Case Study of the Chicago River Watershed: Physical Modeling vs Data-driven Modeling of an Urban Watershed

Naila Mahdi,[1] Haithum Elhadi[1] and Krishna R. Pagilla[2]

[1]Illinois Institute of Technology, Chicago, Illinois; [2]University of Nevada, Reno, Nevada.

Abstract

We developed a water quality model for the highly urbanized Chicago River watershed based on hydrologic simulation using BASINS/HSPF. Appropriate consideration was given to the effective impervious area (EIA). The 5 y water quality simulation resulted in finding total nitrates loadings at both point and nonpoint sources. However, it is always useful to have modeling alternatives to validate the simulation results of a physically based model with a data-driven one. Data-driven modeling has gained a lot of attention in recent decades in both hydrology and water resources research. While physically based models require the description of system inputs, physical laws and boundary and initial conditions, a data-driven model simply extracts knowledge from a large amount of data with only a limited number of assumptions about the physical behaviour of the system. For this case study, both data-driven and physical models were considered to simulate total nitrates. Comparing the performance of the two modeling approaches, the data-driven models show better performance. RMSE for regression models showed an increase in prediction performance of up to 10.7 %. Data-driven models require fewer inputs and can be deployed anywhere in the watershed, while physical models require extensive data inputs and can only be applied to the specific watershed outlets selected in the simulation. These arguments suggest the complementary use of both physical and data-driven models. The physical model can be a planning tool whenever significant physical change takes place in the watershed. The data-driven model can be an operating tool that can be periodically used to inspect the watershed water quality parameters, especially if TMDL and WQS are established for the watershed.

1 Introduction

Watershed models are fundamental to water resources assessment, development and management. They are useful in developing models that describe complex natural processes and complicated systems through sets of equations that explain the problems and solve them. Estimates of nutrient concentrations, loads and yields are useful for analyzing a water body and help to identify source areas and to develop mitigation strategies. Nutrient loads that are transported by a stream during a given period of time are particularly important when considering the quantity of nutrients that enter a lake or reservoir. Load estimates are essential for the establishment and monitoring of total maximum daily loads (TMDLs) as mandated by the Clean Water Act (CWA). Resource and regulatory authorities use yield estimates to help prioritize efforts with regard to land use management and best practices.

Tools, such as hydrological models, that are coupled with geographic information systems (GIS) and remote sensing provide powerful techniques for conducting these kinds of studies (Zoppou 2001; Conway and Lathrop 2005; Wang et al. 2005; Wu et al. 2006; Singh et al. 2011; Yu et al. 2009; Jeon et al. 2007). Other integrated approaches involve the use of statistical and spatial analyses, as well as hydrologic modeling, to examine the effects of land use on water quality (Tong et al. 2007; Tong and Chen 2002). Most researchers depend on field studies that focus on a geographically local scale or that have a small set of land use patterns (Wilson and Weng 2011; Akhavan et al. 2010; Leon et al. 2010). Continuous hydrologic models consider the whole hydrologic cycle and the effects of long term hydrological changes and processes.

BASINS 4.0 was selected to assess the water quality in the watershed. The software is a multi-purpose environmental analysis system that integrates a geographical information system (GIS), national watershed data, state-of-the-art environmental assessment, and modeling tools (such as HSPF, SWAT or SWMM) into one convenient package.

The software promotes better assessment and integration of point and nonpoint sources for management, planning and decision-making. BASINS is a watershed based water quality assessment tool that has been widely accepted in many watershed studies (Tong et al. 2007; Tong and Chen 2002; Tong et al. 2009; Luzio et al. 2002; Fohrer et al. 2001).

HSPF is a watershed scale conceptual model. HSPF performs continuous simulation of hydrology and water quality, and performs flow and water quality routing in the watershed reaches. HSPF is extensively used to model urbanized watersheds

Mahdi, Naila, Haithum Elhadi and Krishna R. Pagilla. 2017. Case Study of the Chicago River Watershed: Physical Modeling vs Data Driven Modeling of an Urban Watershed. Journal of Water Management Modeling 26:C431. https://doi.org/10.14796/JWMM.C431
 www.chijournal.org ISSN 2292-6062.

(Fonseca et al. 2014; Im et al. 2003; Shirinian-Orlando and Uchrin 2007; Wicklein and Schiffer 2008). It is the most comprehensive and flexible hydrology and water quality model available (Bergman et al. 2002; Mohamoud et al. 2010). However, other studies suggest that using the urban land use as a nonpoint source for nutrients can give invalid results because of the impervious cover in urban areas and the way drainage is frequently routed to wastewater treatment plants and then discharged to local rivers as a point source (Ahearn et al. 2005).

Since accurate estimates of runoff volume are important in order to estimate pollutant loads, the effective impervious area (EIA), as a fraction of the total impervious area (TIA), should be determined for use in hydrological models (Sutherland 2000; Deacon et al. 2005; Brabec et al. 2002). Impervious area is a rough measure of the total watershed that is utilized by human activities. EIA is the portion of TIA within a watershed that is partially or totally connected to the drainage collection system. Street surfaces, parking lots, paved driveways and sidewalks, and rooftops that are directly connected to the storm sewer system are all included in EIA (Sutherland 2000). For urban runoff modeling or hydrologic analysis, the EIA for a given basin is usually less than the TIA; however, in highly urbanized basins, EIA values can approach and equal TIA (Deacon et al. 2005). Field measurements, empirical equations, and calibrated computer models are some ways to determine EIA (Sutherland 2000; Deacon et al. 2005).

The huge amounts of data collected daily from monitoring systems and the exponential growth and advances in information systems have resulted in an increased use of data mining to generate models that can explain physical systems. Data-driven modeling is the study of mathematical algorithms that improve automatically through experience and training. Data-driven modeling has developed with contributions from areas such as artificial intelligence, machine learning, data mining, knowledge discovery and pattern recognition. The most-used models are artificial neural networks, fuzzy rule based systems, and statistical methods. Data-driven modeling has gained a lot of attention in the last decades in both hydrology and water resources research (Pries and Ostfeld 2008).

A data-driven model simply extracts knowledge from large amounts of data with only a limited number of assumptions about the physical behaviour of the system. This modeling approach can only be considered feasible if sufficient data is available.

