ENVIRONMENTAL RESEARCH
COMMUNICATIONS

PAPER • OPEN ACCESS

Edge-of-field runoff prediction by a hybrid modeling approach using causal inference

To cite this article: Yao Hu *et al* 2021 *Environ. Res. Commun.* **3** 075003

View the article online for updates and enhancements.

Environmental Research Communications

OPEN ACCESS

RECEIVED
16 April 2021

REVISED
8 June 2021

ACCEPTED FOR PUBLICATION
21 June 2021

PUBLISHED
5 July 2021

PAPER

Edge-of-field runoff prediction by a hybrid modeling approach using causal inference

Yao Hu[1,2,3], Lindsay Fitzpatrick[3], Lauren M Fry[4], Lacey Mason[4], Laura K Read[5] and Dustin C Goering[6]

[1] Department of Geography and Spatial Sciences, University of Delaware, Newark, DE, United States of America
[2] Department of Civil and Environmental Engineering, University of Delaware, Newark, DE, United States of America
[3] Cooperative Institute for Great Lakes Research, University of Michigan, Ann Arbor, MI, United States of America
[4] Great Lakes Environmental Research Laboratory, National Oceanic and Atmospheric Administration, Ann Arbor, MI, United States of America
[5] Research Applications Laboratory, National Center for Atmospheric Research, Boulder, CO, United States of America
[6] North Central River Forecast Center, National Weather Service, National Oceanic and Atmospheric Administration, Chanhassen, MN, United States of America

E-mail: yaohu@udel.edu

Keywords: EOF runoff, hybrid modeling, causal inference, directed information, national water model, XGBoost

Supplementary material for this article is available online

Abstract

Unforeseen runoff events cause nutrient losses that affect crop production, revenue, and contribute to deteriorated water quality, leading to harmful algal blooms and hypoxia in receiving water bodies in the Great Lakes region. To mitigate the negative impacts caused by runoff events, we developed a hybrid modeling approach by combining physics-based and statistical models to predict the occurrence and level of severity of daily runoff events, supporting agricultural producers to avoid nutrient application before significant runoff events. We chose to use the National Oceanic and Atmospheric Administration's National Water Model (NWM) as the physical model given its flexible architecture design, technical robustness, model resolution, data availability, and wide application scale. For the statistical model, we developed a data-driven tool built from *Directed Information* and *eXtreme Gradient Boosting* (XGBoost) to estimate the occurrence and the level of severity of daily edge-of-eld (EOF) runoff events. This data-driven tool ingests a large variety of variables from NWM operational runs and translates them into the EOF runoff predictions on a daily scale in the Great Lakes region. Without calibrating the large-scale NWM for the local runoff prediction, the results show large improvements in the prediction of the occurrence and level of severity of daily EOF runoff using the hybrid physical-statistical modeling approach. Ultimately, the hybrid approach, when integrated into runoff risk decision support tools, is expected to provide dual benefits to agricultural producers and water quality, retaining more nutrients on their fields and lowering nutrient loads to water bodies during runoff events.

1. Introduction

Nutrient losses caused by runoff from agricultural fields (i.e., edge-of-eld (EOF) runoff) negatively affects crop production and profit as well as deteriorates water quality, leading to harmful algal blooms and hypoxia in receiving water bodies in the Great Lakes region (Scavia *et al* 2019, Stackpoole *et al* 2019, Michalak *et al* 2013). Predictions of surface runoff at the EOF scale can alert agricultural producers to avoid nutrient application before significant runoff events. As such, it can save them the costs and efforts associated with nutrient application and mitigate water pollution caused by nutrient losses in the Great Lakes region—dual benefits for both the producers and the Great Lakes.

EOF runoff occurrence and magnitude depends on many physical processes, such as rainfall intensity, soil moisture, and crop growth, to name a few. All these factors vary from farm to farm. To accurately predict the

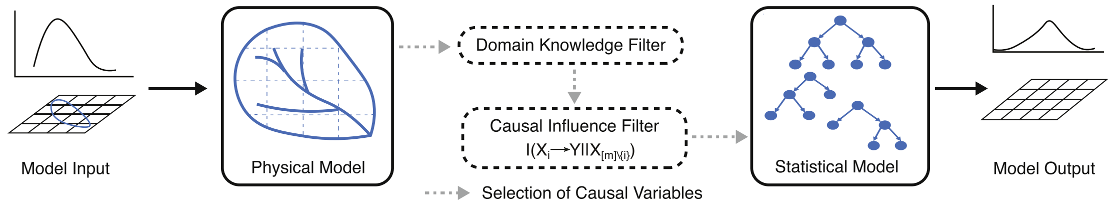


Figure 1. The domain knowledge and *Directed Information* algorithm are used to remove the non-causal variables generated by the physics-based models. Statistical models are used to describe the causal relationship between the selected causal variables and the observations of the Edge-of-Field (EOF) runoff.

occurrence of an EOF runoff event, observations of these factors at the individual farm scale would be necessary. However, due to the high measurement cost and complexity, it is almost impossible in practice to collect data at such a fine scale for large-scale runoff predictions, e.g., in the Great Lakes region. Because of the spatial sparsity of observations and complexity of the runoff generation processes, prediction of EOF runoff events at the field-scale over large domains is quite challenging.

Physics-based models are model representations of physical systems. When they are underpinned by the mechanisms to describe target processes, they can then be used to make predictions of target variables. However, the assessment of model skill relies solely on the goodness of fit (GOF) for a given spatial and temporal resolution. When the GOF is poor, we often resort to model calibration. The conventional approaches based on parameter tuning to calibrate the model against observations are highly computationally expensive (i.e., Curse of dimensionality) and therefore often infeasible (Sun and Sun 2015). In this context, two questions arise: (1) Can the physical model be useful if not being calibrated against the observations of the target process, e.g., EOF runoff in our case? (2) what approach can we use to extract useful information from the model to improve the prediction of the target process?

