

Are all data useful? Inferring causality to predict flows across sewer and drainage systems using directed information and boosted regression trees

Yao Hu ^a, Donald Scavia ^b, Branko Kerkez ^{a,*}

^a Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, United States

^b School for Environment and Sustainability, University of Michigan, Ann Arbor, United States

ARTICLE INFO

Article history:

Received 11 May 2018

Received in revised form

3 August 2018

Accepted 4 September 2018

Available online 4 September 2018

Keywords:

Flow prediction

Causality

Directed information

Boosted regression trees

Data-driven model

ABSTRACT

As more sensor data become available across urban water systems, it is often unclear which of these new measurements are actually useful and how they can be efficiently ingested to improve predictions. We present a data-driven approach for modeling and predicting flows across combined sewer and drainage systems, which fuses sensor measurements with output of a large numerical simulation model. Rather than adjusting the structure and parameters of the numerical model, as is commonly done when new data become available, our approach instead learns causal relationships between the numerically-modeled outputs, distributed rainfall measurements, and measured flows. By treating an existing numerical model – even one that may be outdated – as just another data stream, we illustrate how to automatically select and combine features that best explain flows for any given location. This allows for new sensor measurements to be rapidly fused with existing knowledge of the system without requiring recalibration of the underlying physics. Our approach, based on *Directed Information* (DI) and *Boosted Regression Trees* (BRT), is evaluated by fusing measurements across nearly 30 rain gages, 15 flow locations, and the outputs of a numerical sewer model in the city of Detroit, Michigan: one of the largest combined sewer systems in the world. The results illustrate that the *Boosted Regression Trees* provide skillful predictions of flow, especially when compared to an existing numerical model. The innovation of this paper is the use of the *Directed Information* step, which selects only those inputs that are causal with measurements at locations of interest. Better predictions are achieved when the *Directed Information* step is used because it reduces overfitting during the training phase of the predictive algorithm. In the age of “big water data”, this finding highlights the importance of screening all available data sources before using them as inputs to data-driven models, since more may not always be better. We discuss the generalizability of the case study and the requirements of transferring the approach to other systems.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

The need to understand and predict water flows across cities is important for predicting flash flooding, reducing sewer overflows, and designing infrastructure (Field and Tafuri, 2006; Morales et al., 2017; Paquier et al., 2015). The dynamics of flow across these systems are complicated by the combined influences of hydrology, infrastructure, and highly variable rainfall (Konrad, 2003).

Presently, predictive approaches attempt to capture many of these features explicitly in the form of numerical models. These models, which are underpinned by physical laws and are often derived from first-order principles, represent the urban water systems at high resolutions and capture very specific characteristics, such as pipe dimensions, soil types, orifices, and subcatchment dynamics. For large cities, this can often lead to highly structured models that are difficult to parameterize and calibrate.

Simultaneously, many cities are scaling efforts to monitor assets in real-time, which means that more distributed sensor data are becoming available. For example, flow meters, water level sensors, and water quality sensors are now readily being deployed across urban water systems (Kerkez et al., 2016). The proliferation of

* Corresponding author. Department of Civil and Environmental Engineering, 2350 Hayward, 2044 GG Brown, Ann Arbor, MI, 48109-2125, United States.

E-mail addresses: huya@umich.edu (Y. Hu), scavia@umich.edu (D. Scavia), bkkerkez@umich.edu (B. Kerkez).

sensors seems promising, but more data may not always be helpful, especially if they do not exhibit a causal relationship with states being modeled. In those instances, two questions arise: (1) *Which emerging sources of data are actually useful in explaining flow across large urban water systems?* (2) *For those inputs that are deemed important, what quantity of data is required, and how can these data be rapidly ingested to improve predictions?*

Instead of relying on the recalibration of a numerical model, this paper presents a data-driven approach that combines all available data sources, including the outputs of an existing numerical model, into a holistic and automated prediction of water flows. In this way, the predictive skill embedded in a numerical model is retained when useful, while any additional sources of sensor data are ingested to further improve predictions. The fundamental contribution of the paper is a new method to predict flows in urban drainage systems, which: (1) selects useful (causal) inputs through a *Directed Information* algorithm, and (2) yields flow predictions with the selected inputs using *Boosted Regression Trees*. As more diverse data sources become available to decision makers, this approach will allow for the rapid and automated incorporation of emerging data into holistic predictions of flows. The approach is evaluated by fusing measurements from nearly 30 rain gages and 15 flow sensors with the outputs of a numerical sewer model in the city of Detroit, Michigan.

2. Background

2.1. Predicting through numerical models

A number of popular urban drainage models are presently in use, including the Stormwater Management Model (SWMM), MIKE URBAN and HEC-HMS, to name a few. These models couple hydrology and hydraulics, numerically computing processes such as infiltration and shallow water flow. Once calibrated, these models can be very effective at forecasting and decision making across fine spatiotemporal resolutions. The most common approach to model calibration seeks to adjust the model structure and its parameters so that the model output agrees with the measurements (Sun and Sun, 2015). This often includes a combination of manual parameter tuning that relies on the expertise of modelers or, in some cases, auto-calibration (Doherty, 2015). If knowledge is updated—due to changes in infrastructure, new measurements, or updated information on the watershed—model recalibration is often needed.

It is well known that standard parameter calibration methods are subject to the *curse of dimensionality*, where computational cost increases exponentially with the number of calibrating parameters (Sun and Sun, 2015). Given this complexity and resulting financial cost of recalibration, most water models hardly keep pace with urban change or the emergence of new data sources. In fact, it is not uncommon for many numerical water models in the United States to be over a decade. As such, more streamlined approaches are needed to keep pace with the emergence of new data sources and to ensure forecasts are made using the most relevant and up-to-date information.

2.2. Data-driven forecasting

In lieu of statistical models, a number of data-driven approaches are showing promise to model water systems. Instead of explicitly modeling physics, these approaches rely only on data, such as sensor measurements or features of a system, to make forecasts. While data-driven models cannot always provide insight into system behavior, they can enable streamlined and adaptive toolchains

to rapidly ingest many data sources. Broadly, many of these approaches fall under the umbrella of supervised machine learning (ML) approaches, which use historical data to “learn” complex mappings between inputs and a target variable.