Data-driven modeling has been used in areas such as rainfall–runoff and groundwater modeling (Fallah-Mehdipour et al. 2014; Tokar and Markus 2000; Solomatine and Dulal 2003; Muttil and Liong 2004; Solomatine et al. 2007); flood forecasting (Chen and Yu 2007; Chiang et al. 2007); streamflow prediction, and river management (Preis and Ostfeld 2008; Marsili-Libellia et al. 2013; Mouton et al. 2009; Asefa et al. 2006). Water quality constituents have also been predicted using data-driven models in many studies (Preis and Ostfeld 2008; Solomatine et al. 2007; Ghavidel and Montaseri 2014; Burchard-Levine et al. 2014). They are effective in building knowledge-driven simulations, that are capable of extracting different system states when the nature of complex relationships is unknown, or when the available models are inadequate (Solomatine et al. 2007). It is always useful to have modeling alternatives and to be able validate the simulation results of physically based models with data-driven ones, or vice versa (Preis and Ostfeld 2008; Solomatine and Dulal 2003).

In this project, both physical and data-driven models were developed for the Chicago River watershed to predict total nitrates (NO_3 + NO_2). The performances of the two modeling approaches were compared and assessed based on the results.

2 Study Area

The Chicago River basin (hydrologic unit code 07120003) is the smallest part (6%) of the Upper Illinois River basin (UIRB). UIRB is part of the Mississippi River basin, which is the world's second largest drainage basin and altogether includes more than 40% of the land area in the contiguous United States. The significance of the Chicago River basin lies in its navigable systems, and in particular, the Chicago Sanitary and Ship Canal, which provides a link between Lake Michigan and the Mississippi River. Population in the basin has grown steadily over the years leading to urban and industrial growth. The Chicago River watershed is approximately 82% urban land use. As a result of the urban growth, major changes in the region have taken place and have significantly affected the quality of surface waters. Wastewater disposal and storm runoff became serious issues in the watershed.

Surface water issues related to urbanization include point and nonpoint sources of sediment, nutrients, trace elements and organic compounds; streamflow alterations; and the health and community structure of aquatic biota. Development also alters runoff patterns by changing the lay of the land and thus drainage patterns, which can result in a dramatic increase in the rate and volume of stormwater runoff and a reduction in groundwater recharge.

The changes in land cover, the increase in construction activities that result in compact soils and smooth natural grades, reduced native vegetation, enlarged storm sewer systems, and lined channels all add to the conveyance of greater volumes of runoff downstream at much faster rates (MWRDGC 2007).

Much of the pollutant load in the runoff originates from impervious surfaces, particularly roadways and parking lots. Some of the more common water quality impacts of stormwater runoff are sediment contamination, nutrient enrichment, toxicity to aquatic life, bacterial contamination, salt contamination, impaired aesthetic conditions, and elevated water temperatures.

In general, nutrient loads—nitrogen and phosphorus—were greatest from the urban center of the Chicago metropolitan area, reflecting the effect of wastewater return flows to the Chicago River and the Chicago Sanitary and Ship Canal. The ship canal was also observed to carry the majority of ammonia and phosphorus loads during low flow conditions. It is considered the

main nutrient contributor to the Illinois River and thence the Gulf of Mexico dead zone, the largest hypoxic zone ever measured.

3 Methodology

3.1 Data Sources and Types

For this study a local data warehouse (DW), which aggregates different available data types from various agencies in the watershed, was created as part of the construction of a framework for modeling the Chicago River watershed. The framework includes the data warehouse, data analysis and watershed assessment, physical and data-driven modeling, and optimization. The DW makes it easy to access, retrieve, fill data gaps, analyze and manage data records in the watershed. This helps to integrate the data and to meet different requirements such as watershed assessment and physical modeling. Different water quantity, water quality and land use data were compiled from sources which include the U.S. Geologic Survey (USGS), the Metropolitan Water Reclamation District of Greater Chicago (MWRDG), the Chicago Metropolitan Agency for Planning (CMAP), the U.S. Army Corps of Engineers—Chicago District (USACE), the BASINS data store, and several sources that are permitted through National Pollutant Discharge Elimination System (NPDES) permits, including facilities, treatment plants and combined sewer overflows (CSOs). Figure 1 shows the data sources.

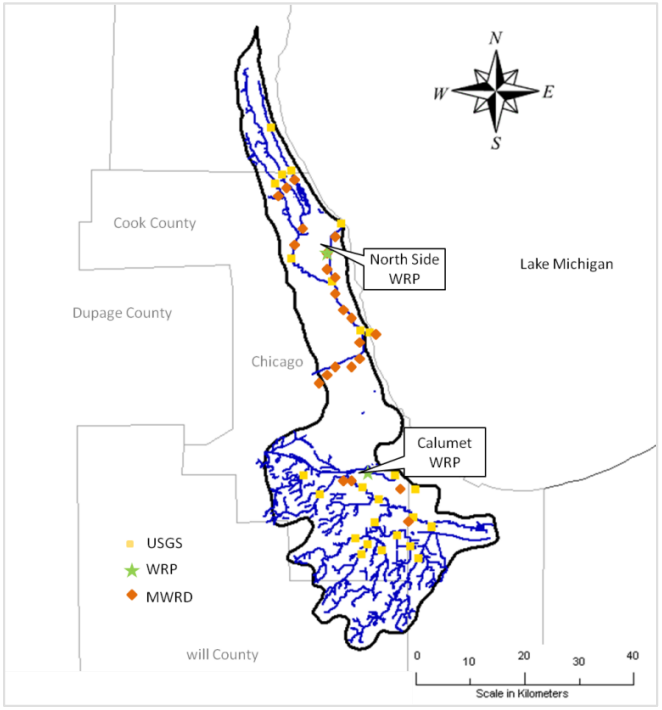

Figure 1 Locations of data sources.

3.2 Physical Modeling

In this section, we describe the development of a water quality model to quantify the effect of detailed (level III) land use on nutrient loading in the Chicago River watershed using BASINS/HSPF.

Nutrient export coefficients that relate the detailed land uses to water quality were obtained from the calibrated and validated model. The following subsections outline the steps carried out to fulfill the objectives of the simulation process. They explain the types and sources of data used and how hydrologic and water quality models were constructed and used in the BASINS/HSPF model environment.

Watershed Simulation

BASINS data layers that can be provided to HSPF include: digital elevation model (DEM) grid data; National Land Cover Data (NLCD or GIRAS) land use data; reach files; permit compliance system (PCS) data; meteorological data; and STORET and USGS data. In order to run HSPF, the observed meteorological data, water quality data and flow data must be formatted to watershed data management (WDM) files that contain time series data required by HSPF. All input data, except for time series, are contained in a user control input (UCI) file. This file contains all the parameter values and control specifications needed to run the HSPF model. For evaluation of the model, all the calibration and validation analysis was performed using the GenScenario tool in the BASINS package.