In this paper, we propose a hybrid modeling approach combining physics-based and statistical models to predict field-scale EOF runoff without incurring computationally expensive approaches to calibrate the physical model. This hybrid approach fuses the useful information of the target physical system embedded in the physical model into statistical models through causal variables identified by the *Directed Information* algorithm. The statistical models then fill the gap in the representation of the EOF runoff process by the physical model and improve the runoff prediction. As such, the fundamental contribution in this work is the development of a hybrid modeling approach based on causal inference to improve the prediction of the target process without recalibrating the physical model. This approach is applied to 13 domains in the Great Lakes region, leading to better predictions of the occurrence and level of severity of daily EOF runoff at all EOF locations within the region.

2. Methodology

In the following, we illustrate the use of hybrid models to predict the occurrence and level of severity of daily EOF runoff. We first ran the physics-based model to generate model outputs across the model domain and then selected model outputs that have statistical causal influence on the EOF runoff, namely cause variables. Finally, we developed the statistical models using the causal variables and runoff measurements to predict EOF runoff at large scales (figure 1).

2.1. Description of the study domain

Thirteen domains in the Great Lakes region were selected as the study region (figure 2(a)). Land use in the region varies geographically, with the cultivated rural areas dominating watershed classification in the southern portions of Minnesota, Wisconsin, and Michigan, and many of the Great Lakes watersheds in Ohio, Pennsylvania, and New York (Mayer *et al* 2014). Agricultural land comprises one-third of the Great Lakes basin area and supports 7% of U.S. farm production (Kerr *et al* 2016). Non-point source pollution has been attributed as the cause of significant surface water quality concerns in the region, including the Great Lakes themselves (Michalak *et al* 2013). Over a hundred EOF runoff observational sites (figure 2(b)), which consist of hydrologic and meteorologic instruments are available at the edge of individual agricultural fields across the states. Water leaving the field, either on the surface or subsurface, is intercepted and channeled through the site to measure and record runoff timing and magnitude. Conservation partners, such as Discovery Farms (Wisconsin and Minnesota), USGS, and USDA-ARS, provided the observational data which was then modified to be used in this study.

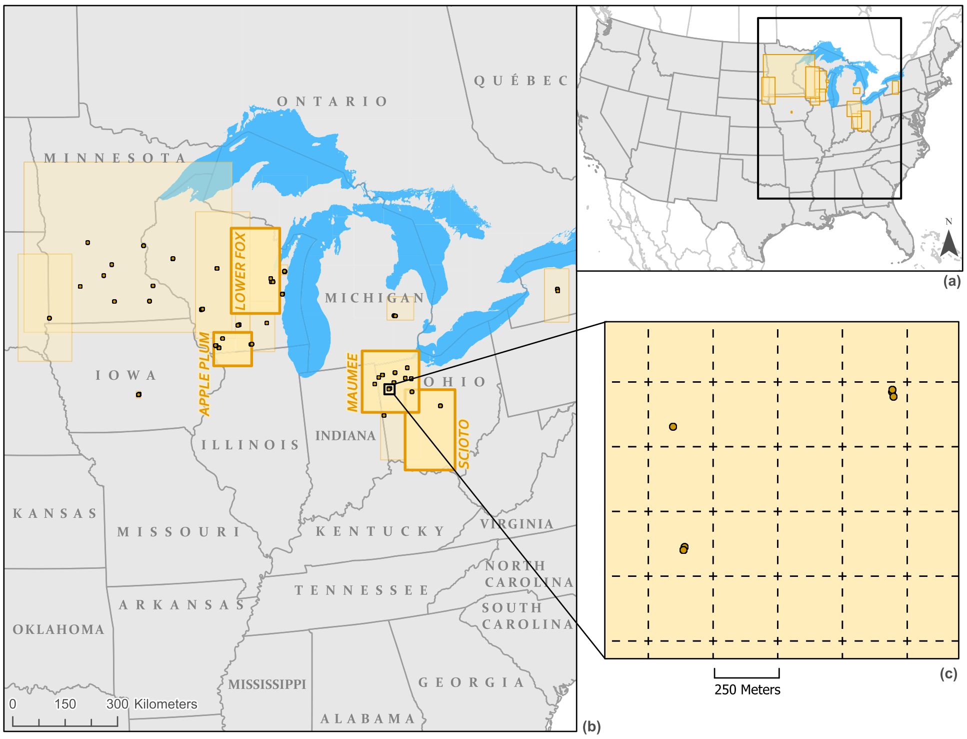


Figure 2. (a) Study area of 13 domains (ranging in area from 49 to 289,765 km^2) in the Great Lakes region with locations of all EOF sites (small orange squares); (b) enlarged version of the study area with two highlighted domains to illustrate the results (i.e., Maumee, OH and LowerFox, WI) and (c) model representation of a region in the domain of Maumee by the 250 m × 250 m routing grids.

2.2. Physics-based model

The choice of physics-based models needs to account for the fact that EOF runoff events occur at the farm scale while their impacts are at a much larger scale. The National Oceanic and Atmospheric Administrations National Water Model (NWM) can be a good candidate since it is a large-scale hydrologic model that provides real-time analysis and forecasts of hydrologic states across the 2.7 million river reaches that span the contiguous United States (CONUS), Hawaii, and Puerto Rico. In addition to streamflow, the NWM produces forecasts of key land surface variables on a 1 km × 1 km grid (e.g., soil moisture, snowpack, evaporation) and of ponded surface water and routed runoff on a 250 m × 250 m grid. The NWM is not specifically designed to predict the EOF runoff, although one of the model outputs from the NWM is the surface runoff (i.e., QQSFC).

