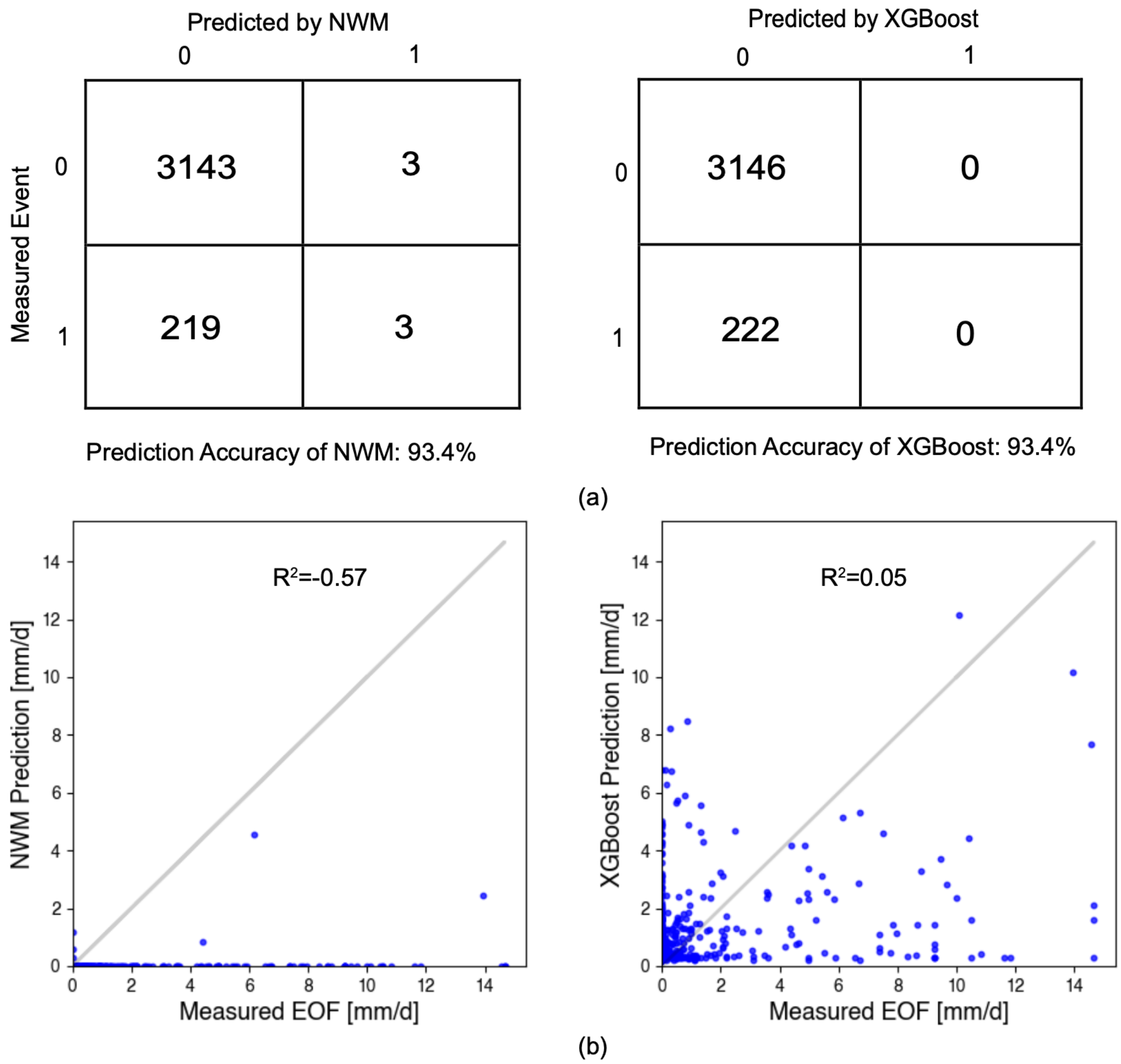
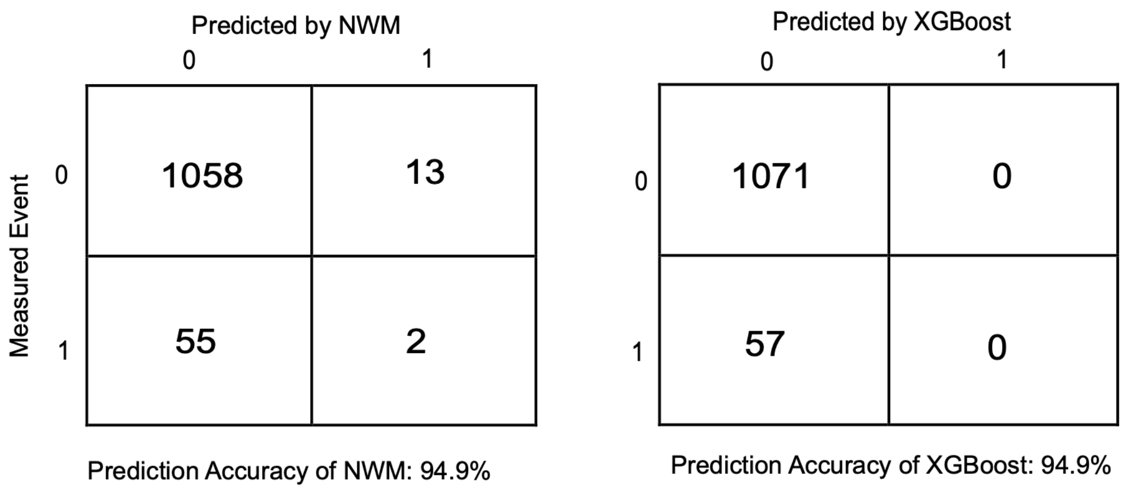
