An evaluation of a data-driven approach to regional scale surface runoff modelling

Ruoyu Zhang

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science
In
Geography

Yang Shao (Chair)
Andrew W. Ellis
Julie Shortridge

April, 16 2018
Blacksburg, VA

Keywords: data-driven modelling; surface runoff simulation; random forest; machine learning; Chesapeake Bay

An evaluation of a data-driven approach to regional scale surface runoff modelling

Ruoyu Zhang

(ACADEMIC ABSTRACT)

Modelling surface runoff can be beneficial to operations within many fields, such as agriculture planning, flood and drought risk assessment, and water resource management. In this study, we built a data-driven model that can reproduce monthly surface runoff at a 4-km grid network covering 13 watersheds in the Chesapeake Bay area. We used a random forest algorithm to build the model, where monthly precipitation, temperature, land cover, and topographic data were used as predictors, and monthly surface runoff generated by the SWAT hydrological model was used as the response. A sub-model was developed for each of 12 monthly surface runoff estimates, independent of one another. Accuracy statistics and variable importance measures from the random forest algorithm reveal that precipitation was the most important variable to the model, but including climatological data from multiple months as predictors significantly improves the model performance. Using 3-month climatological, land cover, and DEM derivatives from 40% of the 4-km grids as the training dataset, our model successfully predicted surface runoff for the remaining 60% of the grids (mean R^2 (RMSE) for the 12 monthly models is 0.83 (6.60 mm)). The lowest R^2 was associated with the model for August, when the surface runoff values are least in a year. In all studied watersheds, the highest predictive errors were found within the watershed with greatest topographic complexity, for which the model tended to underestimate surface runoff. For the other 12 watersheds studied, the data-driven model produced smaller and more spatially consistent predictive errors.

An evaluation of a data-driven approach to regional scale surface runoff modelling

Ruoyu Zhang

(PUBLIC ABSTRACT)

Surface runoff data can be valuable to many fields, such as agriculture planning, water resource management, and flood and drought risk assessment. The traditional approach to acquire the surface runoff data is by simulating hydrological models. However, running such models always requires advanced knowledge to watersheds and computation technologies. In this study, we build a statistical model that can reproduce monthly surface runoff at 4-km grid covering 13 watersheds in Chesapeake Bay area. This model uses publicly accessible climate, land cover, and topographic datasets as predictors, and monthly surface runoff from the SWAT model as the response. We develop 12 monthly models for each month, independent to each other. To test whether the model can be applied to generalize the surface runoff for the entire study area, we use 40% of grid data as the training sample and the remainder as validation. The accuracy statistics, the annual mean R2 and RMSE are 0.83 and 6.60 mm, show our model is capable to accurately reproduce monthly surface runoff of our study area. The statistics for August model are not as satisfying as other months' models. The possible reason is the surface runoff in August is the lowest among the year, thus there is no enough variation for the algorithm to distinguish the minor difference of the response in model building process. When applying the model to watersheds in steep terrain conditions, we need to pay attention to the results in which the error may be relatively large.

ACKNOWLEDGEMENTS

I would like to thank my chair and committee, Dr. Yang Shao, Dr. Andrew Ellis, and Dr. Julie Shortridge for helping and guiding me finish my thesis. I appreciate their patience with my questions and suggestions to my research. It is impossible for me to complete this work without their help. Also, I would like to thank everyone in the Department of Geography, professors, staff, and my peer graduate students. In the past six years studying in Virginia Tech, they expand my knowledge in both academic and everyday life. As an international student, they help me go through many culture shocks and make me feel Virginia Tech my second home. In the end, I would like to thank my family, who provide anything they can to help me pursue my academic goals in Virginia Tech. The past six years are the most unforgettable period in my life. Though I cannot stay in Virginia Tech any longer, I wish everyone who has shared this wonderful period with me the best luck to the future.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

PREFACE

The manuscript has been submitted to *Journal of Hydrology* for peer review with Yang Shao, Andrew W. Ellis, Julie Shortridge, and Jie Ren as co-authors. Ruoyu Zhang, the thesis author and corresponding author of the manuscript, designed the experiment under the guidance of Dr. Yang Shao, wrote the codes for the model, and wrote the manuscript with help from Dr. Andrew Ellis and Dr. Julie Shortridge.

An evaluation of a data-driven approach to regional scale surface runoff modelling

Submitted to *Journal of Hydrology* for review at April 6, 2018

Introduction

Surface runoff information at regional, national, and global scales provides critical information for water resources planning, flood and drought risk assessment, and pollution mitigation (Beven, 2011). Over the past several decades, modelling surface runoff across large geographical scales has received increasing attention. At the global scale, several earth system models (e.g., Community Land Model) include a surface runoff component, although it is generally accepted that runoff outputs can be highly inaccurate (Döll et al., 2003). Döll reviewed several global hydrological models including those developed by Yates (1997), Klepper and Van Drecht (1998), Arnell (1999), and Vörösmarty et al. (1998). Almost all operate at coarse spatial resolution (> 0.5 degree) and model validation studies reported large disagreements between the simulated monthly/annual values and observed discharges (e.g., Meigh et al., 1999). As a key component of the global water assessment model WaterGAP 2 (Alcamo et al., 2003), the Global Hydrology Model (GHM) estimates daily and long-term runoff at 0.5 degree grid resolution (Döll et al., 2003). The GHM essentially calculates the water balance for each grid cell, and the main advantage of GHM is its calibration against observed discharge at gauging stations. However, similar to many other global models, the coarse spatial resolution of the GHM (i.e., 0.5 degree) does not capture the complexity and variability of runoff patterns within each grid cell. It is also unclear how GHM accuracies vary for regions with highly heterogeneous land cover, terrain, and soil properties.

At regional scales, modelling surface runoff has followed two main approaches. The first approach emphasizes the so-called semi-distributed modelling framework. One example of commonly used semi-distributed models is the Soil and Water Assessment Tool (SWAT) (Arnold et al., 1998). Such semi-distributed models describe key physical and hydrological processes, and use various simplified model specifications compared to those of fully distributed hydrological models. The SWAT has been intensively used for catchment-to-regional scale hydrological modelling because of its demonstrated performance in numerous applications and because of its user-friendly interface integrated within GIS software (Di Luzio et al., 2002; Gassman et al., 2014; Jha et al., 2006). Another example of a widely used semi-distributed model is the Variable Infiltration Capacity (VIC) model (Liang et al., 1994; Liang et al., 2003). The VIC model has been integrated into several water budget and land surface modelling frameworks.

Instead of focusing on hydrological processes, the second approach of large-scale surface runoff modelling relies on available observations to build empirical/statistical relationship between input variables of rainfall, temperature, snow and output discharge (e.g., Shamseldin, 1997; Tokar & Johnson, 1999). For example, Tokar and Johnson (1999) employed neural network to predict daily runoff directly from observed daily precipitation, temperature, and snowmelt. Beven (2011) referred to such modelling efforts as a data-based or data-driven approach.