2.3. Data preparation

Based on the location and distribution of the EOF observation sites, we defined 13 domains in the Great Lakes region (figure 2(b)). Each domain encompasses a watershed that contains a few to dozens of EOF monitoring sites. Then, we ran the NWM over these domains, and output 170 model variables at an hourly scale. From these variables, we selected 72 variables (SI: table S1) to test causal influence on the generation of EOF runoff based on the knowledge of the physical processes, including 65 variables from the land surface grid (e.g., precipitation rate [mm/s]) and seven variables from the routing grid (e.g., volumetric soil moisture at different soil layers [−]). There exist non-deterministic relationships between these selected variables because their values are not only determined by the forcing at their own locations but also the ones in the neighborhood, which are not considered for the causal influence test in this study. Because the EOF observational runoff data was already aggregated to daily accumulations, we calculated the daily values for all the variables at the scale of the routing grid where EOF sites are located (figure 2(c)). In addition, to consider the time-lag effects, we also included one- and two-day lag of these variables. In total, we tested 216 variables as candidate variables, from which causal variables were identified for each domain.

2.4. Directed Information

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). Rather than feeding the statistical models with all available data, the pre-selection of causal variables can help mitigate model overfitting (Hu et al 2018). In our case, the target variable is the measurements of EOF runoff, and the causal variables are selected among 216 candidate variables from the NWM using an information-theoretic quantity called *Directed Information* as shown by equations (1) and (2) (Hu *et al* 2018, Kramer 1998, Marko 1973)

$$\frac{1}{\mathsf{T}}\mathbf{I}(X_{1:T} \rightarrow Y_{1:T}\|Z_{1:T}) \coloneqq \frac{1}{\mathsf{T}}\sum_{t=1}^{T}\mathbf{I}(X_{1:t};\ Y_t|Y_{1:t-1},\ Z_{1:t}) \tag{1}$$

$$\coloneqq \frac{1}{\mathsf{T}}\sum_{t=1}^{T}\mathbf{E}_{P_{\mathrm{X,Y,Z}}}\left[\log_2\frac{P(Y_t|Y_{1:t-1},\ Z_{1:t},\ X_{1:t})}{P(Y_t|Y_{1:t-1},\ Z_{1:t})}\right], \tag{2}$$

where T is the time horizon and the boldface capital letter denotes a discrete-time random process, such as $\mathbf{X} = \{X_1, X_2,\ldots,X_i,\ldots\}$ and X_i denotes the random process at time i. For a finite horizon t, we also denote the random process from time 1 up to time t as $X_{1:t} = \{X_1,\ldots,X_t\}$. For equation (1), $\mathbf{I}(X_{1:t};\ Y_t|Y_{1:t-1},\ Z_{1:t})$ denotes the conditional mutual information between $X_{1:t}$ and Y_t, conditioned on $Y_{1:t-1}$ and $Z_{1:t}$. Equation (2) expresses the conditional mutual information as an expected log-likelihood ratio. For random processes $\mathbf{X}$ and $\mathbf{Y}$, the larger the *Directed Information* value calculated by equation (2), the more causal influence $\mathbf{X}$ can have on $\mathbf{Y}$. The random processes can be correlated but deterministic relationships between them need to be avoided when being selected for the *Directed Information* test, i.e., no processes can be deterministically derived from the other processes. In this study, we will use the *Directed Information* approach to identify which of the candidate variables from the NWM have causal influence on the daily EOF runoff for each domain. The Python implementation of the *Directed Information* algorithm to select causal variables will be shared in an open-source code-sharing repository.

2.5. Statistical models

Once the causal variables are selected for each domain, statistical models are then used to represent the causal relationships between the target variable and the causal variables. A highly effective machine learning method, *eXtreme Gradient Boosting* (XGBoost) is applied to the training of the statistical models (Chen and Guestrin 2016). XGBoost uses a type of gradient boosting technique, which generates a new decision tree model at each time step to minimize the residuals between the observations and predictions from the previous step. Each new tree model has its own weight (i.e., the amount of say) depending on how well the individual tree model minimizes the residuals. Gradient Boosting has been known for its effectiveness in predictive performance. For example, Boosted Regression Trees (BRT) implement gradient boosting to learn an ensemble of tree models and weigh them based on their ability to predict the observed data (Hu *et al* 2018, Elith *et al* 2008), and have demonstrated the overall best predictive performance among supervised learning algorithm (Caruana and Niculescu-Mizil 2006).

Different from other tree-based machine learning algorithms (e.g., BRT and Random Forest (RF)) which can be computationally intensive with large-scale datasets, XGBoost is an efficient and scalable tree-based gradient boosting algorithm due to the implementation of several key techniques including weighted quantile sketch to find the optimal splits, sparsity-aware algorithms for parallel tree learning and cache-aware block structure for out-of-core tree learning (Chen and Guestrin 2016). Meanwhile, implementation of these techniques on GPU further enables XGBoost to effectively process large-scale datasets (Mitchell *et al* 2018). In addition, built-in cross-validation and regularization capabilities make the XGBoost models less prone to overfitting (Chen and Guestrin 2016).

2.6. EOF Runoff prediction

Our purpose is to estimate the occurrence and level of severity of daily EOF runoff for each domain using the identified causal variables from the NWM. The NWM provides the information of model outputs at the target locations where causal variables are identified and then used by statistical models to make predictions of daily runoff at the EOF locations on the 250 m × 250 m grid as shown in figure 2. The risk of daily EOF runoff can be decomposed into two parts, the occurrence and the level of severity (None, Low, Medium, High, and Extreme). The first step is to predict the occurrence probability (Pc), i.e., how likely an EOF runoff event is to occur on a given day. If it occurs ($Pc > 0.5$), the second step is to predict its level of severity (Ps). Greater magnitude of runoff (MRF) presents a higher risk of damage resulting from nutrient loss. Five levels of severity were defined based on all daily EOF measurements: Extreme ($MRF \geqslant RF_{80\%}$, 80% of all measured runoff events), High ($RF_{50\%} \leqslant MRF < RF_{80\%}$), Medium ($RF_{20\%} \leqslant MRF < RF_{50\%}$), Low ($0 < MRF < RF_{20\%}$), and None ($MRF = 0$). We tested two approaches to predict Ps: The first approach is to predict the magnitude of daily EOF runoff and then convert it to the level of severity. We noticed that it can be challenging to predict the magnitude of daily EOF runoff since EOF runoff processes are complex and their measurements are often limited and sparse, with zero runoffs in most days of the year. Hence, rather than predict the magnitude, the alternative approach is to directly predict the severity of daily EOF runoff. To train and validate an XGBoost model for each