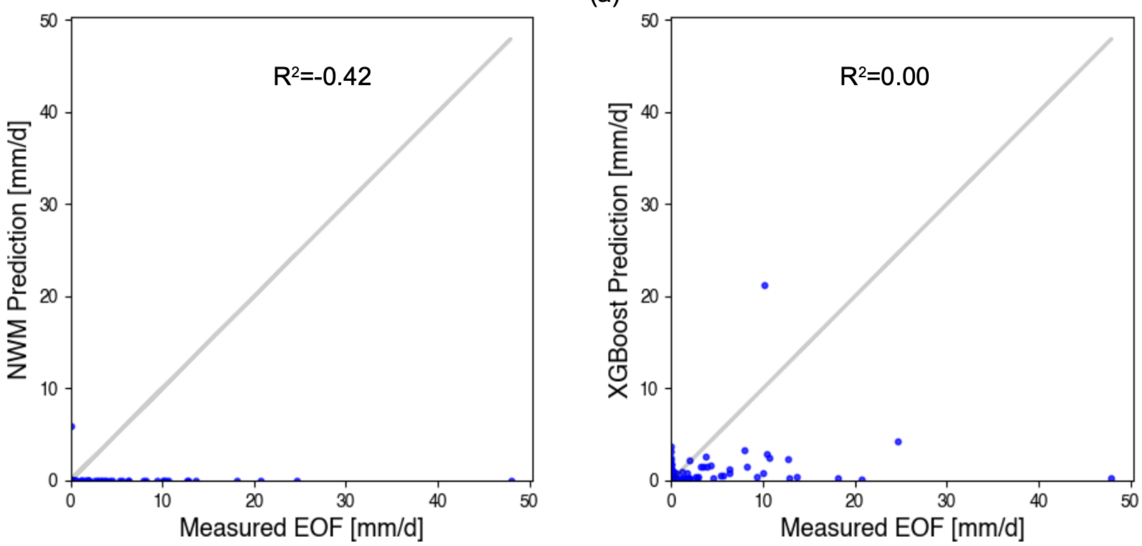