Delineation is part of a segmentation process that is required by HSPF. The watershed is divided into segments which are analyzed. Delineation is used to determine a contributing watershed area for a specific outlet or to divide the watershed into subbasins. The delineation is either automatic, using DEM grids, or manual, where existing streams and basins are empirically selected and used to determine the watershed. For this study automatic delineation was used. The delineation process determined the three GIS layers that are required to run the HSPF model: streams, subbasins, and outlets.

WinHSPF divided the Upper Chicago River subbasin into homogeneous land areas (hydrologic response units, HRUs). The HRUs were used to define 6 reaches and 7 subwatersheds. The hydraulic characteristic of each reach were defined by parameters in the function tables FTABLES, that represent volume–discharge relationships for each reach. A fixed relationship was assumed between water level, surface area, volume and discharge. HRUs can be impervious or pervious areas, which, once determined, are modeled independently.

Each HRU requires input data, such as meteorological data and parameters related to land use, soil characteristics to simulate hydrology, sediments, and nutrients (Donigian et al. 1995). The main simulation modules are PERLND, IMPLND and RCHRES and they simulate pervious land segments, impervious land segments and free flow respectively (Donigian et al. 1995).

Since accurate estimates of runoff volume are essential for the accurate estimation of pollutant loads, the EIA as a percentage of TIA should be determined for basins that are directly connected to the drainage systems (Sutherland 2000). TIA is

determined using two common methods: land use or zoning maps; and aerial photography (Jones et al. 2003). Most current impervious surface studies rely on the methods that correlated percentage impervious surface with land use largely by using estimates of the proportion of imperviousness within each class. Tables 1 and 2 show the TIA and EIA percentages, respectively, adopted for this study based on literature (Brabec et al. 2002).

Table 1 Some of the TIA percentages adopted for this study, based on the literature.

Land Use Category	(TIA)%
Agricultural	0
Commercial	85
Forest	0
Industrial	85
Multi-Family Residential	50
Single-Family Residential	35
Public Open Space	0
Roads	85
Schools	50
Vacant	0
Water	100

Table 2 EIA equations used for calibration and sensitivity analysis of the study area.

Source EIA	Value
Alley et al.	$(0.15 \times TIA^{1.41})$
Laenen	$(3.6 + 0.43 \times TIA)$
Sutherland, Highly connected basins	$(0.4 \times TIA^{1.2})$
Sutherland, Totally connected basins	(TIA)

Flow Simulation

Flow is the first component to be simulated. PWATER and IWATER are the modules used for flow simulation. PWATER calculates the components of the water budget and predicts the total runoff from pervious land segments. IWATER simulates the retention, routing and evaporation of water from impervious land segments. The instream hydraulic behaviour is simulated by HYDR.

For each reach, a fixed relationship is assumed among water level, surface area, volume, and discharge. Instream simulation is based on the assumption of a completely mixed system with unidirectional longitudinal flow simulation. The hydraulic characteristics of reaches in the model are defined by parameters in the function tables (FTABLES) that represent volume discharge relations for reaches. Parameters needed for the simulation such as *nominal upper zone storage*, *nominal lower zone storage*, *soil moisture infiltration rate*, *percent vegetation cover* of each land use type, and *groundwater recession rate* were given BASINS default values or literature values and later adjusted during hydrologic calibration.

Water Quality Simulation

The simulation was done using the HSPF modules PQUAL and IQUAL for pervious and impervious land segmentsrespectively.

PQUAL and IQUAL simulate the pollutants using one of two methods: either by direct washoff by overland flow, where the constituent is simulated using the basic depletion and accumulation rate; or by washoff associated with detached sediments, where the constituent is simulated as a function of sediment removal. The first approach was adopted for all the species since the study area is largely impervious and the nutrients will have washed off with overland flow.

HSPF simulates several physical, chemical and biological processes within a stream reach using the RCHRES module. The reaches are assumed to be completely mixed and the flow is unidirectional. Point sources were added in the HSPF simulation. The two known NPDES that could be added to the watershed are North Side WRP and Calumet Water WRP.

Model Calibration and Validation

Hydrologists need to evaluate model performance to provide a quantitative estimate of the model's performance and predictive ability (Krause et al. 2005). No commonly accepted modeling guidance has yet been established, although the American Society of Civil Engineers (ASCE) has emphasized the need to clearly define model evaluation criteria since in 1993 (Donigian 2002). A weight of evidence approach is commonly accepted and was used to examine and assess model performance. For these reasons multiple model comparisons, both graphical and statistical, are preferred (Donigian 2002).

For this study, model performance and calibration–validation are evaluated through qualitative and quantitative measures, involving both graphical comparisons and statistical tests. The calibration–validation process is a hierarchal process that starts by developing parameters, followed by hydrology calibration–validation and finally water quality calibration–validation. Of the standard regressions, Pearson's coefficients of correlation (r) and determination (r^2) were used. These coefficients describe the degree of co-linearity between simulated and observed data. The regression coefficients are given by the following equations

$$r^2 = \left[\frac{\sum_{i=1}^{n} \left(O_i - \overline{O}\right)\left(S_i - \overline{S}\right)}{\sqrt{\sum_{i=1}^{n} \left(O_i - \overline{O}\right)^2} \sqrt{\sum_{i=1}^{n} \left(S_i - \overline{S}\right)^2}} \right]^2 \quad (1)$$

where:

r = Pearson coefficient,
O_i and S_i = observed and simulated values respectively, and
$\overline{O}$ and $\overline{S}$ = the means of observed and simulated values respectively.

For model performance, r ranges from −1 to 1 and for r^2 the values range from 0 to 1. Generally, a value >0.5 is considered acceptable (Donigian 2002).

The fact that only the dispersion is quantified is one of the major drawbacks of r^2 if it is considered alone. A model which systematically over- or under-predicts will still result in good r^2 values close to 1.0 even if all predictions were wrong. Another model evaluation criterion is the Nash–Sutcliffe efficiency coefficient (Krause et al. 2005). It is calculated as:

$$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - S_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2} \quad (2)$$

The range of *NSE* is from1 (perfect fit) to $-\infty$.

An efficiency <0 indicates that the mean value of the observed time series would have been a better predictor than the model. The largest disadvantage of NSE is the fact that the differences between the observed and simulated values are calculated as squared values. As a result, larger values in a time series are strongly overestimated whereas lower values are neglected (Krause et al. 2005).

Root mean square error (RMSE), normalized root mean square error (NRMSE) and mean absolute error (MAE) are other statistical indices that can be used to evaluate model performance. They are given by the following equations:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i - S_i)^2}{n}} \quad (3)$$

$$NRMSE = \frac{RMSE}{O_{\max} - O_{\min}} \quad (4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|S_i - O_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (5)$$

where O_{max} and O_{min} are the maximum and minimum observed values and e_i is the absolute error.