Semi-distributed models have advantages in model interpretability, especially for understanding of the dominant processes and interactions among model inputs (Abbott et al., 1986; Devia et al., 2015). However, semi-distributed models can be data demanding and computationally-intensive, and thus they can be difficult to implement for large study regions. Necessary model inputs of soil properties, water depth, and vegetation characteristics may not be readily available (Beven, 2011; Pilgrim et al., 1988). Furthermore, a large number of model parameters need to be calibrated – a complicated process that often requires users' knowledge about watershed characteristics (Kim et al., 2014). On the other hand, the data-driven approach has been criticized with respect to model interpretability (Dawson & Wilby, 2001), because researchers often use neural network, support vector machine, and other nonparametric machine learning algorithms to approximate input-output relationships, ignoring the underlying hydrological processes (Beven, 2011), and the data-driven models developed from historical data may not hold the ability to predict future responses in a changed climate (e.g., future climate scenarios). The data-driven approach does have the appealing feature of predictive accuracy. For a given response variable of surface runoff and a set of predictors such as weather data, soil, and land use, a typical data-driven modelling effort often involves data splitting for model development, using a set of training data, and model validation, using a set of testing data. Model developers focus on the selection of approximation algorithm and parameter tuning to achieve the highest possible predictive accuracy (Chapelle et al., 2002; Huang et al., 2006). Therefore, it can be argued that deep knowledge of watershed characteristics/hydrological processes is not essential, while an understanding of data and machine learning algorithms is very important in order to build a high-accuracy runoff model. Another advantage of the data-driven approach is its capability of incorporating various input datasets. The inputs for a data-driven model may include those commonly used for semi-distributed models. Additional input data (e.g., soil moisture data from remote sensing) can be easily integrated to potentially improve model predictive accuracy. Recent studies showed that data-driven modelling and integration of 'non-traditional' inputs can advance our knowledge of underlying functions and processes (Hochachka et al., 2007; Resler et al., 2014).

Most previous data-driven studies focused on predicting stream discharge at certain observation gauges (e.g., Dawson & Wilby, 1998; Hsu et al., 1995; Tokar & Johnson, 1999). Surface runoff values with grid representation are difficult to obtain. In our study, we designed a new modelling framework focused on grid-based surface runoff modelling at monthly intervals. For a large study region, it is feasible to select a subset of watersheds and implement a semi-distributed hydrological model (e.g., SWAT) to obtain surface runoff estimation. The resultant surface runoff (i.e., converted to grid representation) can then be used as a response variable to develop a data-driven runoff model by incorporating a variety of predictors. Within such a data-driven modelling approach, we address the following two research objectives: (1) Evaluate a set of predictors derived from readily available geospatial datasets and determine best predictors for achieving high model accuracy; (2) Examine the spatial generalization capability of the data-driven model. We expect to see good performance of the data-driven model for areas used for training/tuning. However, we are more interested in how the trained data-driven models perform for areas not employed in the training.

Material and Methods

Study Area

Our study area is 44,438 km^2 in size and includes 13 8-digit watersheds in the Chesapeake Bay area of the eastern United States. The 13 selected watersheds include northwestern Virginia and small portions of West Virginia, Maryland, and Pennsylvania (Figure 1). Urban, forest, and agricultural land cover comprise 4.9%, 59.1%, and 26.8% of the study area, respectively. The Potomac River and its tributaries run across the study area and serves as the main source of drinking water for populated regions in Washington D.C. and northern Virginia. Increasing population, expanding urban areas, and intense agricultural practices in the Potomac watersheds continue to be primary contributing factors for water pollution of Potomac river and threat the public health (Pinkney et al., 2001; Yang et al., 2008). Development of watershed simulation models for this region is one important task for effective water resource management. A majority of selected 8-digit watersheds are located in three U.S. eastern level III ecoregions (Omernik, 1987): Northern Piedmont, Blue Ridge, and Ridge and Valley ecoregions. The climate, terrain, land cover, and soil conditions vary substantially across these three ecoregions, thus the study area provides needed variability for examining our data-driving surface runoff models.

Data

We obtained 30m resolution digital elevation model (DEM) data from the U.S. Geological Survey (USGS). Individual DEM tiles were merged to cover the entire study area. Slope values at 30m resolution were then derived from the DEM layer. Climatological data were acquired from the PRISM climate group (http://prism.oregonstate.edu/). The PRISM dataset provides various climatological variables, such as temperature and precipitation, in monthly and daily basis for the continental United States from 1981 to present for recent period, and historical past from 1895 to 1980 (Daly et al., 2000). The spatial resolution of the PRISM data is 4 km by 4 km. Both monthly and daily precipitation and temperature data from 2001 to 2015 were downloaded from PRISM website. The 2006 National Land Cover Dataset (NLCD 2006) was downloaded from the Multi-Resolution Land Characteristics (MRLC) Consortium website (www.mrlc.gov). The 30m resolution NLCD describes the detailed types of land cover in the United States (Fry et al., 2011; Homer et al., 2015). River discharge gauging data are acquired from the United States Geological Survey (USGS) stream flow observation.

Surface runoff from SWAT models

For each 8-digit watershed, we used the SWAT model to estimate monthly surface runoff. The SWAT model was initialized using the ArcSWAT 2012 interface. Using 30m DEM and 1:24,000 stream network data as input, we followed standard SWAT watershed delineation procedures to generate watershed sub-basins. The NLCD2006, STATSGO soil database from ArcSWAT, and slope values were combined to further divide sub-basins into hydrologic response units (HRU). We used PRISM daily precipitation and temperature (max and min) as SWAT weather data input, with the location of the weather station set as the center of each 8-digit watershed. PRISM precipitation and temperature data covering each 8-digit watershed were spatially averaged to represent general weather condition of each sub-basin. Similar approaches were used by Fuka et al. (2014) and Kim et al. (2014), although different weather datasets (e.g., Climate Forecast System Reanalysis by Fuka et al. 2014) were used as forcing data.

The time-period from 2001 to 2002 was used as a "warm up" period for the SWAT model. SWAT calibration was performed through the Generalized Likelihood Uncertainty Estimation (GLUE) algorithm (Beven & Binley, 1992) – an algorithm available through the SWAT-CUP software package. The use of spatially averaged precipitation and temperature data as SWAT input led us to focus on monthly model calibration and subsequent analysis. We calibrated monthly stream flow against the USGS stream flow observation data for each selected watershed. The stream flow calibration for SWAT was based upon years 2003 to 2008; the validation period was 2009 to 2015. Following Kim et al. (2014, 2017), we included a number of SWAT model parameters for model calibration. The Nash and Sutcliffe model efficiency coefficient (NSE) were used as a goodness-of-fit measure. From SWAT model outputs, we focused on surface runoff at the sub-basin level for each watershed, which provides sufficient spatial detail for water yield evaluation.