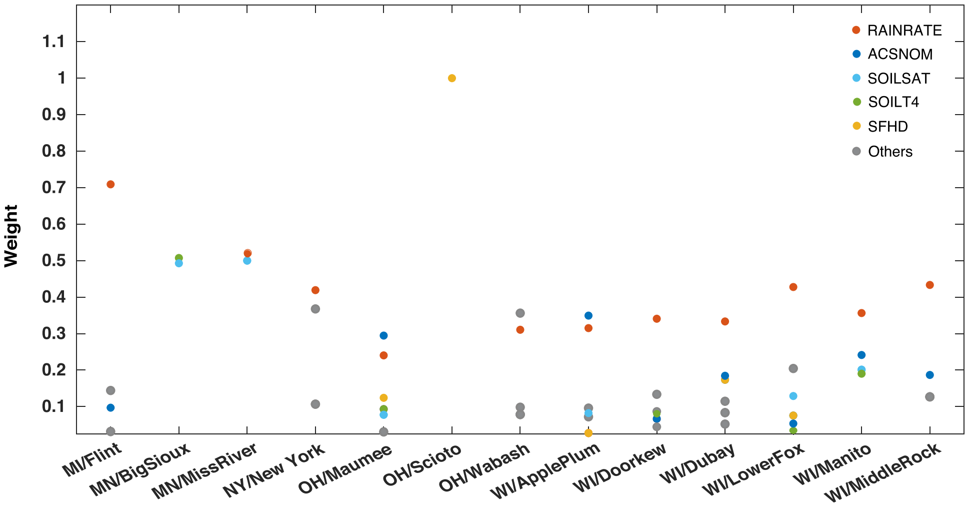


Figure 3. List of selected causal variables from the NWM and the strength of their causal influence (i.e., Weight) for each of the 13 domains in the Great Lakes region shown in figure 2. The top five most selected variables are RAINRATE: Precipitation rate [mm/d]; ACSNOM: Accumulated melting water out of snow bottom [mm/d]; SOILSAT: Fraction of soil saturation, column integrated; SOILT4: Soil temperature [K] at the fourth layer with the thickness equal to 1.0m; SFHD: Instantaneous value of the depth of ponded water on the surface [mm/d] and the rest influential variables are in gray. For each domain, the selected causal variables are ranked by the normalized *Directed Information* values (Weight), measuring the strength of the cause influence they have on the target variable.

domain, we adopted a 60%/40% split of daily EOF measurements for each domain, ranging from 1,046 measurements for the Scioto domain, OH to 22,776 measurements for the ApplePlum domain, WI (figure 2). Five-fold cross-validation and automated hyperparameter tuning are used to ensure the consistent and optimal performance of the XGBoost model.

3. Results

Given the amount of data and types of analysis we conducted, two domains have been selected to illustrate the performance and results of the hybrid modeling approach. The predictive results across all domains are consistent, showing large improvements in the prediction of the occurrence and level of severity of daily EOF runoff. Additional information and figures for all the domains are provided as the Supplementary Information.

3.1. Selection of causal variables

In the presentation of results, for each domain, if one variable and its corresponding lagged variables appear to be influential, they are only counted once regarding the frequency of their appearances. In total, we identified 25 causal variables for 13 domains of interest (figure 3). Among these 25 variables, the daily precipitation (i.e., RAINRATE) was selected in most domains (11 out of 13), followed by the accumulated melting water out of snow bottom (i.e., ACSNOM) for eight out of 13 domains. For each of the domains, the selected causal variables were ranked by the strength of the cause influence (i.e., Weight in figure 3) they have on the target variable. For most domains, RAINRATE was not only the most selected variable but also the most influential variable to the EOF runoff.

3.2. Comparison of statistical models

Figure 4 presents the comparisons of results to predict the occurrence and magnitude of the daily EOF events from two tree-based machine learning algorithms, XGBoost and BRT in terms of contingency statistics, training time, and goodness of fit (GOF). Compared with BRT, XGBoost had slightly better prediction accuracy (91.2% versus 89.5%) and yielded some improvements in the true positive rate (TPR) while at the same time carried a slightly higher false positive rate (FPR). Furthermore, the training time with XGBoost is seven times shorter than BRTs, from 65.7s to 9.2s. Additionally, among all missed runoff events, XGBoost had fewer missed events in all four categories, especially for the events larger than 10.0 mm, only 2.5% of the total missed events (figure 4(b)). Both algorithms predicted more runoff events than the number of observed events and were low-biased with regards to the magnitude of the runoff events (figure 4(c)). In terms of the GOF, the prediction of magnitudes by XGBoost had better agreement with the observed runoff measured by the R^2 (figure 4(d)).

