EVALUATION OF THE SOIL AND WATER ASSESSMENT TOOL AND ARTIFICIAL NEURAL NETWORK AS RAINFALL-RUNOFF MODELS IN A RANGE OF CONDITIONS

A Dissertation

by

XIAOHAN MEI

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Patricia Smith
Committee Members,	Anthony Filippi
	Huilin Gao
	Inci Guneralp
Head of Department,	Ronald Kaiser

December 2021

Major Subject: Water Management and Hydrological Science

ABSTRACT

The Soil and Water Assessment Tool (SWAT) and the Artificial Neural Network (ANN) have been widely used as rainfall-runoff models since the 1990s. The former is a more complex, physically-based model that went through decades of continuous development, while the latter is a simpler data-driven model that focuses on establishing the nonlinear relationship between predictors and targets without considering the physical aspects of hydrological systems. Although both SWAT and ANN are broadly accepted as capable of making successful streamflow estimations, their performance capability had not been adequately compared under various conditions in the past. This dissertation seeks to create watershed level rainfall-runoff models using SWAT and ANN across a range of settings and evaluate their performance capability.

In ANN rainfall-runoff modeling, the three-layered feed-forward neural network is regularly used. Routinely, several neural networks are trained before a model selection process selects the network with the best predictive capability. In study I, two common model selection approaches, including the in-sample approach that is based on Akaike's information criterion (AIC) and Bayesian information criterion (BIC), and the out-of-sample approach that uses blocked cross-validation (BlockedCV), were compared. The results suggested that the BlockedCV is preferable for selecting the rainfall-runoff model with the best predictive capability.

Study II directly compared the SWAT and ANN models' streamflow predictive performance in two small watersheds in the karstified region of San Antonio, Texas. The

paired watershed approach was employed, with one study watershed being highly urbanized and the other primarily covered with evergreen forest and shrub. In addition, the study used the correction factor approach to adjust the goodness-of-fit indicators to incorporate measurement and model uncertainty in the rainfall-runoff modeling process. The results showed that ANN slightly outperformed SWAT in the urban watershed and performed significantly better in the rural watershed. Therefore, suggesting that ANN is a better real-time simulator of streamflow.

Additionally, as gridded precipitation datasets are gaining popularity as a convenient alternative for hydrological modeling during recent decades, Study III evaluated three gridded precipitation datasets, the Tropical Rainfall Measuring Mission (TRMM), the Climate Forecast System Reanalysis (CFSR), and the Parameter-elevation Relationships on Independent Slopes Model (PRISM), against the conventional gauge rainfall observations, and further assessed their capability of driving hydrological simulations in SWAT and ANN. The results of Study III showed that SWAT and ANN simulation outcomes varied in an identical pattern when different precipitation data were applied. Moreover, the PRISM and TRMM driven models were found to have preferable streamflow prediction results than the CFSR and gauge driven models, with the PRISM data produced the best hydrological simulation outcome.

DEDICATION

I would like to dedicate my dissertation to my beloved parents.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Smith, and my committee members, Dr. Filippi, Dr. Gao, and Dr. Guneralp, for their guidance and support throughout the course of this research.

I would like to give a special thanks to my major advisor, Dr. Patricia Smith, for always encouraging me to explore the research topic I am interested in, and always helping me with great patience.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

Finally, thanks to my parents for their unrequited support.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis (or) dissertation committee consisting of Professor Patricia Smith [advisor] of the Department of Biological and Agricultural Engineering, Professor Huilin Gao of the Department of Civil and Environmental Engineering, Professor Anthony Filippi, and Professor Inci Guneralp of the Department of Geography.

Some of the gridded weather datasets in Chapter 4 were collected with the help of my colleague, Dr. Gang Zhao of the Department of Civil and Environmental Engineering.

All other work conducted for the thesis (or) dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by Texas A&M University Department of Water Management & Hydrological Science graduate scholarship and Department of Biological & Agricultural Engineering annual scholarship.

NOMENCLATURE

AIC	Akaike’s information criterion
ANN	Artificial Neural Network
BFZ	Balcones Fault Zone
BIC	Bayesian information criterion
BlockedCV	blocked cross-validation
CC	correlation coefficient
CFSR	Climate Forecast System Reanalysis
Cv	coefficients of variation
DEM	digital elevation model
DO	degree of overlap
GEE	Google Earth Engine
GIS	geographic information system
HRUs	hydrologic response units
HSARB	Headwaters San Antonio River Basin
H/WQ	hydrologic and water quality
LCW	Leon Creed Watershed
i.i.d.	independent and identically distributed
LMRB	Lower Medina River Basin
LOOCV	leave-one-out cross-validation
LULC	land use land cover

MPEG	Multi-Sensor Precipitation Estimate–Geostationary
NED	National Elevation Dataset
NLCD2011	2011 National Land Cover Database
NRCS	Natural Resources Conversation Service
NSE	Nash-Sutcliffe coefficient of efficiency
PBIAS	percent bias
PEC	performance evaluation criteria
PMs	performance measures
PRISM	Parameter-elevation Relationships on Independent Slopes Model\
R^2	coefficient of determination
RMSE	root mean square error
RSR	root mean square error to observation standard deviation ratio
SRPs	satellite rainfall products
STATSGO	state soil geographic database
SWAT	Soil and Water Assessment Tool
SWAT-CUP	SWAT Calibration and Uncertainty Programs
TMDLs	total maximum daily loads
TRMM	Tropical Rainfall Measuring Mission
UMRW	Upper Medina River Watershed
USGS	United States Geological Survey

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

1.1. Background

Modeling of the rainfall-runoff process is of great importance in surface water hydrology. As noted by Beven (2011), its main reason is to extrapolate the available data in space and time because hydrological measurements always have a limited range and often fail to meet what we would like to know about the hydrological systems. The results of rainfall-runoff modeling are often used to support decision-making and serving as the foundation of other more advanced research topics in water resources planning and management. The ultimate purpose of making a model prediction is to aid decision making for a range of hydrological problems, for instance, estimating flood return intervals, assisting river hydraulics modeling and engineering project design, setting total maximum daily loads (TMDLs) standards, and conducting environmental impact analysis (Karunanithi et al., 1994; Wurbs et al., 2002). In the past few decades, different hydrological models have been developed to simulate streamflow on various spatial and temporal scales.

One major issue in rainfall-runoff modeling is to decide the appropriate level of model complexity. Beven (2011) summarized two widely accepted views of modeling. The first suggests that hydrological models are merely tools for extrapolating available data in time and space. Therefore if the model inputs and outputs can be successfully related, it is nonessential to elaborate the details of a watershed. The second view maintains that models should reflect the involved physical processes to the extent possible to ensure confidence when extrapolating beyond the existing observations. The more complex physically-based hydrological models are usually adopted under the second perspective. The Soil and Water Assessment Tool (SWAT) is a physically-based, semi-distributed, deterministic model developed to assess water quality and quantity in large river basins with varying soils, land uses, land cover types, and management practices (Arnold et al., 2012b). In the SWAT model, a study watershed is divided into a few subbasins. To a finer spatial scale, the model further groups lands with homogeneous slopes, soil types, and land cover types into hydrologic response units (HRUs). The HRUs may not be spatially continuous and can be found at different locations within a subbasin. This strategy effectively simplifies the simulation of watershed processes (Gassman et al., 2007; Licciardello et al., 2011; Tuppad et al., 2011). The development of the SWAT model has spanned over the last three decades, with new functions and routines continuously added to the model. The current SWAT model has been widely applied to water, sediment, agricultural chemicals, and contaminant yields in complex systems (Abbaspour et al., 2015; Gassman et al., 2007). Moreover, the SWAT model is closely integrated with the geographic information system (GIS) since most SWAT input data have spatial characters (Jayakrishnan et al., 2005; Olivera et al., 2006; Srinivasan et al., 1994). To date, the ArcGIS and QGIS platforms are integrated with the SWAT model.

A physically-based hydrological model is often referred to as a white-box since its modeling processes are established on known scientific principles of mass and energy fluxes (Moradkhani et al., 2009). Contrary to the complex structure of physically-based models, statistical models also gain popularity for rainfall-runoff modeling due to their simplicity and low computational resource demand. An Artificial Neural Network (ANN) is a computing system that resembles the structure of the human biological neural network. It can identify nonlinear relationships from given patterns and fit nonparametric models on multivariate input data (Govindaraju et al., 2013). ANN has been applied to modeling many components of the hydrological cycle since the 1990s. (ASCE, 2000a). For instance, ANN was applied to deriving rainfall estimates from satellite imagery (Hsu et al., 1997); simulating groundwater recharge in a small-scale watershed (Rogers, 1992). Water quality variables, including various types of nutrients, dissolved oxygen, raw watercolor, and salinity, were successfully estimated using ANN in a few studies (Gazzaz et al., 2012; Kalin et al., 2010; Maier et al., 1996; Sahoo, Ray, et al., 2006; Singh et al., 2009; Zhang et al., 1997); and several studies have reported satisfactory ANN modeling results on streamflow (Ahmed et al., 2007; Birikundavyi et al., 2002; Hu et al., 2001; Humphrey et al., 2016; Isik et al., 2013; Karunanithi et al., 1994; Kişi, 2007; Rezaeianzadeh et al., 2013). When used as a rainfall-runoff model, it only identifies the relationship between historical inputs (meteorological data) and outputs (streamflow) without considering any of the physical processes involved, and therefore fall into the category of lumped model and is often referred to as a black-box by the modelers (ASCE, 2000a). In addition, the ANNs come under the category of stochastic models, given that the fitted parameter values often vary from one training process to another for a fixed training dataset (Jain et al., 2004).

The physically-based SWAT model simulates streamflow by incorporating descriptive mathematical equations designed to conceptualize the hydrological processes. The model requires grid-based geospatial data as input besides the meteorological data and incorporates significant amounts of model parameters (Fatichi et al., 2016; Noori et al., 2016). Solving the equations for the state variables at the level of HRUs for each time step can be time-consuming and computationally demanding (Jimeno-Sáez et al., 2018; Noori et al., 2016). In this regard, statistical models such as ANN hold a clear advantage. ANNs do not require *a priori* knowledge of the watershed physical characteristics as model input, which reduces the model setup procedures. Meanwhile, the time it takes to train and select the best neural network is significantly shorter than calibrating a SWAT model (Jimeno-Sáez et al., 2018; Minns et al., 1996), while is capable of achieving “unreasonably effectiveness” in hydrological applications when sufficient training data is available (Worland et al., 2019). On the other hand, several studies have noted the disadvantages of applying ANN as a rainfall-runoff model. In particular, its lack of physical explanations for the underlying hydrological processes has generated a lot of concern among the hydrologists (ASCE, 2000a; Ha et al., 2003; Jain et al., 2004; Karunanithi et al., 1994; Yaseen et al., 2015). Being incapable of capturing physical dynamics at the watershed level means that ANNs are not suitable for environmental impact studies such as modeling streamflow under changing climate conditions (Humphrey et al., 2016; Milly et al., 2008). Additionally, similar to other statistical models, extrapolating beyond the training data range often undermines ANNs' predictive performance (Minns et al., 1996; Sahoo, Ray, et al., 2006).

A few studies have made the comparison between SWAT and ANN regarding streamflow prediction. Kim et al. (2015) applied SWAT and ANN to impute missing streamflow observations in the Taehwa River Watershed in Korea and found that SWAT was better at simulating low flows, while the neural network model generally performed better at simulating high flows. Jimeno-Sáez et al. (2018) compared the accuracy of streamflow prediction between SWAT and ANN models at the Ladra River Basin and the Segura River Basin in Spain. The authors came to the similar conclusion that ANN is superior at estimating higher flows. Demirel et al. (2009) applied SWAT and ANN models to simulate streamflow in the Pracana River Basin in Portugal and concluded that ANN was more successful at forecasting peak streamflow, whereas the SWAT model performed better on goodness-of-fit indicators. Srivastava et al. (2006) analyzed the performance of streamflow prediction from SWAT and ANN in the agricultural-dominated Honey Brook Watershed in Pennsylvania. They found that the ANN model produced simulation results with the Nash-Sutcliffe coefficient of efficiency (NSE) and coefficient of determination (R^2) better than the SWAT model. In a more recent study, Zakizadeh et al. (2020) used ANN and SWAT to simulate the rainfall-runoff relationship in a small watershed near Tehran city, Iran. The authors concluded that ANN produced simulations with minor error and uncertainty, although both models could achieve excellent predictive performance.

1.2. Research Objectives

While the literature has demonstrated that ANN models can make streamflow successful predictions, and in many cases, even performs better than the more complex SWAT model, the issue of how ANN models would behave in watersheds with different dominant land use land cover types has not been fully addressed. As the process of urbanization continues in Texas and many other parts of the world (Zhang et al., 2018; Zhao et al., 2016), it has become increasingly meaningful to explore if a neural network based model could make reliable predictions for urbanized areas, as well as evaluating how the performance of ANN varies from highly developed urban watersheds to undeveloped rural watersheds. Additionally, it would be meaningful to compare ANN performance with the physically-based SWAT model under these different settings.

Since ANN models tackle the rainfall-runoff system entirely through an input-output manner, deciding the appropriate model structure is essential for accurate streamflow simulation. Routinely, several ANN models are trained, and a model selection phase is applied to find the model with the best generalization capability (Donate et al., 2013). Two main types of model selection approaches are often adopted for this purpose, the out-of-sample approach, which is based on cross-validation that divides the available data into training, validation, and testing sets; and the in-sample approach, which solely relies on in-sample criterion, most notably the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC), for deciding the models' generalization capability (Qi et al., 2001). While it is generally accepted that the performance of statistic models should be assessed using out-of-sample tests rather than in-sample errors (Tashman, 2000), benefits and disadvantages exist for both approaches. Bergmeir et al. (2012) noted that the distinct nature of different time series could cause a model selection method to work well with a specific type of time series but show poor performance on the others. Hence, it is yet to be evaluated if the model selection outcome for these two approaches converges for a hydrological time series.

Additionally, the data quality and input combination of meteorological and hydrological forces are crucial for building an optimal network structure (Maier et al., 2000; Noori et al., 2016) as well as influence the outcome of the SWAT model. In recent decades, grid-based meteorological products, either produced directly from remote sensing platforms or created as a combination of ground and satellite observations, have become more widely available for application in hydrological models. Notable examples of the grid-based weather product include the Parameter-elevation Regressions on Independent Slopes Model (PRISM), the Climate Forecast System Reanalysis (CFSR), and the Tropical Rainfall Measuring Mission (TRMM). In comparison with traditional ground-based measurements, the gridded meteorological datasets provide more extensive spatial and more consistent temporal coverage (Meresa, 2019). However, their use in ANN and SWAT models has not been fully assessed across a range of conditions.

While previous studies have explored the suitability of SWAT and ANN models in various regions, none of these studies were conducted in the state of Texas, where its hot climate condition and growing population have made water sustainability a contentious issue. Besides, the San Antonio region of central south Texas has extensive karst terrain, making hydrological modeling more difficult and rendering model performance patterns vastly different from non-karstified watersheds (Jakada et al., 2020). Furthermore, very few studies on SWAT and ANN considered uncertainty in the measurement and modeling process.

Therefore, the objectives of this dissertation were to (1) compare the efficacy of the in-sample and out-of-sample modeling selection approaches in ANN rainfall-runoff modeling; (2) compare SWAT and ANN streamflow predictions in a pair of watersheds with different dominant land cover type, while to enhance the model evaluation using modified goodness-of-fit indicators that incorporate measurement and model uncertainty; (3) assess the accuracy of three common gridded weather datasets, and evaluate how SWAT and ANN respond to the application of different weather data.

2. A COMPARISON OF IN-SAMPLE AND OUT-OF-SAMPLE MODEL SELECTION APPROACHES FOR ARTIFICIAL NEURAL NETWORK (ANN) DAILY STREAMFLOW SIMULATION*

Abstract. Artificial Neural Networks (ANN) have been widely applied in hydrologic and water quality (H/WQ) modeling in the past three decades. Many studies have demonstrated an ANN's capability to successfully estimate daily streamflow from meteorological data on the watershed level. One major challenge of ANN streamflow modeling is finding the optimal network structure with good generalization capability while ameliorating model overfitting. This study empirically examines two types of model selection approaches for simulating streamflow time series: the out-of-sample approach using blocked cross-validation (BlockedCV) and an in-sample approach that is based on Akaike's information criterion (AIC) and Bayesian information criterion (BIC). A three-layer feed-forward neural network using a back-propagation algorithm is utilized to create the stream-flow models in this[1] study. The rainfall–streamflow relationship of two adjacent, small watersheds in the San Antonio region in south-central Texas are modeled on a daily time scale. The model selection results of the two approaches are compared, and some commonly used performance measures (PMs) are generated on the stand-alone testing datasets to evaluate the models selected by the two

*Mei, X., & Smith, P. K. (2021). A Comparison of In-Sample and Out-of-Sample Model Selection Approaches for Artificial Neural Network (ANN) Daily Streamflow Simulation. Water, 13(18), 2525.

approaches. This study finds that, in general, the out-of-sample and in-sample approaches do not converge to the same model selection results, with AIC and BIC selecting simpler models than BlockedCV. The ANNs were found to have good performance in both study watersheds, with BlockedCV selected models having a Nash–Sutcliffe coefficient of efficiency (NSE) of 0.581 and 0.658, and AIC/BIC selected models having a poorer NSE of 0.574 and 0.310, for the two study watersheds. Overall, out-of-sample BlockedCV selected models with better predictive ability and is preferable to model streamflow time series.

2.1. Introduction

The estimation of streamflow time series on the watershed scale is of great importance in surface water hydrology. Accurate streamflow is the foundation of water resources planning and management, including river hydraulics modeling and engineering project design, water demand assessment and allocation, and water quality studies (Fernandez et al., 2005; Wurbs et al., 2002). Data-driven methods have gained popularity in hydrologic and water quality (H/WQ) modeling in recent years due to their effectiveness in mapping connections between hydrologic inputs and outputs (Worland et al., 2019). Among these methods, the artificial neural network (ANN) has proven to be an effective tool in water resources modeling (Humphrey et al., 2016; Maier et al., 2000).

An ANN is a “parallel-distributed processor” that resembles the biological neural network structure of the human brain. The ANN acquires knowledge or information from a learning process and stores that knowledge in interneuron links using a weighted matrix. The early concept of ANNs as a computational tool was formalized in the 1940s. It went through gradual development in the ensuing decades as computers become more accessible and computational efficiency grew (Govindaraju et al., 2013). ANN-based models hold some clear advantages over conventional conceptual models in H/WQ modeling. ANNs do not require *a priori* knowledge of the physical characteristics of the study watershed as model input, thus significantly reducing the procedures for model setup and simulation (Jimeno-Sáez et al., 2018; Minns et al., 1996). When sufficient data have been provided, ANN models have produced satisfactory results for streamflow forecasting, according to a review provided by Yaseen et al. (2015). However, some believe modeling hydrologic systems with ANN without explaining the underlying physical processes is a significant drawback. For instance, the lack of capability to capture physical dynamics at the watershed level means that ANNs are not suitable for modeling streamflow under changing climate or land use conditions. In addition, the ANNs predictive capability is often unreliable beyond the training data range due to the absence of physical explanation (ASCE, 2000a; Ha et al., 2003; Jain et al., 2004; Karunanithi et al., 1994; Yaseen et al., 2015). Worland et al. (2019) further recommended that machine-learning models such as ANN only be used for making predictions rather than gaining hydrological insights.

The ANN application in hydrology began in the 1990s. Since then, many studies have applied ANN in H/WQ modeling. Several studies have reported satisfactory results using ANN for streamflow estimation. Karunanithi et al. (1994) demonstrated successful streamflow prediction at Huron River in Michigan using neural network models in an early study. The predicted flow closely matches the timing and magnitude of the actual flow. Ahmed et al. (2007) used three data-driven models to generate synthetic streamflow for the Pagladia River in northeast India and concluded that the ANN-based model has the best performance. Birikundavyi et al. (2002) compared ANN to an autoregressive model to forecast daily streamflow in the Mistassibi River in northeastern Quebec. They obtained results showing that the ANN model outperformed the autoregressive model. Similarly, Hu et al. (2001) showed an ANN-based model that simulated daily streamflow and annual reservoir inflow for two watersheds in northern China outperformed an autoregressive model. Humphrey et al. (2016) coupled an ANN model and a conceptual rainfall-runoff model to produce a monthly streamflow forecast for a drainage network in southeast Australia. They reported that the hybrid model outperformed the original conceptual model, especially for high flow periods. Isik et al. (2013) reported that accurate daily streamflow prediction was achieved using a hybrid model based on ANN and the SCS (Soil Conservation Service, the US Department of Agriculture) curve number method. Kişi (2007) compared four different ANN algorithms for streamflow forecasting, all of which reached satisfactory statistical results with correlation coefficients of all four models close to 1. Rezaeianzadeh et al. (2013)

simulated daily watershed outflow at the Khosrow Shirin watershed in Iran using an ANN and HEC-HMS and concluded that the ANN model with a multi-layer perceptron was more efficient in forecasting daily streamflow.

Although the literature has demonstrated that ANN models can make satisfactory streamflow predictions, the ANN-based models often suffer from overfitting problems due to a relatively large number of parameters to be estimated compared with other statistical-based models (Zhang et al., 2005). Model overfitting often refers to ANNs fitting the in-sample data (training set) well but the out-of-sample data (testing set) poorly. Selecting the appropriate model structure is crucial for accurately simulating streamflow while ameliorating overfitting. Routinely, several ANN models are trained, and a model selection phase is applied to find the model with the best generalization capability (Donate et al., 2013). Two main types of model selection approaches are often adopted for this purpose. The out-of-sample approach, based on cross-validation, divides the available data into training, validation, and testing sets. The in-sample approach relies on in-sample criterion calculated on the training dataset, most notably the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC), for deciding the models' generalization capability (Qi et al., 2001). Although it is generally accepted that the performance of statistical models should be assessed using out-of-sample tests rather than in-sample errors (Tashman, 2000), the in-sample model selection approach has the clear advantage of utilizing all available data for modeling training while avoiding data splitting. Previous studies have discussed the benefits and disadvantages of the two approaches. Arlot et al. (2010) argued that out-of-sample cross-validation is more applicable in many practical situations. Qi et al. (2001) showed that the results of a few in-sample model selection criteria were not consistent with the out-of-sample performance for three economic time series. On the other hand, a study conducted by Shao (1997) concluded that AIC and leave-one-out cross-validation (LOOCV) converge to the same model selection result. It is unclear how the out-of-sample and in-sample model selection approaches will perform on hydrological time series without actual experimentation. Additionally, as noted by Bergmeir et al. (2012), the distinct nature of different time series can cause a model selection method to work well with a particular type of time series but show poor performance on others. Hence, it is yet to be evaluated if the model selection outcomes of these two approaches converge for a hydrological time series.

The main objectives of this study are: 1) to create ANN rainfall-streamflow models on the watershed level, 2) determine the optimal model structure of the ANNs using both in-sample and out-of-sample model selection approaches, 3) compare the model selection results of these two approaches, and 4) empirically investigate their efficacy in selecting the optimal neural network. The task is to be accomplished using two small watersheds in the San Antonio Region of south-central Texas, one of which is dominated by a well-developed urban landscape. An agricultural landscape primarily covers the other. The streamflow simulations are to be conducted on a daily time step for the outflows of both watersheds. The following sections will describe the details of the models.

2.2. Materials and Methods

2.2.1. Study Area and Data Acquisition

The city of San Antonio is in the sub-tropic and semi-humid climate zone of south-central Texas. It has long hot summers and warm to cool winters. Snowfall has been reported historically, although it is rare. The San Antonio region is about 210 m above sea level and has an annual precipitation of around 770 mm (Joseph et al., 2013). Combining the rapidly growing total population and a decline of population density at its urban center, San Antonio is one of the fastest-growing metropolitan areas in the US (Kreuter et al., 2001; Zhao et al., 2016).

This study selected two adjacent small watersheds in the San Antonio region for streamflow modeling (Figure 2.1). The Headwaters of the San Antonio River Basin (HSARB, HUC10: 1210030102) is centered at 98.507º west longitude, 29.422º north latitude, and mainly covers the extent of central downtown San Antonio. HSARB has a drainage area of 395.84 km^2, of which 81.66% is classified as developed urban area according to the 2011 National Land Cover Database (NLCD2011) land use land cover (LULC) classification (Yang et al., 2018). The main waterway in the HSARB is the San Antonio River, which originates in the metropolitan area of San Antonio, flows southeast across downtown San Antonio and merges with the Medina River in the city's southern suburbs. Thus, the most downstream point of the drainage basin is located at the southern tip of the HSARB. The streamflow close to the watershed outlet is measured by a USGS surface streamflow gauge (USGS 08178565). The Lower Medina River Basin (LMRB) centered at 98.698º west longitude, 29.319º north latitude, is located west of the San Antonio urban area, and shares a short watershed boundary with HSARB. LMRB has a drainage area of 929.29 km^2 and is much less developed in comparison to the HSARB. The dominant land cover types at LMRB are shrub, pasture, and cultivated crop, covering 25.97%, 15.72%, and 14.38% of the entire LMRB, respectively. The major waterway in LMRB is the Medina River, which flows southeast and merges into the San Antonio River at the outlet of LMRB. The streamflow measurement station closest to the watershed outlet is USGS gauge 08181500, which covers most of the drainage area of the LMRB, located about 7 kilometers from the watershed outlet. The proximity in geographic locations of the two study watersheds helps to reduce uncertainties that may arise in model comparison, as the neighboring watersheds have similar climatology, geology, and hydrology.