RMSE and MAE measure the aggregated difference between simulated values and observed values. Values close to zero indicate better performance.

Percent mean error (PME) is a general calibration–validation measure that has been provided to HSPF model users to be used in model performance evaluation (Donigian 2002). The tabulated values provide general guidance, in terms of the percent mean errors or differences between simulated and observed values, so that users can determine the level of agreement or accuracy (i.e. *very good*, *good*, *fair*) that might be expected from the model application (Donigian 2002). Table 3 shows percent mean error (PME) target values for different modeling processes.

Table 3 General PME calibration and validation target values for HSPF applications (Donigian 2002).

	Very Good	Good	Fair
Hydrology/Flow	<10	10–15	15–25
Sediment	<20	20–30	30–45
Water Temperature	<7	8–12	13–18
Water Quality/Nutrients	<15	15–25	25–35
Pesticides/Toxins	<20	20–30	30–40

3.3 Data-driven Modeling

This section introduces data mining (DM), from the field of artificial intelligence, to estimate total nitrates for the Chicago River watershed. DM models consist of a set of mathematical relationships. DM tasks are divided into two major classes: predictive tasks and descriptive tasks. Predictive tasks are those in which a particular attribute is predicted based on the value of other attributes. The attribute to be predicted is the dependent variable while the attributes used for making the prediction are independent variables. For descriptive tasks, the objective is to develop patterns (e.g. correlations, trends) that summarize the relationships in data, and which are often exploratory in nature. These tasks usually require post-processing techniques to validate and explain the results. Predictive models are divided to classification models, which are used for discrete target variables, and regression models, which are used for continuous target variables (Tan et al. 2006).

There are many methods to construct predictive and classification models such as naive Bayesian, support vector machines, decision tree, neural network, and *k*-nearest neighbor classifications. Regression is the statistical methodology that is most often used for numeric predictions. Both prediction and classification are supervised learning problems where there is an input *X* and an output *Y*, and the model learns the mapping from the input to output. The approach in DM is that a model defined up to a set of parameters, is assumed:

$$y = g(x|\theta) \quad (6)$$

where:

y = prediction or regression,
g = the model,
x = the model input, and
θ = the model parameters.

The DM program optimizes these parameters so that the approximation error is minimized and the estimates are close to the correct values given in the training set. For the Chicago River watershed, we developed data-driven models (using different data mining techniques) to estimate nutrient concentration based on some watershed parameters such as stream flow, precipitation, air temperature, water temperature, dissolved oxygen, turbidity, areas of different land use types, month of year, and others.

DM is part of the knowledge discovery in database (KDD) process. It consists of series of mining steps. For this case study, the open source Waikato Environment for Knowledge Analysis (WEKA) software package was used. It provides a comprehensive collection of DM algorithms and data preprocessing tools that together provide a framework to compare the different algorithms. WEKA has several graphical user interfaces that enable easy access to the underlying processes. The main graphical user interface is the Explorer. It has a panel based interface, where different panels correspond to different data mining tasks such as

preprocess, where data can be loaded from various sources including files and database; and *classify* which gives access to WEKAs different classification and regression algorithms. The panel also provides access to graphical representations of model prediction errors in scatter plots, and allows evaluation via ROC curves and other threshold curves (Hall et al. 2009).

Pre-Processing

Examples of data pre-processing are data cleansing, data integration, and data transformation. Data pre-processing includes the tracking of incomplete data, data that lack certain attributes or values, filling missing or incomplete values, removing errors and outliers, and resolving inconsistencies in data. This process ensures quality data, which will in turn ensure quality results. Descriptive data summarization provides the analytical foundation for data pre-processing.

The basic statistical measures for data summarization include measurements for the central tendency of data such as mean, weighted mean, median and mode; and measurements for data dispersion such as range, quartiles, variance and standard deviation. Graphical representations such as histograms, boxplots, quantile plots, and scatter plots facilitate visual inspection of the data and are useful for data pre-processing and data mining. Data transformation routines are used to convert the data into forms that are suitable for mining. Histograms are highly effective at approximating both sparse and dense data as well as highly skewed and uniform data, and can capture dependencies between attributes. They use binning to approximate data distributions. Datasets for analysis may contain hundreds of attributes, many of which may be irrelevant to the mining task or even redundant. Their inclusion may slow down the mining process and result in the discovery of patterns of poor quality. Various statistical significant tests and techniques, which assume that the attributes are independent of one another, can be performed to select *best attributes* data subsets.

Model Building and Evaluation

This section describes the selection and application of various models that are developed using comparable analytical techniques, and the adjustment of model parameters until optimal values are reached. Input data are randomly partitioned into two independent sets, a training set and a test set. The training set is used in the construction of the model with an accuracy estimated using the test set. This is called the holdout method. The random subsampling method is a variation of the holdout method, in which the method is repeated *k* times and average accuracy is considered. In *k*-fold cross validation, the input data are randomly partitioned into *k* sets, or folds, of approximately equal size. Training and testing are then performed *k* times. Each sample is used the same number of times for training and once for testing. The error is then calculated as the average of the error rates of all the *k* iterations (Han and Kamber 2006). The 10-fold cross validation method was used for building all the models in this study.

Model Attributes

The attributes were selected based on their physical nature and whether they are real time frequently-measured data such as daily flow, air temperature and hourly precipitation; whether they are measurements of specific conductances such as pH, water temperature, dissolved oxygen, turbidity and total chlorophyll; whether they are time-consuming chemical or biological test measurements such as BOD and COD; or whether they are related to the land use of the source. The choice of which of these attributes to select for data-driven models to predict total nitrates was made using the assumption that they would give relevant and useful information and thus good discovered patterns. Table 4 shows the properties and gives descriptive summaries of the predictors. For the Chicago River watershed, a histogram analysis strategy was used to visualize the attributes data; for outliers, 2% of the top and the bottom data were removed. Any missing values were replaced by mean values. The 10-fold cross validation method was used to partition training and testing data sets for all the predictive models used for this study. There were 905 samples, and 154 attributes were investigated.

Table 4 Properties of predictors.