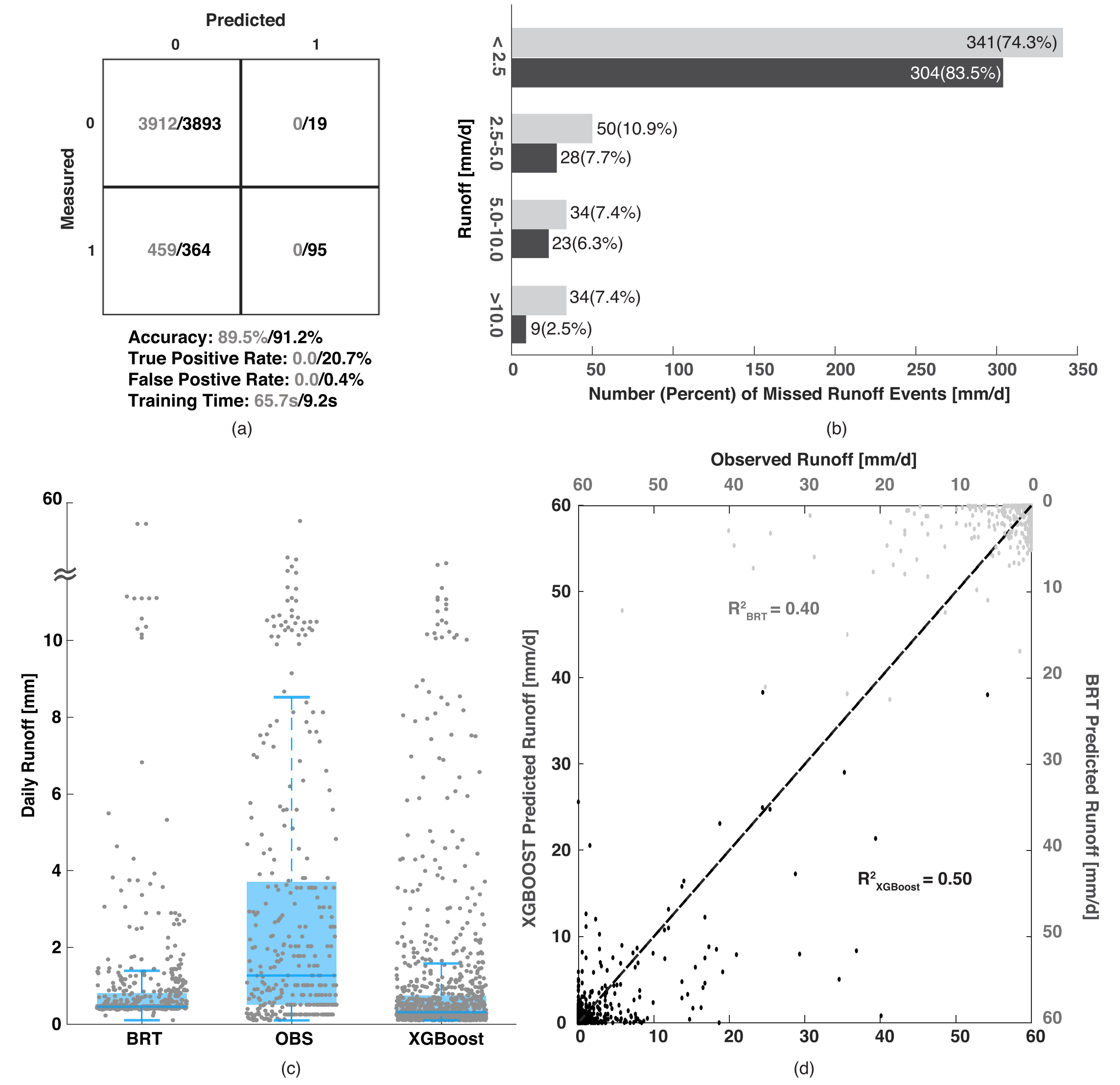


Figure 4. Comparison of results from *Boosted Regression Trees* (BRT, shown in gray) and *eXtreme Gradient Boosting* (XGBoost, shown in black) for the Maumee domain, OH: (a) Comparison of the contingency tables of the occurrence predictions of daily EOF runoff; 0/1: no occurrence/occurrence of EOF runoff. (b) Distribution of the number (percent) of runoff events missed by both approaches; All missed events are divided into four categories: smaller than 2.5 mm; between 2.5 mm and 5.0 mm, between 5.0 mm and 10.0 mm and larger than 10.0 mm. (c) Boxplots of all positive daily runoff events from observations (OBS) and predicted by BRT and XGBoost. (d) Scatter plots of the comparisons between the observed runoff events and the predictions (BRT at the upper right; XGBoost at the lower left) measured by R^2.

3.3. Performance of hybrid models

EOF runoff is a time series with most daily measurements equal or close to zero over years. To better predict positive runoff events, especially large events (e.g., larger than 10.0 mm) shown in figure 4(c), we tested different threshold values (e.g., 0.05 mm or 0.1 mm) to define an EOF runoff event. As a result, the true positive rate (TPR) increases with the negative impacts on the prediction accuracy and the false positive rate (FPR) for the two selected domains, Maumee, OH, and Lower Fox, WI (figures 5(a) and (b)). We also found similar results for the other domains when testing with different thresholds.

We compared the predicted surface runoff by the NWM (i.e., QQSFC) with the observations, as well as the predictions from the XGBoost model. As shown in figures 6(a) and (b), the NWM tended to underestimate the occurrence and magnitude of the runoff, predicting far fewer runoff events than the observed while XGBoost predicted more runoff events but also tended to underestimate the runoff. Overall, compared to NWM, the XGBoost model has largely improved the prediction of EOF runoff as measured by the TPR and FRP (figure 6(a)), and R^2 values (figure 6(c)). The results are similar for the other domains (SI: figures S1–S13 available online at stacks.iop.org/ERC/3/075003/mmedia).

We adopted two approaches to predict the level of severity of daily EOF runoff events. The first approach predicted the magnitudes of the daily EOF runoff, from which we estimated the corresponding levels of severity (figure 7(a)), while the second approach directly predicted the level of severity (figure 7(b)). Compared with the