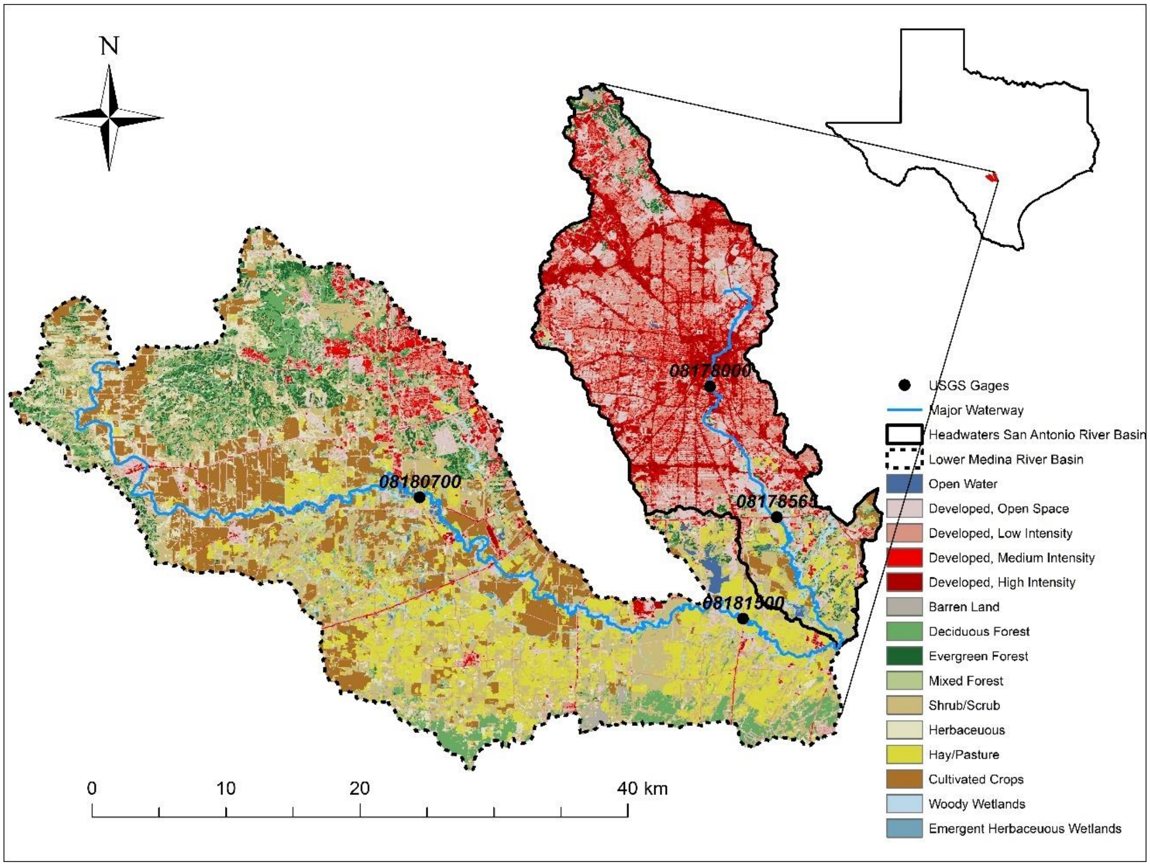

Figure 2.1 Study Area, the Headwaters San Antonio River Basin (HSARB), and the Lower Medina River Basin (LMRB) with NLCD LULC classification and USGS Stream Gages displayed.

Meteorological and hydrological data of the two watersheds were used to train the neural network models. The Parameter-elevation Regressions on Independent Slopes Model (PRISM) daily spatial climate dataset AN81d, produced by the PRISM Climate Group at Oregon State University (Daly et al., 2008), is used for meteorological inputs in this study. PRISM data is accessed using the Google Earth Engine (GEE), a cloud-based geospatial analysis platform that provides access to freely available geospatial data archives produced by multiple government agencies (Gorelick et al., 2017). Polygon masks that cover each watershed are uploaded to the GEE server. The area-averaged daily precipitation and mean temperature are calculated and separately obtained for the two masked areas. The daily discharge observations for the gauges closest to the watershed outlets and their corresponding upstream gauges are obtained from the USGS surface water daily measurement (U.S. Geological Survey, 2016).

2.2.2. ANN Model Description

The purpose of applying ANN as a rainfall-streamflow model is to create a specific model structure that can capture the nonlinear relationships between the input precipitation and target streamflow. A multi-layer feed-forward neural network normally has one input layer, one output layer, and at least one hidden layer that connects the input and output layers. A three-layer feed-forward neural network using a back-propagation algorithm is often employed in hydrological modeling and is generally sufficient for streamflow and water quality simulations (ASCE, 2000a; Gupta et al., 2000; Minns et al., 1996). Each layer possesses at least one node (or neuron) for a standard three-layer feed-forward neural network, and each node is connected to all other nodes in its adjacent layers. The connection links between nodes contain associated weights and biases that represent their connection strength. At each node, a nonlinear transformation often referred to as a transfer function, is applied to the net input of this node to calculate its corresponding output signal. The weights and biases are randomly initialized before training begins and updated using the back-propagation step in every training epoch (ASCE, 2000a). The operation at a node can be defined using equation (1),

$$y = f(\sum_{i=1}^{n} w_i x_i + b) \tag{1}$$

where y is the node's output signal; f is the transfer function; w_i represents the weight vectors associated with the interneuron links; x_i is the input vector, and b is the bias (ASCE, 2000a).

ANN models use the back-propagation algorithm to update the weight and bias terms described in equation (1). The back-propagation algorithm is one of the most popular algorithms for ANN training (ASCE, 2000a; Zhang et al., 2000). Back-propagation minimizes the network loss function, which is usually in the squared error form as described in equation (2), using the gradient descent method (ASCE, 2000a),

$$E = \sum_{N} \sum_{p} (y_i - t_i)^2 \tag{2}$$

where t_i is the observation, y_i is the corresponding ANN prediction, N is the number of observations, and p is the number of output nodes.

The loss computed from equation (2) is propagated backward through the network to each node, and the weights are updated along the steepest descent of the loss function in every training epoch. The weight change of an epoch can be described with equation (3) (ASCE, 2000a),

$$\Delta w_{ij}(n) = -\varepsilon \cdot \frac{\partial E}{\partial w_{ij}} + \alpha \cdot \Delta w_{ij}(n-1) \tag{3}$$

where $\Delta w_{ij}(n)$ and $\Delta w_{ij}(n-1)$ are weight increments between node i and j during the nth and $(n-1)$th epoch, E is the loss function computed using equation (2), and ε and α are learning rate and momentum constant (ASCE, 2000a).

2.2.3. Model Selection Approaches

2.2.3.1. Blocked Cross-Validation

Traditionally, the out-of-sample model selection approach is more often used on ANN models (Tashman, 2000). Cross-validation is the simplest and most widely used method for estimating prediction error according to Hastie et al. (2009). It directly uses the out-of-sample error for model selection. In H/WQ modeling, the cross-validation-based approach was applied in several previous studies (Amiri et al., 2012; Gazzaz et al., 2012; Humphrey et al., 2016; Jimeno-Sáez et al., 2018; Kim et al., 2015; Maier et al., 1996; Srivastava et al., 2006). The model cross-validation procedure splits the available data into three groups: training, validation, and testing. The training set is used for model training, during which the free parameters (i.e., interneuron weights and biases) are estimated for several ANN models with specific model structures. Following the training set, the prediction performance for the models is calculated on the hold-out validation set, and the model that has the best validation performance is selected. The testing set is an independent dataset used for stand-alone measurement of model generalization capability (ASCE, 2000a; Karunanithi et al., 1994). Nevertheless, Bergmeir et al. (2012)

stated that time series data are intrinsically ordered. Therefore, its time dependency and autocorrelation contradict the basic assumption of traditional cross-validation that the data is independent and identically distributed (i.i.d.). The authors have further suggested using blocks of data rather than resampling data randomly in each cross-validation iteration to avoid breaking the data dependency of the studied time series.

To address the issue that the streamflow time series usually has strong autocorrelation, this study applies blocked cross-validation (BlockedCV) as the out-of-sample model selection approach to be evaluated. Unlike the normal k-Fold cross-validation that randomly samples training and validation data, the training/validation data split in BlockedCV maintains the data points' sequential order by grouping them into several data blocks, and all data blocks are fixed throughout the cross-validation process (Figure 2.2).

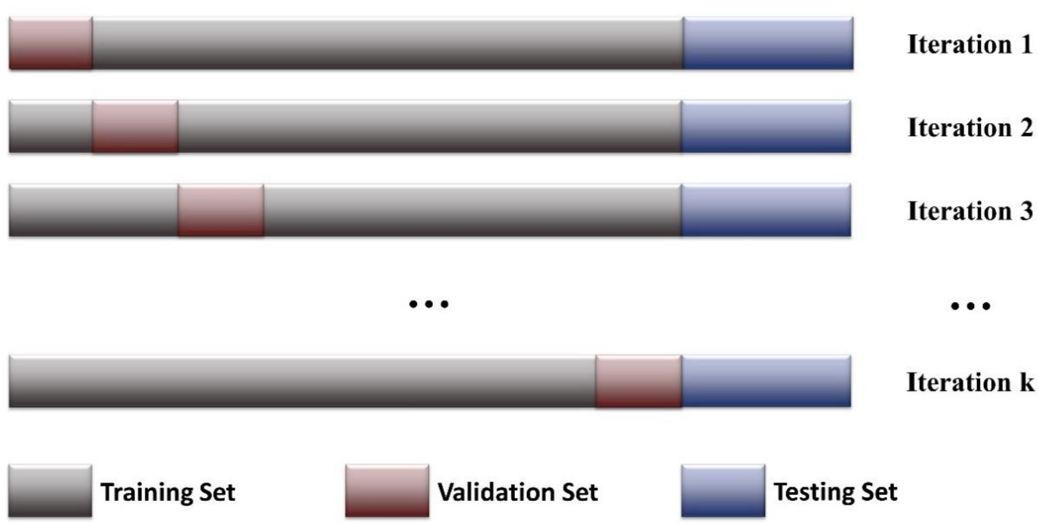

Figure 2.2 Schematic diagram of the blocked cross-validation.

Although the out-of-sample test performance is more often accepted as reliable for model selection, some of its limitations are noted in the literature. For example, James et al. (2013) summarized from past studies that data splitting might significantly increase the variability of the estimates. As the partition of available data is usually done subjectively by the modelers, the performance results can be intensely dependent on where the data splitting takes place between training and validation sets and hence influence the outcome of the selected best model. Moreover, since statistical methods tend to perform better when trained with larger datasets, omitting part of the available data for model validation may compromise model training, especially when the size of the available dataset is small (Bergmeir et al., 2012).

2.2.3.2. AIC and BIC

The in-sample model selection approach is based on the in-sample criterion calculated for the training dataset, which avoids data splitting and enables utilizing all available data for model training. This approach usually considers the in-sample estimation error and model complexity together in its various forms of equations. The size of estimated parameters is often added into the equations as a penalty term since a more complex model is more likely to cause model overfitting. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are two of the most widely used in-sample model selection criteria (Qi et al., 2001). AIC and BIC are founded on information theory and are motivated by the need for balancing goodness-of-fit and model complexity (Sheather, 2009). Lower AIC and BIC values indicate a better model fit. Various forms of AIC and BIC can be found in the literature. This study uses the form proposed by Qi et al. (2001). One common form of AIC (Akaike, 1974) is given in equation (4):

$$\text{AIC} = log(\hat{\sigma}^2_{MLE}) + \frac{2m}{N} \tag{4}$$

where m is the number of model parameters, and N is the number of observations. $\hat{\sigma}^2_{MLE}$ is the maximum likelihood estimate of the variance of the residual term, expressed by equation (5):

$$\hat{\sigma}^2_{MLE} = \frac{SSE}{N} = \frac{\sum(y_i - \hat{y}_i)^2}{N} \tag{5}$$

where y_i is the observation, and $\hat{y}_i$ is the model estimate at time t.

BIC has a common format that resembles AIC, although it imposes a greater penalty for model complexity, which tends to give preference to simpler models (Hastie et al., 2009). BIC can be defined by equation (6) as:

$$\text{BIC} = log(\hat{\sigma}^2_{MLE}) + \frac{mlog(N)}{N} \tag{6}$$

For linear models for which AIC and BIC were originally designed, m represents the number of estimated parameters. For nonlinear and other complex models, m is usually replaced by some measure of model complexity (Hastie et al., 2009). In a three-layer neural network model, the two critical uncertainties associated with the model structure are the number of the input vectors (p) and the number of hidden layer units

(k). Qi et al. (2001) have proposed using $m = k(p + 2) + 1$ to measure the model complexity of a three-layer feed-forward neural network, which this study adopts. One big limitation of the in-sample model selection approach is that the in-sample errors are likely to underestimate forecasting errors. Thus, using the best in-sample fit, the model selected may not make the best prediction of unseen time series (Tashman, 2000).

2.2.4. Model Performance Measures

The performance of the models on the partitioned datasets is evaluated using three goodness-of-fit measures, the Nash–Sutcliffe coefficient of efficiency (NSE), percent bias (PBIAS), and root mean square error to observation standard deviation ratio (RSR). The NSE is a dimensionless index mainly used in H/WQ modeling; it is a normalized statistic determining the magnitude of residual variance compared to observed data variance (Nash et al., 1970). The NSE is expressed in equation (7) as:

$$\mathrm{NSE} = 1 - \frac{\sum_i (Q_{sim} - Q_{obs})_i^2}{\sum_i (Q_{obs,i} - \overline{Q_{obs}})^2} \tag{7}$$

Where Q is the variable, *obs* and *sim* stand for observed and simulated, respectively. NSE ranges from –∞ to 1.0, with NSE = 1.0 representing the optimal fitting. A negative NSE value indicates that the mean observed value is a better fit than the simulated value (Moriasi et al., 2007; Zhang et al., 2010; Zhang et al., 2009).

The PBIAS measures the average tendency of simulated data to be larger or smaller than the observations. The model reaches optimal prediction with a PBIAS of 0.0, and smaller absolute PBIAS indicates a more accurate model prediction. In the following form, positive PBIAS values indicate that the model output overestimates the observation, while negative values indicate underestimation (Moriasi et al., 2007). PBIAS is defined in equation (8) as:

$$\mathrm{PBIAS} = 100 * \frac{\sum_i (Q_{sim} - Q_{obs})_i}{\sum_i Q_{obs,i}} \tag{8}$$

The RSR standardizes the root mean square error (*RMSE*) by dividing it by the standard observation deviation, which facilitates convenient performance evaluation. The RSR is a dimensionless index that ranges from 0 to a substantially large positive value, with the optimal value 0 indicating 0 residual variations and, therefore, perfect model fit (Moriasi et al., 2007). The RSR is defined as:

$$\mathrm{RSR} = \frac{RMSE}{STDEV_{obs}} = \frac{\sqrt{\sum_{i=1}^{N} (Q_{sim} - Q_{obs})_i^{\,2}}}{\sqrt{\sum_{i=1}^{N} (Q_{obs,i} - \overline{Q_{obs}})^2}} \tag{9}$$

Moriasi et al. (2007) proposed performance evaluation criteria (PEC) corresponding to the above performance measures on a monthly time step. Since H/WQ models perform better at coarser time scales, Kalin et al. (2010) developed relaxed performance qualitative ratings on NSE and PBIAS for finer time step models. This study simulates the streamflow on a daily time scale and adopts the performance evaluation criteria from several previous studies (ASABE, 2017; Kalin et al., 2010; Moriasi et al., 2007; Moriasi et al., 2015). The PEC used in this study is summarized in Table 2.1.

Table 2.1 Model performance evaluation criteria

Performance Rating	NSE	PBIAS (%)	RSR
Very good	$NSE \geq 0.7$	$\lvert PBIAS \rvert \leq 25$	$RSR \leq 0.5$
Good	$0.5 \leq NSE < 0.7$	$25 < \lvert PBIAS \rvert \leq 50$	$0.5 < RSR \leq 0.75$
Satisfactory	$0.3 \leq NSE < 0.5$	$50 < \lvert PBIAS \rvert \leq 70$	
Unsatisfactory	$NSE < 0.3$	$\lvert PBIAS \rvert > 70$	$RSR > 0.75$

2.2.5. ANN Models Setup

Determining the best input variable combinations is essential for successfully training an ANN model (Noori et al., 2016). Previous research on ANN streamflow forecasting has mainly applied meteorological variables and discharge from the preceding time steps as model inputs (Demirel et al., 2009; Dorofki et al., 2012; Jimeno-Sáez et al., 2018; Minns et al., 1996; Rezaeianzadeh et al., 2013; Zhang et al., 2000). This study proposes using the discharge measurements from an upstream gauge and meteorological variables as inputs for the ANN models. The proposed input combinations are summarized in Table 2.2. These five model prediction scenarios were applied to both study watersheds. The selected variables include daily precipitation (P_t), precipitation of the previous n days (P_{t-n}), daily mean air temperature (T_t), streamflow measurement from the upstream gauge stations (Q_u, USGS 08178000, USGS 08180700), and total precipitation for the preceding n days (P_n). Precipitation and temperature were selected as inputs mainly because they are the most relevant meteorological variables for hydrological impact studies (Maraun et al., 2010). In addition, the total precipitation of previous time steps is included to represent the antecedent moisture condition in scenarios 4 and 5. The downstream discharge (Q, USGS 08178565, USGS 08181500) close to the watershed outlets are the training targets. The input combinations proposed here avoid using the streamflow of preceding time steps at the estimated site, which allows the application of the modeling approach in regions where streamflow observations are incomplete.

Table 2.2 ANN model input combinations.

Prediction Scenario	Input Combination	Output
1	P_t, P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}, Q_u	Q
2	P_t, P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}, T_t	Q
3	P_t, P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}, T_t, Q_u	Q
4	P_t, P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}, P_n	Q
5	P_t, P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}, P_n, Q_u	Q

Input and output data for setting up the ANN models were obtained from the PRISM database and USGS publicly accessible data. While the PRISM daily spatial climate dataset has complete temporal coverage for the years starting from 1981, the USGS streamflow observations often have long durations of missing data. This study uses a decade-long daily time series for model input and output, the longest available consistent data for the two study watersheds. Simulations for the HSARB were run from

1 October 1987 to 30 September 1997, and for the LMRB, from 1 October 1997 to 30 September 2007. The first 70% of data are used for model training (in-sample approach) or training/validation (out-of-sample approach), and the remaining 30% of data are used as the stand-alone testing set.

The Cox and Stuart test (Cox et al., 1955) was applied to detect whether the precipitation and observed streamflow time series during the study period had a time-dependent trend. The decade-long time series of precipitation, upstream discharge, and target discharge were divided into thirds. The test compared whether the first third of the data was larger or smaller than the last third. As the last third of the time series covered the entire testing period, this test can be used to assess if the testing period data trended away from the training period.

The Cox and Stuart test results showed that in HSARB, the precipitation, up-stream discharge, and target discharge all had extremely small p-values, approximately 0, which indicated a detectable trend in the 10-year time series. Hence, differences between the training and testing data exist in HSARB. Meanwhile, in LMRB, the precipitation data had a p-value close to 0, while the upstream discharge had a p-value of 0.939 and target discharge had a p-value of 0.028, which showed that if the p-value threshold of 1% is applied, the test failed to reject the null hypothesis that no monotonic trend exists in the discharge time series. This result suggested no significant difference between the training and testing discharge data in LMRB, while the precipitation pattern had altered during the study period. The more fluctuating discharge data in HSARB could cause relatively poor predictive outcomes in the testing period, while the stable discharge condition in LMRB is likely to induce better predictive performance in the testing period.

This study applied a three-layer feed-forward neural network with input/output layers and a single hidden layer. The logistic function is used as the transfer function for all hidden nodes, and the popular back-propagation algorithm is used to train the models. The maximum amount of training epoch was set as 200,000 to ensure training coverages, while the learning rate was set as 0.002. In addition to different configurations of input variables, the size of hidden layer units significantly affects the complexity of a neural network and its predictive capability. Unfortunately, there is no unified theory in the literature to determine the optimum number of hidden units (ASCE, 2000a). In this study, the number of hidden units is varied from 1 to 10. Larger hidden layer sizes are excluded from the experimentation since models with larger hidden layer sizes tend to overfit the training data (Gazzaz et al., 2012).

Because hydrologic variables span different magnitudes, prior to their use in neural networks, they should be normalized to a common scale to aid in comparison (Sahoo & Ray, 2006; Starrett et al., 2010). In this study, the input and output variables are normalized on the range of 0 to 1 using the following equation:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{10}$$

Where x_i, min (x), and max(x) denote the observed, minimum, and maximum values of the raw data, respectively, and z_i denotes the normalized values. R software (R Core Team, 2019) was used for data processing and model simulations. The “neuralnet” package was used for training all the ANNs in this work.

2.3. Results and Discussion

The in and out of sample model selection approaches discussed in section 2.2.3 were used to determine the best input combination and the optimum number of hidden layer units of each model scenario in Table 2. Therefore, two groups of models were created for every scenario, one for each approach. The first group of models applies the in-sample model selection criteria (AIC and BIC), splitting the data between training and testing sets as indicated earlier. The BlockedCV is applied to the second group of model scenarios. The first 70% of available data were partitioned into ten fixed blocks. Then, ten training iterations were run using a loop structure with 1/10 of the data serving as the validation set in each iteration. The NSE calculated on the validation sets from all iterations is averaged to obtain the validation statistics used as the selection criterion of the BlockedCV. Moreover, all three performance measures discussed in section 2.2.4 are calculated for the training, validation, and testing datasets to further evaluate the model generalization capability and compare the performance among different ANN models.

2.3.1. Optimum Hidden Layer Size Selection

The AIC and BIC of the training dataset for the in-sample modeling group, the NSE of the validation dataset for the out-of-sample modeling group, and the NSE calculated on the stand-alone testing dataset are displayed in Figure 2.3 for HSARB and Figure 2.4 for LMRB. In Figures 2.3 and 2.4, the AIC and BIC are measured using the left axis, while the NSE is measured using the right axis. From plots a–e of these two figures, no uniform trends are observed for the performance measures as the number of hidden nodes increases across the five prediction scenarios for both watersheds. The best performance measures of each scenario (i.e., the smallest AIC and BIC, the largest NSE) are highlighted in red. Among all ten prediction scenarios for the two watersheds, in 7 out of 10 scenarios, the best AIC and BIC resulted from the same number of nodes in the hidden layer. In the other three scenarios where the best results for AIC and BIC resulted from different nodes in the hidden layer, the BIC criteria resulted in fewer hidden nodes.

In all ten prediction scenarios, the best NSE in the validation data set did not select the same number of nodes in the hidden layer as AIC and BIC criteria, indicating that the out-of-sample BlockedCV approach does not converge to the same hidden nodes as the in-sample information-criteria based approach.

Both approaches' hidden node structure selection results were also compared with the testing NSE for further verification. In only 1 out of 10 prediction scenarios, scenario 4 of HSARB, did using AIC to determine hidden layer size results in the best testing NSE value, and none occurred in those using BIC to determine optimum hidden layer size. However, 3 out of 10 prediction scenarios, using BlockedCV to determine optimum hidden layer size were consistent with the best testing NSE results. Using the testing NSE as an indicator of the predictive ability of the neural networks, from plots a–e of Figures 2.3 and 2.4, none of the criteria used consistently found the optimum hidden layer size.

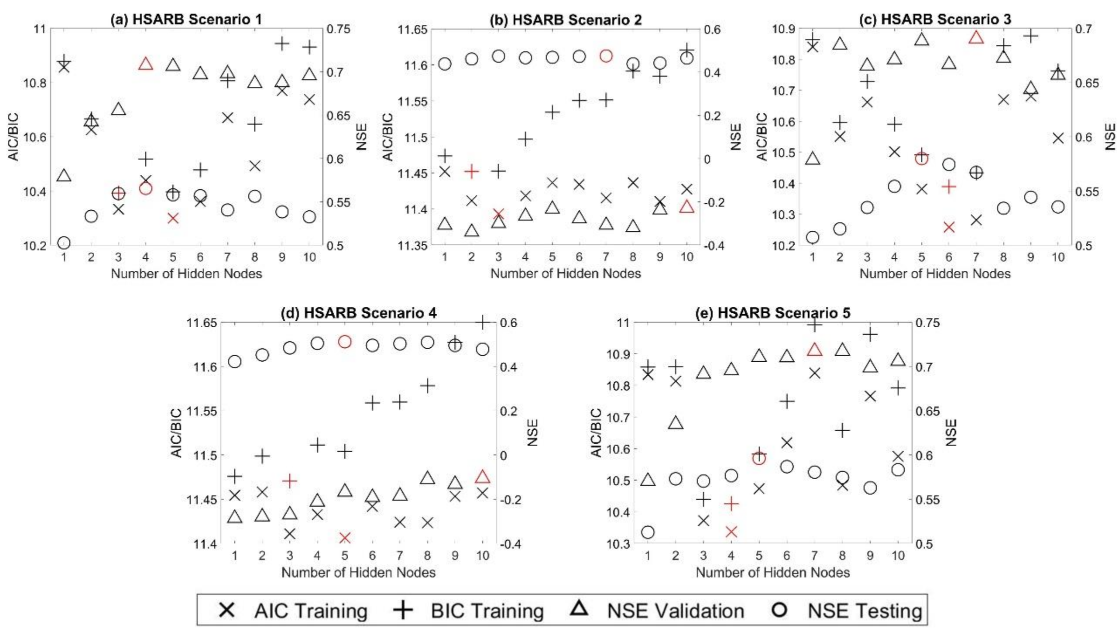

Figure 2.3 Model statistical performance of prediction scenarios 1 through 5 for the HSARB Watershed (a–e). The best performance measures of each scenario (i.e., the smallest AIC and BIC, the largest NSE) are highlighted in red.