Attribute	Description	Unit	Mean	Min	Max	Stdev
MONTH	NUM Number of the month	NA	NA	1	12	NA
DO	Dissolved oxygen	mg/L	7.198	0	15	2.67
NITRATE	Total nitrate	mg/L	2.686	0	11.98	2.903
TOT_P	Total phosphorous	mg/L	0.966	0	74	4.128
TKN	Total Kjeldahl nitrogen	mg/L	1.979	0.2	88	3.741
TURB	Turbidity	NTU	21.28	2.8	312	32.119
TEMP	Water temperature	°C	13.407	-4	33.7	7.674
CHLOROPH	Chlorophyll-A	yll-A	9.054	0	118.4	13.177
BOD	Biochemic oxygen demand	mg/L	4.155	0	46	3.386
COD	Chemical oxygen demand	mg/L	44.466	2	305	37.649
CBOD	Carbonaceous BOD	mg/L	1.653	0	6	1.782
PH	Water pH	NA	7.481	0	9.2	0.661
VSS	Volatile suspended solids	mg/L	137.697	0	916	194.801
ELEV	Elevation	ft	270.976	0.000 07	513.776	137.297
INORG_SS	Inorganic suspended solids	mg/L	29.769	0	428	42.913
MIN_AIR_TEMP	Min. air temperature	°F	43.035	−5.8	79	17.301
AVG_AIR_TEMP	Avg. air temperature	°F	52.608	−0.16	86.16	18.284
MAX_AIR_TEMP	Max. air temperature	°F	61.943	8.1	99	19.942
DAILY_PERC	Daily precipitation	in.	0.093	0	1.82	0.244
FLOW	Daily flow	cfs	67.708	0.02	1450	145.644
TOT_1001	Single family residential area	acre	25 878.139	13 161.3	58 746.6	19 001.896

Prediction Models

This section describes the different regression or classification approaches used in this study. Eight different algorithms were investigated and built as regression or classification models, as applicable, and their merits were compared in the context of performance analysis. The prediction models are: *multiple linear regression*, *artificial neural networks* (ANNs), *decision trees*, *support vector machines* (SVMs), *lazy learners*, and *Gaussian process*.

To use classification models to predict total nitrates, the values were transformed from continuous to three nominal classes. The classes were defined as *low*, *medium* and *high* based on the assessment of watershed data, as shown in Table 5. The classification models are: *artificial neural networks* (ANNs), *model trees*, *support vector machines*, *naive Bayes*, *lazy learners*, and *logistic regression*.

Table 5 Total nitrates classes.

Class	Range
Low	$0 < (NO2 + NO3) \leq 3.99$
Medium	$3.99 < (NO2 + NO3) \leq 7.99$
High	$7.99 < (NO2 + NO3) \leq \infty$

Regression Models Evaluation

This section discusses the criteria used to evaluate the prediction accuracy of the regression models used in the study.

Correlation coefficient is based on the standard correlation coefficient and measures the extent of the linear relationship between predicted (P) and actual (A) values. It is a dimensionless index that ranges from −1 to 1 with 1 corresponding to ideal correlation. The correlation coefficient *C* is given by:

$$C = \frac{Cov(P,A)}{\sigma_P \sigma_A} \tag{7}$$

where:

$Cov(P,A)$ = covariance between the predicted and the actual values, and

σ_P and σ_A = their respective standard deviations.

Root mean squared error (RMSE) measures the confidence intervals. It ranges from 0 to ∞ with 0 corresponding to the ideal situation. It is computed as in Equation 3.

Mean absolute error (MAE) is similar to RMSE, except that it uses absolute error values instead of the squared errors. It is computed as in Equation 5.

*Root relative squared erro*r (RRSE) is relative to whatever is represented by the simple predictor, which is the mean of the actual values. It is computed by normalizing the total squared error, dividing that by the total squared error of the simple predictor, and taking the square root. It is given by:

$$RRSE = \sqrt{\frac{\sum_{i=1}^{N}(p_i - a_i)^2}{\sum_{i=1}^{N}(\hat{a} - a_i)^2}} \tag{8}$$

where â is the actual mean.

Relative absolute error (RAE) is similar to RRSE. The relative absolute error takes the total absolute error and normalizes it, by dividing by the total absolute error of the simple predictor, and taking the square root. The value of this error ranges from 0% to 100% with 0 being the ideal situation. It is given by:

$$RAE = \frac{\sum_{i=1}^{N}|p_i - a_i|}{\sum_{i=1}^{N}|\hat{a} - a_i|} \tag{9}$$

Classification Models Evaluation

This section discusses the criteria used to evaluate the prediction accuracy of the classification models used in the study.

Model accuracy is a criterion that measures the goodness of the model correlation. It refers to the percentage of correct predictions made by the model when compared with the actual classifications in the test data, displayed in a confusion matrix (Han and Kamber 2006). Accuracy is the proportion of total true results to total results. It is given by:

$$Accuracy = \left(T_p + T_n\right) / \left(T_p + T_n + F_p + F_n\right) \tag{10}$$

where:

T_p and F_p = number of true and false positives respectively, and

T_n and F_n = number of true and false negatives respectively.

Precision is the percentage of records that are correct responses and are actually positive or relevant to the positive class, and is given by:

$$Precision = \frac{T_p}{T_p + F_p} \tag{11}$$

Recall is the percentage of positive records that are predicted among all the records predicted by the classifier. It is given by:

$$Recall = \frac{T_p}{T_p + F_n} \tag{12}$$

F-measure is the tradeoff of precision for recall and vice versa. It is the measure that discourages systems from sacrificing to one another excessively. It is given by:

$$F\text{-}measure = \frac{Recall \times Precision}{\left(Recall + Precision\right)/2} \tag{13}$$

Receiver operating characteristic (ROC) is a plot of true positive rate vs false positive rate that compares predicted and actual values. It provides an insight into the decision making ability of a model (sensitivity). That is, how likely the model is to accurately predict the negative or the positive classes. It is a useful metric for evaluating how a model behaves with different probability thresholds.

4 Results and Discussion

4.1 Physical Modeling

The calibration and validation process in HSPF is hierarchical, beginning with the hydrology and ending with water quality constituents (Donigian 2002). The nutrient constituents simulated

were total nitrates (NO_3 + NO_2) as N. Total nitrogen loads were calculated later using scripts provided by HSPF. Various nutrient modeling parameters were added for both pervious and impervious land segments. These parameters include the constituent washoff factor, monthly constituent accumulation factor, and the initial storage for each constituent. These parameters were calibration parameters that were adjusted until reasonable model behaviour was reached.

For the Upper Chicago River subbasin, the results of the simulation were compared with measurements taken from the North Branch of the Chicago River at Grand Ave, Chicago. The location was chosen to represent the outlet for the subbasin. There were two factors that limited the time period for the calibration and validation of the model. First the observed flow was limited to the period 2002–2010 (with some data missing in the period 2003–2004). However, the available meteorological data ended at 2006 so only the period 2002–2006 was used for performing the flow simulation, calibration and validation. The initial simulation trials resulted in values that consistently over-predicted, mostly during the wet season. Calibration parameters, which were adjusted, include the monthly accumulation factors and monthly values for limiting storage for both pervious and impervious land segments. Instream process parameters, nitrification and denitrification parameters (*KNO320*), along with oxidation rate (*KTAM20*) and algal growth rate parameters were all adjusted.