Grid representation and input-output data for data-driven model

For our data-driven runoff simulation model, we preferred a regular grid representation to organize the input-output data. Among the key input data, the PRISM weather data have relatively coarser spatial resolution (4 km) compared to land cover and DEM data (30 m). Therefore, we chose 4-km resolution as the finest analytical unit for our data-driven model. We overlaid the 4-km PRISM grid on top of the study area and removed those located at the edges of study region. The remaining total number of 4km grids is 2,101. For each PRISM grid, we calculated the percent coverage of forest, agricultural land, and urban classes using NLCD2006 as reference. Additionally, we calculated mean elevation, standard deviation of elevation, and standard deviation of slope. The monthly PRISM precipitation and temperature data, percent land cover class, and various DEM-derived statistics were used as input data for our data-driven runoff simulation model (Table 1).

The outputs of the data-driven model were monthly surface runoff values at 4km spatial resolution. The original SWAT runoff values were reported for sub-basins with irregular sizes and shapes. Generally, the size of a sub-basin is larger than the size of the 4-km grid. We overlaid 4km PRISM grids on top of sub-basins and calculated area-weighted monthly surface runoff for each 4km grid as the response variable for the data-driven model.

Data-driven Model

Our data-driven model can be represented using the following input-output approximation:

$$y_i = f(X_i)$$

where y_i is the surface runoff at ith 4km grid, X_i is a vector of predictors, and f is the approximation algorithm. Using this algorithm, monthly models were developed independently of one another. Of the input variables in X that were introduced in the previous section, we assumed that the land cover and topographic variables were temporally static during the study period. Climatological variables were dynamic in both spatial and temporal domains. For each monthly model, the array of input variables tested included monthly climatological data of the current month, and to determine whether including prior climatological data can improve the model, up to four months' climatological data prior to the current month were also tested as the inputs respectively. For example, to model the surface runoff of May, we could include the current month (May) and four previous months' climatological data (January, February, March, and April) as climatological predictors. We named the above example as 5-month May model because a total of five months' climatological data were included. Accordingly, the array of predictors of a 3-month May model would include climatological data from May, April, and March.

In this study, we used the Random Forest algorithm as our input-output approximation algorithm. Random Forest is one of the machine learning algorithms that are widely applied in many fields of study. The Random Forest is developed from classification and regression decision trees and uses a large number of decision trees (i.e., an ensemble approach) to make final predictions (Breiman, 2001). The Random Forest is easy to implement, but its performance is among the best in various machine learning algorithms (Adamowski et al. 2012; Shortridge et al., 2016).

Within this Random Forest based modelling framework, we were particularly interested in how model performance varies through the months of the year. In the initial experiment, we divided the 2,101 4km grids into 40%-60% training and testing datasets for each month. The randomly selected 40% training data points were used to train the Random Forest algorithm and the remaining 60% of testing points were used to assess the performance of each monthly model. We used Root-Mean-Square Error (RMSE) and R-squared (R^2) statistics to measure model performance.

We further examined how model performance varies when different numbers of training sample points are used in Random Forest training. Specifically, we randomly selected data from 5% to 80% of grids as the training sample and the remaining as validation. The amount of data required to train the model can determine whether the data-driven model has the capability of spatial generalization. For example, if the use of a small percentage (e.g., <50%) of training data points leads to reasonable predictive accuracy, we could conclude that our model is capable of spatial generalization.

Results

Surface runoff from SWAT models

For all 13 8-digit watersheds, monthly SWAT stream discharge outputs were validated for the period 2009-2015. The NSE measure ranged from 0.56 to 0.77, suggesting acceptable model performance (Moriasi et al., 2007). To illustrate general spatial pattern of surface runoff across the study area, we present mean monthly runoff for all sub-basins, 2003-2015 (Figure 2). The sub-basins with highest surface runoff (e.g., > 28 mm) are mainly located in the North Branch Potomac watershed and the Middle Potomac-Anacostia-Occoquan watershed. There are more steep valleys in the North Branch Potomac watershed and more urban areas in the Middle-Anacostia-Occoquan watershed, so it was expected to have a larger surface runoff amount from a similar amount of precipitation.

Data-driven model – input data selection and statistics

For each of the 12 monthly models, we started with a training versus validation distribution of grid cells of 40% as training data to build the Random Forest and 60% as validation data to assess model performance. Initially, we only included climatological data for the current month, land cover data, and various DEM derivatives as predictors. N-month (N is up to 5) models were sequentially developed by incorporating multiple months' climatological data. Table 2 shows R^2 and RMSE statistics for the validation datasets when considering data for various lengths of time (months) prior to the month of modelled runoff. The mean R^2 (RMSE) value from the 12 1-month models was 0.664 (9.5 mm). Among all of the 1-month models, those for March and August had the lowest R^2 (0.498 and 0.384 respectively), while models for the remaining months had considerably higher values (around 0.7). As precipitation and temperature data from the months prior to the modelled month were added, the results considerably improved over those generated by the 1-month model (Table 2). The overall trend was that mean R^2 (RMSE) increased (decreased) as prior climatological data were included. The mean R^2 (RMSE) for the monthly models reached 0.862 (6.1 mm) when we used 5-month data as predictors.

The 3-month, 4-month and 5-month models clearly outperformed the 1-month and 2-month models. Further, the 3-month models performed comparably to the 4-month and 5-month models, except for August (Table 2). Therefore, it seems that to simulate the surface runoff for most months, the data-driven model only needs 3-month of climatological, topographic and land cover data as predictors. Though using 5-month climatological data generated better results for August, the model performance was still not as good as other months. Overall, our results suggested that multiple months' climatological data are essential to model performance. The random forest algorithm also reveals the relative importance of each predictor (Breiman, 2001; Ishwaran et al., 2008), and for all 1-month to 5-month models, the most important predictor was always precipitation for the current month.

We examined the relationship between the size of the training data population and the model performance using cross-validation method. Figure 3 shows R^2 and RMSE values for model performance when using a range of training/validation data sample sizes. The general trends for the 1- to 5-month models were similar – the R^2 (RMSE) increased (decreased) with greater training sample size. When the same amount of training data were selected, the 3-, 4- and 5-month models outperformed the 1- and 2-month models, and the 3-month models performed similarly to 4- and 5-month models. When using only 40% of the grid data to train the models, model performance was encouraging (mean $R^2 > 0.8$ and mean RMSE < 7mm for the 3-month model).