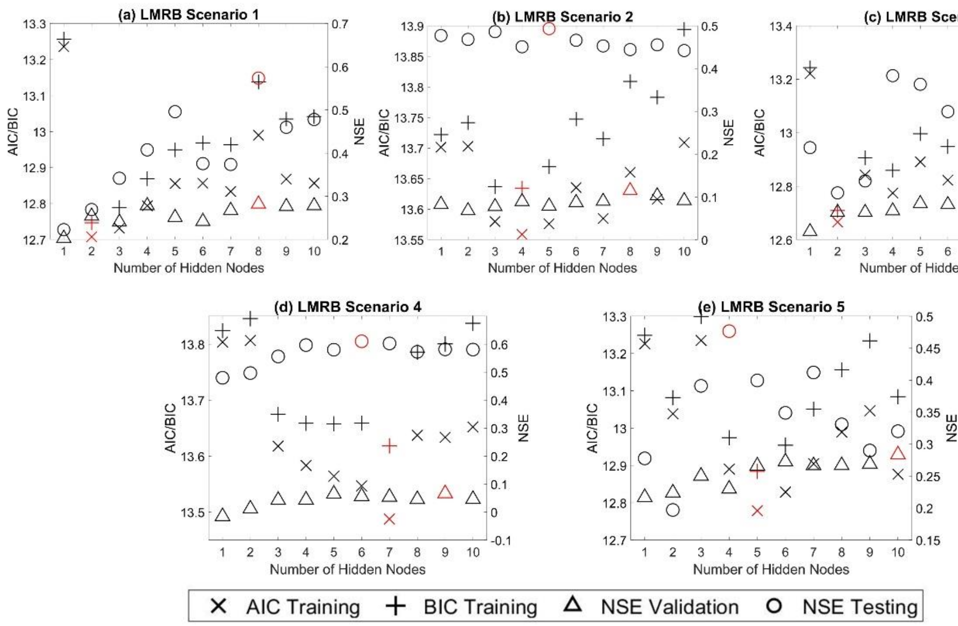

Figure 2.4 Model statistical performance of prediction scenarios 1 through 5 for the LMRB Watershed (a−e). The best performance measures of each scenario (i.e., the smallest AIC and BIC, the largest NSE) are highlighted in red.

2.3.2. Statistical Summary of Model Performance

Table 2.3 presents the AIC selected optimum hidden layer size for each prediction scenario and the corresponding performance data of the urban HSARB Watershed. Based on the criteria defined in Table 1, the training NSE and RSR for scenarios 1, 3, and 5 reached the "very good" level, and scenarios 2 and 4 produced "good" results. The training PBIAS for all five scenarios are numerically close and reached "very good" performance. The testing NSE for scenarios 1, 3, 4, and 5 indicate "good" performance but only "satisfactory" performance for scenario 2. All testing PBIAS values indicate "good" performance, and testing RSR has "satisfactory" to "good" performance. The testing PBIAS of scenarios 2 and 4 is better than that of scenarios 1, 3, and 5. Table 2.4 presents the BlockedCV selected best models for HSARB. Similar to the models selected using AIC, scenarios 1, 3, and 5 achieved better performance in training, validation, and testing datasets for NSE and RSR than scenarios 2 and 4. They are generally in the range of “good” to “very good”. “Unsatisfactory” NSE and RSR performances were observed for the validation dataset of scenarios 2 and 4. However, the PBIAS of scenarios 2 and 4 have achieved better performance in training and testing datasets than scenarios 1, 3, and 5, which contradicts the performance rating from NSE and RSR.

Table 2.3 AIC selected models with the optimum number of hidden nodes, HSARB.

Scenario	Hidden Nodes	Training					Testing		
		AIC	BIC	NSE	PBIAS	RSR	NSE	PBIAS	RSR
1	5	10.299	10.397	0.841	10.8	0.399	0.558	-47.8	0.664
2	3	11.393	11.453	0.519	9.9	0.694	0.474	-36.7	0.725
3	6	10.258	10.389	0.849	8.6	0.388	0.574	-48.5	0.652
4	5	11.406	11.504	0.519	11.1	0.694	0.512	-31.8	0.698
5	4	10.337	10.425	0.834	9.4	0.407	0.577	-45.7	0.650

Table 2.4 BlockedCV selected models with the optimum number of hidden nodes, HSARB.

Scenario	Hidden Nodes	Training			Validation			Testing		
		NSE	PBIAS	RSR	NSE	PBIAS	RSR	NSE	PBIAS	RSR
1	4	0.774	11.3	0.474	0.707	29.6	0.538	0.566	-48.8	0.659
2	10	0.537	8.8	0.680	-0.229	65.7	1.105	0.465	-36.8	0.731
3	7	0.764	11.5	0.486	0.690	33.7	0.555	0.567	-47.6	0.658
4	10	0.525	10.6	0.689	-0.106	62.3	1.049	0.478	-30.5	0.723
5	7	0.790	11.4	0.457	0.717	31.3	0.529	0.581	-40.0	0.647

Table 2.5 presents the AIC selected optimum hidden layer size for each prediction scenario and their corresponding performance data for the nonurban LMRB watershed. All three performance measures for the training dataset have "very good" performance. However, "unsatisfactory" testing NSE and RSR is observed for scenario 1. Meanwhile, most of the testing NSE and RSR of scenarios 2, 3, and 5 only reach a "satisfactory" level. The exception occurs with scenario 4, where the testing NSE and RSR have "good" performance. Results of the testing PBIAS appears to contradict the results from NSE and RSR again, where PBIAS for scenarios 1, 3, and 5 have better performance than that of scenario 2 and 4. Table 2.6 shows the BlockedCV selected best models for LMRB. Overall, the performance measures for the training dataset reached the "very good" level, while the validation performances fall in the range of mostly "unsatisfactory". The testing NSE and RSR show that scenarios 1, 3, and 4 have "good" performance, while scenarios 2 and 5 have "satisfactory" and "unsatisfactory" results. Performance patterns of PBIAS are identical to that of the AIC selected models, where scenarios 1, 3, and 5 have "very good" and scenarios 2 and 4 have worse PBIAS performance.

Table 2.5 AIC selected models with the optimum number of hidden nodes, LMRB.

Scenario	Hidden Nodes	Training					Testing		
		AIC	BIC	NSE	PBIAS	RSR	NSE	PBIAS	RSR
1	2	12.707	12.746	0.920	16.2	0.282	0.269	18.8	0.854
2	4	13.559	13.634	0.816	-10.8	0.429	0.451	-58.6	0.741
3	2	12.666	12.710	0.924	15.2	0.276	0.310	20.4	0.830
4	7	13.488	13.618	0.831	-15.2	0.411	0.603	-47.9	0.630
5	5	12.779	12.884	0.916	10.4	0.289	0.400	4.6	0.775

Table 2.6 BlockedCV selected models with the optimum number of hidden nodes, LMRB.

Scenario	Hidden Nodes	Training			Validation			Testing		
		NSE	PBIAS	RSR	NSE	PBIAS	RSR	NSE	PBIAS	RSR
1	8	0.920	14.0	0.283	0.282	-35.0	0.845	0.574	3.5	0.653
2	8	0.817	-11.9	0.427	0.115	-53.0	0.939	0.444	-61.5	0.745
3	9	0.916	16.6	0.290	0.290	-32.8	0.841	0.658	-7.3	0.585
4	9	0.817	-11.7	0.427	0.066	-54.6	0.964	0.582	-48.7	0.647
5	10	0.916	13.8	0.289	0.283	-35.2	0.845	0.320	11.6	0.824

Taken together, in HSARB, the testing NSE and RSR performance show that scenarios 1, 3, and 5 have better performance than scenario 2 and 4, which indicate that the inclusion of upstream discharge as one of the input variables improve the model performance, while in LMRB, the testing NSE and RSR performance did not reveal the same pattern. More mixed statistical results are observed as scenarios 2 and 4, in some cases, have better testing NSE performance than the other three scenarios. The reason for this is not apparent, but it may be caused by water allocation at the segment between the upstream gauge and the watershed outlet of the Lower Medina River. Moreover, contradictory testing results were found between the PBIAS and the other two performance measures in both HSARB and LMRB, as shown in Table 2.3 to 2.6, which might indicate that the models capture the high flow periods significantly better than the low flows, as NSE and RSR give more weight to high values when compared with low values because their error terms are squared (Moriasi et al., 2015).

Comparing the model selection results of AIC and BlockedCV of all prediction scenarios of the two study watersheds (Tables 2.3 to 2.6), in all ten prediction scenarios, the AIC and BlockedCV approaches selected different optimum hidden layer sizes, and in nine out of ten prediction scenarios the AIC approach selected a simpler hidden layer structure than BlockedCV, with the exception of scenario 1 of HSARB. This result may be explained by the addition of a penalty term on the number of model parameters by AIC, whereas the NSE merely measures the deviation between paired observed and predicted values. Best Model Structure

Table 2.7 summarizes the findings from Tables 3 to 6 and presents the best model structure across all prediction scenarios selected using different criteria. The notations in Table 2.7 are as follows, “Si-j” denotes scenario “i” with “j” hidden nodes. In HSARB, the AIC and BIC both selected S3-6 as the best model, the BlockedCV selected model S5-7, and the NSE calculated from the testing dataset indicate model S5-5 has the best predictive performance. The BlockedCV selected the scenario consistent with the testing NSE, although it did not choose the same model. In LMRB, the AIC and BIC both selected S3-2 as the best model, while BlockedCV selected model S3-9, which agrees with the testing NSE performance. The final model selections at the two study watersheds show that BlockedCV achieves better selection results regarding the predictive model performance. However, it should be noted that limitations exist when using testing NSE as the indicator of the model's predictive ability, since the model testing performance can be strongly affected by the subjective training/testing data partition, and very different testing NSE can occur when calculated from a different testing dataset.

Table 2.7 Selected best model structure with different criteria.

Study Watershed	Selection Criteria	Best Model	Study Watershed	Selection Criteria	Best Model
HSARB	AIC	S3-6	LMRB	AIC	S3-2
	BIC	S3-6		BIC	S3-2
	BlockedCV	S5-7		BlockedCV	S3-9
	Predictive Performance	S5-5		Predictive Performance	S3-9

To better understand the difference between the models selected from different criteria and further assess their predictive performance, Figures 2.5 and 2.6 present the testing phase streamflow time series and scatter plots for the best models selected by AIC, BIC, and BlockedCV. A closer inspection of the hydrographs (Figure 2.5) shows that the selected models can capture the timing of major peaks in both watersheds. However, the simulated and observed time series show apparent deviations in discharge magnitude for all models. For example, in HSARB, both S3-6 and S5-7 underestimate the peak flows, whereas the fit of low flows is barely discernable due to the large vertical scales of the time-series graphs. Conversely, both models S3-2 and S3-9 of LMRB overestimated the peak flows. In the meantime, the time-series graphs fail to provide valuable insights for the low flows as well.

As noted by Moriasi et al. (2015), although a time series plot is an effective graphical measure for evaluating event-specific prediction issues and allows the modeler to find possible temporal mismatches, it can become cluttered with too many data points. Hence, scatter plots should be applied for analyzing longer-duration datasets. Plots a,b of Figure 2.6 show the simulated against observed streamflow data of models S3-6 and S5-7 of HSARB, and both plots suggest the models underestimate the observed streamflow with least-square regression lines that have slopes smaller than 1. Meanwhile, the significant negative PBIAS values that were observed for both model S3-6 and S5-7, strengthen the graphic result that the ANN models in HSARB underestimated streamflow. Plots c,d provide the scatter plots of models S3-2 and S3-9 of LMRB. For model S3-2, the regression line has a slope greater than 1 and a positive PBIAS value, which indicates an apparent overestimation of streamflow. Much better performance is observed for model S3-9 of LMRB, with a regression line is the closest to the 1:1 reference line among the four models, and a slight negative PBIAS value, which indicates the streamflow is slightly underestimated overall.

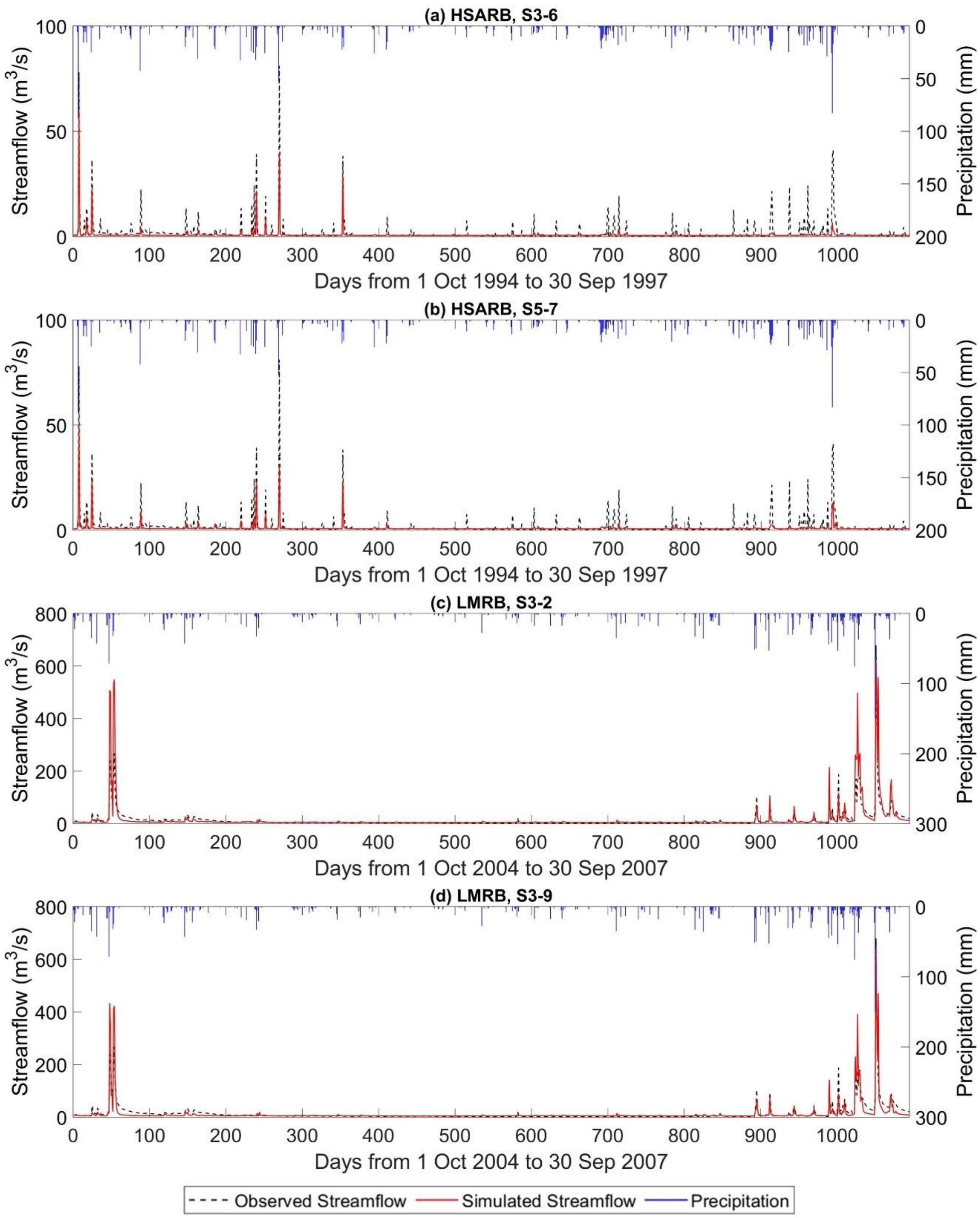

Figure 2.5 Daily precipitation, observed, and simulated streamflow of testing phase for (a) HSARB, S3-6 ;(b) HSARB, S5-7; (c) LMRB S3-2; (d) LMRB S3-9.

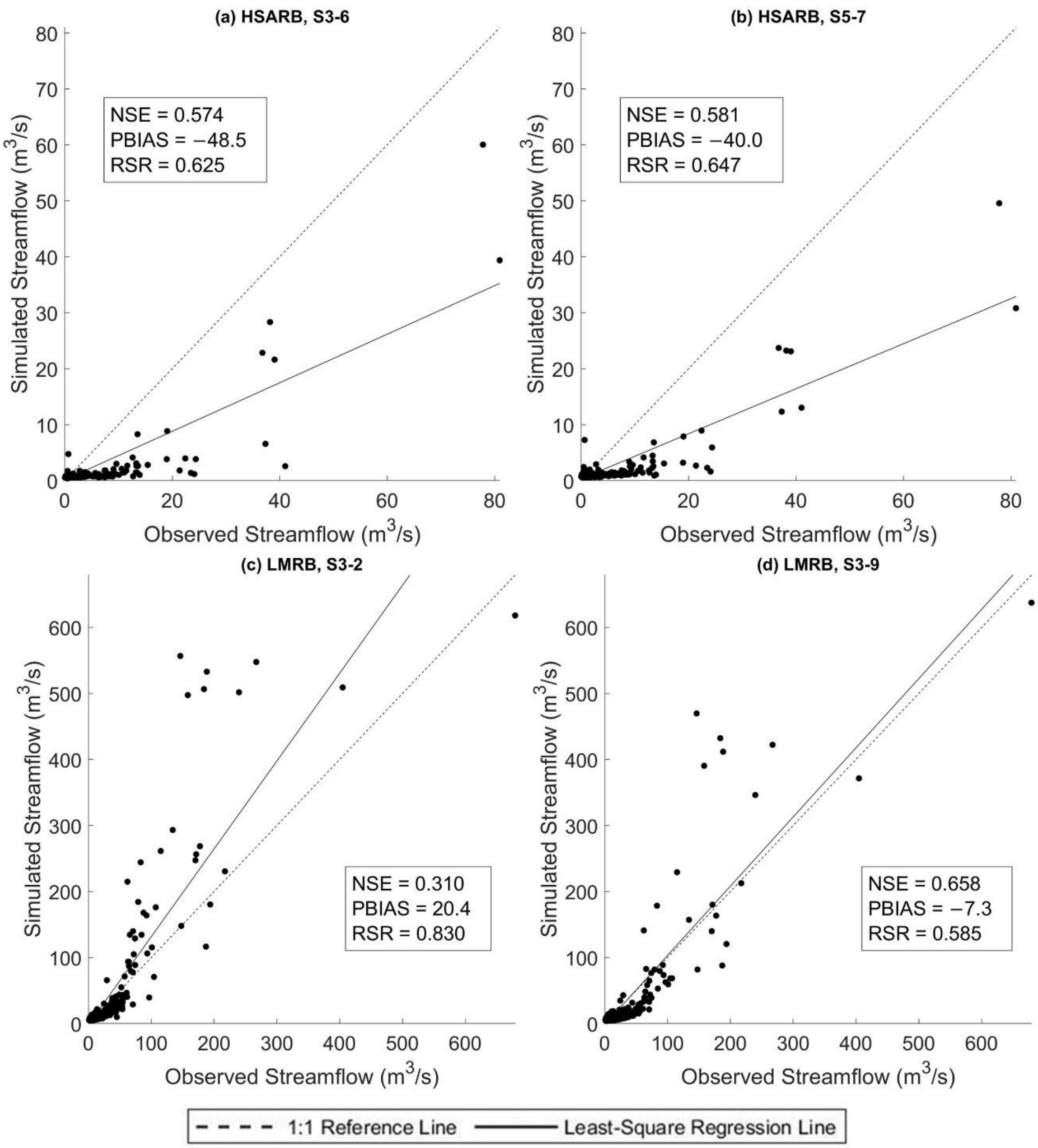

Figure 2.6 Scatter plots of testing phase daily streamflow of (a) HSARB, S3-6; (b) HSARB, S5-7; (c) LMRB S3-2; (d) LMRB S3-9.

The testing phase performance results of the four selected models are displayed in Figure 6 as well. In HSARB, the models selected from the two approaches have close performance results. More specifically, the BlockedCV selected model S5-7 obtained slightly better NSE and PBIAS and slightly worse RSR results than the AIC and BIC selected model S3-6. However, in LMRB, all three performance measures indicate that the BlockedCV selected model S3-9 has significantly better performance than the AIC and BIC selected model S3-2. In both study watersheds, the best models selected by the in-sample AIC and BIC criteria are simpler than the models selected by the out-of-sample BlockedCV approach. As mentioned in section 2.3.2, the in-sample approach tends to select smaller hidden layer sizes than the out-of-sample approach. These results corroborate the ideas of Qi et al. (2001), who suggested that the in-sample model selection criteria may over-penalize the model complexity and select models that underfit the data.

An interesting finding on the selected best models is that the model S3-9 of LMRB has better performance in statistical and graphical measures than the model S5-7 of HSARB. This result is contrary to the expectation that ANN would have the better predictive ability in an urban watershed, where its dominant impervious surface causes more direct rainfall-runoff relation than in a nonurban watershed where the rainfall-runoff process is more complex. A possible explanation might be that the LMRB has a much higher average outflow volume than the HSARB (Figures 5 and 6). Prior studies have noted that the ANN models produce better simulation results estimating high flows (Demirel et al., 2009; Humphrey et al., 2016; Jimeno-Sáez et al., 2018).

2.4. Summary

In this study, daily streamflow simulation models were developed for two small watersheds with distinctive land cover types in south-central Texas using a three-layer feed-forward neural network. Five prediction scenarios using different combinations of meteorological and hydrological variables were considered, and for each prediction scenario, a range of nodes in the hidden layer is evaluated. The results show that the best networks could produce “satisfactory” to “good” daily streamflow prediction performances for most of the considered input combinations. While plenty of studies have reported applying ANN to H/WQ modeling, the model overfitting problem and model selection method for a hydrological time series simulation have not been adequately addressed.

This study empirically investigated two main approaches for selecting the best ANN rainfall-streamflow models: the in-sample approach using AIC and BIC as criteria; and the out-of-sample approach using BlockedCV. The evidence from this study suggests that none of the proposed approaches consistently selects the model that has the best testing dataset predictive ability based on the criteria of optimum hidden layer size. However, when considering selecting among predictive scenarios where model structure difference is more notable, it is found that BlockedCV is more capable of identifying the best predictive model. Furthermore, the AIC and BIC are also found to select a simpler model structure than BlockedCV. Overall, this study strengthens the idea that the in-sample model selection criteria may over-penalize model complexity and select models that underfit the data for modeling a streamflow time series. The final best models in both study watersheds selected through BlockedCV are found to have "good" performance on the testing data. However, a closer inspection of the scatter plots and corresponding PBIAS values indicate that the models can perform very differently on low and high flow data, especially in the LMRB. This finding is also supported by the counter-intuitive result that the largely rural LMRB has better model performance than the urban HSARB, which could be explained by the fact that the HSARB has a much smaller average outflow discharge than the LMRB. Further studies could assess the model selection criteria separately on different quantiles and magnitudes of the flow data while choosing watersheds with closer average flow volume for comparison.

3. A COMPARISON OF DAILY STREAMFLOW PREDICTED BY AN ARTIFICIAL NEURAL NETWORK AND THE SOIL AND WATER ASSESSMENT TOOL (SWAT) IN TWO SMALL WATERSHEDS IN CENTRAL SOUTH TEXAS

Abstract. Currently available rainfall-runoff models vary from simple, lumped, data-driven models that depend on the observed inputs and outputs of a watershed alone, to the more complex physically-based models that represent the process using mathematical equations that describe important physical laws of conservation of mass, energy, and momentum. In this study, the accuracy of streamflow estimated by a data-driven Artificial Neural Network (ANN) and the physically-based Soil and Water Assessment Tool (SWAT) are compared. The models were applied in two small watersheds, one highly urbanized and the other primarily covered with evergreen forest and shrub, in the San Antonio Region of central south Texas, where karst geologic features are prevalent. Both models predicted daily streamflow in the urbanized watershed very well with the ANN and SWAT have the Nash–Sutcliffe coefficient of efficiency (NSE) values of 0.76 and 0.72 in the validation period, respectively. However, both models predicted streamflow poorly in the nonurban watershed. The NSE values of the ANNs significantly improved when a time series autoregressive model structure using historical streamflow data was implemented in the nonurban watershed. The SWAT model achieved minimal improvement through model calibration with the current model structure. This result suggests that an ANN model may be more suitable

for short-term streamflow forecasting in watersheds heavily affected by karst features where surface water flow is strongly influenced by the complex processes of rapid groundwater recharge and discharge.

3.1. Introduction

The prediction of streamflow using rainfall-runoff models is vital in water resources management, with an ultimate purpose of improving decision-making for a wide range of hydrological problems. Streamflow is the result of complex natural processes at the watershed scale. In the past several decades, various computer-based hydrological models have been developed to simulate streamflow, most of which focus on capturing the rainfall-runoff process since precipitation is usually the primary driving force of a hydrological system (Moradkhani et al., 2009). As more and more models became available, hydrologists started to classify the models into different categories based on their structure. One common classification divides hydrological models into lumped or distributed models. A lumped hydrological model considers the study watershed as a single unit. The parameters representing spatial characteristics related to the rainfall-runoff process are averaged or ignored for the entire watershed (Brirhet et al., 2016). In a distributed model, the variation of watershed characteristics and hydrological processes in space are explicitly considered, usually through discretizing the watershed into a large number of rectangular grid cells or a limited number of subbasins based on the drainage and topographic features (Islam, 2011). Another frequently used classification approach considers models as deterministic or stochastic. In a deterministic model, only a single model output value is generated with a given set of input data and model parameters (Beven, 2011); whereas stochastic hydrological models usually provide probability distributions of the target variables (Beaumont, 1979). Exploring the suitability and analyzing limitations of different models is one of the popular topics in modern hydrology.

One major issue in hydrological modeling is deciding the appropriate level of model complexity. Beven (2011) summarized two widely accepted views of modeling. The first suggests that hydrological models are merely tools for extrapolating available data in time and space. Therefore, if the model inputs and outputs can be successfully related, it is not essential to elaborate the details of a watershed. The second view maintains that models should reflect the physical processes involved to the extent possible to ensure confidence when extrapolating beyond the existing observations. More complex physically-based hydrological models are usually adopted under the second perspective.