When modeling flows across urban water systems, traditional ML approaches, such as generalized linear regression, may not work well due to the nonlinearities and collinearities inherent in complex systems (Dormann et al., 2013). Nonlinear mappings can be learned through methods such as *Artificial Neural Networks* (ANN), but these often require a large amount of data and are computationally expensive (Schalkoff, 1997). For many applications, a supervised ML approach known as *Boosted Regression Trees*, has recently shown promise in balancing computational intensity with performance. According to Caruana and Niculescu-Mizil (2006), *Boosted Regression Trees* have shown the best overall predictive performance among supervised learning algorithms, while remaining immune to collinearity. Regardless of the choice of algorithm, it is well known that many data-driven approaches may be sensitive to overfitting due to inadequate input data selection, especially if irrelevant inputs are used during training (Ng, 1998). As such, the prospects associated with access to many new sensor measurements may be hampered by the realization that more is not always better. That is, not all data may be useful, and brute-force use of all available data may actually lead to worse forecasts.

3. Methods

Instead of forcing the choice between a numerical or data-driven approach, as is commonly done, our method seeks to strike a balance by combining the benefits of both. The toolchain uses *Boosted Regression Trees* to make forecasts by ingesting a large number of input features (Fig. 1). However, instead of using sensor data as the only input, we also treat an existing numerical model as another input data source — the idea being that even an out-of-date or poorly calibrated numerical model may still embed a significant amount of information, which may not be captured by sensor data alone. Depending on the location of interest, various combinations of input features may provide the best forecast. As such, an innovation in our approach is a preliminary step, which uses the criterion of *Directed Information*, to determine which inputs may lead to the best prediction. This forms a holistic and automated toolchain, which ingests all available data but reduces the risk of overfitting by selecting the most “useful” data for forecasting.

3.1. Feature selection using directed information

Given a set of random processes $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m\}$ and a target random process \mathbf{Y} , we seek to predict the process \mathbf{Y} using the processes in \mathbf{X} . In our case, \mathbf{Y} represents a time series of flow measurements at the location for which we would like to generate future predictions (namely, the target variable). The processes \mathbf{X} include time series that could be used to derive a predictive model for \mathbf{Y} , such as rainfall measurements, as well as the outputs of a numerical model. In many cases, it may not be advantageous to use all processes from \mathbf{X} as inputs to a predictive model because this may lead to a model that performs well in the training phase, but one that performs poorly in prediction (i.e. overfitting). Instead, the goal is to select a subset of processes that are “useful” in describing \mathbf{Y} . Statistically, this can be captured using *Granger Causality* (Granger, 1969). If predictions of \mathbf{Y} are improved by using a process $\mathbf{X}_i \in \mathbf{X}$, we say \mathbf{X}_i statistically causes \mathbf{Y} . More formally, \mathbf{X}_i causes \mathbf{Y} , as measured by Granger Causality, if the past of \mathbf{X}_i can help predict the future of \mathbf{Y} .

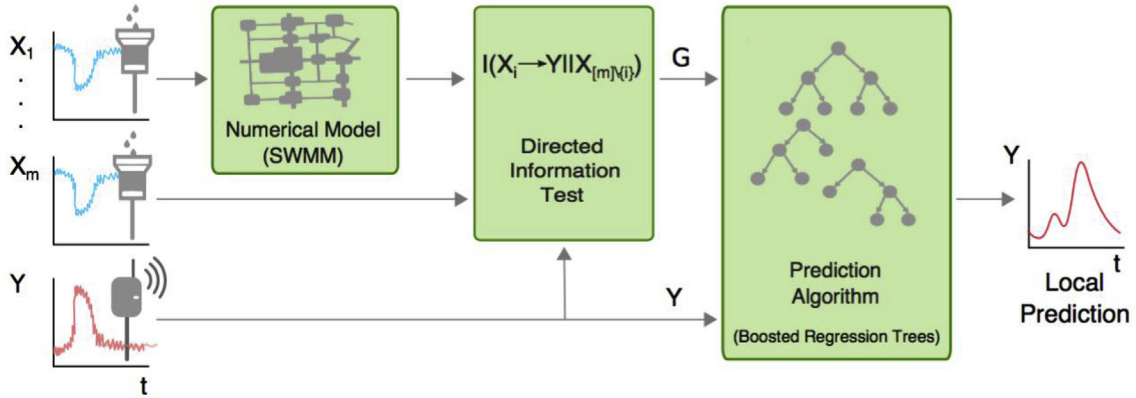


Fig. 1. Predicting flows (Y) by combining inputs features (sensor data) and the outputs of a numerical model (X_1, \dots, X_m). The Directed Information check is used to select only those input features that are statistically causal with the flow measurements (Y). The selected features are then used as inputs to the Boosted Regression Trees prediction algorithm.

The metric of causality, known as *Directed Information* (DI), is an information theoretic quantity that measures the statistical causation. Given a set of random processes $\mathbf{X} = \{X_1, X_2, \dots, X_m\}$ and Y , the *Directed Information* I from X_1 to Y is defined as the time-averaged expected log-likelihood ratio between two conditional probability distributions of Y at time step t , Y^t (Equation (1)). This ratio is also known as Kullback–Leibler Divergence (Kullback and Leibler, 1951). For the conditional probability in the numerator, Y^t is conditioned on the past of X_1 , $X_1^{1:t-1}$ (Marko, 1973; Kramer, 1998):

$$I(X_1 \rightarrow Y | X_2, \dots, X_m) : \\ = \frac{1}{n} \sum_{t=1}^n E_{P_{Y, X_1, X_2, \dots, X_m}} \left[\log \frac{P_{Y^t | X_1^{1:t-1}, X_2^{1:t-1}, \dots, X_m^{1:t-1}}}{P_{Y^t | X_2^{1:t-1}, \dots, X_m^{1:t-1}}} \right], \quad (1)$$

where $X_i^{t_1:t_2}$ denotes the process X_i from time step t_1 to t_2 . If the past of X_1 can help predict the future of Y , then the conditional probability with $X_1^{1:t-1}$, $P_{Y^t | X_1^{1:t-1}, X_2^{1:t-1}, \dots, X_m^{1:t-1}}$ is larger than the conditional probability without $X_1^{1:t-1}$, $P_{Y^t | X_2^{1:t-1}, \dots, X_m^{1:t-1}}$, and the expected log-likelihood of their ratio will be positive. Otherwise, if the past of X_1 cannot help predict the future of Y , then the two conditional probabilities are equal, in which case the expected log-likelihood of their ratio is zero. In other words, the future value of Y is conditionally independent from the past value of X_1 , $X_1^{1:t-1}$ given the past value of the rest processes in \mathbf{X} , $X_2^{1:t-1}, \dots, X_m^{1:t-1}$. For random processes Y and X_1 , the larger the *Directed Information* value, the more causal influence X_1 has on Y – and hence the more “useful” X_1 is on predicting Y .