To examine whether the predictive errors had characteristic spatial patterns, we computed the mean seasonal absolute prediction error for each 4km grid – the Random Forest-predicted and SWAT-generated monthly surface runoff values were compared for years from 2003 to 2015 (Dec – Feb as winter, Mar – May as spring, Jun – Aug as summer, and Sep – Nov as fall). For simplicity, we focused on results from the 3-month model trained by 40% of the 4km grids. Figure 4 shows the distribution of time-domain RMSE for the study area, the mean value of error for each gird is acquired from the absolute errors from 10 random iterations. The general trend was that the absolute error was small for grids in the center of each 8-digit watershed and increased near the margin. Comparing these results to the mean surface runoff generated by the SWAT model (Figure 2), areas with large absolute error were associated with areas with large surface runoff. Particularly in winter and spring, there was a cluster of high-error grids in the northwest of our study area (i.e., North Branch Potomac watershed). One possible reason was that this watershed's terrain and climate conditions were significantly different from those of other watersheds. Thus, we think that our model should be applied for areas with similar climate and ecological conditions, and applying our model to areas having significant variation of terrain and climate properties may not be ideal.

Discussion

The data-driven modelling framework, combined with machine learning algorithms such as random forest, are now widely applied in many fields. In hydrological research, there is an increasing number of applications of machine learning models to predict the rainfall-runoff process, and the statistical results are generally satisfying (Smith & Eli, 1995; Sudheer et al., 2002). Such models are capable of approximating the underlying relationships in the rainfall-runoff process, which can be an efficient alternative to physical-based, distributed hydrological models. In addition, they are relatively easy to implement and quite adaptive when a large number of predictors are involved and relations among predictors are complex. The biggest concern with distributed hydrological models is that the calibration process can be time-consuming, especially when a relatively large study area is of interest (e.g., 13 8-digit watersheds for our study). The machine learning model can provide a relatively simple and reliable approach to supplement the traditionally intricate hydrological models.

With climatological, land and topographic variables, and SWAT simulations of surface runoff, our study demonstrated that a data-driven model could achieve high predictive accuracy (e.g., $R^2 > 0.8$) using limited training data points. The most accurate monthly models (February, April, September) were associated with R^2 values near 0.9. The model for August yielded the lowest R^2 value (0.69). We checked the variable importance values within the random forest algorithm and found that, for the 3-month August model, August precipitation was the most important predictor but not significant compared to other months. Figure 5 shows relative variable importance values for the 3-month February and August models. We calculated the relative variable importance by normalizing the variable importance values with the variable that has the greatest importance in the model. For the February model, the importance of February precipitation weighted 61% among all predictors, while, for August model, the importance of August precipitation was only 38%. This indicates that the surface runoff mechanism in August is not as precipitation-dominant as that of February. Additionally, the mean surface runoff in August, 6.46 mm, was the lowest among the 12 months and the variation of surface runoff is small. February had the highest annual mean surface runoff (28.06 mm), which provided enough variation for the random forest algorithm to train the data, and their models had higher R^2 than others. Though the August model is incapable to perform as accurate as other monthly models, the importance of the August model is less significant because the August surface runoff contributes least to the annual surface runoff. As forest and agriculture land are dominant in our study area, predictors explaining complex forest ecosystem functions and intense irrigation in agricultural activities possibly need to be included in the August model.

Therefore, we concluded that the accuracy of our model to retrieve the surface runoff was not satisfying enough when the variation of runoff was small. To further improve the monthly models, future study could include more geospatial information, such as soil moisture, soil infiltration rate, and leaf area index, as predictors. Additionally, other machine learning algorithms, such as neural network and support vector machine, need to be evaluated for predictive performance comparison (e.g., Shao & Lunetta, 2012). We also note that the use of 4- or 5-month models may take longer time in model training, although computing effort is not a significant concern under now commonly used high performance parallel computing environment.

Conclusions

Summary

Random forest and other machine learning models are recognized as powerful tools to approximate complex and non-linear relations between predictors and response. We used the random forest algorithm to build a data-driven model to estimate monthly surface runoff on a 4 km resolution for 13 watersheds in the Chesapeake Bay area of the eastern United States. The model was driven by 4 km resolution monthly precipitation, monthly temperature, DEM, slope, and land cover data as predictors, and monthly SWAT model-generated surface runoff as the response variable. Within the data-driven modelling framework, we found that it was important to include multiple months (i.e., current month and 2 months prior to the current month) of climatological data as model input. For our study area, the accuracies of 3-month models yielded very good results that were not considerably improved upon by extending the climatological data to prior than three months. The 12 monthly 3-month models produced a mean R^2 (RMSE) value of 0.83 (6.60 mm) when 40% of the grids across the study region were randomly selected as the training sample. Model performance could be further improved by using more training samples, but this risks creating models that are overly trained to the data for the region rather than more robust models that are transferable spatially. The spatial distribution of predictive errors suggested that such a data-driven model can be applied to a large study area with similar terrain, land use, and climate conditions. However, spatial generalization of the data-driven model to areas with significantly different ecohydrological conditions (e.g., across ecoregions) may need to be applied with caution.

Limitations and Future Works

Currently, our surface runoff model demonstrated good spatial generalization capability by using 40% grid data as the training set and the remaining grid data points as validation. This suggests that our model can potentially be extended to the surrounding watersheds with similar climate, topographic and land conditions. However, we have not examined whether our model is capable of temporal generalization. It is not clear whether our model could predict future runoff scenarios using recent or historical data as input. The temporal generalization capability can be assessed using a cross-validation approach. For example, for a specific subbasin or grid, we can use the data from 2003 to 2010 to build the surface runoff model and then evaluate the model performance for years from 2011 to 2015. Such cross-validation procedures may be repeated by randomly selecting training time periods (and validation time periods) to further assess the model robustness. In the process of temporal runoff modeling, additional climate variables (e.g., temperature variations, rainfall intensity and duration within each month) may need to be considered as model inputs to improve the performance. Overall, the temporal generalization is probably more important than spatial generalization, especially in studies where future climate projections are considered for predicting water yield and quality. For a future study, we will fully explore both the model's spatial and temporal generalization capacity using various input variables. A model with robust spatial-temporal generalization can be a cost-effective tool supporting water resource managers' tasks in water resource regulation and planning.

Our surface runoff model operates at the monthly scale at present, thus it can be beneficial to agencies and water resource managers who are interested in analyzing the long-term surface runoff patterns. For applications in flash flood prediction and monitoring, a daily scale runoff model is needed. Limited by the fact that there is a lack of daily observation or reliable simulation, we could not develop a data-driven surface runoff model at the daily scale. Initially, we considered using the daily surface runoff from SWAT model as our response variable for a potential daily model. However, SWAT model was not designed for the event-based purpose and the accuracy of daily surface runoff simulation from the SWAT model may not be trusted. Additional 'event-based' hydrological models may need to be examined to derive high quality surface runoff data at daily scale.