The Soil and Water Assessment Tool (SWAT) is a physically-based, semi-distributed, deterministic model developed to assess water quality and quantity in large river basins with varying soils, land uses, land cover types, and management practices (Arnold et al., 2012b). In the SWAT model, a study watershed is divided into a few subbasins. At a finer spatial scale, the model further groups lands with homogeneous slopes, soil types, and land cover types into hydrologic response units (HRUs). The

HRUs may not be spatially continuous and can be found at different locations within a subbasin. This strategy effectively simplifies the simulation of watershed processes (Gassman et al., 2007; Licciardello et al., 2011; Tuppad et al., 2011). The development of the SWAT model spans the last three decades, with new functions and routines continuously added to the model. The current SWAT model has been widely applied to water, sediment, agricultural chemicals, and contaminant yields in complex systems (Abbaspour et al., 2015; Gassman et al., 2007). Moreover, the SWAT model is closely integrated with a geographic information system (GIS) since most of the SWAT input data have spatial characteristic (Jayakrishnan et al., 2005; Olivera et al., 2006; Srinivasan et al., 1994). To date, the ArcGIS and QGIS platforms are integrated with the SWAT model.

A physically-based hydrological model is often referred to as a white-box since its modeling processes are established on known scientific principles of mass and energy fluxes (Moradkhani et al., 2009). In contrast to the complex structure of physically-based models, statistical models have gained popularity for rainfall-runoff modeling based on their simplicity and low computational resource demand. An Artificial Neural Network (ANN) is a computing system that resembles the structure of the human biological neural network. It can identify nonlinear relationships from given patterns and fit nonparametric models on multivariate input data (Govindaraju et al., 2013). ANNs have been applied to rainfall-runoff modeling since the 1990s. When used as a rainfall-runoff model, an ANN only identifies the relationship between historical inputs (meteorological data) and outputs (streamflow) without considering any of the physical processes involved, therefore falling into the category of a lumped model and is often referred to as a black-box model (ASCE, 2000a). In addition, the ANNs are stochastic models, given that the fitted parameter values often vary from one training process to another for a fixed training dataset (Jain et al., 2004), creating a distribution of outputs, rather than one value.

The physically-based SWAT model simulates streamflow by incorporating descriptive mathematical equations designed to conceptualize the hydrological processes in a watershed. The model requires grid-based geospatial data as inputs in addition to meteorological data and incorporates a significant number of model parameters (Fatichi et al., 2016; Noori et al., 2016). Solving the equations for the state variables at the level of HRUs for each time step can be time-consuming and computationally demanding (Jimeno-Sáez et al., 2018; Noori et al., 2016). In this regard, statistical models such as ANN hold a clear advantage. ANNs do not require *a priori* knowledge of the watershed physical characteristics as model input, which reduces the model setup procedures. Meanwhile, the time it takes to train and select the best neural network is significantly shorter than calibrating a SWAT model (Jimeno-Sáez et al., 2018; Minns et al., 1996), capable of achieving "unreasonable effectiveness" in hydrological applications when sufficient training data is available (Worland et al., 2019). On the other hand, several studies have noted the disadvantages of applying ANN as a rainfall-runoff model. In particular, ANNs' lack of explanations for the underlying physical hydrological processes has generated a lot of concern among hydrologists (ASCE, 2000a; Ha et al., 2003; Jain et al., 2004; Karunanithi et al., 1994; Yaseen et al., 2015). The inability to capture physical dynamics at the watershed level means that ANNs are not suitable for environmental impact studies such as modeling streamflow under changing climate conditions (Humphrey et al., 2016; Milly et al., 2008). Additionally, similar to other statistical models, extrapolating beyond the training data range often undermines ANNs' predictive performance (Minns et al., 1996; Sahoo, Ray, et al., 2006).

A few studies have compared SWAT and ANN regarding streamflow prediction. Kim et al. (2015) applied SWAT and ANN to impute missing streamflow observations in the Taehwa River Watershed in Korea and found that SWAT was better at simulating low flows, while the neural network model generally performed better at simulating high flows. Jimeno-Sáez et al. (2018) compared the accuracy of streamflow prediction between SWAT and ANN models at the Ladra River Basin and the Segura River Basin in Spain. The authors also came to the conclusion that ANN is better at estimating higher flows. Demirel et al. (2009) applied SWAT and ANN models to simulate streamflow in the Pracana River Basin in Portugal and concluded that ANN was more successful at forecasting peak streamflow, whereas the SWAT model performed better in terms of overall goodness-of-fit indicators. Srivastava et al. (2006) analyzed the performance of streamflow prediction from SWAT and ANN in the agricultural-dominated Honey Brook Watershed in Pennsylvania. They found that the ANN model produced simulation results with the Nash-Sutcliffe coefficient of efficiency (NSE) and coefficient of determination (R^2) better than the SWAT model. In a more recent study, Zakizadeh et al. (2020) used ANN and SWAT to simulate the rainfall-runoff relationship in a small watershed near Tehran city, Iran. The authors concluded that ANN produced simulations with minor error and uncertainty, although both models could achieve excellent predictive performance.

While previous studies have explored the suitability of SWAT and ANN models in various regions, none of these studies were conducted in the state of Texas, where its hot climate condition and growing population have made water sustainability a contentious issue. Additionally, the San Antonio region of central south Texas has extensive karst terrain, making hydrological modeling more difficult and rendering model performance patterns vastly different from non-karstified watersheds (Jakada et al., 2020). Furthermore, very few studies on SWAT and ANN have considered uncertainty in the measurement and modeling process. Hence the objectives of this study are (a) to parameterize SWAT and an ANN model to simulate streamflow in two small watersheds in the San Antonio Region, one rural and one urban; (b) to compare SWAT and ANN model performance in karstic watersheds; (c) to analyze SWAT and ANN performance under different dominant land-use type; and (d) to enhance the model evaluation using modified goodness-of-fit indicators that incorporate measurement and model uncertainty.

3.2. Materials and Methods

3.2.1. Study Area

Two small watersheds in the San Antonio region of Texas are selected for this study, both located within the Medina River Basin (HUC8: 12100302), which covers part of the San Antonio urban area and a large rural area to the west of the city (Figure 3.1a). The study area has a subtropical and semi-humid climate with long hot summers and short warm winters. A part of the Edwards Balcones Fault Zone (BFZ) aquifer lies under the study area, where karst geologic features are prevalent (Loáiciga et al., 2000). Processes including diversion of surface runoff into sinkholes, fast groundwater movement through subsurface conduits, and recharge to surface water from springs linked directly to the aquifers are often found in such terrain (Jakada et al., 2020).

The two study watersheds were delineated in ArcSWAT using the digital elevation model (DEM). The outlets of both study watersheds were selected at the USGS gages that have long-term consistent streamflow records. The Leon Creek Watershed (LCW) covers the western part of the city of San Antonio. It is centered at 98.67º west longitude, 29.56º north latitude, and has a drainage area of 535.76 km^2. The drainage area of the LCW is situated across the contributing, recharge, and artesian zones of the Edwards (BFZ) aquifer (Figure 1a). Elevation of the LCW decreases from north to south, ranging from 546 m in the northern part to 177 m near the watershed outlet (Figure 3.1c). The LCW is heavily urbanized. According to the 2011 National Land Cover Database (NLCD, 2011) land use land cover (LULC) classification (Yang et al., 2018), 47.2% of the LCW is classified as developed urban area with different levels of development intensity. Besides the extensive impervious surface of the city, urban afforestation has made about 34.2% of the LCW evergreen or deciduous forest. Leon Creek is the main waterway in the LCW. It originates from multiple smaller creeks in the northern part of LCW and flows south before merging into the lower part of the Medina River.

The Upper Medina River Watershed (UMRW) is about 30 kilometers northwest of San Antonio. It is centered at 99.32º west longitude, 29.82º north latitude. The drainage area of the UMRW is 847.03 km2 and is located entirely within the contributing zone of the Edwards BFZ aquifer (Figure 3.1a). The UMRW is primarily rural, and its dominant land cover types are forest and shrub. Deciduous and evergreen forests combined cover 48.5% of the drainage area, while shrubland alone covers 38.9%, according to the NLCD2011 classification. Elevation of the UMRW declines from 727 m at its highest point in the northwest to 364 m at the lowest point in the southeast watershed outlet (Figure 3.1b). The upper Medina River is the main waterway in the UMRW. It follows the topographic decline of the UMRW and flows southeastwards. The Medina River is a tributary of the San Antonio River, and it eventually merges into the San Antonio River further south outside the Medina River Basin.

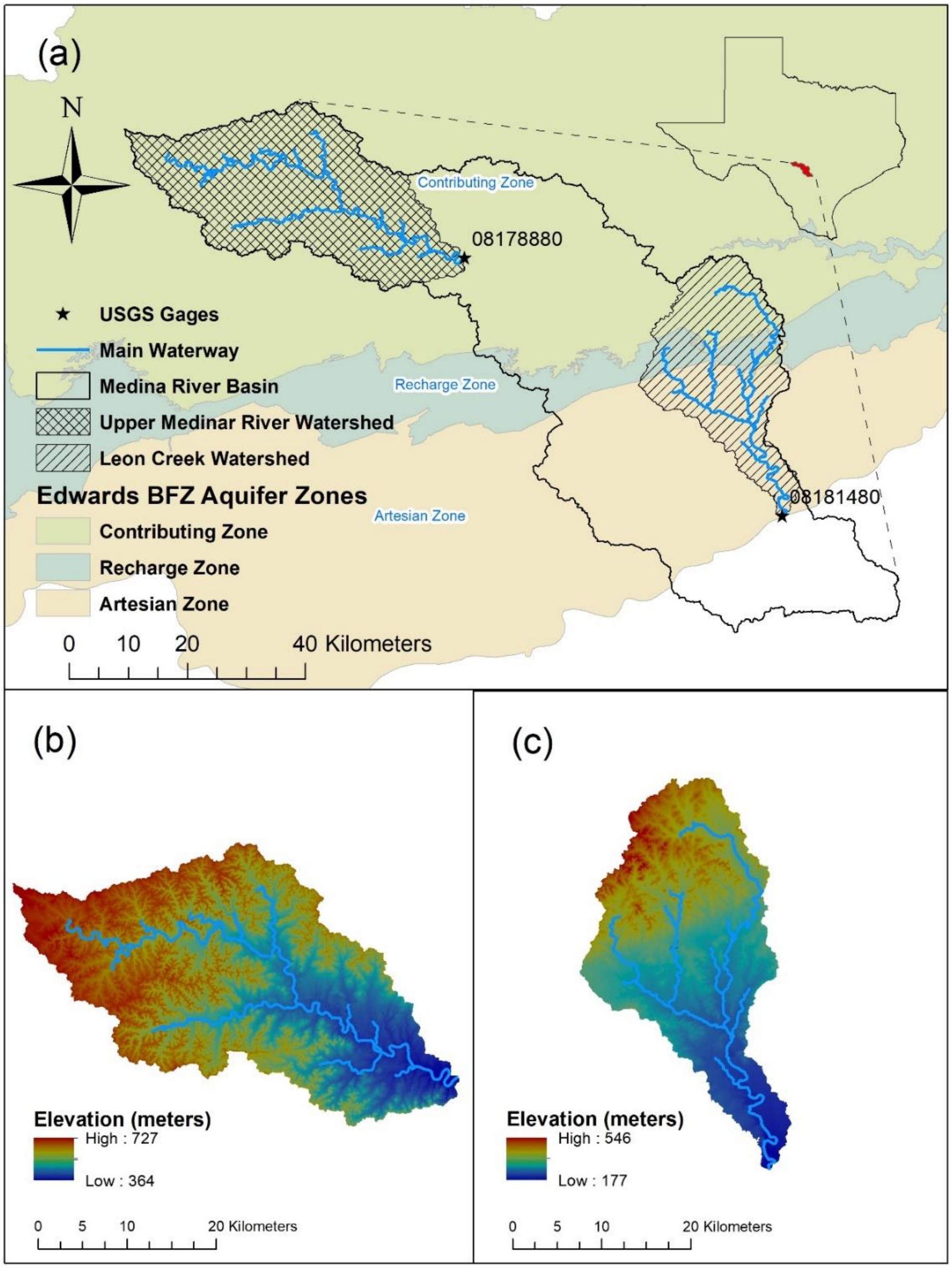

Figure 3.1 (a) Location of the UMRW and LCW in the state of Texas, with zone illustration of the Edwards BFZ Aquifer; (b) DEM of the UMRW; (c) DEM of the LCW.

3.2.2. Data Acquisition

All the data used for constructing hydrological models in this study was obtained from publicly accessible data sources. SWAT requires four main data files to set up model simulations, including the DEM data, LULC data, soil data, and meteorological data. The National Elevation Dataset (NED) which has a 30 m resolution was used for the DEM data. The 2011 National Land Cover Data Set (NLCD2011) was collected as the LULC data. The NED and NLCD2011 datasets for the study area were obtained from the USDA Natural Resources Conversation Service (NRCS) geospatial data gateway (USDA-NRCS, 2014). The state soil geographic (STATSGO) database preloaded with the ArcSWAT interface was used as the soil data. Most importantly, the Parameter-elevation Regressions on Independent Slopes Model (PRISM) daily spatial climate dataset AN81d (Daly et al., 2008) was acquired as meteorological inputs using the Google Earth Engine (GEE), a cloud-based geospatial analysis platform that allows easy access to many governments supported free geospatial data archives (Gorelick et al., 2017). The area-averaged daily precipitation, minimum, maximum, and mean temperature of the PRISM dataset were collected by applying shapefile masks of the two study watersheds on GEE.

The USGS daily streamflow observations at Leon Creek (USGS 08181480) and Upper Medina River (USGS 08178880) from the years 2000 to 2010 were used (U.S. Geological Survey, 2016) in this study, observations from 2000 to 2006 were used to calibrate the SWAT model and provided a training target for the ANN model.

Observations from 2007 to 2009 were used as standalone testing data. The statistical summary (i.e., mean, median, standard deviation, Std.Dev, coefficient of variation, Cv, maximum, Q_{max}, minimum, Q_{min}, first-, second-, third-, and fourth-order autocorrelation coefficients, r_1, r_2, r_3, r_4) of the observed streamflow for both study watersheds are displayed in Table 3.1. In LCW, the overall streamflow difference between the calibration and validation periods is smaller than the difference in UMRW. This study proposes to use streamflow of previous time steps as one of the predictors in the ANN models. Hence, autocorrelation coefficients of the streamflow up to the fourth-order were calculated and are presented in Table 3.1. The autocorrelation decreases quite significantly as lag increases for all modeling periods and watersheds. The autocorrelation is more substantial in the LCW during the calibration period. In the UMRW, the autocorrelation of the calibration period is much weaker than that of the validation period.

Table 3.1 Statistical Summary of Daily Streamflow Observations.

Watershed	Time Period	Streamflow Data (m^3/s)									
		Mean	Median	Std.Dev	Cv	Q_{max}	Q_{min}	r_1	r_2	r_3	r_4
LCW	Calibration	1.581	0.185	16.419	10.386	580.150	0.025	0.621	0.490	0.296	0.112
	Validation	1.486	0.198	10.474	7.046	277.623	0.018	0.390	0.071	0.076	0.079
UMRW	Calibration	6.358	1.837	65.366	10.280	2943.200	0.009	0.552	0.342	0.267	0.094
	Validation	3.992	0.819	10.420	2.610	163.291	0.000	0.707	0.607	0.538	0.499

3.2.3. SWAT Modeling Approach

SWAT formulates hydrological processes in a watershed using mathematical equations that describe important physical laws of conservation of mass, energy, and momentum. Water balance within the system at each time step is calculated to produce simulation results of hydrological and water quality (H/WQ) variables. A detailed description of the SWAT model process can be found in the SWAT theoretical documentation (Neitsch et al., 2011). In this study, a DEM covering a much larger spatial extent than the study area was used to define the two study watersheds. Based on the topography provided by the DEM, the subbasin thresholds (minimum area for initiating stream networks) were applied to define the number and location of subbasins (Her et al., 2015). The subbasins were further discretized into HRUs. A 10% threshold was applied to remove minor slope, soil, and land use classes to restrict the total number of HRUs for improving computational efficiency. As a result, 25 subbasins and 298 HRUs were defined for LCW, while 23 subbasins and 169 HRUs were defined for UMRW.

While SWAT adopts the more traditional approach of utilizing gauge weather data as inputs, the PRISM dataset for the contiguous United States is only available in a gridded format. In this study, GEE was used to calculate the area-averaged meteorological data for the two study watersheds. The watershed centroids were used as "virtual rain gauges" (Elhassan et al., 2016). Daily precipitation and maximum and minimum temperature from 1998 to 2009 were used in the SWAT model simulation. The model simulations were set up on a daily time step for the 12-year simulation period. The calendar years of 1998 to 1999 were used for model warm-up, and 2000 to 2006, was used for model calibration. The purpose of model calibration is to minimize the difference between model simulation and observation through adjusting model parameters. The models are then validated using observations from 2007 to 2009 without further change to the calibrated parameters. The SWAT model calibration was conducted in SWAT Calibration and Uncertainty Programs (SWAT-CUP) using the SUFI-2 procedure (Abbaspour, 2011). Table 3.2 summarized 15 parameters selected for calibration, all of which are considered sensitive for streamflow simulation according to the literature (Arabi et al., 2007; Chen et al., 2020; Koycegiz et al., 2019; Qi et al., 2017). Snow-melt parameters were left at default values since snow rarely occurs in the San Antonio region. Meanwhile, multiple groundwater parameters were adjusted due to their significant impact on modeling the recharge and discharge of the Edwards BFZ aquifer. More details of the adjusted parameters are discussed in section 3.3.1.

Table 3.2 Description of the calibrated SWAT parameters.

Hydrology Input Parameter	Description	File Extension	Type of Change	Initial Value Range
CN2	SCS runoff curve number for antecedent moisture condition II	.mgt	Relative	(-10%, 10%)
ALPHA_BF	Base flow alpha factor (days)	.gw	Replace	(0, 1)
GW_DELAY	Delay time for aquifer recharge (days)	.gw	Replace	(0, 500)
GWQMN	Threshold depth of water in the shallow aquifer required for return flow to occur (mm H2O)	.gw	Replace	(0, 5000)
GW_REVAP	Groundwater "revap" coefficient	.gw	Replace	(0.02, 0.2)
REVAPMN	Threshold depth of water in the shallow aquifer for "revap" or percolation to the deep aquifer to occur (mm H2O)	.gw	Replace	(0, 500)
RCHRG_DP	Deep aquifer percolation fraction	.gw	Replace	(0, 1)
SOL_AWC	Available water capacity of the soil layer (mm H2O/mm soil)	.sol	Relative	(-5%, 5%)
SOL_K	Soil saturated hydraulic conductivity (mm/h)	.sol	Relative	(-5%, 5%)
ESCO	Soil Evaporation compensation factor	.hru	Replace	(0.6, 0.95)
CANMX	Maximum canopy storage (mm H2O)	.hru	Replace	(0, 100)
CH_K1	Effective hydraulic conductivity in tributary channel alluvium (mm/hr)	.sub	Replace	(5, 130)
CH_K2	Main channel hydraulic conductivity (mm/h)	.rte	Replace	(5, 130)
CH_N2	Manning's "n" value for the main channel	.rte	Replace	(0.01, 0.3)
SURLAG	Surface runoff lag coefficient (days)	.bsn	Replace	(1, 24)

Relative means the existing parameter value is multiplied by 1 plus the given value; Replace means the given value replaces the existing parameter value.

3.2.4. ANN Modeling Approach

The primary purpose of applying ANN as a rainfall-runoff model is to determine a model structure that best captures the nonlinear relationships between the input meteorological variables and the target streamflow. A comprehensive review of ANN application in hydrology can be found at ASCE (2000a). A three-layered feed-forward neural network using a back-propagation algorithm was employed in this study. Several neural networks were trained before a model selection phase that selected the model with the best generalization capability (Donate et al., 2013). Table 3.3 summarizes six model structures examined for both LCW and UMRW, including scenarios 1 through 3, which only used meteorological variables as predictors, and scenarios 4 through 6 included precipitation data and streamflow from previous time steps as predictors. The considered predictors include daily precipitation (P_t), precipitation of the previous n days (P_{t-n}), daily mean air temperature (T_t), streamflow observation of the previous n days (Q_{t-n}), and total precipitation for the preceding n days (P_n). All input variables are normalized to the range of 0 to 1 to speed up model training.

Table 3.3 ANN model input combinations.

Model Scenario	Input Combination	Output
1	P_t, P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}, T_t	Q
2	P_t, P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}, P_n	Q
3	P_t, P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}, P_n, T_t	Q
4	P_t, Q_{t-1}, Q_{t-2}	Q
5	P_t, Q_{t-1}, Q_{t-2}, Q_{t-3}	Q
6	P_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}	Q

Cross-validation is considered the simplest and most widely used method for estimating prediction error in statistical modeling (Hastie et al., 2009). Since the time dependency and autocorrelation of a streamflow time series contradict the basic assumption of traditional cross-validation that the data is independent and identically distributed (Bergmeir et al., 2012), this study proposes to use the blocked cross-validation (BlockedCV) for determining the best model structure. Using BlockedCV, the data points are grouped into consistent blocks, and their sequential order is preserved in the training/validation data split process. To match the SWAT model simulation period discussed in section 3.2.3, the data from 2000 to 2006 was used for modeling training and validation, while data from 2007 to 2009 was used for testing. Unlike the notations commonly used in physically-based hydrological model analysis, the validation dataset in statistical cross-validation often refers to the dataset of which statistical outcome is used for model selection. In contrast, the testing dataset is used for standalone model verification without further change the model parameters.

There is, in general, no commonly accepted rule for determining the number of hidden units for a three-layered neural network. However, some studies have discussed the size of hidden units to explore. Ha et al. (2003) evaluated 1 to 9 hidden units in their study of water quality estimation. In another similar study for water quality prediction, Kalin et al. (2010) searched from 1 to 10 hidden units. Demirel et al. (2009) supported the number of hidden units be two-thirds of the sum of the number of input and output nodes, while Gazzaz et al. (2012) suggested that the hidden units size fall between i and 2i + 1, where i represents the number of input nodes. In this study, the number of hidden layer units was investigated from 1 to 10 for each prediction scenarios displayed in Table 3.3. The root mean square error (RMSE) of the validation dataset of all trained models was calculated to find the best model structure. The RMSE is commonly adopted for evaluating statistical models and represents how far the residuals are from 0 on average (Kuhn et al., 2013). The R software (R Core Team, 2019) was used for all ANN model training and model structure selection in this study.

3.2.5. Model Performance Measures

Previous studies have suggested that no single metric is sufficient to verify a hydrological model (Harmel et al., 2014; Yaseen et al., 2018). This study used two goodness-of-fit indicators to evaluate model performance, including the Nash–Sutcliffe coefficient of efficiency (NSE) and percent bias (PBIAS). The NSE is a normalized statistic that determines the magnitude of residual variance compared to the observed data variance. It ranges from $-\infty$ to 1.0, with an NSE = 1.0 representing an optimal fit (Nash et al., 1970). The PBIAS measures the average tendency of model overestimation or underestimation. A smaller absolute PBIAS indicates better model fit. To evaluate the model performance, we adopted the evaluation criteria proposed by Moriasi et al. (2007). The rating criteria are slightly relaxed since the simulations in this study are conducted on the finer daily time step. A more detailed rating guideline is available at ASABE (2017). Table 3.4 displays the form of NSE and PBIAS and their corresponding model performance criteria.

Table 3.4 Goodness-of-fit indicators and model performance evaluation criteria.

Performance Rating	NSE $NSE = 1 - \frac{\sum_i (S_i - O_i)^2}{\sum_i (O_i - \bar{O})^2}$	PBIAS (%) $PBIAS = 100 * \frac{\sum_i (S_i - O_i)}{\sum_i O_i}$
Very good	$\text{NSE} \geq 0.7$	$\lvert\text{PBIAS}\rvert \leq 25$
Good	$0.5 \leq \text{NSE} < 0.7$	$25 < \lvert\text{PBIAS}\rvert \leq 50$
Satisfactory	$0.3 \leq \text{NSE} < 0.5$	$50 < \lvert\text{PBIAS}\rvert \leq 70$
Unsatisfactory	$\text{NSE} < 0.3$	$\lvert\text{PBIAS}\rvert > 70$

S_i is the i^{th} simulated data; O_i is the i^{th} observed data; $\bar{O}$ is the mean of the observed data.

3.2.6. Incorporating Measurement and Model Uncertainty

Uncertainty should always be accounted for in a hydrological model application since decisions on water resources management are increasingly based on H/WQ modeling (Beven, 2011; Harmel et al., 2007). In order to incorporate both measurement and model uncertainty, Harmel et al. (2010) proposed using a correction factor to modify the error term ($e_i = S_i - O_i$, as shown in Table 3.4) in the traditional statistical indicators. The theoretical basis of the correction factor is that the more the uncertainty distribution of the simulated and observed data pairs overlap, the closer the simulated and observed values are to one another (Harmel et al., 2010). The degree of overlap is represented by the joint probability of the two uncertainty distributions and can be expressed by (Harmel et al., 2010):

$$DO_i = \int_{S_i min}^{S_i max} P_S(s_i)ds \cdot \int_{O_i min}^{O_i max} P_O(o_i)do \qquad (4)$$

where DO_i is the degree of overlap of distributions for each simulated (S_i) and observed (O_i) pair, $P_S(s_i)ds$ is the probability density function of the simulated value (S_i), and $P_O(o_i)do$ is the probability density function of the observed value (O_i).