To reduce potential overfitting, we use a model complexity penalty known as minimum description length (MDL; Grünwald, 2007):

$$MDL = h \frac{\log_2(\mathbf{n})}{2\mathbf{n}}, \quad (2)$$

where h denotes the Markov order and \mathbf{n} denotes the sample size of X_i used for model fitting. The use of this penalty term ensures that only the random processes with *Directed Information* values larger than the MDL are considered as causal. This is summarized in Algorithm 1, which seeks to store all these causal processes in a new subset G , which will be subsequently used as input to a prediction algorithm (modified from Quinn et al., 2015). When selecting candidate features for the *Directed Information* test, non-deterministic relationships among all features need to be

guaranteed—that is, no feature can be derived directly from the others.

Algorithm 1: Assessment of causal influence with directed information

Input : m random processes: $\mathbf{X} = \{X_1, \dots, X_m\}$
 Target random process: Y

Output: Causal influence processes: G

```

begin
  for each  $i \in [m]$  do
     $G(i) \leftarrow \emptyset$ ;
  end
  for each  $i \in [m]$  do
    if  $I(X_i \rightarrow Y | X_{[m] \setminus \{i\}}) \geq MDL$  then
       $G(i) \leftarrow G(i) \cup \{i\}$ ;
    else
      continue;
    end
  end
  return  $G$ 
end
    
```

3.2. Prediction using boosted regression trees

Once the causal features are selected using the *Directed Information* (DI) approach, they can be used to train a predictive model, which in our case takes the form of *Boosted Regression Trees* (BRT). Instead of learning one regression tree, *Boosted Regression Trees* learn multiple trees and weigh them to describe the relationship between the target variable and the features. *Boosted Regression Trees* rely on boosting methods that create an ensemble of regression models to improve the accuracy of model fitting (Elith et al., 2008). Given a regression problem, *Boosted Regression Trees* assign individual weights to every sample point of the training data set. A single regression tree is then constructed and evaluated using the data. A loss function (Wald, 1950), which describes the deviance between the measurements and predicted values, is used to update the individual weights on the tree. The data points with larger deviance are assigned larger weights in the next step. After the updated weights are assigned to individual data points, a new regression tree is constructed. The procedure repeats until the number of the iterations M is reached (Algorithm 2).

Through a forward and additive fashion, *Boosted Regression Trees* gradually optimize predictive performance by using linear combination of all individual trees. Like many supervised ML approaches, *Boosted Regression Trees* can still be subject to overfitting if too many features are used. To this end, the *Directed Information* is used to select only causal features, thereby reducing the potential for overfitting in the final *Boosted Regression Trees*.

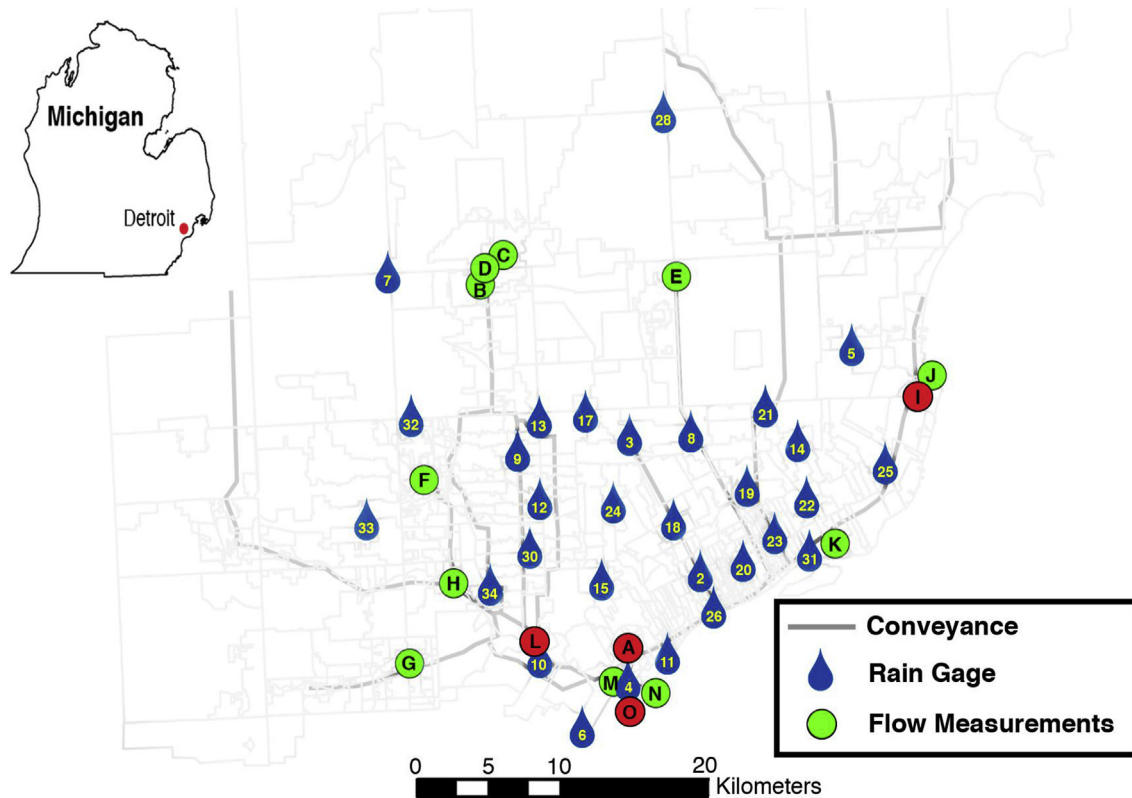


Fig. 2. Detroit sewer collection system service area, showing location of 30 rain gages and 15 prediction points of interest. Site A measures inflow volume to the Wastewater Treatment Plant and sites B–O measure combined sewer overflow volume.

Algorithm 2: Boosted Regression Trees

Input : Features: $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m\}$
 Target variable: \mathbf{Y}
 Number of iterations: M

Output: Ensemble of regression trees: \mathbf{T}

begin

Initialize weights w_i^0 for the measurements of \mathbf{Y} , $\{Y_i, i = 1, \dots, n\}$;

for each $j \in [1, \dots, M]$ **do**

1. Use features \mathbf{X} to fit a regression tree, T_j to the weighted measurements using weights, w_i^{j-1} and yield the predicted values, g^j ;

2. Compute the loss between g^j and the measurements of \mathbf{Y} ;

3. Update the weights based on the loss from step 2, w_i^j .

end

Build \mathbf{T} through the linear combination of T_j ;

return \mathbf{T}

end

4. Case study and implementation

4.1. Study area

Our case study concerns the prediction of flows in large combined sewer systems. Specifically, we focus on the city of Detroit, Michigan, one of the largest combined sewer collection systems in the United States (Fig. 2). We seek to predict daily flows at locations of interest using a large number of available data sources. Detroit and its surrounding suburbs are the largest urban source of total phosphorus to the river system connecting lakes Huron and Erie (Maccoux et al., 2016), and because phosphorus load from this system is driven primarily by flow, predicting and controlling the

occurrence of combined sewer overflows is important. While dozens of new measurements (e.g., rain gages) have become available across the city, they have not yet been used in a predictive model. A physically-based numerical model is available but was updated over 6 years ago. As such, this case study presents a great opportunity to apply our approach to fusing new sensor data with the expertise embedded in the existing numerical model.