Figures 2 and 3 show, graphically, the calibration and validation results simultaneously for flow simulation. Figures 4 and 5 show the calibration and validation results for total nitrates simulation. Table 6 summarizes the calibration and validation statistics for the nutrients simulated. The results of the calibration show that there is an acceptable agreement between the observed and simulated data. Statistical results for best fit calibration of total nitrates and the percent mean error between the simulated and observed data for nitrates show that the model performance criterion PME was very good for all the constituents as it fell within the accepted tolerances suggested by Donigian (Table 3 above).

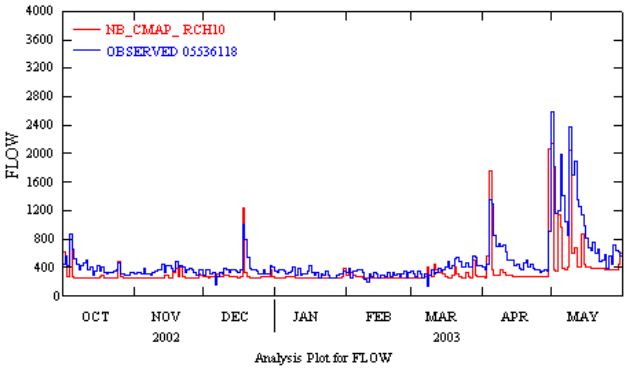

Figure 2 Simulation of flow for calibration period.

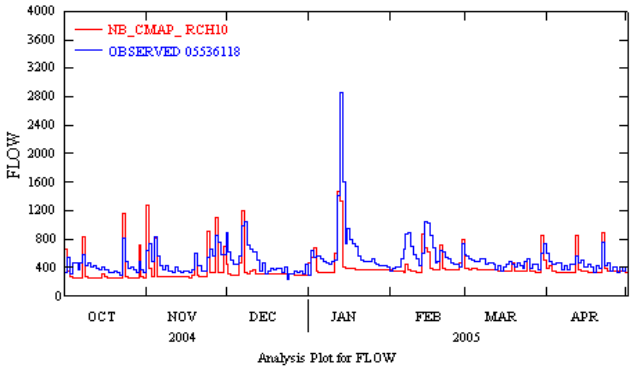

Figure 3 Simulation of flow for validation period.

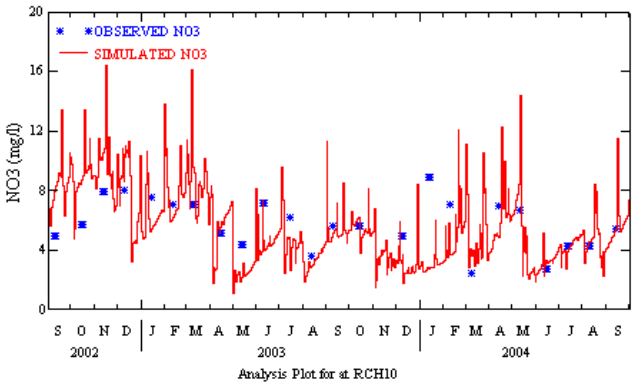

Figure 4 Simulation of total nitrates for calibration period.

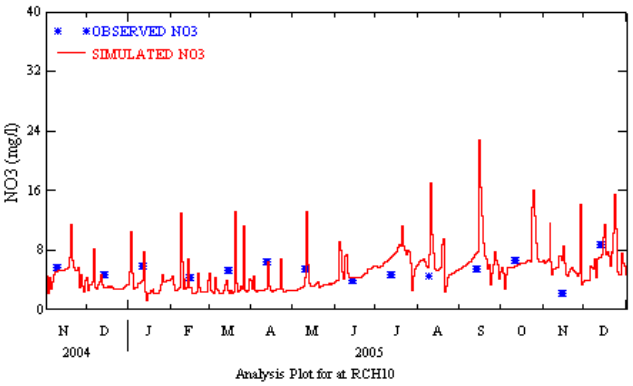

Figure 5 Simulation of total nitrates for validation period.

Table 6 Statistical results of total nitrates calibration and validation.

	Mean observed value	Mean simulated value	ME	PME	MAE	RMSE	NSE
Calibration	5.81	5.50	0.23	3.93	1.74	2.21	0.13
Validation	5.25	5.01	0.40	7.61	1.86	2.16	−0.64

According to the results obtained from the validation process period, the model performance is considered very good. The model can be successfully applied to the watershed with appropriate consideration given to the EIA values considered in Table 2.

4.2 Data-driven Modeling

Table 7 compares the prediction accuracy of the six regression models. It shows that ANN, decision tree and Gaussian process performed better than SVM and lazy learner. They each showed similar performance, with very close values, to RMSE and MAE, with correlation coefficients of 74.49%, 74.48% and 74.41% respectively. Table 8 shows the results of the classification models. The results show that ANN is the best classification model to predict total nitrates followed by decision tree. The worst is the naive Bayes. However, the decision tree provides a clear logical model that can be easily understood.

Table 7 Evaluation of regression models for total nitrates.

Model	Correlation	RMSE	MAE	RRSE	RAE
Multi-linear regression	0.6759	2.1306	1.4842	73.68%	60.73%
ANN	0.7449	1.9469	1.2686	67.32%	51.91%
Decision tree	0.7448	1.9279	1.2217	49.99%	66.65%
SVM	0.6331	2.3431	1.3042	81.02%	53.36%
Lazy learner	0.6295	2.2450	1.5583	77.63%	63.76%
Gaussian process	0.7441	1.9368	1.2731	66.97%	52.09%

Table 8 Evaluation of classification models for total nitrates.

Model	Accuracy	Precision	Recall	F-measure	ROC Area
Naive Bayes	80.77%	0.759	0.808	0.78	0.862
ANN	83.32%	0.819	0.833	0.82	0.918
Logistic regression	81.99%	0.808	0.820	0.81	0.904
SVM	81.55%	0.760	0.815	0.79	0.775
Decision tree	82.32%	0.817	0.823	0.82	0.823
Lazy learner	81.66%	0.761	0.817	0.79	0.854

4.3 Physical Modeling vs Data-driven Modeling

For the proposed framework for the Chicago River watershed, both data-driven and physical models were developed. Results comparing the performance of the two model approaches are shown in Table 9. It shows that data-driven models show better performance. RMSE for regression model vs physical model showed up to a 10.7% increase in prediction performance.