Figure S3: Comparison of predictions by the National Water Model (i.e., Predicted by NWM) and XGBoost Model (i.e., Predicted by XGBoost) for Cluster1 under the split scenario with the minimum R^2 value ($R^2_{min} = 0.05; R^2_{med} = 0.12; R^2_{max} = 0.18$): (a) Confusion matrices of the occurrence predictions of daily runoff events: 0/1: no/yes for a runoff event. (b) Scatter plots of the comparisons between the observed runoff events and the predictions by the NWM and XGBoost model measured by R^2.



(a)



(b)

Figure S4: Comparison of predictions by the National Water Model (i.e., Predicted by NWM) and XGBoost Model (i.e., Predicted by XGBoost) for Cluster2 under the split scenario with the minimum R^2 value ($R^2_{min} = 0.00$; $R^2_{med} = 0.58$;$R^2_{max} = 0.72$): (a) Confusion matrices of the occurrence predictions of daily runoff events: 0/1: no/yes for a runoff event. (b) Scatter plots of the comparisons between the observed runoff events and the predictions by the NWM and XGBoost model measured by R^2.

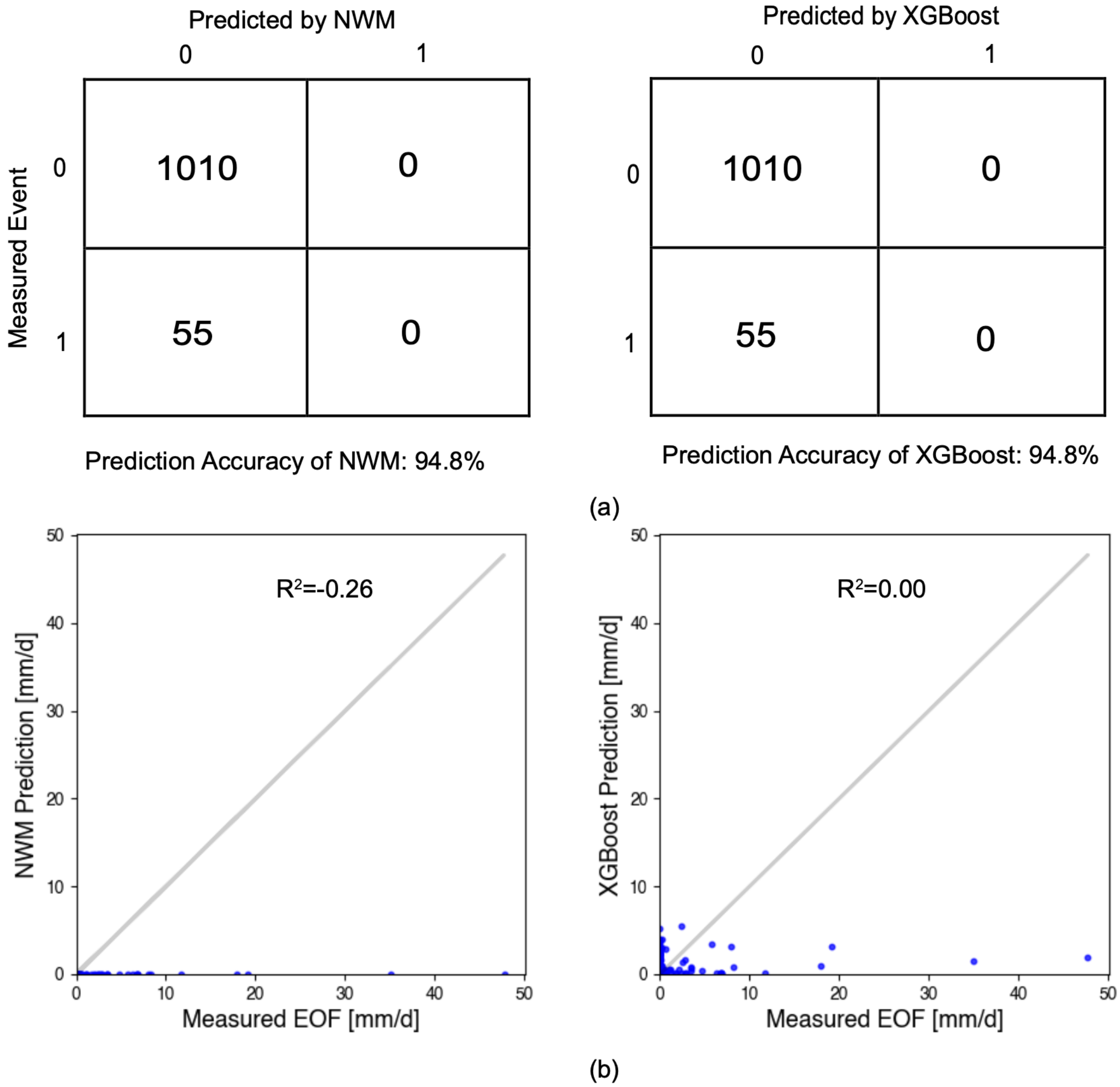


Figure S5: Comparison of predictions by the National Water Model (i.e., Predicted by NWM) and XGBoost Model (i.e., Predicted by XGBoost) for Cluster3 under the split scenario with the minimum R^2 value ($R^2_{min} = 0.00$; $R^2_{med} = 0.11$; $R^2_{max} = 0.20$): (a) Confusion matrices of the occurrence predictions of daily runoff events: 0/1: no/yes for a runoff event. (b) Scatter plots of the comparisons between the observed runoff events and the predictions by the NWM and XGBoost model measured by R^2.