4.2. Data source: numerical model

In 1998, a physically-based hydrological and hydraulic model was developed using the EPA Storm Water Management Model (SWMM) for the Detroit sewer collection system's 1963.2 km² service area (Tenbroek et al., 1999) (Fig. 2). The first version of the model was initially calibrated with available flow data (Santini et al., 2001). Since then, it has been updated several times to reflect new facilities, including 14 major combined sewer overflow outfalls (sites B through O in Fig. 2). The latest version of the model was released in 2012, which is the model that has been shared with the authors. Initial inspection revealed that while the model did represent some of the larger, downstream flows adequately (e.g. flows at the final outlet of the system), it generally overestimated daily flows across smaller, upstream locations.

4.3. Data source: sensor measurements

Hourly flow measurements were made by sensors at the terminal node of the system (Fig. 2, site A from April, 2014 to July, 2014), representing the inflows into the Wastewater Treatment Plant. Event-based, combined sewer overflow volume

measurements from May 2013 to October 2015 were also obtained from the Michigan Department of Environmental Quality¹ for 38 storm events during this period. Most combined sewer overflow events occurred within one day; however, when an event spanned multiple days, a daily average was obtained by dividing the total discharge by the number of days of the event. Hourly precipitation data from 2013 to 2015, which served as input into the numerical model, were also obtained from 30 distributed rain gages in the service area (Fig. 2).

4.4. Implementation

The objective of the evaluation was to predict daily inflows to the wastewater treatment plant and sewer overflows at 15 sites. For the purpose of this study, predicting inflows to the treatment plant tests the ability of the approach to describe large-scale, continuous flows, while predicting combined sewer overflows captures the ability to predict smaller-scale, more dynamic events. Data from all rain gages were used as inputs to the SWMM model, after which SWMM outputs were used as input into the *Directed Information* test. The rainfall data were manually quality controlled by gap filling measurements or saturation points through interpolation with neighboring gages. This was intended to ensure that the highest possible quality inputs were used as inputs to the SWMM. For each site, the *Directed Information* test was first used to select causal input features from available rain gages and co-located SWMM outputs. While feasible, upstream flow measurements were not used as inputs to downstream predictions since it was assumed that upstream dynamics are implicitly captured by SWMM. To meet the non-determinism criterion of the *Directed Information* test, one of the gages (site 2) was randomly removed from the data set before the *Directed Information* algorithm was executed. Once causal features were selected, they were forwarded to the *Boosted Regression Trees* algorithm (60/40% split for training and validation). Through iteration, the number of trees was set to 500, while the tree depth was set equal to 4 and the learning rate equal to 0.1.

To promote transparency, experimental repeatability, and broader adoption, all of the source code for this paper is shared in an open source web repository (<http://github.com/kLabUM/DIBRT>). The entire approach has been implemented in MATLAB. Due to security considerations and data agreement with the Great Lakes Water Authority (owner of the data), the authors are unable to share the SWMM model and sensor data. However, an anonymized dataset has been provided in the same web repository to allow others to evaluate the general functionality of our approach.

4.5. Evaluation

To evaluate performance, two fit metrics were used, R-squared (R^2) and Nash-Sutcliffe efficiency (NSE):

$$R^2 = \frac{\left(\sum_{i=1}^n \hat{Y}_i Y_i - n \bar{\hat{Y}} \bar{Y} \right)^2}{\left(\sum_{i=1}^n \hat{Y}_i^2 - n \bar{\hat{Y}}^2 \right) \left(\sum_{i=1}^n Y_i^2 - n \bar{Y}^2 \right)} \quad \text{where } R^2 \in [0, 1]$$

$$\text{NSE} = 1 - \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \times 100 \quad \text{where } \text{NSE} \in (-\infty, 1], \quad (3)$$

where n is the sample size; Y_i and \hat{Y}_i are the measurements and predicted values, respectively; \bar{Y} and $\bar{\hat{Y}}$ are the mean value of Y_i and \hat{Y}_i . For interpretation, an R^2 value closer to 1 indicates a good model fit, while a NSE of 1 indicates perfect fit. To determine if the *Directed Information* step of our approach actually improves final forecasts made by the *Boosted Regression Trees*, two scenarios were evaluated. The first used the *Directed Information* step, as described previously, and the other did not, instead directly feeding all available inputs to the regression tree algorithm. The improvement in R^2 and NSE scores was calculated to determine the benefits of using *Directed Information*.

5. Results

Given the amount of data in the study, four sites have been selected to illustrate and visualize the performance of the approach. The predictive performance across all sites is summarized below, while detailed information and additional figures are provided in the [Supplementary Information](#) of this paper.

5.1. Performance

The performance of the algorithm at site A is shown in Fig. 3, from April 2014 to July 2014. Measured flows are compared to those predicted by the numerical model (SWMM), as well as those predicted by our algorithm. While the SWMM model performed relatively well at this location compared to other sites (NSE = 0.17 and $R^2 = 0.59$), it nonetheless had a positive bias, tending to over-predict peak flows. For this location, which corresponds to the largest conduit in the system (inflow to the treatment plant), the *Directed Information* algorithm selected 7 of the 30 features as inputs to the *Boosted Regression Trees*. Out of these features, SWMM model output was selected as the major source of information, followed by six rain gages (Table 1). There was no clear correspondence between the causal influence of gages and their proximity to the modeled site. Once trained on two months of data, the *Directed Information Boosted Regression Trees* algorithm predicted flows well (NSE = 0.52/ $R^2 = 0.58$). The predictive ability was quite pronounced especially during rainfall events, during which the *Boosted Regression Trees* were able to accurately reconstruct both the magnitude and dynamics of the flows.

Site O is one of the most downstream combined sewer overflows. The frequency of the difference between modeled and predicted overflow volumes for 2013–2015 is compared for SWMM and *Boosted Regression Trees* (Fig. 4a). Since many storms did not result in overflow events, we compare the frequency of prediction residuals (difference between measurements and prediction) rather than a time series comparison. This site had more overflows compared to other locations, even during small storm events. For this location, the SWMM model vastly over-predicted the overflow volumes, by nearly an order of magnitude (NSE = -24.1 and $R^2 = 0.39$). However, the *Boosted Regression Trees* predictions showed much better agreement with the measurements (NSE = 0.62 and $R^2 = 0.61$). For this location, the *Directed Information* algorithm selected 10 total features as inputs to the *Boosted Regression Trees* Algorithm. Interestingly, even though SWMM alone performed poorly, its outputs were still selected as the most informative feature for the predictive model (Table 1). Compared to SWMM, the algorithm reduced the overestimation bias by nearly a factor of 8.