Table 9 Comparing physical and data-driven models for total nitrates.

Model	RMSE
Physical model (HSPF)	2.1600
ANN	1.9469
Gaussian process	1.9368
Decision tree	1.9279

5 Conclusion

Although the use of a data-driven approach for modeling complex physical systems is receiving increasing interest, it is not easy to precisely link the data-driven technique to the most important physical variables that govern the natural processes of a watershed system. The continuous calibrated and validated physical model allows for the evaluation of the behaviour of the watershed under possible future conditions. This property of the physical model would benefit from the analysis of different scenarios that the watershed may face such as climate change, population change, or the inclusion or removal of certain physical variables. This would provide a planning tool for regulatory environmental agencies in the Chicago River watershed and its use would allow them to develop better management programs. However, the data-driven models require fewer inputs and can be deployed anywhere in the watershed, while the physical model requires extensive data inputs and can only be applied to the specific watershed outlets selected in the simulation. The data-driven models can be used as operational tools to maintain the water quality parameters especially if total maximum daily loads (TMDL) and water quality standards (WQS) are developed for the Chicago River watershed. We suggest that the use of both physical and data-driven models is essential for the health of the watershed. The physical model can be a planning tool whenever significant physical change takes place in the watershed while the data-driven model can be an operating tool that can be used periodically to inspect the watershed water quality parameters, especially if TMDL and WQS are established for the watershed.

Although, the modeling approach and the methodology were implemented for highly urbanized watershed, it is not restricted and can be used without modification for any other watershed, provided that data is available and proper models were selected.

References

Ahearn, D. S., R. W. Sheibley, R. A. Dahlgren, M. Anderson, J. Johnson and K. W. Tate. 2005. "Land Use and Land Cover Influence on Water Quality in the Last Free-Flowing River Draining the Western Sierra Nevada, California." *Journal of Hydrology* 313:234–47. https://doi.org/10.1016/j.jhydrol.2005.02.038

Akhavan, S., J. Abedi-Koupai, S.-F. Mousavi, M. Afyuni, S. S. Eslamian and K. C. Abbaspour. 2010. "Application of SWAT Model to Investigate Nitrate Leaching in Hamadan-Bahar Watershed, Iran." *Agriculture, Ecosystems and Environment* 139 (4): 675–88.

Asefa, T., M. Kemblowski, M. McKee and A. Khalil. 2006. "Multi-Time Scale Stream Flow Predictions: The Support Vector Machines Approach." *Journal of Hydrology* 318:7–16.

Bergman, M. J., W. Green and L. J. Donnangelo. 2002. "Calibration of Storm Loads in the South Prong Watershed, Florida,

Using Basins/HSPF." *JAWRA Journal of the American Water Resources Association* 38:1423–36.

Brabec E., S. Schulte and P. L. Richards. 2002. "Impervious Surfaces and Water Quality: A Review of Current Literature and Its Implications for Watershed Planning." *Journal of Planning Literature* 16:499.

Burchard-Levine, A., S. Liu, F. Vince, M. Li and A. Ostfeld. 2014. "A Hybrid Evolutionary Data-driven Model for River Water Quality Early Warning." *Journal of Environmental Management* 143:8–16. https://doi.org/10.1016/j.jenvman.2014.04.017

Chen, S. T. and P. S. Yu. 2007. "Real-Time Probabilistic Forecasting of Flood Stages." *Journal of Hydrology* 340:63–77.

Chiang, Y. M., K. L. Hsu, F. J. Chang, Y. Hong and S. Sorooshian. 2007. "Merging Multiple Precipitation Sources for Flash Flood Forecasting." *Journal of Hydrology* 340:183–96.

Conway, T. M. and L. G. Lathrop. 2005. "Alternative Land Use Regulations and Environmental Impacts: Assessing Future Land Use in an Urbanizing Watershed." *Landscape and Urban Planning* 70 (1): 1–15.

Deacon, J. R., S. A. Soule and T. E. Smith. 2005. *Effects of Urbanization on Stream Quality at Selected Sites in the Seacoast Region in New Hampshire, 2001–3*. Reston, VA: U.S. Geological Survey. U.S. Geological Survey Scientific Investigations Report 2005–5103. https://pubs.usgs.gov/sir/2005/5103/SIR2005-5103_report.pdf

Donigian, A. S. 2002. *HSPF Watershed Model Calibration and Validation*. Mountain View, CA: Aquaterra Consultants.

Donigian, A. S., B. R. Bicknell and J. C. Imhoff. 1995. "Hydrological Simulation Program Fortran (HSPF)." In *Computer Models of Watershed Hydrology*, edited by V. P. Singh, 395–442. Littleton, CO: Water Resources Publications.

Fallah-Mehdipour, E., O. Haddad and M. Mario. 2014. "Genetic Programming in Groundwater Modeling." *Journal of Hydrologic Engineering* 19 (12). https://doi.org/10.1061/(ASCE)HE.1943-5584.0000987

Fohrer, N., S. Haverkamp, K. Eckhardt and H. G. Frede. 2001. "Hydrologic Response to Land Use Changes on the Catchment Scale." *Physics and Chemistry of the Earth (B)* 26 (7–8): 577–82.

Fonseca, A., D. P. Ames, P. Yang, C. Botelhoa, R. Boaventura and V. Vilar. 2014. "Watershed Model Parameter Estimation and Uncertainty in Data-Limited Environments." *Environmental Modelling & Software* 51:84–93.

Ghavidel, S. Z. Z. and M. Montaseri. 2014. "Application of Different Data-Driven Methods for the Prediction of Total Dissolved Solids in the Zarinehroud Basin." *Stochastic Environmental Research and Risk Assessment* 28 (8): 2101–8. https://doi.org/10.1007/s00477-014-0899-y

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten. 2009. "The WEKA Data Mining Software: An Update." *ACM SIGKDD Explorations Newsletter* 11 (1): 10–18.

Han, J. and M. Kamber. 2006. *Data Mining: Concepts and Techniques,* 2nd ed. Waltham, MA: Morgan Kaufmann Publishers.

Im, S., K. M. Brannan and S. Mostaghimi. 2003. "Simulating Hydrologic and Water Quality Impacts in an Urbanizing Watershed." *JAWRA Journal of the American Water Resources Association* 39 (6): 1465–79.

Jeon, J., C. G. Yoon, A. S. Donigian Jr. and W. Jung. 2007. "Development of the HSPF Paddy Model to Estimate Watershed Pollutant Loads in Paddy Farming Regions." *Agricultural Water Management* 90 (1–2): 75–86.