For our surface runoff model, the number of predictors can be large as we apply the 4-month or 5-month approach to build the model. The overfitting and collinearity should be taking into our consideration. Based on the results of our model, we did not observe the decrease (increase) of R2 (RMSE) as we include more predictors (e.g., 1-month to 5-month model) using 40% grid data as training sample, though the statistics are saturated starting from 3-month models. We thus do not think the overfitting should be a major concern in this study area. However, when the same approach is applied to a different study area and the inconsistency of statistics is observed, it is necessary to further investigate the impacts from overfitting. Generally, collinearity of the input variables is not a major issue for a predictive model. As long as a predictor improves model prediction (evaluated through cross-validation), we could include the specific predictor. However, collinearity does affect model interpretation because it makes more difficult for the users in assessing individual predictor's impacts.

The data-driven model is always criticized by its neglecting of physical processes in many fields. Though our model has acceptable accuracy statistics, our model does not explain the exact relationship between rainfall and runoff, nor the connections between precipitation and other variables. In addition, since the initial objective of building the model is to keep it simple and user-friendly, we only used climate and some land data as the input variables. The data of soil property, which are considered as one of the most important components in the runoff generation mechanism, such as the infiltration and soil water capacity, are not included as predictors in the model. To further explore the relationship between predictors and the response and the interactions among predictors, we need to conduct thorough analysis of variable importance statistics. For example, we can detect how the variable importance of individual predictor changes when we include additional variables, such as soil property data, into the model. Such changes can potentially reveal the roles the new variables play within the rainfall-runoff modeling framework. The interpretation of variable importance and their changes could potentially generate meaningful physical explanations.

Table 1. Summary of primary input data (predictors) for the data-driven model

Data	Spatial Resolution	Temporal Information
NLCD Forest, urban, and agriculture (%)	30 m	2006
DEM Mean and standard deviation of elevation Standard deviation of slope	30 m	-
PRISM precipitation (monthly)	4 km	2003 – 2015
PRISM maximum temperature (monthly)	4 km	2003 – 2015
PRISM minimum temperature (monthly)	4 km	2003 – 2015

Table 2. Statistics for monthly runoff generated by the data-driven models using 1 to 5 months of climatological data (RMSE in mm).

	1 Month		**2 Month**		**3 Month**		**4 Month**		**5 Month**	
Month	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Jan	0.675	10.9	0.796	8.9	0.864	7.0	0.886	6.5	0.879	6.7
Feb	0.770	11.4	0.870	8.8	0.908	7.4	0.906	7.3	0.904	7.4
Mar	0.498	16.9	0.772	12.0	0.848	9.6	0.839	10.3	0.871	8.8
Apr	0.796	6.4	0.867	5.2	0.893	4.7	0.890	4.7	0.903	4.4
May	0.709	10.3	0.826	8.0	0.868	7.0	0.879	6.7	0.887	6.3
Jun	0.669	8.6	0.800	7.0	0.822	6.5	0.845	6.1	0.861	5.7
Jul	0.522	7.4	0.638	6.6	0.747	5.7	0.779	5.3	0.832	4.7
Aug	0.384	5.8	0.514	5.2	0.585	4.8	0.643	4.5	0.690	4.3
Sep	0.792	11.3	0.851	9.5	0.875	8.9	0.879	8.7	0.896	7.9
Oct	0.722	9.5	0.841	6.6	0.851	6.2	0.862	5.9	0.878	5.8
Nov	0.718	5.1	0.819	4.0	0.844	3.8	0.860	3.6	0.863	3.6
Dec	0.715	11.1	0.824	9.0	0.870	7.6	0.888	7.3	0.874	7.4
Mean	0.664	9.5	0.785	7.6	0.831	6.6	0.846	6.4	0.862	6.1

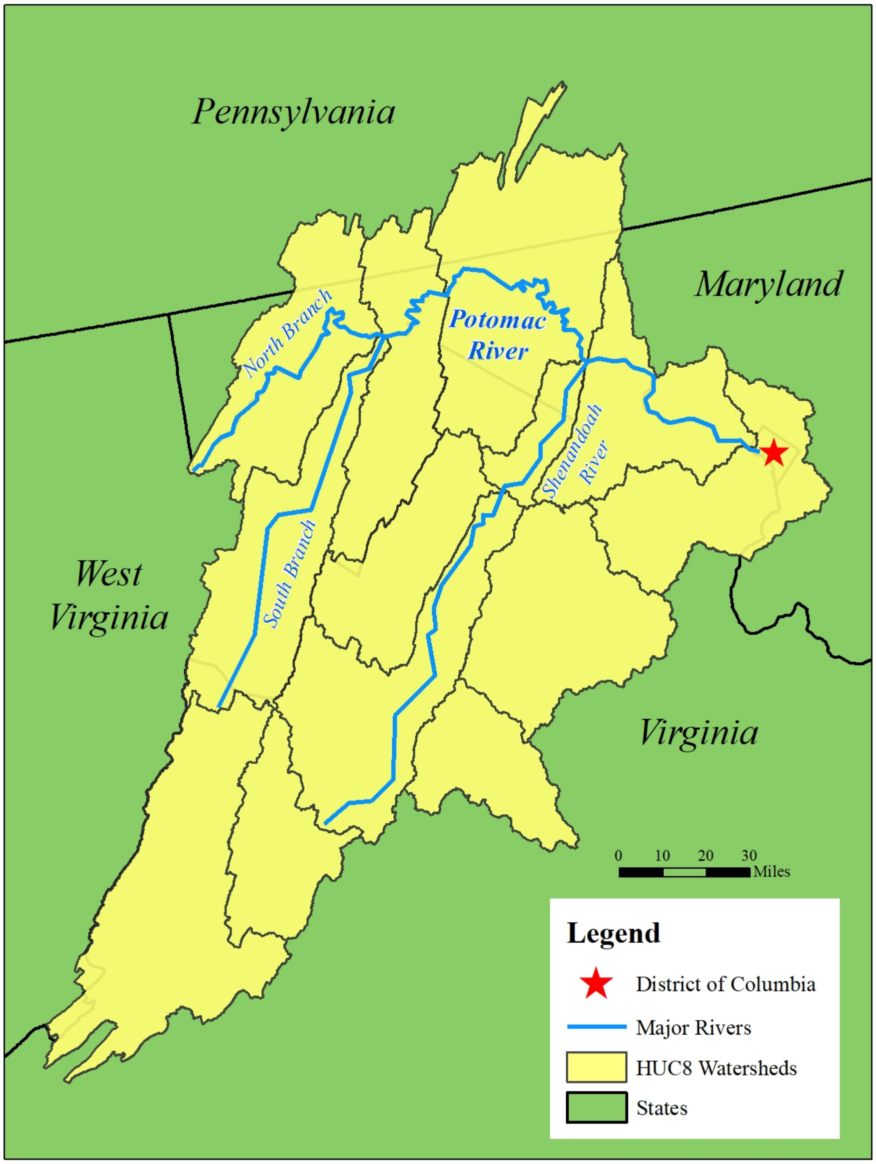

Figure 1. The study area includes 13 8-digit watersheds in the Chesapeake Bay area encompassing portions of the states of Maryland (MD), Pennsylvania (PA), Virginia (VA), and West Virginia (WV).