The performance of the algorithm at site I is illustrated by Fig. 4b, which corresponds to a sewer overflow location in the system. For site I, the SWMM model generally over-predicted flows (NSE = -2.71 and $R^2 = 0.65$) and the *Directed Information* step did

¹ Michigan Department of Environmental Quality (MDEQ): <http://www.michigan.gov/deq/>.

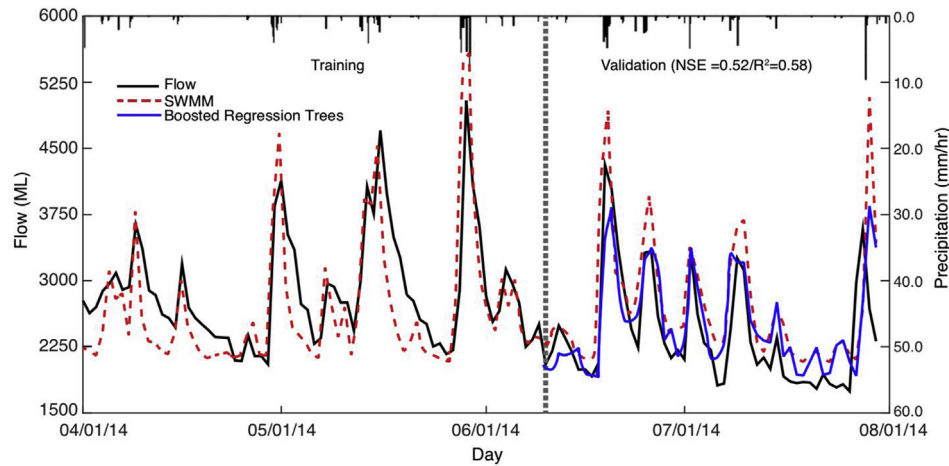


Fig. 3. Comparison of daily flows (million liters, ML) between April, 2014 and July, 2014 at site A (black-solid: measured inflow; red-dashed: SWMM prediction; blue-solid: Boosted Regression Trees prediction). The gray dashed line separates the training and validation phases for the Boosted Regression Trees. The upper part of the figure shows the average precipitation (mm per hour; mm/hr) during this period. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1
List of influential features on the flow measurements and their DI values.

Site	Influential Feature	DI Value ^a
Site A	SWMM-modeled flow	0.99
	Gage 10	0.08
	Gage 12	0.06
	Gage 19	0.06
	Gage 25	0.06
	Gage 9	0.04
Site O	SWMM-modeled flow	0.72
	Gage 30	0.05
	Gage 13	0.04
	Gage 26	0.03
	Gage 34	0.03
	Gage 3	0.02
	Gage 10	0.02
	Gage 21	0.02
	Gage 8	0.01
	Gage 24	0.01
Site I	Gage 19	0.14
	Gage 14	0.04
	Gage 32	0.03
	Gage 18	0.02
	Gage 9	0.01
Site L	Gage 12	0.01
	Gage 30	0.12
	SWMM-modeled flow	0.03
	Gage 3	0.03
	Gage 7	0.01
	Gage 11	0.01

^a For site A, features with DI value no less than MDL = 0.03 were considered as influential; for sites O, I and L, features with DI value no less than MDL = 0.01 are considered as influential.

not select it as an influential input feature. As such, the *Boosted Regression Trees* were trained only on six selected gages, yielding an improved predictive performance (NSE = 0.69 and $R^2 = 0.8$) compared to SWMM.

The measurements at site L (Fig. 4c) showed very few overflows across the 2013–2015 study period. Here, neither the SWMM model (NSE = -0.07 and $R^2 = 0.0$), nor the *Boosted Regression Trees* (NSE = 0.14 and $R^2 = 0.16$) performed well. While the *Directed Information* algorithm selected the SWMM outputs as one informative feature, a number of rain gages were deemed much more informative (Table 1).

5.2. Selection of informative inputs

Overall, the *Boosted Regression Trees* approach, when combined with *Directed Information* feature selection, was able to predict flows at 10 of the 15 sites well, as measured by NSE or R^2 scores (>0.4). The SWMM model was selected as an informative input for 11 of the 15 sites (Fig. 5). The number of inputs selected varied from site to site, with no clear relationships to physical features, such as distance to the input rain gages. Performance was mainly related to the magnitude and variability of flows at the target location. The algorithm generally performed better at locations with more non-zero measurements (e.g. active flows or overflows during every storm). Many of the lower-performing sites generally had mostly no flows or overflows during storms. The variability of flows also played a role in predictive performance. Flow at sites with highly variable flows or overflows (measured by deviation from mean) was more difficult to predict.

The performance of the approach across all sites is summarized in Table 2 as a comparison of the quality of the predictions with and without *Directed Information*. Overall, the use of the *Directed Information* step reduced the fit during the training phase of the *Boosted Regression Trees*, but improved its performance during validation, as quantified by an improvement in NSE and R^2 scores. The use of the *Directed Information* step improved the predictive performance at almost all locations and improved the performance significantly (increased NSE or R^2 by at least 0.05) for more than half of the sites.

6. Discussion

The use of data-driven prediction techniques, such as *Boosted Regression Trees*, shows a good potential for predicting complex and nonlinear flows across large water systems. As more data become available, these methods will offer an automated and efficient way to rapidly ingest and adapt to new sources of information. As shown here, new data sources are not limited to new sensors, such as rain gages. Rather, existing numerical models can serve as valuable inputs. For a given location, if the underlying numerical model already captures flows accurately, the *Boosted Regression Trees* will still improve predictions, but may not outperform the numerical model. This was the case for site A (Fig. 3) where the SWMM model already had fairly strong performance. In such instances, *Boosted Regression Trees* offer a rapid way to ingest new