Jones, T., C. Johnston and C. Kipkie. 2003. "Using Annual Hydrographs to Determine Effective Impervious Area." *Journal of Water Management Modeling* R215. https://doi.org/10.14796/JWMM.R215-14

Krause, P., D. Boyle and F. Base. 2005. "Comparison of Different Efficiency Criteria for Hydrological Model Assessment." *Advanced Geosciences* 5:89–97.

Leon, L. F., W. Booty, I. Wong, C. McCrimmon, S. Melles, G. Benoy and J. Vanrobaeys. 2010. "Advances in the Integration of Watershed and Lake Modeling in the Lake Winnipeg Basin." In *Modeling for Environment's Sake: Proceedings of the 5th Biennial Conference of the International Environmental Modelling and Software Society, iEMSs 2010*, 860–7.

Luzio, M. D., R. Srinivian and J. G. Arnold. 2002. "Integration of Watershed Tools and Swat Model to Basins." *JAWRA Journal of the American Water Resources Association* 38 (4): 1127–42.

Marsili-Libellia, S., E. Giustia and A. Nocitab. 2013. "A New Instream Flow Assessment Method Based on Fuzzy Habitat Suitability and Large-Scale River Modeling." *Environmental Modelling & Software* 41:27–38.

Mohamoud, Y. M., R. Parmar and K. Wolfe. 2010. "Modeling Best Management Practices (BMPs) with HSPF." In *Proceedings, Watershed Management Conference 2010, Madison, WI, August 23–27, 2010*. Reston, VA: American Society of Civil Engineers. https://doi.org/10.1061/41143(394)81

Mouton, A. M., B. De Baets and P. L. M. Goethals. 2009. "Knowledge-Based Versus Data-Driven Fuzzy Habitat Suitability Models for River Management." *Environmental Modelling & Software* 24 (8): 982–993.

Muttil, N. and S. Y. Liong. 2004. "Physically Interpretable Rainfall Runoff Models Using Genetic Programming." In *Hydroinformatics: Proceedings of the 6th International Conference, Singapore, 21–24 June 2004*, 2 vols., edited by S.-Y. Liong, K.-K. Phoon and V. Babovic. London: World Scientific. https://doi.org/10.1142/9789812702838_0204

MWRDGC (Metropolitan Water Reclamation District of Greater Chicago). 2007. *Cook County Stormwater Management Plan*.

Chicago, IL: Metropolitan Water Reclamation District of Greater Chicago. http://www.mwrd.org/

Preis, A. and A. Ostfeld. 2008. "A Coupled Model Tree Genetic Algorithm Scheme for Flow and Water Quality Predictions in Watersheds." *Journal of Hydrology* 349:364–75.

Shirinian-Orlando, A. A. and C. G. Uchrin. 2007. "Modeling the Hydrology and Water Quality Using BASINS/HSPF for the Upper Maurice River Watershed, New Jersey." *Journal of Environmental Science & Health, Part A: Toxic/Hazardous Substances & Environmental Engineering* 42 (3): 289–303.

Singh R. K., R. K. Panda, K. K. Satapathy and S. V. Ngachan. 2011. "Simulation of Runoff and Sediment Yield from a Hilly Watershed in the Eastern Himalaya, India Using the WEPP Model." *Journal of Hydrology* 405 (3-4): 261–76.

Solomatine, D. P. and K. N. Dulal. 2003. "Model Tree as an Alternative to Neural Network in Rainfall–Runoff Modeling." *Hydrological Sciences Journal* 48 (3): 399–411.

Solomatine, D. P., M. Maskey and D. L. Shrestha. 2007. "Instance-Based Learning Compared to Other Data-driven Methods in Hydrologic Forecasting." *Hydrological Processes* 21. https://doi.org/10.1002/hyp.6592

Sutherland, R. C. 2000. "Methods for Estimating the Effective Impervious Area of Urban Watersheds." *The Practice of Watershed Protection* 32:193–5.

Tan, P. N., M. Steinbach and V. Kumar. 2006. *Introduction to Data Mining*. Boston, MA: Addison–Wesley.

Tokar, A. S. and M. Markus. 2000. "Precipitation–Runoff Modeling Using Artificial Neural Networks and Conceptual Models." *Journal of Hydrologic Engineering* 5 (2): 156–61.

Tong, S. T. Y. and W. Chen. 2002. "Modeling the Relationship Between Land Use and Surface Water Quality." *Journal of Environmental Management* 66 (4): 377–93.

Tong, S. T. Y., A. J. Liu and J. A. Goodrich. 2007. "Climate Change Impacts on Nutrient and Sediment Loads in a Midwestern Agricultural Watershed." *Journal of Environmental Informatics* 9 (1): 18–28.

Tong, S. T. Y., A. J. Liu and J. A. Goodrich. 2009. "Assessing the Water Quality Impacts of Future Land-Use Changes in an Urbanising Watershed." *Civil Engineering and Environmental Systems* 26 (1): 3–18.

Wang, S. H., D. G. Huggins, L. Frees, C. G. Volkman, N. C. Lim, D. S. Baker, V. Smith and F. DeNoyelles, Jr. 2005. "An Integrated Modeling Approach to Total Watershed Management: Water Quality and Watershed Management of Cheney Reservoir." *Water, Air, & Soil Pollution* 164:1–19.

Wicklein, S. M. and D. M. Schiffer. 2008. *Simulation of Runoff and Water Quality for 1990 and 2008 Land Use Conditions in the Reedy Creek Watershed, East-Central Florida*. Reston, VA: U.S. Geological Survey. Water-Resources Investigations Report 02-4018. https://fl.water.usgs.gov/publications/Abstracts/wri02_4018_wicklein.html

Wilson, C. O. and Q. Weng. 2011. "Simulating the Impacts of Future Land Use and Climate Changes on Surface Water Quality in the Des Plaines River Watershed, Chicago Metropolitan Statistical Area, Illinois." *Science of the Total Environment* 409 (20): 4387–405.

Wu, Q., H. Li, R. Wang, J. Paulussen, Y. He, M. Wang, B. Wang and Z. Wang. 2006. "Monitoring and Predicting Land Use Change in Beijing Using Remote Sensing and GIS." *Landscape and Urban Planning* 78:322–33. https://doi.org/10.1016/j.landurbplan.2005.10.002

Yu, X., X. Zhang and L. Niu. 2009. "Simulated Multi-Scale Watershed Runoff and Sediment Production Based on GeoWEPP Model." *International Journal of Sediment Research* 24 (4): 465–78.

Zoppou, C. 2001. "Review of Urban Storm Water Models." *Environmental Modelling & Software* 16 (3): 195–231.