The degree of overlap ranges from 0 to 1, and it is then used to calculate the correction factor and modify the error term, which can be expressed by (Harmel et al., 2010):

$$CF(sim + obs)_i = 1 - DO_i \qquad (2)$$

$$e(sim + obs)_i = CF(sim + obs)_i \cdot (S_i - O_i) \qquad (3)$$

where $CF(sim + obs)_i$ is the correction factor that incorporates measurement and model uncertainty for each simulated (S_i) and observed (O_i) pair. The $e(sim + obs)_i$ term is the modified error term for each simulated (S_i) and observed (O_i) pair. The modified error term is then used to substitute the simple error term ($e_i = S_i - O_i$) in the NSE and PBIAS presented in Table 3.4.

In addition, we assume the uncertainty distributions of the observed and simulated data points to be normally distributed in this study. The normal distribution has two parameters, the mean and standard deviation. The means for each simulated (S_i) and observed (O_i) uncertainty distribution were set at each simulated and observed value, respectively. The standard deviations (Std.Dev) were obtained from coefficients of variation (Cv). Harmel et al. (2010) proposed using four Cv values ranging from low to high. This study adopted two Cv values (0.026 and 0.256) recommended by Harmel et al. (2010). The standard deviation is derived from equation 4 (Haan, 2002):

$$Cv = \frac{Std.Dev}{\bar{x}} \tag{4}$$

where std is the sample standard deviation, $\bar{x}$ is the sample mean of the uncertainty distribution.

3.3. Results and Discussion

3.3.1. SWAT Model Calibration and Validation

The SUFI-2 algorithm is an iterative procedure, which requires updating the parameter ranges after each iteration until model performance stabilizes. For both LCW and UMRW, five calibration iterations, each with 500 simulations, were run. The parameter range from the last calibration iteration was used without further change in the validation iteration, which was composed of 500 simulations as well. The calibrated parameter value range and the best-fitted parameter values for the validation period are summarized in Table 3.5.

Table 3.5 Calibrated SWAT parameter range and the best-fitted validation value

Hydrology Input Parameter	Default Value in SWAT	Calibrated Parameter Range		Best Fitted Validation Value	
		LCW	UMRW	LCW	UMRW
CN2*	35 to 98	(-15.5%, -8.3%)	(12.0%, 16.4%)	-14.7%	12.0%
ALPHA_BF	0.048	(0.075, 0.224)	(0.751, 0.882)	0.086	0.774
GW_DELAY	31	(228, 420)	(334, 411)	363	370
GWQMN	1000	(3712, 4378)	(2521, 3496)	4268	3302
GW_REVAP	0.02	(0.125, 0.150)	(0.037, 0.076)	0.136	0.056
REVAPMN	750	(331, 412)	(85, 225)	348	154.2
RCHRG_DP	0.05	(0.049, 0.159)	(0.152, 0.286)	0.157	0.179
SOL_AWC*	0.01 to 0.42	(-5.3%, 1.5%)	(-3.4%, -1.4%)	-2.6%	-2.8%
SOL_K*	0 to 2000	(-3.2%, -1.6%)	(-4.4% -1.9%)	-2.1%	-4.4%
ESCO	0.95	(0.862, 0.921)	(0.745, 0.797)	0.872	0.796
CANMX	0	(65, 82)	(79, 100)	79	81
CH_K1	0	(77, 114)	(9, 24)	107	23
CH_K2	0	(43, 60)	(111, 130)	51	114
CH_N2	0.014	(0.010, 0.031)	(0.247, 0.300)	0.014	0.273
SURLAG	4	(3.81, 9.75)	(19.41, 24.00)	7.94	23.42

* Indicate percent change to existing parameter values.

The SCS runoff curve number (CN2) and soil parameters are adjusted through a percentage change due to their spatial heterogeneity. Similarly, groundwater (.gw) and general HRU parameters (.hru) can have spatial heterogeneity to the level of HRUs, and subbasin (.sub) and routing parameters (.rte) can have spatial heterogeneity to the level of subbasins. However, all other parameters were adjusted through direct replacement in this study since their default values are fixed across the entire watershed in SWAT 2012. The groundwater (.gw) parameters were the focus of this study as they control the speed of recharge into the Edwards Aquifer. The baseflow alpha factor (ALPHA_BF) is slightly increased from the default in LCW, and significantly increased in UMRW. The larger ALPHA_BF indicates that the groundwater flow response to the changes in recharge is more rapid in UMRW. The groundwater delay time (GW_DELAY) was extended in LCW and UMRW, which indicate that water exits the soil profile and enters the shallow aquifer relatively slow in both watersheds. Meanwhile, the threshold depth of water in the shallow aquifer required for return flow to occur (GWQMN) is significantly increased from its default value, possibly reflecting a large storage capacity of the shallow aquifer in both study watersheds, which is very common in karstic regions. A larger groundwater "revap" coefficient was found in LCW than UMRW, suggesting that water movement from the shallow aquifer to the root zone is faster in LCW, the fast water discharge is likely occurring through sinkholes in the region when hydraulic head of groundwater is high. In addition, a relatively small deep aquifer percolation fraction (RCHRG_DP) was found in both LCW and UMRW, indicating that only a tiny proportion of water in the root zone was recharged into the deep aquifer for our calibrated models. A possible explanation for this might be that the deep aquifer in this karstic region mainly receives fast recharge through the sinkholes. The deep aquifer has a relatively high hydraulic head, which reduces the amount of water recharge by percolating the soil layers.

3.3.2. ANN Model Selection

The three-layer feed-forward neural network structure contains one input layer, one hidden layer, and one output layer. All nodes are fully connected to nodes in their adjacent layers with links that contain weight and bias information. As mentioned in section 3.2.4, for all six prediction scenarios presented in Table 3.3, the hidden layer size was explored from 1 to 10. The RMSE of the cross-validation dataset is calculated to select the best model structure. The models with the smallest RMSE for each prediction scenario are displayed in Table 3.6 for LCW and Table 3.7 for UMRW. Among the six prediction scenarios, scenario 6 of LCW and scenario 5 of UMRW had the smallest cross-validation RMSE and were therefore selected as the best model for each study watershed.

Table 3.6 Best ANN model structure and performance result for the LCW.

Prediction Scenario	Hidden Nodes	Training		Testing		Validation RMSE (m3/s)
		NSE	PBIAS (%)	NSE	PBIAS (%)	
1	7	0.901	60.7	0.736	-16.9	2.247
2	5	0.891	59.4	0.671	1.3	2.178
3	6	0.908	68.3	0.760	-6.3	2.231
4	6	0.882	73.2	0.759	-2.0	2.183
5	8	0.901	57.7	0.770	-11.2	2.149
6	3	0.887	56.2	0.756	-13.4	2.111

Table 3.7 Best ANN model structure and performance result for the UMRW.

Prediction Scenario	Hidden Nodes	Training		Testing		Validation RMSE (m^3/s)
		NSE	PBIAS (%)	NSE	PBIAS (%)	
1	8	0.949	40.5	-0.076	-89.7	4.150
2	6	0.961	8.8	-0.029	-90.3	4.238
3	9	0.950	44.7	-0.077	-89.5	3.984
4	1	0.670	3.2	0.294	-72.8	3.518
5	7	0.892	3.7	0.316	-79.8	2.360
6	8	0.863	-11.7	0.355	-79.1	2.658

The trained ANN models overall had very good performance in LCW (Table 3.6). In the training period, all six scenarios had NSE values around 0.9, although the

PBIAS values were all above 50%, which indicates that streamflow is continuously overestimated during the training process. In the testing period, the NSE values ranged between 0.67 and 0.77, while all PBIAS values are below 25%, overall indicating very good streamflow estimation performance.

In UMRW (Table 3.7), the NSE and PBIAS performance for all six prediction scenarios in the training period is in the range of "good" to "very good", with NSE ranged from 0.67 to 0.95 and PBIAS ranged from -11.7% to 44.7%. However, the predictive performance of the testing period is much worse. The NSE values were below 0 for scenarios 1 through 3, in which cases the observed mean is a better predictor than the model. The NSE values are near 0.3 for scenarios 4 through 6, near the boundary of unsatisfactory and satisfactory performance. The PBIAS of all six scenarios were below -70%, which suggests severe underestimation of the streamflow. As mentioned in section 3.2.4, prediction scenarios 4 through 6 included the streamflow of previous time steps as predictors, while scenarios 1 through 3 only used the meteorological data. It is apparent from Tables 3.6 and 3.7 that the inclusion of streamflow as one of the predictors did not cause a clear difference for the predictive performance in LCW but significantly improved the predictive performance in UMRW.

3.3.3. SWAT and ANN Model Performance Comparison

The goodness-of-fit indicators as traditionally calculated and as modified with the correction factor for SWAT and ANN models are presented in Table 3.8. Similar to the results in Harmel et al. (2010), when the uncertainty level is low ($Cv = 0.026$), there was almost no noticeable improvement in the indicator values. When the more significant uncertainty level was applied (Cv = 0.256), both NSE and PBIAS values show different levels of improvement in the two study watersheds.

In the urban LCW, both SWAT-LCW and ANN-LCW models produced very good simulation results, with the calibration (training) NSE reaching approximately 0.90 and validation (testing) NSE above 0.70. In addition, both SWAT-LCW and ANN-LCW models overestimated streamflow in the calibration (training) period. The SWAT-LCW model further overestimated the streamflow in the validation period, while the ANN model underestimated the streamflow in the testing period. Overall, the statistical indicators suggest that the ANN model slightly outperformed the SWAT in the urban LCW.

In the rural UMRW, the ANN model performed significantly better than the SWAT model. The NSE performance of the training period was classified as very good for the ANN-UMRW model, using either the traditional calculation or the modified indicators. The testing period NSE of the ANN-UMRW model was classified as satisfactory. In contrast, the calibration period SWAT-UMRW model only had good performance with NSE ranging from 0.54 to 0.56 from different uncertainty levels, and the validation period SWAT-UMRW model had unsatisfactory NSE performance on all uncertainty levels. Interestingly, the PBIAS statistics of the calibration (training) period for both SWAT-UMRW and ANN-UMRW models were classified as very good.

However, the validation(testing) PBIAS was similar to that in LCW, with ANN-UMRW underestimated streamflow and SWAT-UMRW made overestimation.

Table 3.8 Statistical performance of SWAT and ANN models

Watershed	Model	Indicator	Calibration (Training) (2000-2006)			Validation (Testing) (2007-2009)		
			Traditional Calculation	Cv = 0.026	Cv = 0.256	Traditional Calculation	Cv = 0.026	Cv = 0.256
LCW	ANN	NSE	0.89^{vg}	0.89^{vg}	0.91^{vg}	0.76^{vg}	0.76^{vg}	0.79^{vg}
		PBIAS (%)	56.2^{s}	56.2^{s}	56.4^{s}	-13.4^{vg}	-13.3^{vg}	-10.7^{vg}
	SWAT	NSE	0.90^{vg}	0.90^{vg}	0.92^{vg}	0.72^{vg}	0.72^{vg}	0.74^{vg}
		PBIAS (%)	49.6^{g}	49.6^{g}	48.9^{g}	36.8^{g}	36.8^{g}	33.5^{g}
UMRW	ANN	NSE	0.89^{vg}	0.89^{vg}	0.96^{vg}	0.32^{s}	0.32^{s}	0.34^{s}
		PBIAS (%)	3.7^{vg}	3.7^{vg}	5.8^{vg}	-79.8^{u}	-79.8^{u}	-78.4^{u}
	SWAT	NSE	0.54^{g}	0.54^{g}	0.56^{g}	-0.02^{u}	-0.02^{u}	0.07^{u}
		PBIAS (%)	4.3^{vg}	4.3^{vg}	3.5^{vg}	27.3^{g}	27.3^{g}	25.3^{g}

Superscripts represent the performance levels, “vg” - “very good”, “g” - “good”, “s” - “satisfactory”, “u” - “unsatisfactory”.

The hydrographs with precipitation records (Figure 3.2) for the validation(testing) period (2007 to 2009) were created to further analyze the difference between observed daily streamflow and SWAT and ANN simulation results. In Figure 3.2, the annual validation(testing) period hydrograph for LCW was displayed in plots (a) through (c), and for UMRW was displayed in plots (d) through (f). In LCW, both SWAT and ANN models captured the timing of most major streamflow peaks except for one significant storm event in the late summer of 2008, prior to which a wet period with relatively large precipitation volume was recorded. The SWAT-LCW model performed better during 2007 and 2009 and less accurately during 2008, which has extended periods of deficient observed flow records. The ANN-LCW model had more consistent predictive performance throughout the entire testing period than the SWAT-LCW model. Additionally, the hydrographs in LCW show that the SWAT model overestimated the magnitude of almost every major streamflow peak in 2007 and 2009, while the magnitude of the ANN model predicted peak flow was much smaller. In the rural UMRW, both SWAT and ANN models failed to capture the timing and magnitude of the significant storm events. More specifically, the ANN-UMRW model caught the timing of the storm events from 2007 to 2009 but often severely underestimated their magnitudes. On the other hand, the SWAT-UMRW model failed to capture the timing and magnitude of the storm events and falsely simulated a few nonexistent storm flow events. Additionally, the recession limbs simulated by SWAT-UMRW were much longer than that from the actual flow.

Figure 3.3 presented the validation/testing scatter plots from 2007 to 2009 for LCW and UMRW. The observed high flow and low flow were divided at the 5% probability of exceedance. Statistical indicators using traditional calculation were also displayed in the corresponding plots. In LCW, the SWAT and ANN models performed well predicting the top 5% of observed streamflow with a slight underestimation (Figure 3.3a). However, for the bottom 95% of streamflow, both SWAT and ANN performed poorly with very large overestimations of the predicted streamflow (Figure 3.3b). In the rural UMRW, both the SWAT-UMRW and ANN-UMRW models performed poorly for low and high flow conditions resulting in negative NSE values. In addition, the scatter plots (Figure 3.3c and 3.3d) suggest that SWAT-UMRW overestimated streamflow for low and high conditions, while the ANN-UMRW models underestimated predicted streamflow.

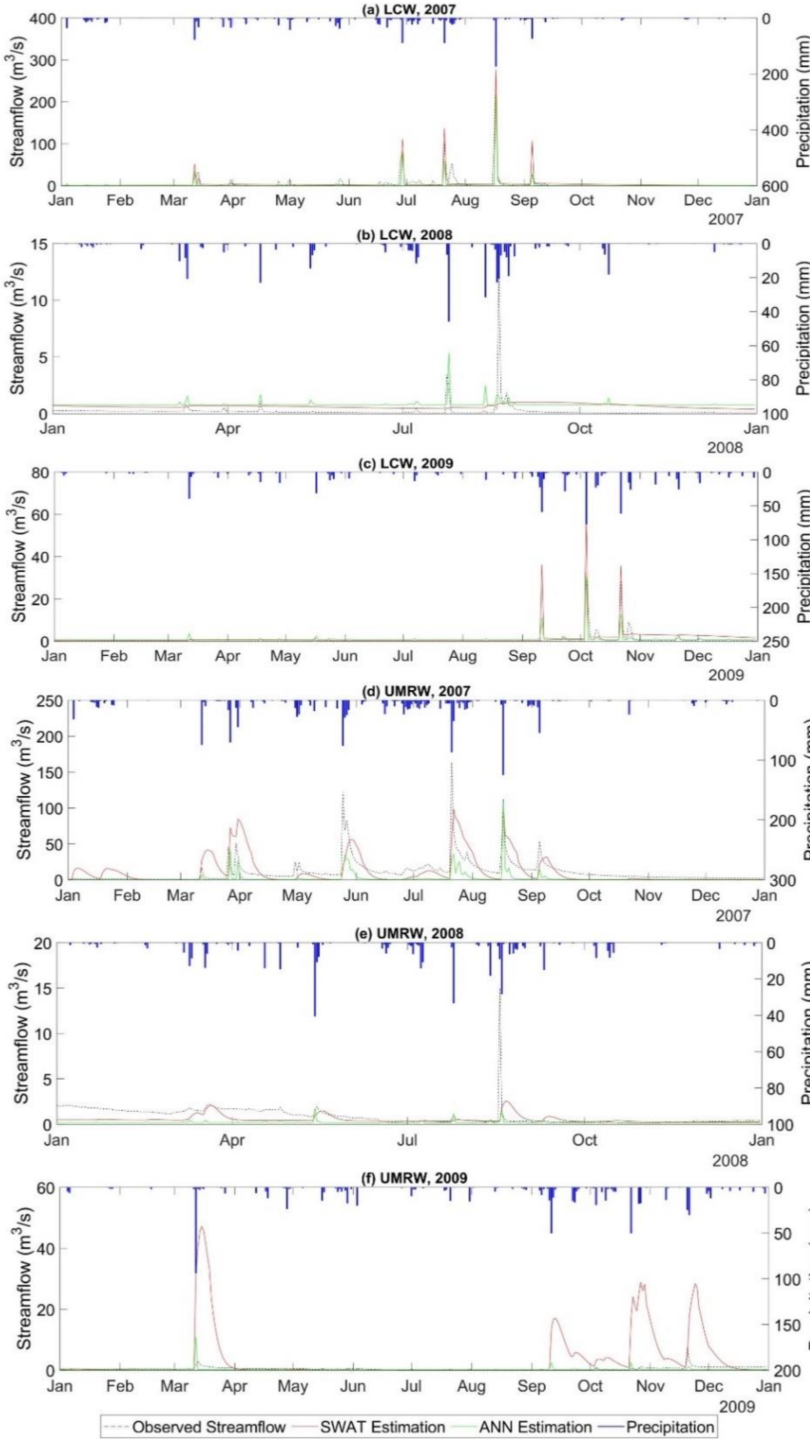

Figure 3.2 Daily precipitation, observed, and simulated streamflow of year (a) 2007, (b) 2008, (c) 2009 for the LCW, and year (d) 2007, (e) 2008, (f) 2009 for the UMRW.

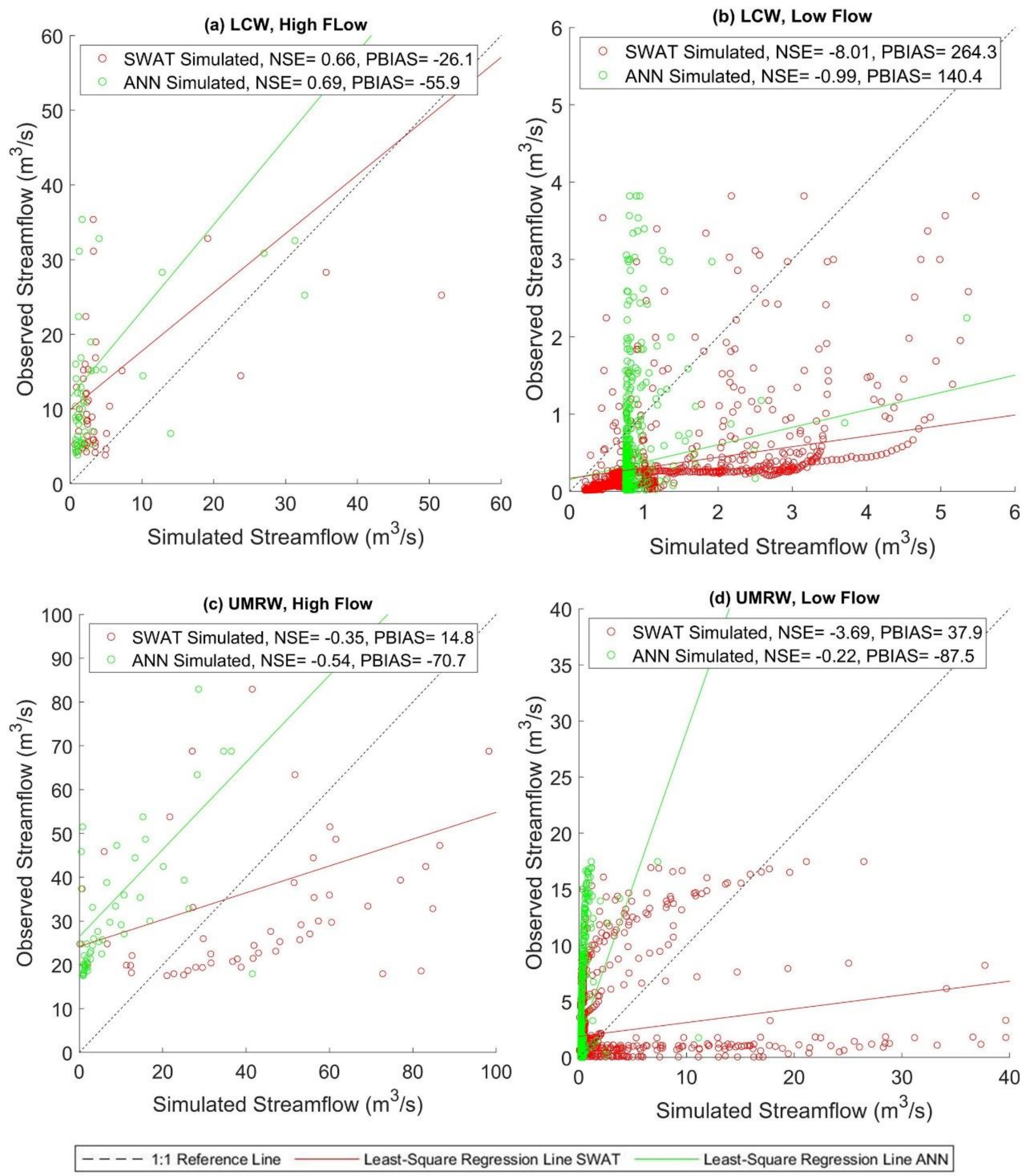

Figure 3.3 Scatter plots of validation/testing periods of daily streamflow for (a) LCW high flow, (b) LCW low flow, (c) UMRW high flow, and (d) UMRW low flow.

The LCW and UMRW are situated above the Edwards BFZ Aquifer, with UMRW entirely located within the contribution zone and LCW located through the contributing, recharge, and artesian zone (Figure 3.1a). Karst geological features are found prevalent in the Edwards BFZ Aquifer region, and a significant amount of surface runoff in the contributing zone infiltrates into the water table aquifer (Loáiciga et al., 2000). In the rural UMRW, the karst terrain makes rainfall-runoff modeling particularly difficult. The observed streamflow time series did not correspond well with precipitation mostly likely due to the fast infiltration or direct recharge through sinkholes. The water recharging into the Edwards BFZ Aquifer generally moves from west to east through large subterranean conduits (Loáiciga et al., 2000), and eventually discharges back to the surface through a few large springs far east outside of the watershed boundary of UMRW. In the meantime, the two hydrological models employed in this study, especially the SWAT model, heavily relied on the input precipitation for estimating the watershed outflow. The daily streamflow simulation results suggest that the calibrated SWAT model failed to quantify the amount of groundwater recharge and discharge in UMRW, hence overestimated the watershed outflow. On the contrary, the estimation of watershed outflow in the urban LCW was much more successful for SWAT and ANN. These results are likely due to the extensive impervious surface in LCW, causing more direct surface runoff into the stream instead of recharging into the water table aquifer. A few previous studies pointed out that the traditional SWAT models have generally not yielded satisfactory runoff estimates in karst watersheds, and some suggested applying additional modules or using trace test method to more accurately quantify streamflow and infiltration (Jakada et al., 2020; Shao et al., 2019; Spruill et al., 2000; Wang et al., 2019), which are beyond the scope of this study.

3.4. Summary

This study investigated watershed level daily streamflow simulation using SWAT and ANN in two small watersheds in the San Antonio Region, where karst terrain is prevalent. This study set out to compare the efficacy of the SWAT and ANN models in estimating streamflow in the karstic watershed in central south Texas. The paired watershed approach is employed to assess SWAT and ANN performance under different dominant land cover types. Additionally, we applied the correction factor approach to enhance model evaluation using modified goodness-of-fit indicators incorporating measurement and model uncertainty. The conclusions of this study are summarized as follows:

(1) Six ANN prediction scenarios were set up for the two study watersheds, and blocked cross-validation was applied to select the best neural network structure for each watershed. The model selection results (Table 3.6 and 3.7) show that in urban LCW, the inclusion of previous time step streamflow as one of the predictors did not noticeably improve model performance; whereas in the rural UMRW, the model performance significantly improved when a time series

autoregressive model structure using historical streamflow data was implemented.

(2) Considering both the statistical performance and graphical comparison between the observed and simulated time series, both SWAT and ANN models made very good daily streamflow estimations in the urban LCW, while their performance in the rural UMRW was much less satisfactory. This discrepancy could be attributed to the extensive impervious surface in the urban watershed, causing more direct surface runoff into stream networks than the rural watershed.

(3) The calibrated SWAT model has inferior performance in the rural UMRW, which is located within the contributing zone of Edwards BFZ Aquifer, where karst geological features are found prevalent. This finding is consistent with a few previous studies, which suggested that the conventional SWAT model is less capable in a karst environment.

(4) The statistical indicators performance results for the standalone validation (testing) period suggest that the ANN model performed slightly better than the SWAT model in LCW and significantly better in UMRW. These findings, taken together with the fact that the data-driven ANN model has a short response time compared with SWAT, suggest that ANN is a better real-time simulator of streamflow although not addressing the physical aspect of a hydrological system.

(5) Applying the correction factor approach to modify goodness-of-fit indicators to incorporate measurement and model uncertainty yielded similar results in the two

study watersheds. When the uncertainty level is low (Cv = 0.026), there was almost no visible improvement in the indicator values. When the more significant uncertainty level was applied (Cv = 0.256), both NSE and PBIAS values show different levels of improvement for the ANN and SWAT models in both study watersheds.