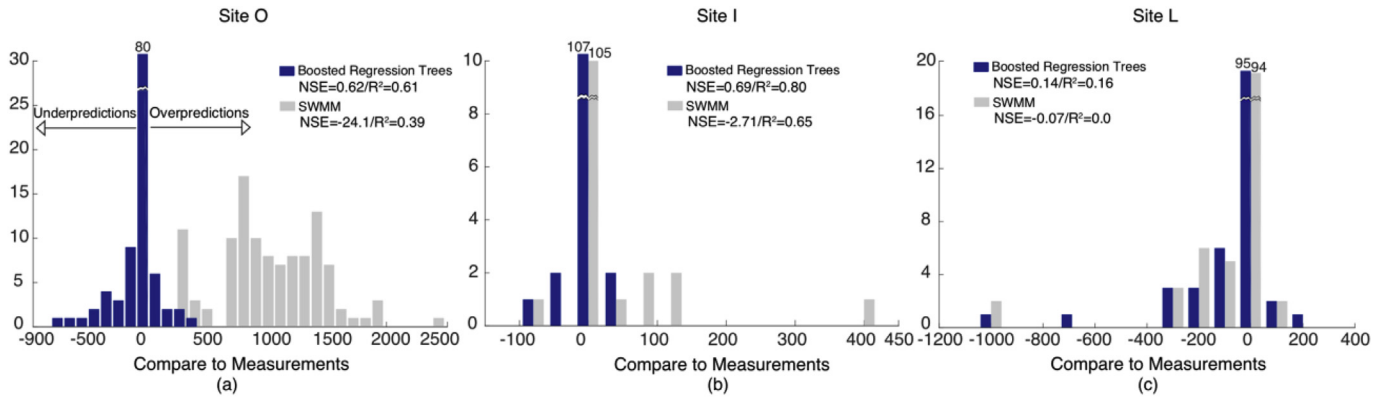


Fig. 4. Histogram of the difference between the combined sewer overflow volume measurements (million liters, ML) to predictions made by the Boosted Regression Trees and the numerical model for (a) site O, (b) site I, and (c) site L, from May and October from 2013 to 2015. Values were obtained by calculating the difference between each prediction and measurement. Similar plots for all other sites are included in the [supplementary information](#) section of this paper.

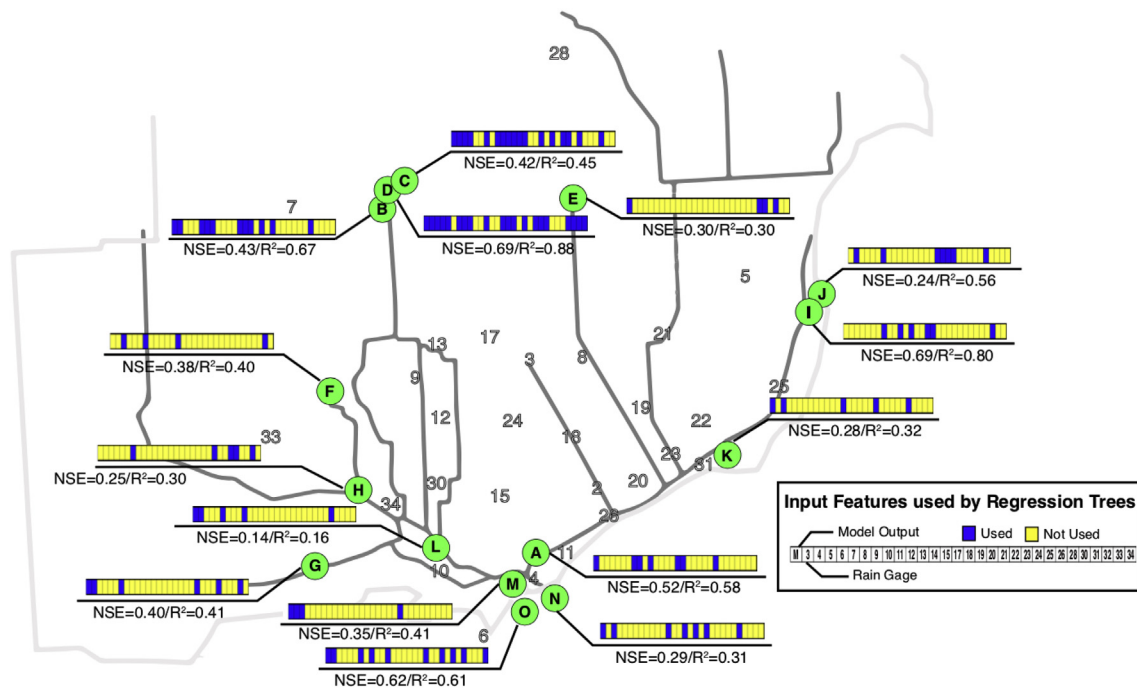


Fig. 5. Evaluation of Boosted Regression Trees performance, showing the fit metrics obtained for each site. The boundary of the service area is outlined in light gray, the Detroit sewer collection system is marked in dark gray, the locations of rain gauges are marked by numbers. The input features used by the algorithm (selected by Directed Information criterion) are indexed in the bar connected to each site. The color-coded bar indicates which features are selected (blue) and which are not (yellow). The first element (M) indicates if the output of the SWMM model was used as an input to the regression trees, while the other elements indicate the number of the rain gauge. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

data sources to “nudge” the outputs of the numerical model to match observations more closely.

Even when the outputs of the numerical model show strong bias or inaccuracy, they may still prove useful when used as inputs to *Boosted Regression Trees* approach. Numerical models, even those that could be considered “out of date”, still embed a lot of information and domain expertise. For example, a numerical model that may not be correct in regard to absolute flow values, may still be correct in regard to timing of flows and their relative magnitudes. This was the case for site O (Fig. 4a), where the outputs of the SWMM model were heavily biased, but were nonetheless selected as the most causal feature. In such cases, the role of *Boosted Regression Trees* is analogous to correcting these biases by using additional sources of sensor data. By extension, if the underlying

numerical model improves as the result of better model inputs or improvement of model structures, or another model becomes available in the future, the data-driven approach should immediately benefit since it does not need to be altered to account for these changes.

As illustrated, the use of more data does not necessarily lead to better predictions. This important point appears to run counter to conventional wisdom on water data, which often assumes that data-driven techniques can arrive at the best answer by ingesting and optimizing around as much data as possible. Rather, ensuring statistical causality between inputs and outputs is important. The use of *Directed Information* provides a reliable and automated way to accomplish this. In our case study, when input features were selected using the *Directed Information* criterion, the performance

Table 2
Measures of model fitting and overfitting by NSE and R² values for all outfalls.