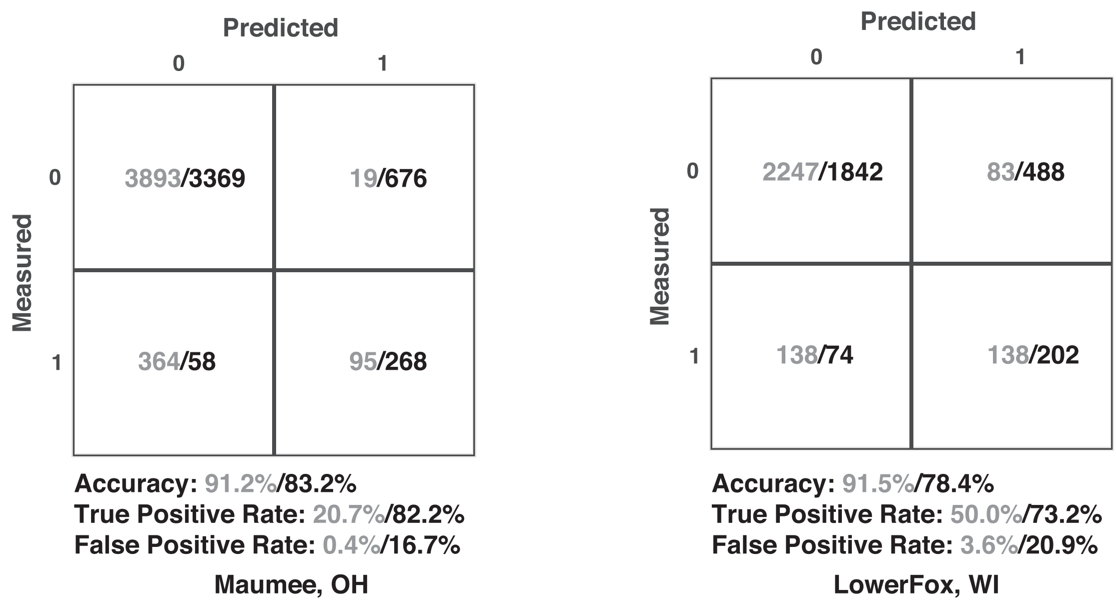


Figure 5. Comparison of the contingency tables of the occurrence predictions of EOF runoff with different thresholds (gray for 0.0 mm and black for 0.1 mm) used to define an EOF runoff event for two selected domains, Maumee, OH (a) and LowerFox, WI (b). Three metrics are calculated for the two scenarios: prediction accuracy, true positive rate, and false positive rate.

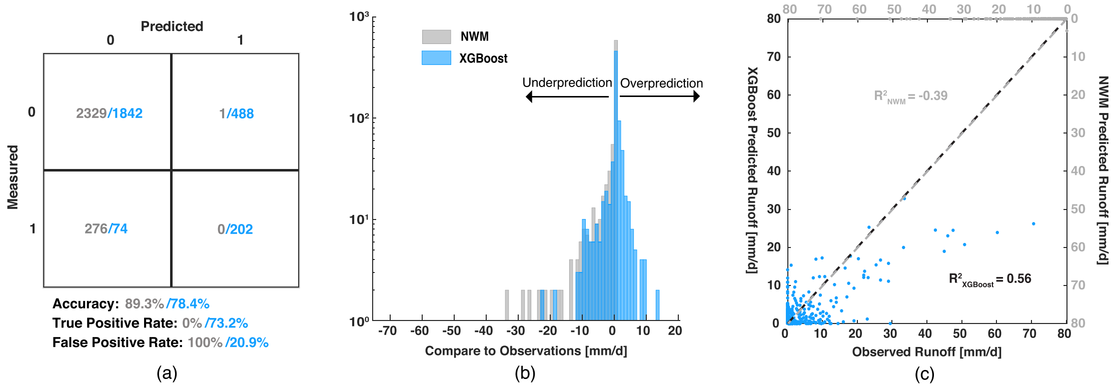


Figure 6. Comparison of predictions by the National Water Model (NWM, shown in gray) and *eXtreme Gradient Boosting* (XGBoost, shown in blue) for the LowerFox domain, WI: (a) Comparison of the contingency tables of the occurrence predictions of daily EOF runoff; 0/1: no occurrence/occurrence of EOF runoff. (b) Differences between all positive daily runoff events from observations (OBS) and predictions by the NWM and XGBoost. (c) Scatter plots of the comparisons between the observed runoff events and the predictions by the NWM (upper right) and XGBoost (lower left) measured by R^2.

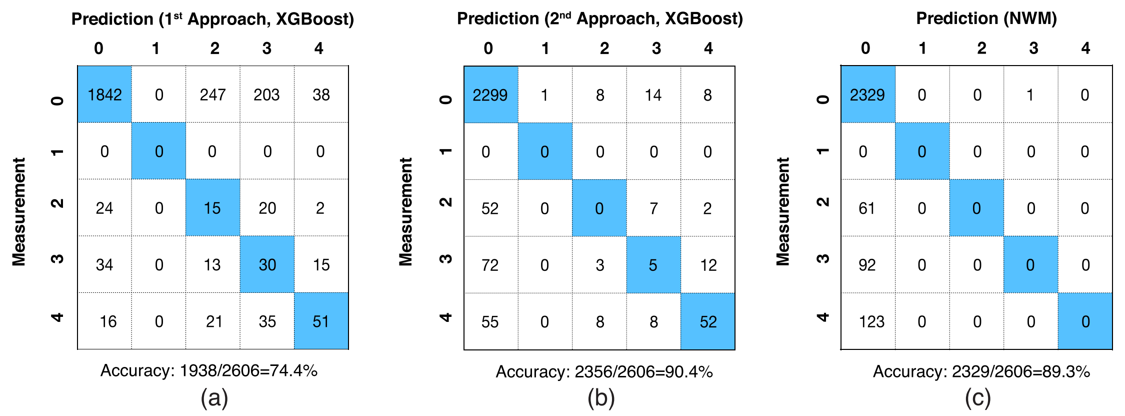
indirect approach, the direct approach achieves better prediction accuracy. We also compared these two approaches for the remaining domains in the Great Lakes region and similar results were obtained (SI: Table S2). In addition, we converted the predictions by the NWM into different levels of severity and found that the XGBoost model using the direct approach outperformed the NWM in terms of prediction accuracy (figures 7(b) and (c)).

4. Discussion

In this section, we propose to discuss the added values to identify causal variables for the development of statistical models, the choices we made to use the XGBoost model, and predict both the occurrence and level of severity of daily EOF runoff, as well as the limitations of hybrid modeling.