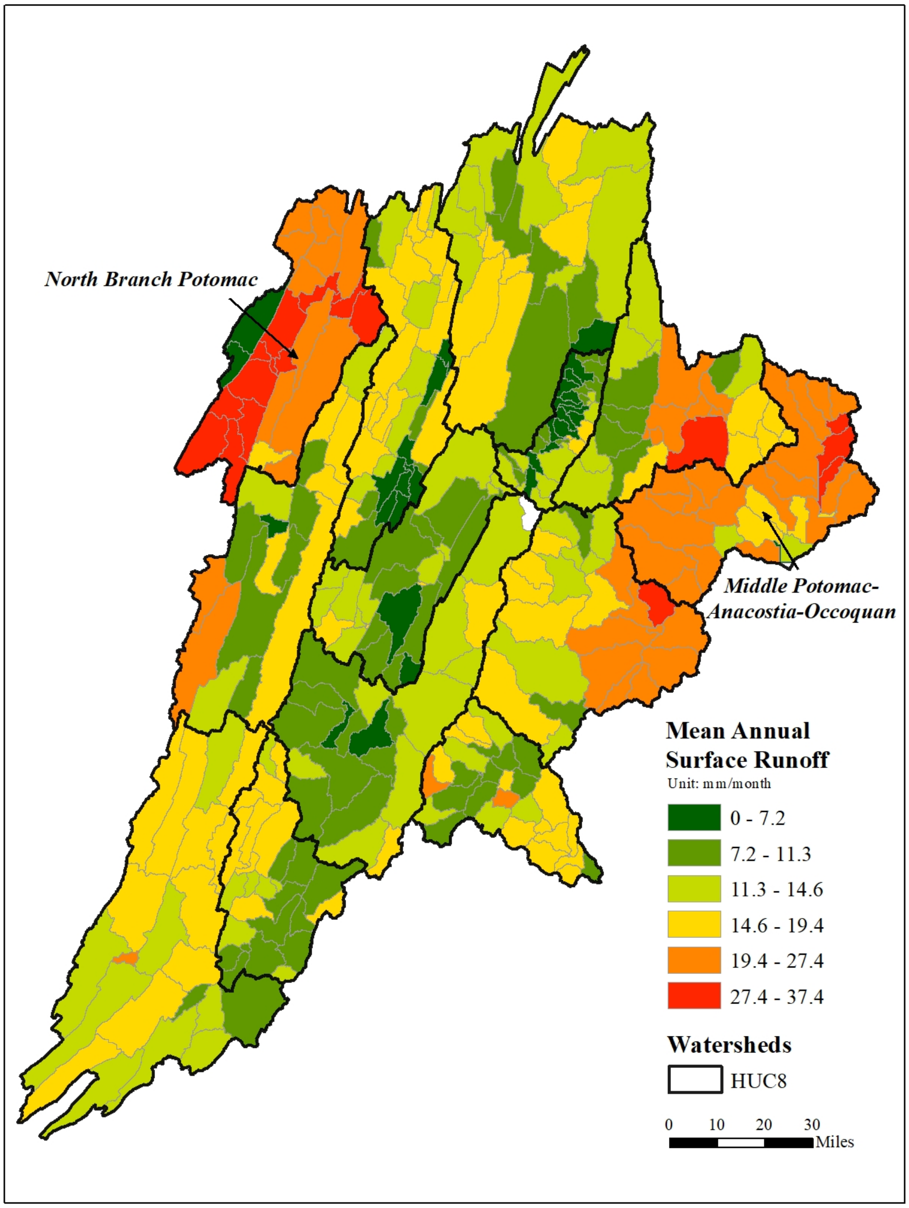

Figure 2. Mean monthly surface runoff generated by the SWAT model at the sub-basin level for each of the 13 8-digit watersheds. Two watersheds with the highest surface runoff are highlighted.

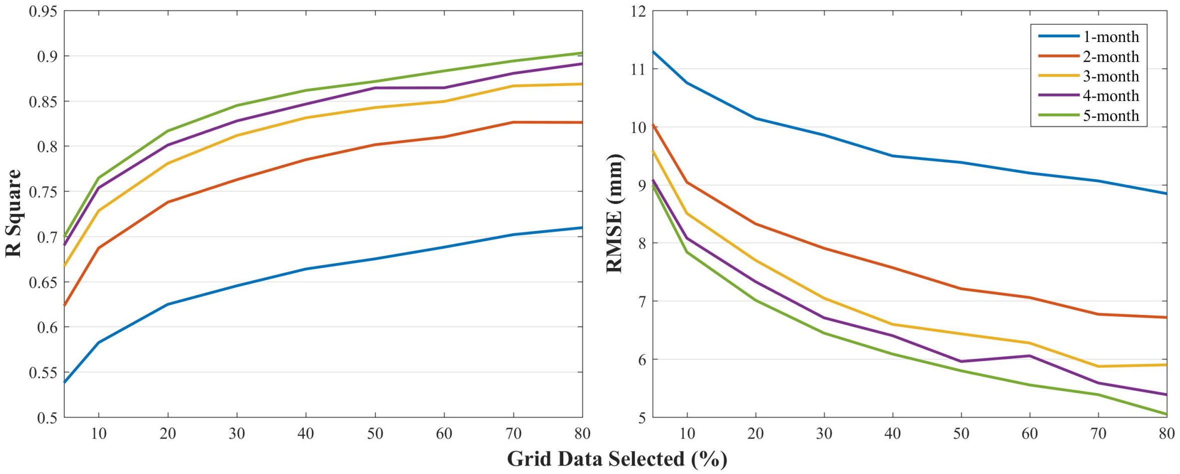

Figure 3. Mean R^2 and RMSE (mm) values for 12 monthly models for varying percentages of data used for training, with the remainder used for validation.