Site	DI Test	NSE(tr)	NSE(val)	R ² (tr)	R ² (val)
A ^a	Yes	0.55	0.52	0.93	0.58
	No	0.56	0.5	0.94	0.57
	Improvement	-0.01	0.02	-0.01	0.01
B ^a	Yes	0.92	0.43	0.94	0.67
	No	0.97	0.32	0.98	0.33
	Improvement	-0.05	0.11	-0.04	0.34
C ^a	Yes	0.76	0.42	0.82	0.45
	No	0.82	0.41	0.88	0.47
	Improvement	-0.06	0.01	-0.06	-0.02
D ^a	Yes	0.94	0.69	0.95	0.88
	No	0.93	0.68	0.95	0.88
	Improvement	0.01	0.01	0	0
E ^a	Yes	0.39	0.3	0.43	0.3
	No	0.45	0.19	0.53	0.34
	Improvement	-0.06	0.11	-0.1	-0.04
F	Yes	1	0.38	1	0.4
	No	1	-0.38	1	0
	Improvement	0	0.76	0	0.4
G ^a	Yes	0.65	0.4	0.75	0.41
	No	0.71	0	0.83	0.11
	Improvement	-0.06	0.4	-0.08	0.3
H	Yes	0.67	0.25	0.83	0.3
	No	0.73	-0.37	0.91	0.09
	Improvement	-0.06	0.62	-0.08	0.21
I	Yes	1	0.69	1	0.8
	No	1	0.59	1	0.76
	Improvement	0	0.1	0	0.04
J	Yes	0.46	0.24	0.72	0.56
	No	0.49	0.2	0.74	0.62
	Improvement	-0.03	0.04	-0.02	-0.06
K ^a	Yes	0.56	0.28	0.74	0.32
	No	0.64	0.35	0.84	0.38
	Improvement	-0.08	-0.07	-0.1	-0.06
L ^a	Yes	0.69	0.14	0.77	0.16
	No	0.73	0.15	0.81	0.17
	Improvement	-0.04	-0.01	-0.04	-0.01
M ^a	Yes	0.98	0.35	0.98	0.41
	No	1	0.28	1	0.31
	Improvement	-0.02	0.07	-0.02	0.1
N ^a	Yes	0.58	0.29	0.76	0.31
	No	0.62	0.26	0.82	0.28
	Improvement	-0.04	0.03	-0.06	0.03
O ^a	Yes	0.84	0.62	0.88	0.61
	No	0.87	0.52	0.9	0.52
	Improvement	-0.03	0.1	-0.02	0.09

^a Model output from the site is selected as an influential feature by DI test; DI test (NO) means all the candidate features are used for training and validation rather than the influential ones selected by DI test (Yes); NSE(R²)(tr) and NSE(R²)(val) are NSE(R²) values for model training and validation; Improvement indicates the difference of NSE(R²) values w and w/o DI test.

of the *Boosted Regression Trees* improved. In many cases, only half or fewer of the available data sources were actually selected for use in the predictions.

The role of the *Directed Information* step in improving predictive performance may be best explained when interpreting the results of the training phase of the *Boosted Regression Trees* (Table 2). The use of more input features may lead to an improved fit during the training phase since more data are available to explain the variability in the target variable (Ng, 1998). However, some of this variability may only be temporary or the inputs may not exhibit causality with the target variable. As such, strong fit during training may lead to worse predictive performance during the validation phase since the predictive algorithm becomes sensitive to non-informative inputs. As opposed to selecting all possible input features, the chance of overfitting during training will thus be reduced when using only informative inputs. While this will lead to a reduction of fit during the training phase, it will often translate to an improvement in fit during the validation phase, as seen in our

study.

This result suggests that the concept of model complexity should be considered more broadly. Complexity of a model is often tied to notions of model structure. As suggested by our case study, when modeling water systems the amount of input data used should also be considered, where more input data may lead to overfitting if not screened ahead of time. As such, the temptation to use all available data when training data-driven water models should be accompanied with a keen appreciation of unintended overfitting.

The SWMM model output was selected as an influential feature by the *Directed Information* algorithm in the majority of the study locations. This is intuitive since the numerical model does embedded a significant amount of information regarding the connectivity and nonlinearities of the system. However, aside from *Directed Information*-based causality, no clear physiographic features explain why some rain gages were selected over others for use in the prediction (Fig. 5). Neither gage proximity to the modeled site nor connectivity via the drainage system were identified as factors that could explain why one gage may have been selected over the others. The challenge in identifying informative gages without the use of a tool such as *Directed Information* may be rooted in the operational complexity of the Detroit sewer collection system. As one of the largest combined sewers in the world, this system contains a large number of control points, in the form of pumps, gates, and valves, which are represented in the numerical model but often operated based on operators' discretion. As such, stochastic uncertainty is embedded in measurements of flow, which limits the ability to deterministically trace the inputs of any given rain input. As such, many observations of flow may thus often be explained by statistical relationships between the input and output data. This, however, plays to be the strength of our approach, which blends statistically-, physically-, or numerically-based mappings.

Through this case study, a number of requirements become apparent when assessing the ability of our approach to work across other systems. First and foremost, the approach will benefit from as many input data as possible – not all, of course, of which will be used. This will improve the likelihood of finding locations that will be causal with the output. Since the *Directed Information* pre-processing step is computationally efficient, ingesting many data inputs can be conducted seamlessly. Once the most informative features are selected, the length of the time record and variability in the output measurements will become important. In general, the time record is a proxy for number of available training storms. While having more storm observations is always better to capture any statistical variability, the size of the storms plays an important role as well. A short time record (a few months or less) will suffice in training the algorithm if the output signal shows a proportional response to a broad range of inputs. For example, site A (Fig. 3) shows a proportional response to a large number of storms, which allowed the *Boosted Regression Trees* to explore a broad output space. The length of the time record will become important especially when predicting sewer overflows. Unlike continually measured sites (e.g. site A), which generally exhibit many non-zero flows, measurements of overflow will primarily be populated with many zero-flow observations. This highly nonlinear behavior challenges a data-driven prediction algorithm because many rain inputs may not be large enough to cause any response. For some sites, large storm events may be less frequent and may thus results in few, but highly variable outputs (e.g. site L). In these instances, the *Boosted Regression Trees* does not have enough relevant training data unless a longer time record is available. For very few sites, this may require years of observations, which were not available in our study. This data requirement does not, however, change the

implementation of our toolchain, as more data can simply be ingested as they are measured.

Another important consideration when applying this approach relates to the temporal granularity of predictions. In our study, daily flow and volume measurements were available. This placed a bound on the temporal resolution of the predictions. This daily resolution still has utility in our case study since Detroit's sewer system is one of the largest in the world. The system dynamics play out over relatively long time scales, as can be seen in Fig. 3. Storms often lead to flow responses or overflows that can last multiple days, which means that daily forecasts have utility in treatment planning, collection system dewatering, and overflow operations. Higher-order dynamics are obscured or averaged at such resolutions, which may be important for some smaller sewer system or other applications. If higher resolution forecasts are desired, our approach would require higher resolution flow data. While an additional analysis would be required to assess performance across these time scales, the toolchain could be applied to these data without requiring modifications. This is particularly true about the *Directed Information* step of the approach, which can still be used to determine which input–output relationships are causal. Depending on the dynamics in higher resolution measurements, the boosted regression trees could be replaced with more dynamically-based predictive approaches. This presents good opportunities for future studies, which will be carried out across smaller and more rapidly changing systems.