4. HYDROLOGICAL EVALUATION OF GRIDDED CLIMATE DATASETS IN A TEXAS URBAN WATERSHED USING SWAT AND ANN

Abstract. Precipitation is a vital component of the hydrologic cycle, and successful hydrological modeling largely depends on the quality of precipitation input. Gridded precipitation datasets are gaining popularity as a convenient alternative for hydrological modeling. However, many of the gridded precipitation data have not been adequately assessed across a range of conditions. This study compared three gridded precipitation datasets, Tropical Rainfall Measuring Mission (TRMM), Climate Forecast System Reanalysis (CFSR), and Parameter-elevation Relationships on Independent Slopes Model (PRISM). This study used the conventional gauge observation as reference data and evaluated the suitability of the three sources of gridded rainfall data to drive rainfall-runoff simulations. The Soil and Water Assessment Tool (SWAT) and Artificial Neural Network (ANN) were used to create daily streamflow simulations in the Leon Creek Watershed (LCW) in San Antonio, Texas, with the TRMM, CFSR, PRISM, and gauge rainfall data used as inputs. A direct comparison of the gridded data sources showed that the TRMM data underestimates the volume of rainfall, while PRISM data most closely matches the volume of rainfall when compared to the gauge rainfall observations. The hydrological simulation results showed that the PRISM and TRMM rainfall data driven models had preferable results to the CFSR and gauge driven models, in terms of both graphical comparison and goodness-of-fit indicator values. Additionally, no significant

discrepancy was found between SWAT and ANN simulation results when the same precipitation data source was used, while SWAT and ANN simulation results varied in an identical pattern when different precipitation data sources were applied.

4.1. Introduction

Precipitation is a critical input variable in hydrological modeling. In the past, records from rain gauges have been the primary data sources used to drive watershed level rainfall-runoff models (Beven, 2011) and are often recognized as the most accurate surface precipitation measurement (Stampoulis et al., 2012). However, some apparent limitations exist when gauge rainfall data is applied. Most notably, the rain gauge data are point measurements that may have a poor representation of precipitation across a watershed. Worqlul et al. (2014) pointed out that capturing the spatial variation of precipitation in a moderate-sized watershed can be difficult unless a large number of rain gauges is available. In addition, precipitation records from rain gauges are often incomplete both spatially and temporarily (Fuka et al., 2014), especially in remote regions where maintaining a rain gauge network can be challenging and expensive.

In recent decades, alternative precipitation datasets using different measurement approaches have become available. In particular, the availability of satellite rainfall products (SRPs) has vastly improved in the past few years, providing new opportunities for hydrologists to obtain efficient precipitation data in remote regions where ground-based rain gauges are sparse (Worqlul et al., 2014). The Tropical Rainfall Measuring

Mission (TRMM) is one of the freely available SRPs. It was designed by NASA and the Japan Aerospace Exploration Agency (JAXA) to monitor and study tropical rainfall (Adler et al., 2003). The TRMM 3B42 product contains a merged microwave/infrared (IR) precipitation estimate band with a 3-hour temporal resolution and a 0.25-degree spatial resolution. The TRMM 3B43 dataset is gauge-adjusted and covers the global latitude belt from 50ºS to 50ºN (Li et al., 2018). Recent studies have evaluated the performance of TRMM products in different regions of the world. Ochoa et al. (2014) compared TRMM data with an interpolated gauge dataset in the Pacific–Andean region in western South America. They concluded that TRMM could capture the seasonal features of precipitation but suggested that TRMM systematically overestimated precipitation in some parts of the study area. Stampoulis et al. (2012) compared TRMM 3B42 version 6 data against a network of rain gauges over continental Europe, and the authors came to a similar conclusion that TRMM generally overestimated rainfall. Worqlul et al. (2014) compared TRMM 3B42 dataset with two other gridded rainfall products, Multi-Sensor Precipitation Estimate–Geostationary (MPEG) and Climate Forecast System Reanalysis (CFSR), in the Lake Tana Basin in Ethiopia. Their analysis found that MPEG and CFSR have a lower root mean square error (RMSE) with ground observations than TRMM, whereas TRMM had an overall lower logarithm bias over the ground observations than the other two. Li et al. (2018) conducted a study in a large watershed in southern China using TRMM and gauge data to drive the SWAT model. They found that TRMM rainfall data showed superior performance at monthly and annual time steps in terms of the Nash-Sutcliffe Coefficient of Efficiency (NSE) and relative bias ratio (BIAS). Furthermore, Himanshu et al. (2018) investigated the TRMM 3B42 dataset over an agricultural watershed in Krishna River Basin of India using the SWAT model and found that the TRMM driven model always performed worse than that gauge driven model on daily and monthly simulation time steps. To date, the accuracy of TRMM rainfall estimates when used for hydrological modeling is questionable. As pointed out by Li et al. (2018), the satellite may fail to detect the ground-based precipitation event. Therefore, it should be verified in more regions with different geological and climatological conditions before its extensive application in hydrological problems.

Climate Forecast System Reanalysis (CFSR) also provides freely available spatially distributed rainfall estimates widely used in hydrological modeling. CFSR was developed based on surface and satellite observations with a 38-km resolution. It covers a 32-year period from January 1979 to March 2011 and has complete global coverage at 6-hourly and monthly time steps. (Saha et al., 2014). Several studies have used the CFSR dataset for driving hydrological model. Radcliffe et al. (2017) compared the effects of CFSR and the Parameter-elevation Relationships on Independent Slopes Model (PRISM) data on SWAT model streamflow prediction in two small watersheds in the southern U.S., and concluded that the PRISM data produced better streamflow prediction. Roth et al. (2016) applied the CFSR and rain gauge data to streamflow and soil loss modeling using SWAT in Ethiopia and concluded that conventional rain gauges produce much better simulation results than the CFSR data. The authors also pointed out that the CFSR data could not sufficiently represent the spatial variability of regional climate in some of their study watersheds. However, in another study conducted by Fuka et al. (2014), which applied the CFSR data to a few small to moderate-sized watersheds in the US and Ethiopia using the SWAT model, the authors found that the CFSR data produced streamflow simulations that are as good or better than models using rain gauge data. In a more recent study, Mararakanye et al. (2020) compared CFSR data with rain gauge measurement and used both for streamflow simulation in an agricultural watershed in South Africa. Their results suggested that the statistical agreement between CFSR and gauge rainfall data is low, and the model using gauge data slightly outperformed the model using CFSR data.

Two ground-based precipitation measurement sources are compared with the TRMM and CFSR datasets in this study, including conventional gauge data and the Parameter-elevation Regressions on Independent Slopes Model (PRISM) data. The PRISM datasets are gridded climate datasets that cover the conterminous US. In particular, the PRISM AN81d daily spatial climate dataset covers the period from 1981 to the current date. It has 2.5 arc minutes spatial resolution and multiple bands, including precipitation, temperature, and vapor pressure deficit. The PRISM datasets were developed by interpolating available ground-based weather observations using routines that simulate how weather changes with elevation (Daly et al., 2008). Given their comprehensive coverage over the continental US, the PRISM datasets have been widely applied in previous hydrological modeling studies and were proven to be a reliable source of weather input. Chen et al. (2020) used PRISM climate data to drive the SWAT model for predicting monthly streamflow for the Upper Mississippi River Basin in the U.S.; they reported satisfactory results of NSE values ranging between 0.50 and 0.79 of ten sites in their study area. Muche et al. (2019) compared four gridded datasets using the SWAT model. The authors set up streamflow simulations in a Kansas Agricultural Watershed and found that the PRISM-driven model performed better during dry years than wet years. Yen et al. (2016) used Hydrologic and Water Quality System (HAWQS) for watershed modeling at the Illinois River Basin in the U.S., the PRISM data was used as the climate input, and the monthly streamflow prediction result was at a very good level with an NSE value of 0.70. Gao et al. (2017) compared SWAT streamflow prediction driven by PRISM, Next Generation Weather Radar (NEXRAD), and a network land-based National Climatic Data Center (NCDC) weather stations. They concluded that the PRISM-based model generated a smaller bias than the models utilizing NEXRAD and land-based weather stations.

In addition to direct comparison, hydrological models are often used to evaluate the accuracy of different weather products (Guo et al., 2004). The Soil and Water Assessment Tool (SWAT) is one of the most widely used rainfall-runoff models. It is a physically-based, semi-distributed, deterministic model developed to assess water quality and quantity at the watershed level (Arnold et al., 2012a). The climatic inputs of the SWAT model can be measured records or generated by the model itself (Gassman et al., 2007). The measured weather data can be input into the SWAT model in a point source data format, thus giving modelers significant flexibility in manipulating the weather data.

Artificial Neural Networks have become a popular rainfall-runoff modeling tool in the past three decades (ASCE, 2000a). An ANN model identifies nonlinear relationships from given patterns and fits nonparametric models on multivariate input data without considering any of the physical processes involved, typically referred to as a data-driven model (Govindaraju et al., 2013). Compared to the SWAT model, ANN models have straightforward setup and execution procedures, while the modelers have ample flexibility to determine the model inputs (Minns et al., 1996). Both SWAT and ANN models are found to have excellent performance producing streamflow estimation when accurate meteorological data were provided in many previous studies (Ahmed et al., 2007; Demirel et al., 2009; Jimeno-Sáez et al., 2018; Kim et al., 2015; Srivastava et al., 2006; Tuppad et al., 2011; Yaseen et al., 2015; Zakizadeh et al., 2020).

While plenty of previous studies have explored the hydrologic application of the weather products mentioned above, the applicability of the CFSR, TRMM, and PRISM datasets have not been adequately investigated in central Texas. In addition, there has been no detailed investigation of the effect that the alternative weather products have on streamflow simulation outcomes in SWAT and ANN. Therefore, this study seeks to use these two hydrological models to evaluate the suitability of the aforementioned gridded weather products. Specifically, the objectives of this study are to: (1) directly compare the TRMM, CFSR, PRISM, and conventional gauge rainfall datasets, (2) use the four rainfall data sources to separately calibrate/train the SWAT and ANN models for the same evaluation period, (3) compare the hydrological model performance when using each rainfall data source.

4.2. Methods and Materials

4.2.1. Study Area

The Leon Creek Watershed (LCW) in the San Antonio region of central south Texas was chosen as the study watershed due to the authors' familiarity with the area. The San Antonio region in central south Texas has a subtropical, semi-humid climate, with an average annual precipitation of near 750 mm (Cepeda, 2017). The study watershed was delineated using ArcSWAT by selecting the watershed outlet at USGS surface water gage 08181480 (U.S. Geological Survey, 2016). The delineated watershed has a drainage area of 535.76 km^2. It covers the western part of downtown San Antonio and centers at 98.67º west longitude, 29.56º north latitude. The LCW is heavily urbanized with extensive impervious covers. 47.2% of the LCW is classified as developed urban land according to the 2011 National Land Cover Database (NLCD2011). The elevation of LCW declines from its highest point of 548 m in the northern part of the watershed to the lowest point of 176 m in the south near the watershed outlet (Figure 4.1). Leon Creek is the main waterway in LCW, which originates from multiple smaller creeks in the northern part of the study area and flows southward. Leon Creek is a tributary of the Medina River, and it merges into the Medina River further south outside of the delineated study watershed.

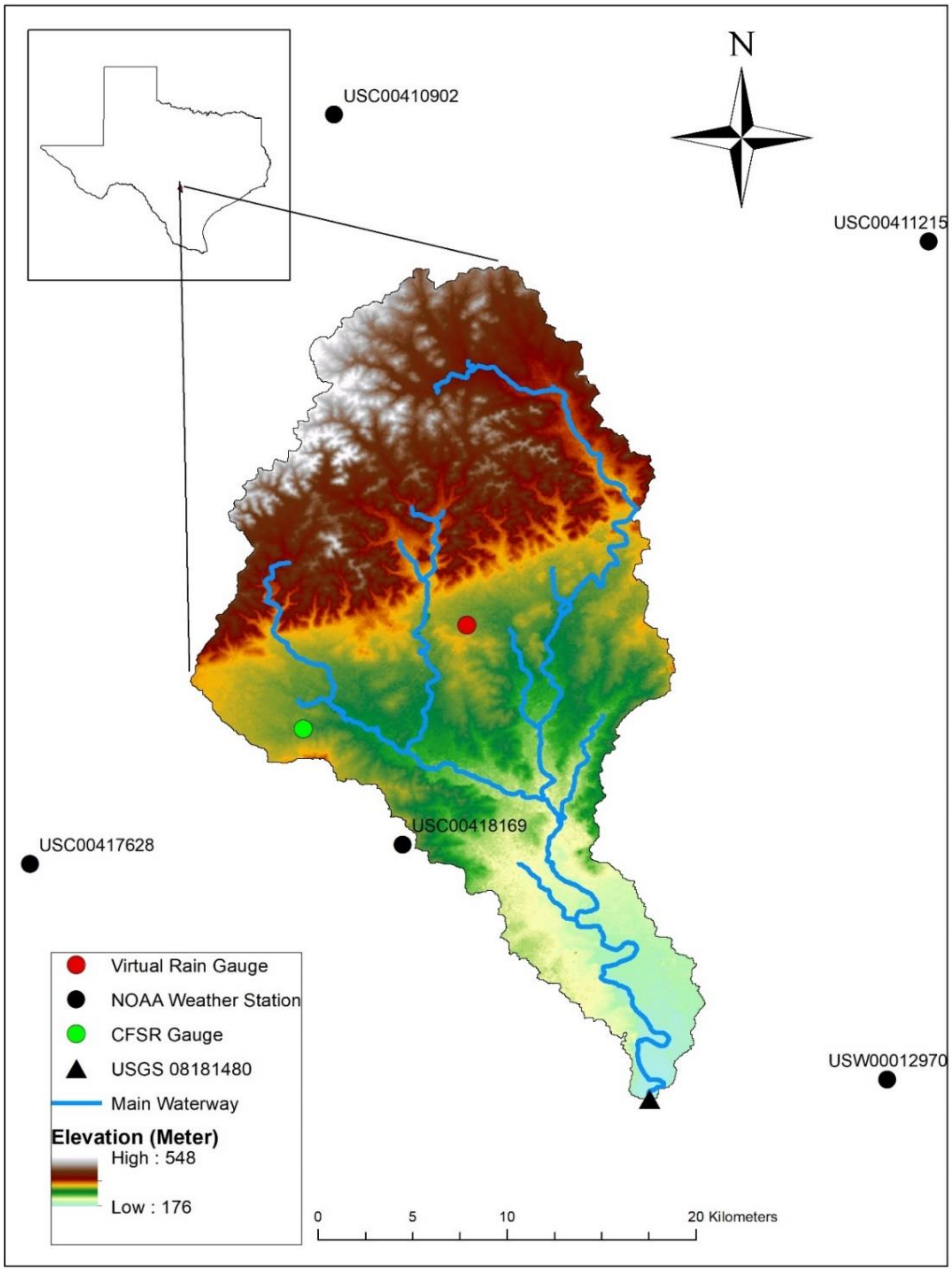

Figure 4.1 Location and digital elevation model of the Leon Creek Watershed (LCW) in Texas.

4.2.2. Data Acquisition

Weather records from 1998 to 2009 of TRMM, CFSR, PRISM, and conventional rain gauges were collected to drive the hydrological models. The conventional weather gauge station data was obtained from the National Oceanic and Atmospheric Administration's (NOAA) National Centers for Environmental Information (NCEI) climate data archive (https://www.ncdc.noaa.gov/cdo-web/search). A large number of weather stations have operated in the San Antonio region in the past. Nevertheless, only five stations proximate to LCW were found to have long-term precipitation records on a daily basis, none of which is physically located within the study watershed. The locations and station IDs of these five NOAA stations are shown in Figure 4.1. The precipitation, maximum, and minimum temperatures of the five stations were collected. The conventional gauge data was found to have multiple missing values during the study period, therefore, days with missing data were removed.

The CFSR weather data was downloaded from the Texas A&M University Global Weather Data for SWAT website (https://globalweather.tamu.edu/). The website provides CFSR data aggregated to a daily time step and interpolated to a SWAT input file format. The rectangular extent of the study watershed was used to extract the CFSR data. Within the study watershed, one CFSR gauge was available (Figure 4.1).

The TRMM and PRISM datasets were accessed using Google Earth Engine (GEE), a cloud-based geospatial analysis platform that provides easy access to many free geospatial data archives (Gorelick et al., 2017). The shapefile of the LCW was uploaded to the GEE platform to retrieve the pixels within the study watershed. The merged microwave/IR precipitation band of the TRMM 3B42 product at a 3-hour temporal resolution was downloaded. The TRMM data was further aggregated into daily time steps for comparison with the other precipitation datasets at the same temporal resolution. In this study, the daily precipitation, mean, minimum, and maximum temperature are collected for the PRISM data. The GEE platform was used to conduct map algebra that calculates the areal-averaged weather data from TRMM and PRISM of the study watershed.

Other data required for this study was obtained from multiple sources. ANNs require only metrological data and streamflow observation for model training, whereas SWAT requires additional spatial characteristics data, including the digital elevation model (DEM), land use land cover (LULC) map, and soil map. In this work, the state soil geographic (STATSGO) database preloaded with the ArcSWAT interface was used as the soil map (Schwarz et al., 1995). The National Elevation Dataset (NED) with 30 m resolution was used as the input DEM, and the 2011 National Land Cover Data Set (NLCD2011) was used as the LULC map. The NED and NLCD2011 datasets were accessed from the USDA Natural Resources Conversation Service (NRCS) geospatial data gateway (USDA-NRCS, 2014). In addition, the daily streamflow used for model calibration/training was obtained from USGS surface water gage 08181480 (U.S. Geological Survey, 2016) in Leon Creek from 2000 to 2009.

4.2.3. Hydrological Simulations

4.2.3.1. SWAT Modeling Approach

This study used the ArcSWAT 2012 built for ArcGIS 10.5 to construct the rainfall-runoff model for LCW. First, a threshold of 1500 ha was applied for stream definition. The threshold determines the minimum area for initiating stream networks. As a result, 25 subbasins were created. The study area was further discretized into 298 hydrological response units (HRUs) by applying a 10% threshold to remove minor slope, soil, and land cover classes. This procedure reduced the total number of HRUs, which improves computational efficiency. A detailed description of the SWAT modeling process can be found in the SWAT theoretical documentation (Neitsch et al., 2011).

The weather data sources discussed in section 4.2.2 were used as the SWAT weather input. The CFSR and conventional gauge data were in point source format, the format of SWAT weather input files (Arnold et al., 2012a). The longitude/latitude coordinates and elevation of the CFSR gauge and conventional gauges were directly used to create the precipitation and temperature files. The TRMM and PRISM data were original in gridded format and converted into areal-averaged point source files using GEE. The location and elevation of the watershed centroid were obtained using ArcGIS and set as the “virtual rain gauge” (Elhassan et al., 2016), as displayed in Figure 4.1. In total, four SWAT modeling scenarios were created, SWAT-CFSR, SWAT-GAUGE, SWAT-TRMM, and SWAT-PRISM. The TRMM dataset only provides rainfall estimates; hence the temperature data from PRISM was used to drive the SWAT-TRMM model. Meanwhile, the temperature data from the other three sources were used to drive their corresponding modeling scenarios.

The four SWAT modeling scenarios were run for a 12-year simulation period on a daily time step. The year 1998 to 1999 was used for model warm-up, 2000 to 2006 was used for calibration, and 2007 to 2009 for model validation. The model calibration and validation processes were carried out in the SWAT Calibration and Uncertainty Programs (SWAT-CUP) using the SUFI-2 procedure. This study selected 15 parameters that are considered sensitive for streamflow simulation according to the literature (Arabi et al., 2007; Chen et al., 2020; Jimeno-Sáez et al., 2018; Kim et al., 2015; Koycegiz et al., 2019; Qi et al., 2017). Their description and corresponding error range are summarized in Table 4.1.

Table 4.1 Description of the calibrated SWAT parameters.

Hydrology Parameter	Description	File Extension	Value Range
CN2*	SCS runoff curve number for antecedent moisture condition II	.mgt	(-10%, 10%)
ALPHA_BF	Base flow alpha factor (days)	.gw	(0, 1)
GW_DELAY	Delay time for aquifer recharge (days)	.gw	(0, 500)
GWQMN	Threshold depth of water in the shallow aquifer required for return flow to occur (mm H2O)	.gw	(0, 5000)
GW_REVAP	Groundwater "revap" coefficient	.gw	(0.02, 0.2)
REVAPMN	Threshold depth of water in the shallow aquifer for "revap" or percolation to the deep aquifer to occur (mm H2O)	.gw	(0, 500)
RCHRG_DP	Deep aquifer percolation fraction	.gw	(0, 1)
SOL_AWC*	Available water capacity of the soil layer (mm H2O/mm soil)	.sol	(-5%, 5%)
SOL_K*	Soil saturated hydraulic conductivity (mm/h)	.sol	(-5%, 5%)
ESCO	Soil Evaporation compensation factor	.hru	(0.6, 0.95)
CANMX	Maximum canopy storage (mm H2O)	.hru	(0, 100)
CH_K1	Effective hydraulic conductivity in tributary channel alluvium (mm/hr)	.sub	(5, 130)
CH_K2	Main channel hydraulic conductivity (mm/h)	.rte	(5, 130)
CH_N2	Manning's "n" value for the main channel	.rte	(0.01, 0.3)
SURLAG	Surface runoff lag coefficient (days)	.bsn	(1, 24)

Parameters using relative Change are marked by *, indicating parameter value is multiplied by 1 plus the given value.

4.2.3.2. ANN Modeling Approach

A comprehensive review of the conception and application of ANNs as rainfall-runoff models can be found in ASCE (2000a) and ASCE (2000b). Three-layered feed-forward neural networks are widely applied in hydrological modeling and were used in this study. The ANNs use a training process to estimate free model parameters. Routinely, a range of neural networks with different structures is trained, after which a model selection process is implemented to determine the model that makes the best prediction outcome. In this study, three model structures that only utilize meteorological data as input were explored (Table 4.2). The ANN-TRMM, ANN-CFSR, ANN-PRISM, and ANN-GAUGE models were trained using weather inputs corresponding to each of the rainfall datasets. The Thiessen polygon method was applied to interpolate the weather observations from the five gauging stations to the areal-averaged data of LCW, which was used as input to the ANN models. Multiple missing dates were removed from model training for the ANN-GUAGE model. As for SWAT, the temperature data from the PRISM dataset was used in the ANN-TRMM model. The predictors included daily precipitation (P_t), precipitation of the previous n days (P_{t-n}), daily mean air temperature (T_t), and total precipitation for the preceding n days (P_n). The training target was observed streamflow (Q) at the watershed outlet. The back-propagation algorithm was used for model training, and the logistic function was set as the transfer function at the hidden layer units. All input variables and the training target are normalized to the range of 0 to 1 to speed up model training.

Table 4.2 ANN model input combinations.

Prediction Scenario	Input Combination	Output
1	P_t, P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}, T_t	Q
2	P_t, P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}, P_n	Q
3	P_t, P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}, P_n, T_t	Q

In ANN rainfall-runoff modeling, too few hidden neurons may cause the model to fail to capture the complex nonlinear relationship between the predictors and targets, while too many hidden neurons can cause model overfitting (Demirel et al., 2009). In this work, the number of hidden layer units of all three input combinations was explored from 1 to 10, close to the experimental procedure of previous studies (Ha et al., 2003; Kalin et al., 2010; Noori et al., 2016). To select the best model among the trained models with different input combinations and hidden layer size, the Blocked Cross-Validation (BlockedCV) approach was applied. Cross-validation is the most widely used method for estimating prediction error in statistical modeling (Hastie et al., 2009). In rainfall-runoff modeling, the meteorological and hydrological data usually have strong autocorrelation and time dependency. BlockedCV groups the data points into sequentially consistent blocks and maintains sequential order within the blocks in the data splitting and cross-validation process (Bergmeir et al., 2012). Data from 2000 to 2006 was used to train the model, and data from 2007 to 2009 was used for standalone model testing. The R software (R Core Team, 2019) was used for all ANN simulations in this study.

4.2.4. Precipitation and Hydrological Models Evaluation

4.2.4.1. Rainfall Products Evaluation

The rainfall products comparison was conducted on the areal-averaged value of the study watershed, with calibration and validation phases evaluated independently. The Thiessen Polygon method areal-averaged precipitation from the conventional gauges, was compared with the other three gridded precipitation products. The relationship between the daily time series was evaluated using the Pearson correlation coefficient (CC) and percent bias (PBIAS), which mathematical formulation can be expressed in equation 1 and 2:

$$CC = \frac{COV(P_{grid}, P_{gauge})}{\sigma(P_{grid})\sigma(P_{gauge})} \tag{1}$$

$$PBIAS = 100 \cdot \frac{\sum_i (P_{grid} - P_{gauge})_i}{\sum_i P_{gauge,i}} \tag{2}$$

where P_{grid} and P_{gauge} denote the daily precipitation from the gridded weather products and conventional gauge data, respectively. The precipitation data were also aggregated to monthly resolution and evaluated graphically using box plots and scatter plots.

4.2.4.2. Hydrological Models Evaluation

The hydrological modeling results were evaluated using the Nash–Sutcliffe coefficient of efficiency (NSE) and percent bias (PBIAS). The NSE is a normalized statistic that determines the magnitude of residual variance compared to observed data variance. NSE ranges from –∞ to 1.0, with NSE = 1.0 representing the optimal fitting (Nash et al., 1970). The PBIAS measures the average tendency of model overestimation or underestimation. A smaller absolute PBIAS indicates better model fit to observed data. NSE and PBIAS were adopted in this study because they are commonly used in the literature, and extensive information regarding these two indicators is available from previous studies (Moriasi et al., 2015). The model performance evaluation criteria were adopted from (Moriasi et al., 2007), which recommended performance ratings for monthly time step hydrological simulations. Hydrological models are known to typically perform better at coarser temporal resolutions; hence, the performance ratings were slightly relaxed in this study (Kalin et al., 2010). The mathematical formulations of NSE and PBIAS and their corresponding performance criteria for daily streamflow simulation are presented in Table 4.3.

Table 4.3 Goodness-of-fit indicators and model performance evaluation criteria for the hydrological models.