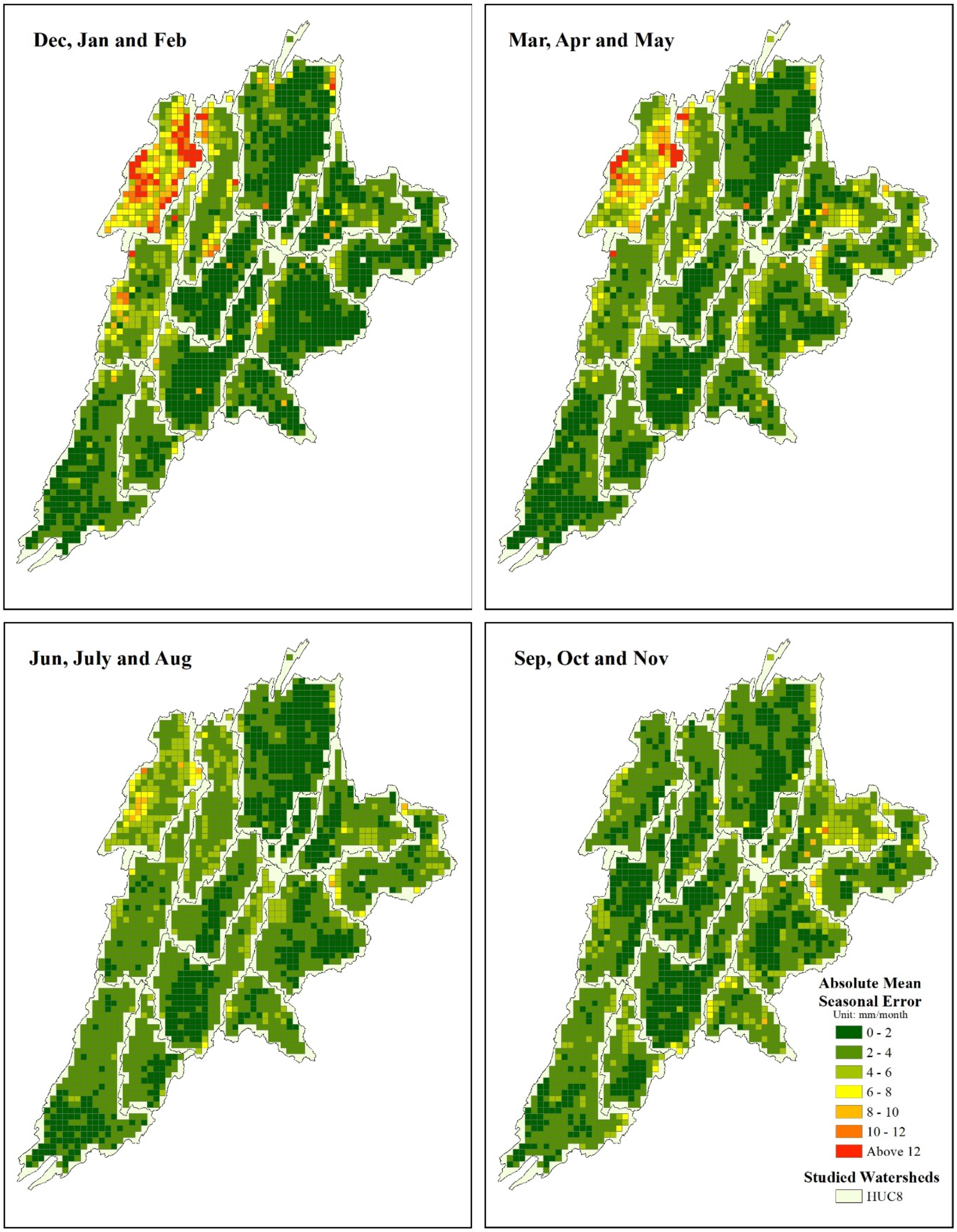

Figure 4. Absolute Mean Seasonal Error (mm at monthly level) from the 12 monthly models across 13 watersheds studied at 4km grid network. A commonly used season separation is used: Dec – Feb as winter, Mar – May as spring, Jun – Aug as summer, and Sep – Nov as fall.

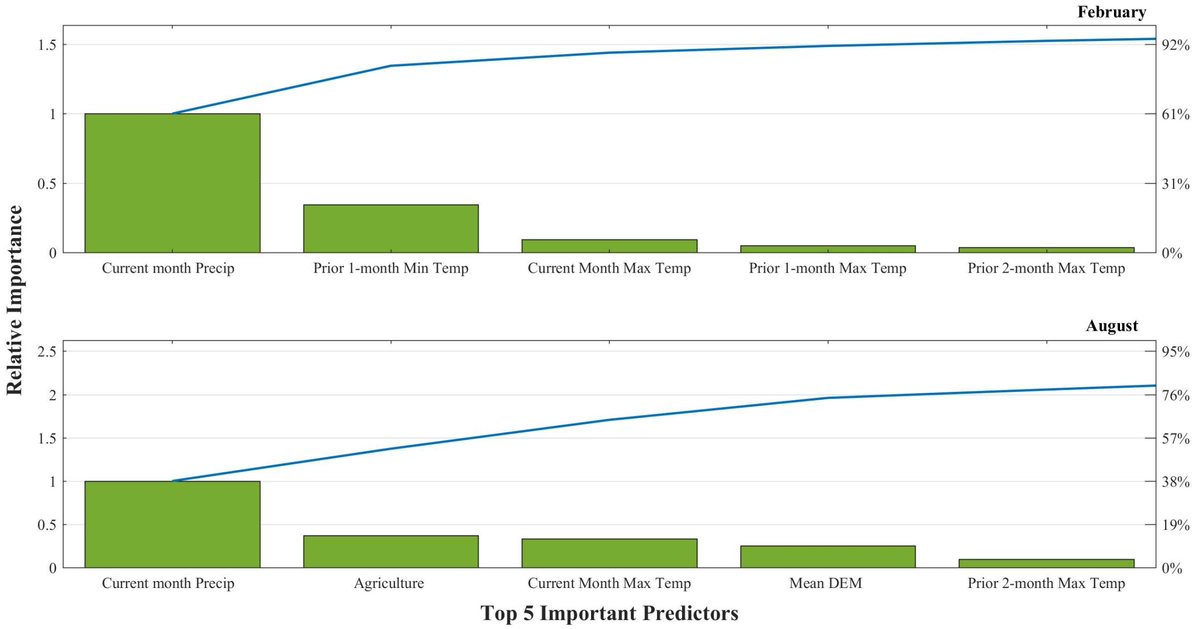

Figure 5. Top 5 important predictors in February (highest R^2) and August (lowest R^2) and their relative importance to all predictors. We normalized every predictor's variable importance values by dividing the most important predictor's importance value to get the relative importance.

References

Abbott, M. B., Bathurst, J. C., Cunge, J. A., O'connell, P. E., & Rasmussen, J. (1986). An introduction to the European Hydrological System—Systeme Hydrologique Europeen,"SHE", 2: Structure of a physically-based, distributed modelling system. *Journal of Hydrology*, *87*(1–2), 61–77.

Adamowski, J., Fung Chan, H., Prasher, S. O., Ozga-Zielinski, B., & Sliusarieva, A. (2012). Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resources Research*, *48*(1).

Alcamo, J., Döll, P., Henrichs, T., Kaspar, F., Lehner, B., Rösch, T., & Siebert, S. (2003). Development and testing of the WaterGAP 2 global model of water use and availability. *Hydrological Sciences Journal*, *48*(3), 317–337.

Arnell, N. W. (1999). Climate change and global water resources. *Global Environmental Change*, *9*, S31–S49.

Arnold, J. G., Srinivasan, R., Muttiah, R. S., & Williams, J. R. (1998). Large area hydrologic modeling and assessment part I: model development. *JAWRA Journal of the American Water Resources Association*, *34*(1), 73–89.