4.1. Causal inference for model diagnosis

The generation of EOF runoff can be driven by a variety of factors. Among all the selected causal variables, rainfall appears to be the most influential, followed by the accumulated melting water from snow bottom which



Prediction (1st Approach, XGBoost)

Measurement	0	1	2	3	4
0	1842	0	247	203	38
1	0	0	0	0	0
2	24	0	15	20	2
3	34	0	13	30	15
4	16	0	21	35	51

Accuracy: 1938/2606=74.4%

(a)

Prediction (2nd Approach, XGBoost)

Measurement	0	1	2	3	4
0	2299	1	8	14	8
1	0	0	0	0	0
2	52	0	0	7	2
3	72	0	3	5	12
4	55	0	8	8	52

Accuracy: 2356/2606=90.4%

(b)

Prediction (NWM)

Measurement	0	1	2	3	4
0	2329	0	0	1	0
1	0	0	0	0	0
2	61	0	0	0	0
3	92	0	0	0	0
4	123	0	0	0	0

Accuracy: 2329/2606=89.3%

(c)

Figure 7. Contingency tables of the level of severity predictions of daily EOF runoff events by the XGBoost model using two approaches and the NWM for the LowerFox domain, WI: (a) Indirect approach: predict the magnitudes of daily runoff events by the XGBoost model and then converting them to different levels of severity. (b) Direct approach: predict the level of severity of daily runoff directly by the XGBoost model. (c) Converting the predictions of magnitudes of daily runoff by the NWM into different levels of severity.

is often the main cause of the EOF runoff during winter. In addition, the other causal variables (e.g., soil saturation, temperature, and moisture) selected by the *Directed Information* approach generally agree with the expectation from the domain knowledge. In this sense, the results shown in figure 3 justify the use of the causal inference approach based on the *Directed Information* in our case study.

We also noticed that not a single variable appears to be influential for all the 13 domains in the Great Lakes region, even for the rainfall. For the domains (e.g., Scioto, OH) where rainfall is not a causal variable, it can be possible that the generation of runoff is indeed not affected by the rainfall or the rainfall as a data product derived for these domains from the Analysis of Record for Calibration (AORC) was not accurate (Kitzmiller *et al* 2018), e.g., due to the lack of sufficient measurements in the neighborhood to make the interpolation, provided that the data quality of the EOF measurements in these domains can be assured. In this sense, causal influence helps narrow down the potential causes of the non-selection of rainfall as an influential variable for some domains.

Furthermore, the results of causal inference can be used to diagnose the problems with model inputs, model parameters, and its representation of physical processes, especially when certain model outputs are supposed to have causal influence on the target variable. For example, in our case study, among all NWM outputs, predicted surface runoff (i.e., QQSFC) by the NWM should have the best agreement with the observed EOF runoff for each domain. If that would be the case, QQSFC should appear as the most influential variable in all domains. In fact, QQSFC is not selected as an influential variable for a single domain due to its large discrepancies in predictions of the EOF runoff (figure 6), which mainly arise from the uncertainties associated with model input (i.e., rainfall), model parameters, and its representation of the runoff process. It is likely that the model representation and parameters are the main cause, especially when rainfall is selected as the influential variable for most domains. This can be further justified by the fact that the current version of NWM is not developed and calibrated to simulate surface runoff. Meanwhile, if the NWM can improve its representation of the mechanisms to simulate runoff, we can expect that more runoff predictions by the NWM in the Great Lakes region will be selected as the influential variable using the *Directed Information* algorithm.

4.2. Statistical models and hybrid modeling

The causal relationships between EOF runoff events and the selected causal variables are represented using statistical models. In our case,the NWM is a large-scale hydrologic model, providing hourly streamflow predictions and other hydrologic information at 1 km and 250 m grids over CONUS, and the amount of data generated by NWM simulations easily falls into the realm of Big Data. Meanwhile, to ensure the prediction accuracy of the risk level, we will need to continuously assimilate the stream of EOF measurements and newly generated model outputs from the NWM to update the statistical models. Hence, it will require the machine learning algorithms to process new data feeds and train the statistical models in a timely manner. Compared with the BRT, it is more computationally effective to train statistical models using XGBoost, which also generated better predictions of the occurrence and magnitude of daily EOF runoff. Hence, XGBoost is selected for the training of the statistical models that make predictions of the occurrence and level of severity of daily EOF runoff events.

Measurements of EOF runoff events are imbalanced with most values equal or close to zeros. As such, the prediction accuracy mainly reflects how well they match the observed zero runoffs. In contrast, the ability to predict positive runoffs, as measured by the true positive rate (TPR), is more critical, since they cause actual economic and environmental damages. By setting the threshold (e.g., 0.1 mm) to define an EOF runoff event, XGBoost models can capture more runoff events, which gives rise to higher TPR but lower prediction accuracy (PA). Meanwhile, the impacts of the threshold on TPR and PA vary with domains, depending on the actual EOF measurements. We thus need to conduct individual tests to seek the optimal threshold to balance TPR and PA for individual domains.

As mentioned above, at the current stage of development, the NWM was not able to fully represent the underlying mechanisms to generate EOF runoff, leading to large discrepancies between the observed runoff events and the NWM predictions (figure 6). Instead of relying on the NWM alone, the hybrid modeling approach built upon both the NWM and the XGBoost model was able to improve the predictions of the occurrence and magnitude of daily EOF runoff (figure 6). This is mainly because the XGBoost model can help fill the gap in the representation of the surface runoff generation processes by the NWM while taking advantage of its outputs to predict EOF runoff. As such, the XGBoost model outperforms the NWM in the prediction of both the occurrence and level of severity of EOF runoff (figures 6 and 7).

However, we need to be aware that the XGBoost model is not developed to replace but assist the NWM to improve the predictions of EOF runoff. When the NWM improves its predictions of EOF runoff events, we will likely see the following outcomes: (1) the selection of predicted runoff by the NWM as an influential variable by the *Directed Information* algorithm; the better the NWM predictions, the more influence they can have on the observations as measured by the *Directed Information* value; and (2) the improvement of the XGBoost model in runoff predictions, since the XGBoost model is built upon the model outputs from the NWM. When the quality of the model outputs is improved in terms of runoff predictions, as a data-driven model, the predictions from the XGBoost model will therefore improve accordingly.