Overall, the approach presented in this study stands to provide a number of benefits to decision makers and modelers. From an operational perspective, the toolchain provides an automated method by which cities and municipalities can leverage all their existing and emerging data sources to improve forecasts of flows. This will be particularly useful in operational situations where an existing numerical model may not provide sufficiently fine-grained warnings of floods or impending overflows. However, the resulting predictive model should not be used for infrastructure planning purposes (changing pipe diameters, evaluating new designs, etc.) since all relationships are statistical and inherent to the data of the existing system. For these purposes, the numerical model will rather need to be updated and recalibrated. To this end, our approach can also serve as a tool to guide model calibration. The use of *Directed Information* can serve as an alternative to traditional metrics, such as NSE or R^2 , providing insight not only on fit but rather the causality between the model structure and observation. The *Directed Information* criterion could then be used to help determine which inputs may be most important to the numerical model, which may reduce amount of inputs and time spent on calibration.

Finally, the *usefulness* of any particular data source must be viewed holistically. A non-causal relationship may suggest that an input data source is not relevant or of sufficient quality to explain a particular output. However, utility and causality are two-way properties. Namely, a flow measurement (the output) may not be informative to begin with. If this is the case, the inputs may still be useful for forecasting at other locations. A level of user discretion should thus be exercised when evaluating input and output pairings. The *Directed Information* step will help in this regard, as it will provide a first check to determine if certain input–output pairings should even be considered before a predictive model is constructed. Ultimately, the quality of the final prediction, which can be evaluated using more classic fit metrics or specific requirements of an application, will remain a good proxy for utility of any particular forecast and thus, implicitly, the utility of the underlying data. While not conducted in this case study due to the temporal granularity of data and the assumption that SWMM model captures upstream dynamics, future studies could also consider the value of

flow measurements as predictors for other flow measurements.

7. Conclusions

This paper introduced a holistic, data-driven toolchain based on *Directed Information* and *Boosted Regression Trees* to provide flow forecasts across urban drainage systems. More broadly, this methodology should also work well for other types of water systems where many data or numerical models are available. It was demonstrated that the use of more data is not always advantageous and may often lead to worse predictions. Rather, a computationally-efficient pre-processing step (*Directed Information*) will be important in selecting only those input data that are informative to the overall prediction. The approach based on *Boosted Regression Trees* was also shown to be effective at learning complex and non-linear mappings between rainfall inputs and flow. More importantly, it was demonstrated that the outputs of a numerical model could also be used as an important input to the data-driven approach. Even if a numerical model is no longer fully calibrated due to aging or changes in the system, it still embeds valuable information that can improve the predictive performance of the regression trees. This will provide a rapid and automated way for city managers to use a diverse set of information, which may be at their disposal, without requiring the often-expensive recalibration of numerical models. Naturally, if a numerical model does improve, so will the predictions of our approach. This discovery will be important as the push for “smart” water systems and “big water data” continues. Future work should be carried out to determine how consistent our findings are across other study areas and other types of water systems.

Acknowledgements

This work was funded by the Fred A and Barbara M Erb Family Foundation grant number 903 and the University of Michigan Graham Sustainability Institute. We appreciate the SWMM model and data provided by the Great Lakes Water Authority.

Appendix A. Supplementary data

Supplementary data related to this article can be found at <https://doi.org/10.1016/j.watres.2018.09.009>.

References

- Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 161–168.
- Doherty, J., 2015. Calibration and Uncertainty Analysis for Complex Environmental Models. Watermark Numerical Computing.
- Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., et al., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36 (1), 27–46.
- Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. *J. Anim. Ecol.* 77 (4), 802–813.
- Field, R., Tafuri, A.N., 2006. The Use of Best Management Practices (BMPs) in Urban Watersheds. DEStech Publications, Inc.
- Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* J. Econometric Soc. 424–438.
- Grünwald, P.D., 2007. The Minimum Description Length Principle. MIT press.
- Kerkez, B., Gruden, C., Lewis, M., Montestruque, L., Quigley, M., Wong, B., et al., 2016. Smarter Stormwater Systems. ACS Publications.
- Konrad, C.P., 2003. Effects of Urban Development on Floods. US Geological Survey.
- Kramer, G., 1998. Directed Information for Channels with Feedback. Eidgenössische Technische Hochschule Zurich.
- Kullback, S., Leibler, R.A., 1951. On information and sufficiency. *Ann. Math. Stat.* 22 (1), 79–86.
- Maccoux, M.J., Dove, A., Backus, S.M., Dolan, D.M., 2016. Total and soluble reactive phosphorus loadings to Lake Erie: a detailed accounting by year, basin, country, and tributary. *J. Great Lake. Res.* 42 (6), 1151–1165.
- Marko, H., 1973. The bidirectional communication theory—a generalization of

- information theory. *IEEE Trans. Commun.* 21 (12), 1345–1351.
- Morales, V.M., Mier, J.M., Garcia, M.H., 2017. Innovative modeling framework for combined sewer overflows prediction. *Urban Water J.* 14 (1), 97–111.
- Ng, A.Y., 1998. On Feature Selection: Learning with Exponentially Many Irrelevant Features as Training Examples. Massachusetts Institute of Technology.
- Paquier, A., Mignot, E., Bazin, P.-H., 2015. From hydraulic modelling to urban flood risk. *Procedia Eng.* 115, 37–44.
- Quinn, C.J., Kiyavash, N., Coleman, T.P., 2015. Directed information graphs. *IEEE Trans. Inf. Theor.* 61 (12), 6887–6909.
- Santini, A., Brink, P., Sherman, B., TenBroek, M., 2001. An equivalent rainfall technique to identify dry weather flow data for city of Detroit flow balance analysis. *Models and Applications to urban water systems*. Monograph 9, 259.
- Schalkoff, R.J., 1997. *Artificial Neural Networks*, vol. 1. McGraw-Hill, New York.
- Sun, N.-Z., Sun, A., 2015. *Model Calibration and Parameter Estimation: for Environmental and Water Resource Systems*. Springer.
- Tenbroek, B.M., Bunyan, R.J.T., Whiting, G., Redman-White, W., Uren, M.J., Brunson, K.M., Edwards, C.F., 1999. Measurement of buried oxide thermal conductivity for accurate electrothermal simulation of SOI device. *IEEE Trans. Electron. Dev.* 46 (1), 251–253.
- Wald, A., 1950. *Statistical Decision Functions*.