Performance Rating	NSE $NSE = 1 - \frac{\sum_i (S_i - O_i)^2}{\sum_i (O_i - \bar{O})^2}$	PBIAS (%) $PBIAS = 100 * \frac{\sum_i (S_i - O_i)}{\sum_i O_i}$
Very good	$\text{NSE} \geq 0.7$	$\lvert\text{PBIAS}\rvert \leq 25$
Good	$0.5 \leq \text{NSE} < 0.7$	$25 < \lvert\text{PBIAS}\rvert \leq 50$
Satisfactory	$0.3 \leq \text{NSE} < 0.5$	$50 < \lvert\text{PBIAS}\rvert \leq 70$
Unsatisfactory	$\text{NSE} < 0.3$	$\lvert\text{PBIAS}\rvert > 70$

S_i is the i^{th} simulated data; O_i is the i^{th} observed data; $\bar{O}$ is the mean of the observed data.

4.3. Results and Discussion

4.3.1. Precipitation Data Analysis

The areal-averaged gauge rainfall data from NOAA was used as the reference to analyze the precipitation of the three gridded weather datasets (TRMM, CFSR, and PRISM). Table 4.4 summarized the daily average precipitation depth (Mean), the standard deviation (Std), and the maximum daily precipitation (Max) of the four data sources. The correlation coefficient (CC) and percent bias (PBIAS) between the gridded rainfall data and the reference data are also presented. Since the conventional gauge precipitation records were incomplete during the study period, the missing dates were removed from all datasets for the calculation of CC and PBIAS.

The statistical summary shows that the daily rainfall during the calibration and validation periods are close in magnitude. The TRMM data had the lowest mean, maximum, and standard deviation of the rainfall, significantly lower than the estimates from the CFSR, PRISM, and conventional gauge data. However, the mean daily rainfall values from the CFSR, PRISM, and conventional gauge datasets were relatively close. The CFSR data had the highest average daily rainfall estimation (2.56 mm/d) for the calibration period, while in the validation period, the gauge data has the highest average daily value (2.44 mm/d). The PRISM data had the highest estimates of the maximum daily rainfall for both the calibration (180.92 mm/d) and validation periods (173.79 mm/d), and the largest standard deviations (8.75 mm for calibration and 9.52 mm for validation period). The CC values indicated that the PRISM data has the strongest correlation with the conventional gauge data among the three gridded rainfall datasets, and the CFSR data has the weakest correlation. The PBIAS values agree with the daily mean estimates that the TRMM estimation of daily precipitation was significantly lower than that from gauge observation. Meanwhile, the CFSR and PRISM estimation of daily rainfall was slightly higher than the gauge observation.

Table 4.4 Statistical summary of all precipitation data and comparison between the areal-averaged gridded rainfall with conventional gauges data.

Weather Data	Calibration Period Precipitation					Validation Period Precipitation				
	Mean (mm/d)	Std (mm)	Max (mm/d)	CC	PBIAS	Mean (mm/d)	Std (mm)	Max (mm/d)	CC	PBIAS
TRMM	0.84	3.15	39.92	0.65	-65.4	0.79	3.01	44.28	0.65	-63.7
CFSR	2.56	8.37	130.42	0.55	5.5	2.32	6.28	104.71	0.55	5.6
PRISM	2.46	8.75	180.92	0.71	2.4	2.36	9.52	173.79	0.65	16.6
GAUGE	2.35	7.50	80.50			2.44	8.49	96.22		

Furthermore, the precipitation data were aggregated to the monthly time step to make graphical comparisons. The box plots of the aggregated monthly precipitation value of the four precipitation data sources are displayed in Figure 4.2. Some of the extremely high values were removed when creating the box plots to make the figure more readable. In both the calibration and validation periods, the TRMM 3B42 product had the lowest estimate of the median, lower and upper quartiles, and maximum values. In the calibration period, the monthly median rainfall estimation from CFSR (45.15 mm), PRISM (52.64 mm), and conventional gauge (52.51 mm) were relatively close compared with that from TRMM (16.98 mm). In the validation period, the PRISM data provided the highest estimates of median monthly precipitation of 46.68 mm, while the CFSR (31.14 mm) and conventional gauge (33.63 mm) had similar but smaller estimates.

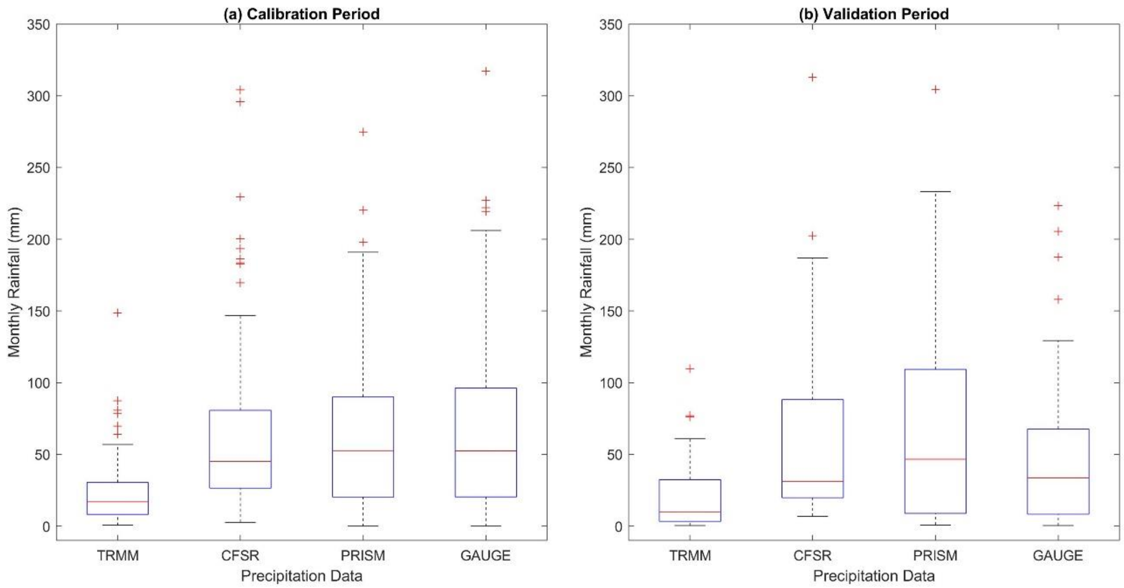

Figure 4.2 Monthly precipitation of TRMM, CFSR, PRISM, and conventional gauge for the (a) calibration and (b) validation period.

The scatter plots that compare aggregated monthly TRMM, CFSR, and PRISM precipitation data with the conventional gauge reference data are presented in Figure 4.3. In agreement with the results suggested by Table 4.4 and Figure 4.2, the least square regression lines for the TRMM data (Figure 3a and 3d) have slopes that are significantly lower than that for the CFSR and PRISM data (Figure 4.3b, 4.3c, 4.3e, and 4.3f), which indicates substantial underestimation of precipitation. Meanwhile, the least square regression lines for the CFSR and PRISM data were closer to the 1:1 reference line, indicating a closer approximation between these two datasets with the reference data. In particular, the PRISM data points (Figure 4.3c and 4.3f) were distributed nearer to the 1:1 reference line, while the CFSR data points (Figure 4.3b and 4.3e) were spread further apart, suggesting the PRISM data better approximates the gauge observations.

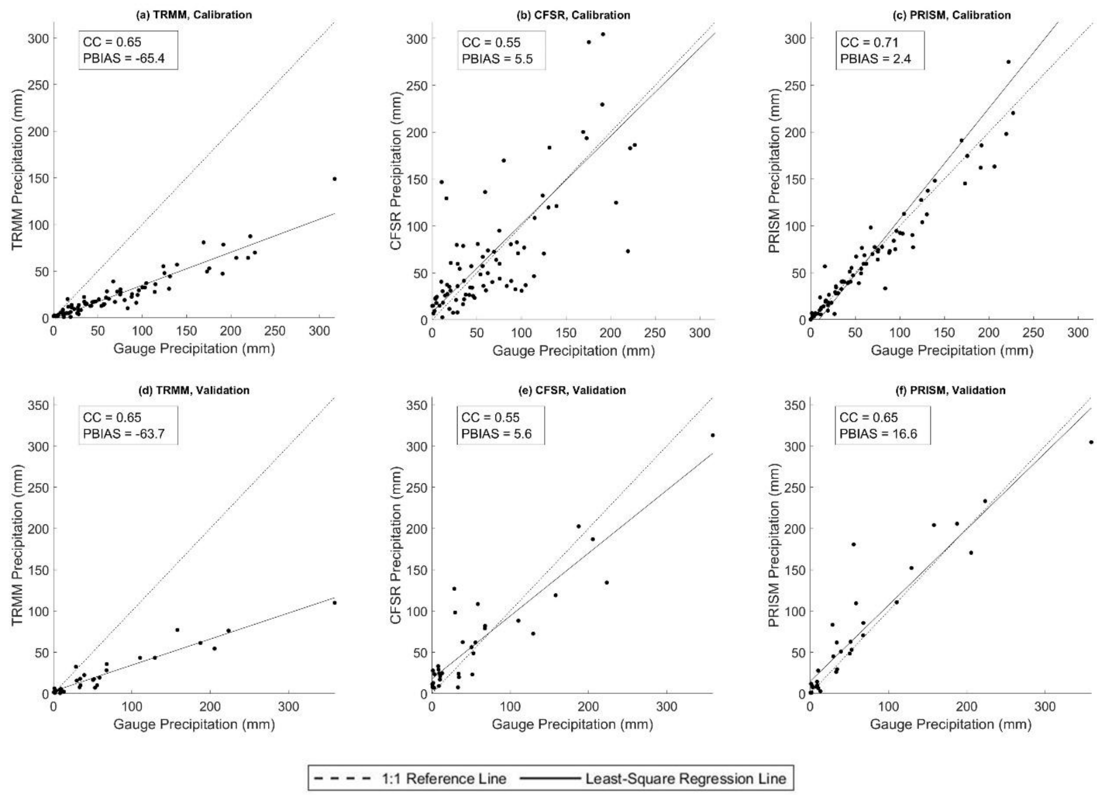

Figure 4.3 Comparison of the gridded monthly precipitation estimates with the conventional gauge data of the calibration period (a) TRMM, (b) CFSR, and (c) PRISM data; and the validation period (d) TRMM, (e) CFSR, and (f) PRISM data.

4.3.2. Hydrological Simulations

4.3.2.1. SWAT Calibration and Validation

SWAT calibration results are conditioned on the selected procedure, objective function, and availability of data over the time period (Abbaspour, 2011). The iterative SUFI-2 procedure was used in this study, and 500 simulations were run in each calibration iteration. The parameter values were updated after each iteration, and the iterations ended after the objective function ceased to improve. The Nash-Sutcliffe Coefficient of Efficiency (NSE) was used as the objective function. The four modeling scenarios were calibrated against the observed daily streamflow at the watershed outlet, and separate calibrations were made for each scenario. The calibration process minimized the difference between simulated and observed streamflow (Abbaspour et al., 2018). The last iteration of the calibration was used to define the parameter output ranges, which were used in the corresponding validation iterations without further modification. The calibration outcome and the best-fitted value in the validation iteration were presented in Table 4.5.

Table 4.5 Calibrated SWAT parameter ranges and the best-fitted validation values.

Hydrology Parameter	Default Value in SWAT	Calibrated Parameter Range				Best Fitted Validation Value			
		TRMM	CFSR	PRISM	GAUGE	TRMM	CFSR	PRISM	GAUGE
CN2.mgt*	35 to 98	(20.7%, 30.5%)	(-18.3%, -12.3%)	(-15.5%, -8.3%)	(-8.2%, -5.6%)	27.5%	-13.2%	-14.7%	-7.5%
ALPHA_BF.gw	0.048	(0.958, 1.000)	(0.377, 0.550)	(0.075, 0.224)	(0.649, 0.820)	0.968	0.537	0.086	0.671
GW_DELAY.gw	31	(123, 166)	(422, 461)	(228, 420)	(144, 235)	137	426	363	163
GWQMN.gw	1000	(2964, 3536)	(2239, 3753)	(3712, 4378)	(2463, 3442)	3227	3348	4268	2660
GW_REVAP.gw	0.02	(0.170, 0.190)	(0.129, 0.162)	(0.125, 0.150)	(0.158, 0.187)	0.185	0.155	0.136	0.168
REVAPMN.gw	750	(175, 249)	(236, 324)	(331, 412)	(124, 206)	225	259	348	171
RCHRG_DP.gw	0.05	(0.003, 0.085)	(0.055, 0.159)	(0.049, 0.159)	(0.000, 0.169)	0.016	0.150	0.157	0.032
SOL_AWC.sol*	0.01 to 0.42	(-1.2%, 0.7%)	(-1.3%, 0.4%)	(-5.3%, 1.5%)	(-5.2%, -1.1%)	0.4%	-1.2%	-2.6%	-1.1%
SOL_K.sol*	0 to 2000	(-2.6%, -1.4%)	(-0.9%, 0.4%)	(-3.2%, -1.6%)	(-1.6%, 1.8%)	-1.5%	-0.7%	-2.1%	-1.3%
ESCO.hru	0.95	(0.911, 0.950)	(0.713, 0.764)	(0.862, 0.921)	(0.600, 0.653)	0.938	0.740	0.872	0.614
CANMX.hru	0	(0, 4)	(51, 65)	(65, 82)	(40, 57)	0	62	79	50
CH_K1.sub	0	(11, 24)	(85, 113)	(77, 114)	(116, 130)	17	86	107	129
CH_K2.rte	0	(16, 29)	(116, 130)	(43, 60)	(8, 26)	28	120	51	21
CH_N2.rte	0.014	(0.114, 0.190)	(0.013, 0.032)	(0.010, 0.031)	(0.112, 0.180)	0.126	0.015	0.014	0.156
SURLAG.bsn	4	(6.09, 12.67)	(19.93, 23.16)	(3.81, 9.75)	(8.25, 12.73)	7.30	20.91	7.94	12.62

Parameters using percent change to existing values are marked by *.

The SCS runoff curve number (CN2.mgt) was reduced in the SWAT-CFSR, SWAT-PRISM, and SWAT-GAUGE models but significantly increased in the SWAT-TRMM model, compensating its lower precipitation i. Similarly, the maximum canopy storage (CANMX.hru) of SWAT-TRMM was kept at 0, while in other models, the CANMX.hru value was increased by different extents from the default. The hydraulic conductivity of the tributary channel (CH_K1.sub) and main channel (CH_K2.rte) for the SWAT-TRMM model was notably lower than the other models, while the Manning's n, thecoefficient for the main channel (CH_N2.rte), which represents the roughness of the channel, was higher in the SWAT-TRMM model. The adjustments conducted on the channel-related parameters reduce streamflow velocity in the SWAT-TRMM model compared to the other models. The soil parameters were adjusted through a percent change due to their spatial heterogeneity. In this study, the available water capacity of the soil layer (SOL_AWC.sol) and the soil saturated hydraulic conductivity (SOL_K.sol) were only slightly adjusted for all four SWAT modeling scenarios, which indicated that the streamflow was not sensitive to soil parameters in the study watershed.

The groundwater parameters (.gw) govern the speed and volume of groundwater recharge and discharge in the study watershed, which can be vital to the simulation performance since the San Antonio region is situated above Edwards Aquifer, which stores an enormous amount of groundwater and supplies much of the municipal consumption for San Antonio (Elhassan et al., 2016; Loáiciga et al., 2000). The relatively large base flow alpha factor (ALPHA_BF.gw) for the SWAT-TRMM, SWAT-CFSR, and SWAT-GAUGE models suggests the study area’s groundwater has a rapid response to recharge. Furthermore, the higher than default groundwater "revap" coefficient (GW_REVAP.gw) indicates the water transfer from the shallow aquifer to the root zone occurs at a relatively high rate. Meanwhile, the threshold depth of water in the shallow aquifer required for return flow to occur (GWQMN.gw) was markedly increased from the default value for all SWAT model scenarios, which likely suggested the study area has a large water storage capacity in its shallow aquifer, typically of karstic geology. In addition, the delay time for aquifer recharge (GW_DELAY.gw) was increased from the default value for all modeling scenarios, suggesting a longer time for water to exit the soil profile and enters the shallow aquifer in the study area.

4.3.2.2. ANN Training and Model Selection

All nodes in the neural networks were fully connected to nodes in their adjacent layers in this study. The links connecting the nodes contain weight and bias information which were optimized in the training process (ASCE, 2000a). The three-layer feed-forward neural network structure only contains one hidden layer besides the input and output layers. Thus, the primary purpose of the model selection process was to decide the number of hidden layer units that produces the best simulation outcome. As mentioned in section 4.2.3.2, the available data was split 70/30 ratio into the training and testing groups, respectively. The training data was further divided into ten blocks. In each BlockedCV iteration, nine blocks were used for model training, while the other block was used to calculate cross-validation statistics. The root mean square error (RMSE) of the standalone block was calculated in each training cross-validation iteration. The RMSE values were averaged after the training iterations finished. The model with the smallest averaged RMSE was selected as the best model. The models that produce the smallest cross-validation RMSE for each prediction scenario are displayed in Table 4.6.

Table 4.6 Best model structure and performance results of all prediction scenarios.

Model	Prediction Scenario	Hidden Nodes	Training Period		Testing Period		Cross-Validation RMSE (m3/s)
			NSE	PBIAS	NSE	PBIAS	
ANN-TRMM	1	9	0.835	31.9	0.188	-58.2	1.908
	2	6	0.825	40.0	0.491	-20.4	**1.888**
	3	9	0.829	46.6	0.371	-37.0	1.923
ANN-CFSR	1	7	0.835	-45.3	-0.010	-93.4	1.855
	2	8	0.819	-36.8	-0.003	-80.3	1.853
	3	6	0.796	-23.7	-0.005	-76.1	**1.852**
ANN-PRISM	1	7	0.901	60.7	0.736	-16.9	2.247
	2	5	0.891	59.4	0.671	1.3	**2.178**
	3	6	0.908	68.3	0.760	-6.3	2.231
ANN-GAUGE	1	8	0.699	-40.5	0.030	-75.7	2.094
	2	9	0.787	-44.2	0.091	-87.5	**2.059**
	3	8	0.748	-47.3	-0.016	-75.1	2.199

The best models selected using the cross-validation RMSE as criteria had their hidden unit sizes that fell between 5 and 9. This finding is consistent with that of Wu et al. (2005), which found the size of the hidden units to be near two-thirds of the sum of the number of input and output neurons. The smallest cross-validation RMSE values for each model were highlighted with an underscore in Table 4.6. The ANN-TRMM, ANN-PRISM, and ANN-GAUGE models selected scenario 2 as having the best model input combination, which only used precipitation data as predictors. The ANN-CFSR model selected scenario 3 as the best input combination. The inclusion of temperature as one of the predictors had slightly improved the cross-validation performance of the ANN-CFSR model. Overall, the NSE and PBIAS values were close among the different prediction scenarios of a particular ANN model.

4.3.2.3. Comparison of Model Performance

The ANN models summarized in Table 4.6 were further screened based on the cross-validation RMSE, in which only one prediction scenario for each model was chosen as the best model. The best ANN models were compared with the calibrated SWAT models, and their goodness-of-fit indicators were summarized in Table 4.7. In the calibration period, the SWAT models' NSE performance ranged from satisfactory to very good (0.48 to 0.90), and the ANN models' NSE performance was all on the very good level (NSE $\geq$ 0.7). However, the PBIAS values of the calibration period suggested that the SWAT models, with the exception of the SWAT-TRMM model, overestimated streamflow., SWAT-TRMM, however, was the model which had the much smaller rainfall input. Similarly, the ANN models also showed notable forecasting bias. The ANN-TRMM and ANN-PRISM model overestimated the streamflow by over 40%, while the ANN-GAUGE model underestimated the streamflow by 44.2%.

Table 4.7 Statistical performance of SWAT and ANN models driven by different weather data.

Model	Weather Data	Calibration/Training (2000-2006)		Validation/Testing (2007-2009)	
		NSE	PBIAS	NSE	PBIAS
SWAT	TRMM	0.48^{s}	10.3^{vg}	0.37^{s}	1.1^{vg}
	CFSR	0.56^{g}	52.4^{s}	0.22^{u}	0.9^{vg}
	PRISM	0.90^{vg}	49.6^{g}	0.72^{vg}	36.8^{g}
	GAUGE	0.61^{g}	29.2^{g}	0.07^{u}	35.5^{g}
ANN	TRMM	0.83^{vg}	40.0^{g}	0.49^{s}	-20.4^{vg}
	CFSR	0.80^{vg}	-23.7^{vg}	-0.01^{u}	-76.1^{u}
	PRISM	0.89^{vg}	56.2^{s}	0.76^{vg}	-13.4^{vg}
	GAUGE	0.79^{vg}	-44.2^{g}	0.09^{u}	-87.5^{u}

Superscripts represent the performance levels, “vg” - “very good”, “g” - “good”, “s” - “satisfactory”, “u” - “unsatisfactory”.

The hydrological models’ performance were worse in the validation period, during which only the TRMM and PRISM driven models reached at least a satisfactory level NSE performance. The SWAT-TRMM had a satisfactory validation NSE performance of 0.37, and the SWAT-PRISM model had a validation NSE value of 0.72, which was considered very good for daily streamflow simulation. Meanwhile, the validation NSE performance of SWAT-CFSR and SWAT-GAUGE models was below satisfactory level. Surprisingly, the SWAT-TRMM and SWAT-CFSR models had very minimal PBIAS values in the validation period, although not performing well based on the NSE criterion. Comparably, the ANN-TRMM model had a satisfactory performance of 0.49 NSE value, and the ANN-PRISM model had a very good performance of 0.76 NSE value, while the ANN-CFSR and ANN-GAUGE models performed poorly. Additionally, all ANN based models underestimated the streamflow according to the PBIAS values in the validation period, with ANN-CFSR and ANN-GAUGE severely underestimating streamflow, and ANN-TRMM and ANN-PRISM having a relatively a lower magnitude of underestimation.

The hydrographs of the validation period with precipitation records are displayed in Figure 4.4. Since gauge observations for much of the validation period were missing, there were several discontinuities in the ANN-GAUGE time series (Figure 4.4d) as no streamflow predictions were made on the missing dates. The observed streamflow time series suggests that the Leon Creek had close to 0 discharge volume for most of the validation period with only occasional moderate to high flows caused by intense storm events. In general, the TRMM and PRISM driven models captured the timing of major storm events rather well but had different levels of bias in flow magnitude (Figure 4.4a and 4.4c). The CFSR and conventional gauge driven models predicted the streamflow peaks poorly. The simulated streamflow time series also showed that the SWAT models generally made higher peak discharge estimations than the ANN models.

A scatter plot comparison of the simulated versus observed streamflow for the models driven by the different weather data sources is shown in Figure 4.5. The deviation of streamflow prediction increased with increasing discharge magnitude for all ANN and SWAT models. The SWAT-PRISM and ANN-PRISM models had least square regression lines comparably close to the 1:1 reference line, suggesting a smaller deviation between the simulated and observed data than other models. Meanwhile, the ANN-CFSR and ANN-GAUGE models were found to severely underpredict the streamflow with an extremely small regression line slope, which is in agreement with the PBIAS findings presented in Table 4.7. The regression line slope for all SWAT models was below that of the 1:1 reference line, which contradicts the PBIAS results that SWAT models overpredicted the streamflow. A very few underpredicted high flow values could be the cause of the small regression slope of the SWAT models, indicating the SWAT models overestimated low flows but underestimated high flows.

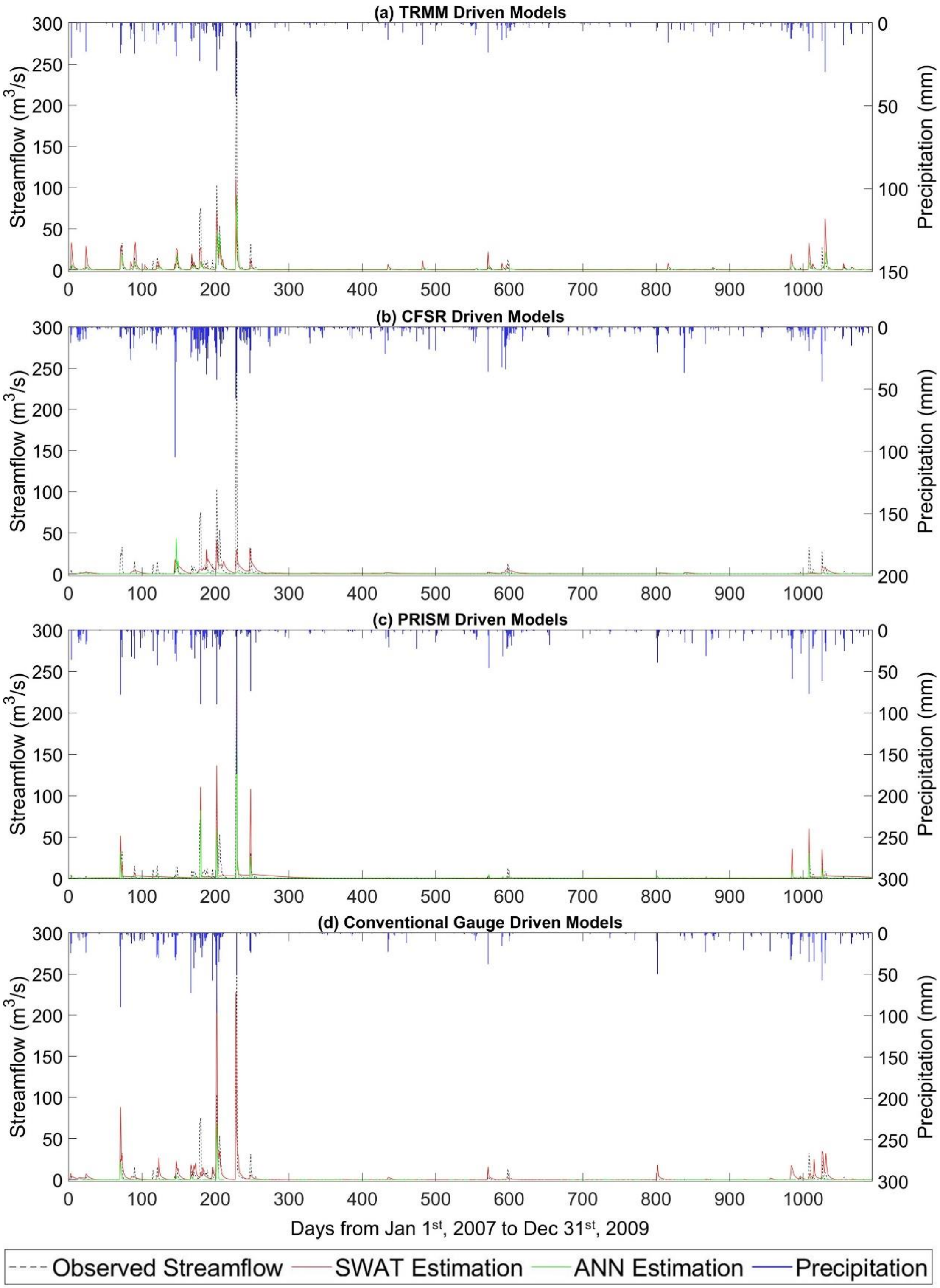

Figure 4.4 Hydrograph of the validation period for (a) TRMM, (b) CFSR, (c) PRISM, and (d) conventional gauge-driven SWAT and ANN models.