In this study, two different approaches were developed to predict the level of severity for each domain using the XGBoost model. Overall, the direct approach outperforms the indirect approach. This is mainly because measurements of daily EOF runoff are zero-inflated. As such, it is difficult to predict the magnitude of EOF runoff with high accuracy, and prediction errors further propagate through the conversion from the magnitude to the level of severity. In addition, from users' perspective, they are more concerned about the potential damage a runoff event can cause rather than the prediction of the magnitude with high accuracy. Thus, it makes better sense to directly predict the level of severity of EOF runoff.

4.3. Limitations and outlook

Compared with the conventional approaches to calibrate model parameters, the hybrid modeling approach shows its advantage in significantly reducing the computational intensity. While using the hybrid approach, we need to be aware of its limitations: (1) Because of the nature of data-driven models, it requires constant updates to improve their prediction accuracy when more training data becomes available. For this reason, computationally efficient statistical models with built-in parallel mechanisms stand out from the rest. (2) When selecting the causal variables, the *Directed Information* requires the candidate variables to exhibit a certain degree of variation. As such, constant variables or variables with small variations, which may be critical to the physical processes cannot be identified as causal variables. For example, geophysical parameters, which are critical to the runoff process, cannot be selected as causal variables by the algorithm. (3) Statistical models are often trained with specific datasets and locations. We need to be alert when applying the statistical models with the data out of the training data range and/or in a different location. In our case, when statistical models are trained and validated at the locations where we have the EOF runoff, they will be only applied to the watersheds with similar hydrologic responses in the Great Lakes region, which are characterized by the geophysical parameters.

5. Conclusion

This paper introduced a hybrid modeling approach combining the physics-based model, NWM with the statistical model, XGBoost model to achieve better predictions of the occurrence and level of severity of EOF runoff in the Great Lakes region. Through the domain knowledge and *Directed Information*, we can identify the valuable information embedded in the NWM outputs, which can then be used by the XGBoost model to fill the gap in the representation of the runoff generation processes in the NWM. As a result, the hybrid approach effectively improves the predictions of EOF runoff. More broadly, this approach is suitable for the improvement of model predictions when there exist discrepancies between observations and predictions by physics-based models and calibration of these models is not feasible due to high computational costs. Naturally, as the predictions by the physical models improve, so will the hybrid approach. In addition, future work is needed to determine how the statistical models trained and validated for one location can be applied to other locations.

Acknowledgments

The dataset for this research is available in this in-text data citation reference: Hu (2021) [shared under the Creative Commons Attribution CC BY. http://creativecommons.org/licenses/by/4.0/]. Any additional information can be obtained from the corresponding author. This work was supported by the Great Lakes Restoration Initiative through the U.S. Environmental Protection Agency and National Oceanic and Atmospheric Administration. An award is granted to Cooperative Institute for Great Lakes Research (CIGLR) through the NOAA Cooperative Agreement with the University of Michigan (NA17OAR4320152). The NOAA-GLERL contribution is No. 1983. We want to thank Arezoo RafieeiNasab and Aubrey Dugger for assisting with code development and questions related to the WRF-Hydro modeling system. We also want to thank the following agencies for providing us with daily EOF measurements, including USGS, USDA-ARS, Discovery Farms Minnesota, and Discovery Farms Wisconsin.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI:https://doi.org/http://www.hydroshare.org/resource/9460830270ec4d8b9d9c4260cca2114d.

ORCID iDs

Yao Hu https://orcid.org/0000-0002-0199-6044

References

Caruana R and Niculescu-Mizil A 2006 An empirical comparison of supervised learning algorithms *Proceedings of The XXIII International Conference on Machine Learning* pp 161–8

Chen T and Guestrin C 2016 Xgboost: a scalable tree boosting system *Proceedings of The XXII ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* pp 785–94

Elith J, Leathwick J R and Hastie T 2008 A working guide to boosted regression trees *Journal of Animal Ecology* **77** 802–13

Granger C W 1969 Investigating causal relations by econometric models and cross-spectral methods *Econometrica: journal of the Econometric Society* **37** 424–38

Hu Y, Scavia D and Kerkez B 2018 Are all data useful? inferring causality to predict flows across sewer and drainage systems using directed information and boosted regression trees *Water Research.* **145** 697–706

Hu Yao 2021 Edge of field runoff for the Great Lakes Region HydroShare https://www.hydroshare.org/resource/9460830270ec4d8b9d9c4260cca2114d/

Kerr J M, DePinto J V, McGrath D, Sowa S P and Swinton S M 2016 Sustainable management of Great Lakes watersheds dominated by agricultural land use *J. Great Lakes Res.* **42** 1252–9

Kitzmiller D H, Wu W, Zhang Z and Patrick N 2018 The analysis of record for calibration: a high-resolution precipitation and surface weather dataset for the united states *AGU Fall Meeting Abstracts* **2018** H41H–06

Kramer G 1998 *Directed Information for Channels with Feedback* (Germany: Hartung-Gorre)

Marko H 1973 The bidirectional communication theory-a generalization of information theory *IEEE Trans. Commun.* **21** 1345–51

Mayer A, Winkler R and Fry L 2014 Classification of watersheds into integrated social and biophysical indicators with clustering analysis *Ecological Indicators.* **45** 340–9

Michalak A M *et al* 2013 Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions *Proc. Natl Acad. Sci.* **110** 6448–52

Mitchell R, Adinets A, Rao T and Frank E 2018 Xgboost: scalable GPU accelerated learning arXiv:1806.11248

Scavia D, Bocaniov S A, Dagnew A, Hu Y, Kerkez B, Long C M and Wang Y C 2019 Detroit river phosphorus loads: anatomy of a binational watershed *J. Great Lakes Res.* **45** 1150–61

Stackpoole S M, Stets E G and Sprague L A 2019 Variable impacts of contemporary versus legacy agricultural phosphorus on us river water quality *Proc. Natl Acad. Sci.* **116** 20562–7

Sun N Z and Sun A 2015 *Model Calibration and Parameter Estimation: for Environmental and Water Resource Systems* (Berlin: Springer)