Beven, K., & Binley, A. (1992). The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298.

Beven, K. J. (2011). *Rainfall-runoff modelling: the primer*. John Wiley & Sons.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chapelle, O., Vapnik, V., Bousquet, O., & Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning*, *46*(1), 131–159.

Daly, C., Taylor, G. H., Gibson, W. P., Parzybok, T. W., Johnson, G. L., & Pasteris, P. A. (2000). High-quality spatial climate data sets for the United States and beyond. *Transactions of the ASAE*, *43*(6), 1957.

Dawson, C. W., & Wilby, R. (1998). An artificial neural network approach to rainfall-runoff modelling. *Hydrological Sciences Journal*, *43*(1), 47–66.

Dawson, C. W., & Wilby, R. L. (2001). Hydrological modelling using artificial neural networks. *Progress in Physical Geography*, *25*(1), 80–108.

Devia, G. K., Ganasri, B. P., & Dwarakish, G. S. (2015). A review on hydrological models. *Aquatic Procedia*, *4*, 1001–1007.

Di Luzio, M., Srinivasan, R., Arnold, J. G., & Neitsch, S. L. (2002). ArcView interface for SWAT2000. *BRC Report*, 2–7.

Döll, P., Kaspar, F., & Lehner, B. (2003). A global hydrological model for deriving water availability indicators: model tuning and validation. *Journal of Hydrology*, *270*(1), 105–134.

Fry, J. A., Xian, G., Jin, S., Dewitz, J. A., Homer, C. G., Limin, Y., … Wickham, J. D. (2011). Completion of the 2006 national land cover database for the conterminous United States. *Photogrammetric Engineering and Remote Sensing*, *77*(9), 858–864.

Fuka, D. R., Walter, M. T., MacAlister, C., Degaetano, A. T., Steenhuis, T. S., & Easton, Z. M. (2014). Using the Climate Forecast System Reanalysis as weather input data for watershed models. *Hydrological Processes*, *28*(22), 5613–5623.

Gassman, P. W., Sadeghi, A. M., & Srinivasan, R. (2014). Applications of the SWAT model special section: overview and insights. *Journal of Environmental Quality*, *43*(1), 1–8.

Hochachka, W. M., Caruana, R., Fink, D., Munson, A. R. T., Riedewald, M., Sorokina, D., & Kelling, S. (2007). Data-mining discovery of pattern and process in ecological systems. *Journal of Wildlife Management*, *71*(7), 2427–2437.

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., … Megown, K. (2015). Completion of the 2011 National Land Cover Database for the conterminous United States–representing a decade of land cover change information. *Photogrammetric Engineering & Remote Sensing*, *81*(5), 345–354.

Hsu, K., Gupta, H. V., & Sorooshian, S. (1995). Artificial neural network modeling of the rainfall-runoff process. *Water Resources Research*, *31*(10), 2517–2530.

Huang, G.-B., Chen, L., & Siew, C. K. (2006). Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Networks*, *17*(4), 879–892.

Jha, M. K., & Peiffer, S. (2006). *Applications of remote sensing and GIS technologies in groundwater hydrology: past, present and future*. Bayreuth: BayCEER

Kim, Y., Band, L. E., & Ficklin, D. L. (2017). Projected hydrological changes in the North Carolina piedmont using bias-corrected North American Regional Climate Change Assessment Program (NARCCAP) data. *Journal of Hydrology: Regional Studies*, *12*, 273–288.

Kim, Y., Band, L. E., & Song, C. (2014). The influence of forest regrowth on the stream discharge in the North Carolina Piedmont watersheds. *JAWRA Journal of the American Water Resources Association*, *50*(1), 57–73.

Klepper, O., & Van Drecht, G. (1998). WARiBaS, Water Assessment on a River Basin Scale. A computer program for calculating water demand and water satisfaction on a catchment basin level; to be used for global scale water stress analysis. *RIVM Rapport 402001009*.

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research: Atmospheres*, *99*(D7), 14415–14428.

Liang, X., Xie, Z., & Huang, M. (2003). A new parameterization for surface and groundwater interactions and its impact on water budgets with the variable infiltration capacity (VIC) land surface model. *Journal of Geophysical Research: Atmospheres*, *108*(D16).

Meigh, J. R., McKenzie, A. A., & Sene, K. J. (1999). A grid-based approach to water scarcity estimates for eastern and southern Africa. *Water Resources Management*, *13*(2), 85–115.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885–900.

Omernik, J. M. (1987). Ecoregions of the conterminous United States. *Annals of the Association of American Geographers*, *77*(1), 118–125.

Pilgrim, D. H., Chapman, T. G., & Doran, D. G. (1988). Problems of rainfall-runoff modelling in arid and semiarid regions. *Hydrological Sciences Journal*, *33*(4), 379–400.

Pinkney, A. E., Harshbarger, J. C., May, E. B., & Melancon, M. J. (2001). Tumor prevalence and biomarkers of exposure in brown bullheads (Ameiurus nebulosus) from the tidal Potomac River, USA, watershed. *Environmental Toxicology and Chemistry*, *20*(6), 1196–1205.

Shamseldin, A. Y. (1997). Application of a neural network technique to rainfall-runoff modelling. *Journal of Hydrology*, *199*(3–4), 272–294.

Shao, Y., & Lunetta, R. S. (2012). Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points. *ISPRS*

Journal of Photogrammetry and Remote Sensing, *70*, 78–87.

Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences*, *20*(7), 2611.

Smith, J., & Eli, R. N. (1995). Neural-network models of rainfall-runoff process. *Journal of Water Resources Planning and Management*, *121*(6), 499–508.

Sudheer, K. P., Gosain, A. K., & Ramasastri, K. S. (2002). A data-driven algorithm for constructing artificial neural network rainfall-runoff models. *Hydrological Processes*, *16*(6), 1325–1330.

Tokar, A. S., & Johnson, P. A. (1999). Rainfall-runoff modeling using artificial neural networks. *Journal of Hydrologic Engineering*, *4*(3), 232–239.

Vörösmarty, C. J., Federer, C. A., & Schloss, A. L. (1998). Potential evaporation functions compared on US watersheds: Possible implications for global-scale water balance and terrestrial ecosystem modeling. *Journal of Hydrology*, *207*(3–4), 147–169.

Yang, W., Chen, P., Villegas, E. N., Landy, R. B., Kanetsky, C., Cama, V., … Prelewicz, G. J. (2008). Cryptosporidium source tracking in the Potomac River watershed. *Applied and Environmental Microbiology*, *74*(21), 6495–6504.

Yates, D. N. (1997). Approaches to continental scale runoff for integrated assessment models. *Journal of Hydrology*, *201*(1), 289–310. https://doi.org/https://doi.org/10.1016/S0022-1694(97)00044-9