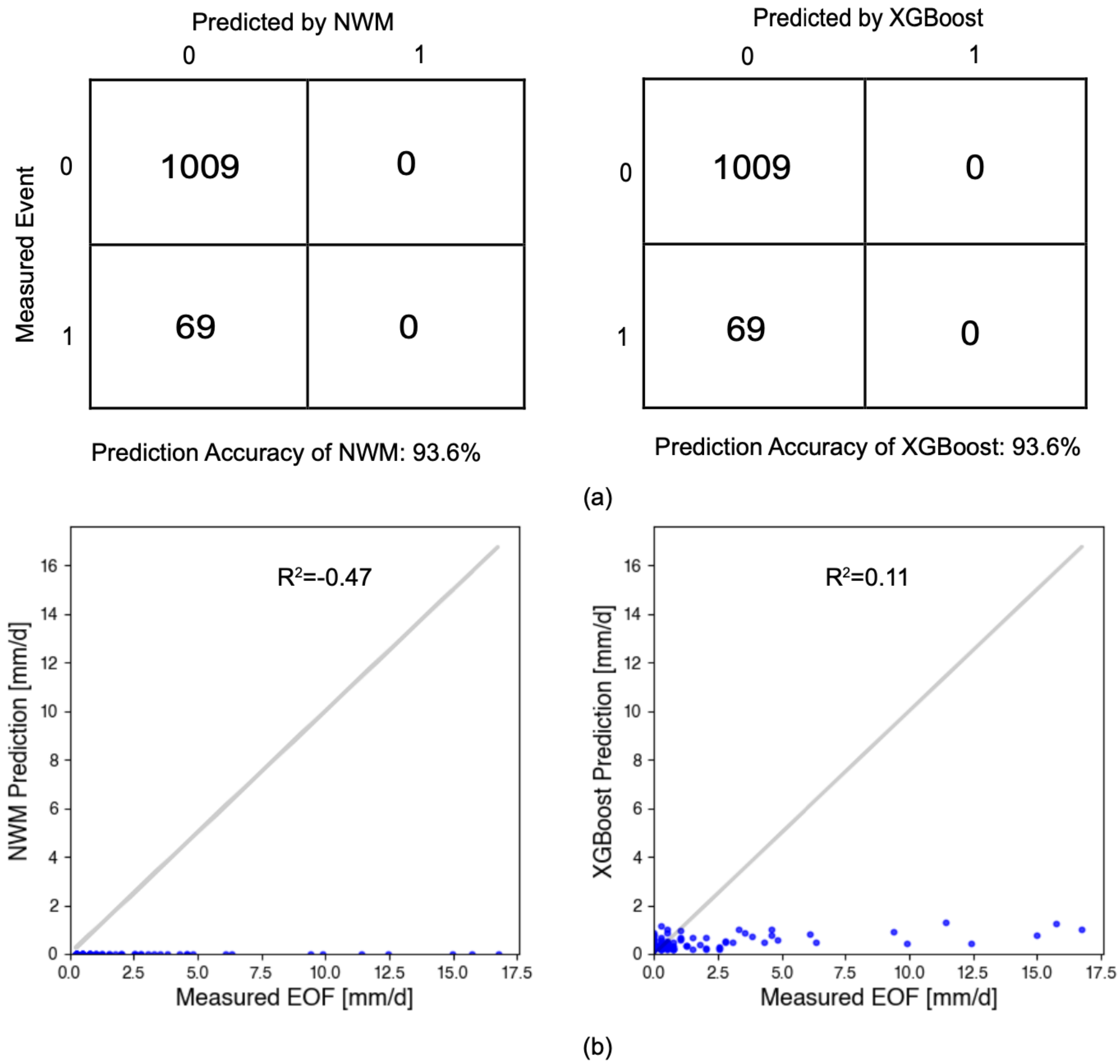


Figure S6: Comparison of predictions by the National Water Model (i.e., Predicted by NWM) and XGBoost Model (i.e., Predicted by XGBoost) for Cluste4 under the split scenario with the minimum R^2 value ($R^2_{min} = 0.11$; $R^2_{med} = 0.21$;$R^2_{max} = 0.32$): (a) Confusion matrices of the occurrence predictions of daily runoff events: 0/1: no/yes for a runoff event. (b) Scatter plots of the comparisons between the observed runoff events and the predictions by the NWM and XGBoost model measured by R^2.