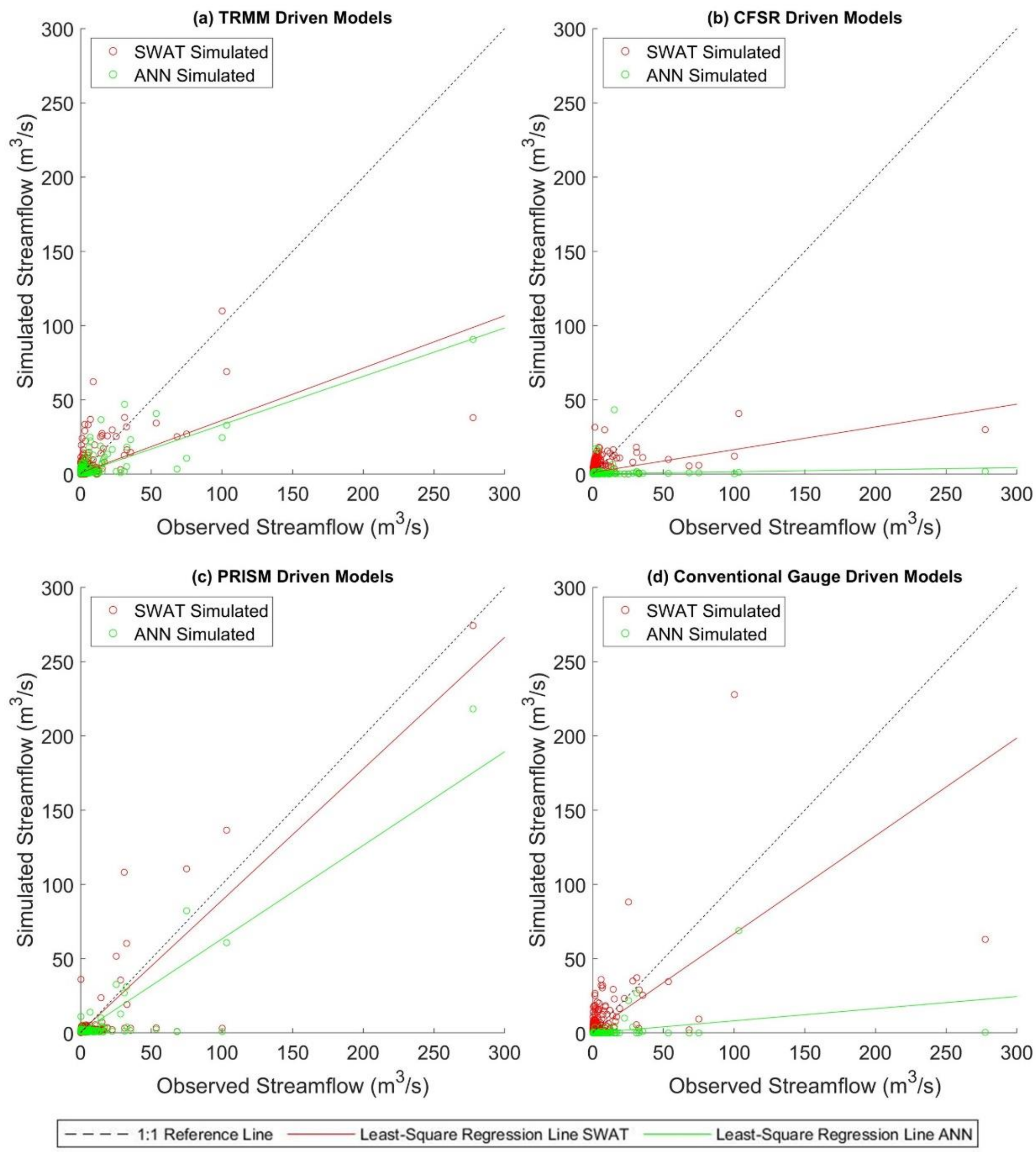

Figure 4.5 Comparison of validation period simulated and observed streamflow of the (a) TRMM, (b) CFSR, (c) PRISM, and (d) conventional gauge-driven models.

Overall, the performance of hydrological models in the standalone validation/testing period showed that the SWAT model overpredicted the streamflow and the ANN models underpredicted the streamflow for all evaluated weather data sources. The PRISM data was found to provide the most accurate hydrological simulations for both SWAT and ANN. The TRMM data also had satisfactory hydrological simulation performance, although NSE values were not as good as the PRISM-driven models.

The CFSR and conventional gauge driven models performed poorly according to the goodness-of-fit indicators and graphical comparison. This finding is unexpected given the common knowledge that rain gauges provide the most accurate precipitation measurement. These results are likely due to the lack of spatial representation of rain gauges and CFSR data in the study watershed. The five conventional rain gauges that had usable data in the region were spread outside the LCW boundary. Interpolating the rainfall data to the study area may fail to produce precise spatial representation. Another possible factor for the rain gauge-driven models' failure is that temporal inconsistencies of gauge rainfall observations restrict the hydrological models' prediction capability. This is particularly true for the ANN-GAUGE model, in which the temporal inconsistencies in input data undermine the time-dependent nature of streamflow series forecasting simulations. The CFSR data was initially in gridded format but automatically interpolated to the centroid point of the grid cell (Dile et al., 2014), which also lacks an accurate representation of the study watershed. On the other hand, the TRMM and PRISM data were precisely extracted and averaged for the study watershed. By comparing the models driven by different weather data sources, it can therefore be assumed that the areal-averaged rainfall data input at the "virtual rain gauge" (watershed centroid) is a viable option for making good streamflow simulations in SWAT and

ANN. Moreover, no significant discrepancy was found between SWAT and ANN simulation results when the same weather data source was used in the two models. At the same time, both models were greatly affected by the quality of precipitation inputs, and the results from the two models varied in the same way when different precipitation data sources were used.

4.4. Summary

SWAT and ANN are two widely used tools for streamflow prediction in the hydrological science community. This study evaluated the ability of four weather data sources to represent precipitation and drive hydrological simulations in a small urban watershed in central south Texas. The four different weather sources were directly compared on daily and monthly time steps. Furthermore, four SWAT models were calibrated and validated, and a number of ANN models were trained and selected to assess the relative performance of these different weather sources. Finally, goodness-of-fit indicators and graphical comparisons were employed to evaluate the results of hydrological simulations and further evaluate the different weather data source performance. The conclusions of this study can be summarized as follows:

(1) The Thiessen polygon method was adopted to interpolate areal-averaged gauge rainfall for the study watershed. Using the interpolated gauge rainfall as reference data, the TRMM data was found to severely underestimate rainfall, while the PRISM data most closely approximated the gauge observations.

(2) Only meteorological data was applied as the ANN model inputs. The ANN model selection results suggest that the precipitation data is adequate to make satisfactory streamflow prediction, except for the ANN-CFSR model, in which the addition of temperature as a predictor slightly improved the cross-validation RMSE performance.

(3) The calibrated SWAT models and the selected best ANN models had satisfactory to very good model performance during the calibration/training period, while the model performance significantly reduced in the validation/testing period, with the exception of both PRISM driven models.

(4) In the stand-alone validation/testing period, the PRISM data was found to provide the most accurate hydrological simulations for both SWAT and ANN. The TRMM data also had satisfactory level hydrological simulation performance. However, the CFSR and conventional gauge driven models performed poorly. The most likely explanation is that the interpolated CFSR and gauge rainfall data lacks spatial representation in the study watershed. Hence, the areal-averaged PRISM and TRMM data can offer a viable alternative for rainfall-runoff modeling when ground-based rainfall observation is limited.

(5) The input of precipitation is vital for hydrological simulations, and both the SWAT and ANN models were strongly affected by the quality of precipitation inputs. Specifically, the SWAT and ANN models varied in an identical pattern when different precipitation data sources were used as inputs, and there was no

significant discrepancy found between SWAT and ANN simulation results when the same weather data source was applied.

5. CONCLUSIONS

The motivation of this dissertation was to explore the capability of SWAT and ANN as rainfall-runoff models across a range of conditions. All studies were conducted in the Edwards Aquifer region of Texas, using publicly accessible geological, meteorological, and hydrological data. The study in chapter 2 evaluated two major approaches to model selection for the ANN based rainfall-runoff model. This study also acted as a precursor to examine if the ANN model could produce successful streamflow prediction in the San Antonio region. The study in chapter 3 directly compared SWAT and ANN by using each as rainfall-runoff models in a pair of small watersheds, one primarily urban and the other primarily rural. A correction factor approach was used to adjust the goodness-of-fit indicators to incorporate measurement and model uncertainty in the rainfall-runoff modeling evaluation process. The urban watershed in this study was found to have better streamflow prediction for both the SWAT and ANN approaches; hence it was used as the study watershed in chapter 4, which focused on assessing three common gridded precipitation datasets and their impacts on hydrological modeling using both SWAT and ANN.

The study in chapter 2 was conducted using ANN in two 10-digit watersheds, the Headwaters San Antonio River Basin (HSARB), which is heavily urban, and the Lower Medina River Basin (LMRB), which is largely rural. In three of the five model prediction scenarios, the discharge from the upstream gauge station was used as a predictor. The modeling results show that the AIC and BlockedCV selected networks with the optimum number of hidden nodes could produce good daily streamflow prediction for most of the input scenarios. While several studies in the past showed that the cross-validation-based approach could be successfully applied to ANN hydrological modeling (Humphrey et al., 2016; Jimeno-Sáez et al., 2018; Kim et al., 2015; Maier et al., 1996; Srivastava et al., 2006), none of these studies specifically used the BlockedCV. Through empirical investigation, the study in chapter 2 found that the out-of-sample approach (BlockedCV) was more desirable for identifying the neural network that best predicted streamflow. The in-sample approaches (AIC and BIC) tended to select simpler models that underfit the training data due to their penalty on the number of free parameters. These results corroborate the findings of a great deal of the previous work in Qi et al. (2001), which reached similar conclusions on a few economic time series.

Contrary to expectations, the selected best model of the rural LMRB performed better statistically than that of the urban HSARB. This result may be explained by the much larger average outflow from LMRB than HSARB, as ANN models usually perform better when predictors and targets have larger values. Additionally, a Cox and Stuart test showed no significant difference between the training and testing discharge datasets in LMRB, while the testing period discharge trended away from the training period in the HSARB. This significant difference in the training and testing datasets in HSARB may be another reason for the relatively poorer predictive performance of the urban watershed.

The study in chapter 3 employed a paired watershed experimental design similar to that of the chapter 2 study. In the chapter 3 study, two different 10-digit watersheds were selected, the urban Leon Creek Watershed (LCW) and the rural Upper Medina

River Watershed (UMRW). This study briefly discussed the conceptual distinction between SWAT and ANN and empirically assessed the two models' performance in the karstic San Antonio region. The ANN input scenarios in this study are slightly different from the study in chapter 2, as the upstream discharge observation were not available for LCW and UMRW. In contrast to the findings in the previous study, both SWAT and ANN models successfully predicted streamflow in the urban watershed but did not perform well in the rural watershed. In particular, the SWAT model had an unsatisfactory performance with a negative NSE value in the rural UMRW. For the ANN models, the inclusion of previous time step streamflow as a predictor did not noticeably improve model performance in the urban LCW, whereas the model performance significantly improved when a time series autoregressive model structure using historical streamflow data was implemented in the rural UMRW. It is difficult to explain these contradictory results between these two studies, but it could be related to the level of similarity between training and testing data. While the two studies were conducted in the same geological region, their study periods were completely different. This finding suggests that training/testing data split can strongly affect the performance results of ANNs.

Overall, the ANN models outperformed the SWAT models for both high and low flow conditions by different margins in the chapter 3 study. This finding is different from that of Kim et al. (2015) and Jimeno-Sáez et al. (2018), which both suggest that ANN models generally performed better at simulating high flows while SWAT had better performance simulating low flows. However, it is in accordance with the results of

Srivastava et al. (2006), who reported that the ANN model produced simulation results with NSE values better than the SWAT model. Additionally, applying probability based correction factors to the goodness-of-fit indicators did not result in a visible improvement when a lower uncertainty level was assumed (Cv = 0.026), and only a slight improvement when a higher uncertainty level was assumed (Cv = 0.256). These results reflect those of Harmel et al. (2010), who also observed a noticeable improvement in the goodness-of-fit indicators. In addition, the chapter 3 study also validated that the results and opinions of a few previous studies (Malagò et al., 2016; Shao et al., 2019; Spruill et al., 2000; Wang et al., 2019), which pointed out that the conventional SWAT model is not capable of accurately modeling hydrological variables in karst watershed.

Both of the studies in chapters one and two used weather data from the Parameter-elevation Relationships on Independent Slopes Model (PRISM). The study in chapter 4 expanded on the work done in the chapter 3 study to evaluate alternatives to conventional ground-based weather data. The gridded-based PRISM data, the ground-based NOAA data, and two additionally gridded precipitation datasets, Tropical Rainfall Measuring Mission (TRMM) and Climate Forecast System Reanalysis (CFSR), were compared and used to drive hydrological simulations in the Leon Creek Watershed. Surprisingly, the conventional gauge driven hydrological models were found to perform poorly in the chapter 4 study. This finding was unexpected but supported the idea that the hydrological modeling performance can be undermined when the ground-based rainfall observation network is unable to capture the spatial variation of precipitation (Li et al., 2018; Worqlul et al., 2014).

Additionally, while the PRISM datasets have been widely applied in previous hydrological modeling studies and were proven to be a reliable source of weather input (Chen et al., 2020; Muche et al., 2019; Radcliffe et al., 2017; Tobin et al., 2013; Yen et al., 2016), the accuracy of TRMM and CFSR and the performance of using them in hydrological modeling were more mixed (Fuka et al., 2014; Himanshu et al., 2018; Li et al., 2018; Mararakanye et al., 2020; Ochoa et al., 2014; Radcliffe et al., 2017; Roth et al., 2016; Stampoulis et al., 2012; Worqlul et al., 2014). This study showed that using areal-average rainfall from multiple sources as input to SWAT and ANN can make satisfactory streamflow predictions. Among the evaluated rainfall products, the PRISM data produced the best hydrological simulation outcome. The TRMM precipitation data was found to significantly underestimate the volume of rainfall compared with the other three rainfall data sources. However, the TRMM driven hydrological models still achieved satisfactory performance results. In contrast, the CFSR and conventional gauge data performed poorly, most likely caused by their poor spatial representation in the study watershed. Hence, the areal-averaged PRISM and TRMM data can offer a viable alternative for rainfall-runoff modeling when ground-based rainfall observation is limited. More importantly, the SWAT and ANN models varied in an identical pattern when different precipitation data sources were used as inputs.

In general, this research shows that ANN models can be a reliable real-time simulator of streamflow, outperforming the physically-based SWAT model in several cases. However, this work was limited to hydrological simulations in small-sized watersheds in the San Antonio region. The study area was unique with its vast karstic groundwater aquifer that has rapid recharge and discharge capabilities. An issue that was not addressed was whether the paired watershed study between urban and rural watersheds would yield closer model performance in an area without similar karst geological features. Furthermore, the scope of this research was limited in terms of the temporal resolution of the hydrological simulations. Only daily time step simulations were conducted in all sections of this dissertation to enable sufficient training data for the ANNs.

A natural progression of this work is to apply similar experiments to larger watersheds and regions with different geological and climatological conditions to further verify the conclusions. More broadly, further investigations into different temporal resolutions of ANN simulations are also recommended. A future study could investigate coarser temporal resolutions if sufficient long climate and streamflow records are available, with the aim of providing long-term streamflow trend prediction and water availability analysis using an ANN model. Shorter time steps could also be considered for the purpose of flood warning, for example, if accurate weather data at finer temporal resolution becomes available. In the meantime, the ANN models in this research were created as lumped models with a single point format weather input. Another interesting topic regarding hydrological modeling using ANN is to explore the application of a distributed format of weather input. Such work can be done by discretizing the study watershed into smaller subbasins.

This research also tested and verified the method of converting gridded format weather data into a point format by calculating their areal-averaged value. Converted meteorological data provided reliable inputs with the suitable format for the SWAT and ANN models and produced satisfactory streamflow simulation results. This method can be expanded into hydrological simulations using other lumped or semi-distributed models in the future, as more gridded weather products, either produced from satellite remote sensing techniques alone or created as hybrid ground-based measurement and remotely sensed estimates, are becoming publicly accessible. Moreover, due to the limitation of time and scope, this research only evaluated three common gridded-based precipitation datasets. Further research could also be conducted using radar estimated precipitation data that are gradually becoming available on a more refined spatial scale.

REFERENCE

Abbaspour, K. C. (2011). SWAT-CUP4: SWAT calibration and uncertainty programs–a user manual. *Swiss Federal Institute of Aquatic Science and Technology, Eawag, 106*.

Abbaspour, K. C., Rouholahnejad, E., Vaghefi, S., Srinivasan, R., Yang, H., & Kløve, B. (2015). A continental-scale hydrology and water quality model for Europe: Calibration and uncertainty of a high-resolution large-scale SWAT model. *Journal of Hydrology, 524*, 733-752.

Abbaspour, K. C., Vaghefi, S. A., & Srinivasan, R. (2018). A guideline for successful calibration and uncertainty analysis for soil and water assessment: A review of papers from the 2016 International SWAT Conference. In: Multidisciplinary Digital Publishing Institute.

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., . . . Bolvin, D. (2003). The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). *Journal of hydrometeorology, 4*(6), 1147-1167.

Ahmed, J. A., & Sarma, A. K. (2007). Artificial neural network model for synthetic streamflow generation. *Water resources management, 21*(6), 1015-1029.

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike* (pp. 215-222): Springer.

Amiri, B., Sudheer, K., & Fohrer, N. (2012). Linkage between in-stream total phosphorus and land cover in Chugoku district, Japan: an ANN approach. *Journal of Hydrology and Hydromechanics*, *60*(1), 33-44.

Arabi, M., Govindaraju, R. S., & Hantush, M. M. (2007). A probabilistic approach for analysis of uncertainty in the evaluation of watershed management practices. *Journal of Hydrology, 333*(2-4), 459-471.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys, 4*, 40-79.

Arnold, J., Kiniry, J., Srinivasan, R., Williams, J., Haney, E., & Neitsch, S. (2012a). Soil and Water Assessment Tool, Input/Output Documentation Version 2012. Texas Water Resources Institute. In.

Arnold, J. G., Kiniry, J., Srinivasan, R., Williams, J., Haney, E., & Neitsch, S. J. T. W. R. I. (2012b). Soil and water assessment tool input/output documentation version 2012. *7*.

ASABE. (2017). Guidelines for Calibrating, Validating, and Evaluating Hydrologic and Water Quality (H/WQ) Models.

ASCE. (2000a). Artificial neural networks in hydrology. I: Preliminary concepts. *Journal of Hydrologic Engineering, 5*(2), 115-123.

ASCE. (2000b). Artificial neural networks in hydrology. II: Hydrologic applications. *Journal of Hydrologic Engineering, 5*(2), 124-137.

Beaumont, C. (1979). Stochastic models in hydrology. *Progress in Physical Geography, 3*(3), 363-391.

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences, 191*, 192-213.

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

Birikundavyi, S., Labib, R., Trung, H., & Rousselle, J. (2002). Performance of neural networks in daily streamflow forecasting. *Journal of Hydrologic Engineering, 7*(5), 392-398.

Brirhet, H., & Benaabidate, L. (2016). Comparison of two hydrological models (lumped and distributed) over a pilot area of the Issen watershed in the Souss basin, Morocco. *European Scientific Journal, 12*(18).

Cepeda, J. C. (2017). INFLUENCE OF PACIFIC SEA SURFACE TEMPERATURES ON PRECIPITATION IN TEXAS: DATA FROM AMARILLO AND SAN ANTONIO, 1900-2013. *Texas Journal of Science, 69*(1).

Chen, M., Gassman, P. W., Srinivasan, R., Cui, Y., & Arritt, R. (2020). Analysis of alternative climate datasets and evapotranspiration methods for the Upper Mississippi River Basin using SWAT within HAWQS. *Science of the Total Environment*, 137562.

Cox, D. R., & Stuart, A. (1955). Some quick sign tests for trend in location and dispersion. *Biometrika, 42*(1/2), 80-95.

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., . . . Pasteris, P. P. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology, 28*(15), 2031-2064.

Demirel, M. C., Venancio, A., & Kahya, E. (2009). Flow forecast by SWAT model and ANN in Pracana basin, Portugal. *Advances in Engineering Software, 40*(7), 467-473.

Dile, Y. T., & Srinivasan, R. (2014). Evaluation of CFSR climate data for hydrologic prediction in data-scarce watersheds: an application in the Blue Nile River Basin. *JAWRA Journal of the American Water Resources Association, 50*(5), 1226-1241.

Donate, J. P., Cortez, P., SáNchez, G. G., & De Miguel, A. S. (2013). Time series forecasting using a weighted cross-validation evolutionary artificial neural network ensemble. *Neurocomputing, 109*, 27-32.

Dorofki, M., Elshafie, A. H., Jaafar, O., Karim, O. A., & Mastura, S. (2012). Comparison of artificial neural network transfer functions abilities to simulate extreme runoff data. *International Proceedings of Chemical, Biological Environmental Engineering, 33*, 39-44.

Elhassan, A., Xie, H., Al-othman, A. A., Mcclelland, J., & Sharif, H. O. (2016). Water quality modelling in the San Antonio River Basin driven by radar rainfall data. *Geomatics, Natural Hazards and Risk, 7*(3), 953-970.

Fatichi, S., Vivoni, E. R., Ogden, F. L., Ivanov, V. Y., Mirus, B., Gochis, D., . . . Ebel, B. (2016). An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. *Journal of Hydrology, 537*, 45-60.

Fernandez, G. P., Chescheir, G. M., Skaggs, R. W., & Amatya, D. M. (2005). Development and testing of watershed-scale models for poorly drained soils. *Transactions of the ASAE, 48*(2), 639-652.

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.

Gao, J., Sheshukov, A. Y., Yen, H., & White, M. J. (2017). Impacts of alternative climate information on hydrologic processes with SWAT: A comparison of NCDC, PRISM and NEXRAD datasets. *Catena, 156*, 353-364.

Gassman, P. W., Reyes, M. R., Green, C. H., & Arnold, J. G. (2007). The soil and water assessment tool: historical development, applications, and future research directions. *Transactions of the ASABE, 50*(4), 1211-1250.

Gazzaz, N. M., Yusoff, M. K., Aris, A. Z., Juahir, H., & Ramli, M. F. (2012). Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. *Marine pollution bulletin, 64*(11), 2409-2420.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment, 202*, 18-27.

Govindaraju, R. S., & Rao, A. R. (2013). *Artificial neural networks in hydrology* (Vol. 36): Springer Science & Business Media.

Guo, J., Liang, X., & Leung, L. R. (2004). Impacts of different precipitation data sources on water budgets. *Journal of Hydrology, 298*(1-4), 311-334.

Gupta, H., Hsu, K., & Sorooshian, S. (2000). Effective and efficient modeling for streamflow forecasting. In *Artificial neural networks in hydrology* (pp. 7-22): Springer.

Ha, H., & Stenstrom, M. K. (2003). Identification of land use with water quality data in stormwater using a neural network. *Water Research, 37*(17), 4222-4230.

Haan, C. (2002). Statistical Methods in Hydrology, 2nd Edn., 496 pp. In: Iowa State Press, Ames, IA.

Harmel, D., Smith, P., Migliaccio, K., Chaubey, I., Douglas-Mankin, K., Benham, B., . . . Robson, B. J. (2014). Evaluating, interpreting, and communicating performance of hydrologic/water quality models considering intended use: A review and recommendations. *Environmental modelling & software, 57*, 40-51.

Harmel, D., & Smith, P. K. (2007). Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling. *Journal of Hydrology, 337*(3-4), 326-336.

Harmel, D., Smith, P. K., & Migliaccio, K. W. (2010). Modifying goodness-of-fit indicators to incorporate both measurement and model uncertainty in model calibration and validation. *Transactions of the ASABE, 53*(1), 55-63.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*: Springer Science & Business Media.

Her, Y., Frankenberger, J., Chaubey, I., & Srinivasan, R. (2015). Threshold effects in HRU definition ofthe soil and water assessment tool. *Transactions of the ASABE, 58*(2), 367-378.

Himanshu, S. K., Pandey, A., & Patil, A. (2018). Hydrologic evaluation of the TMPA-3B42V7 precipitation data set over an agricultural watershed using the SWAT model. *Journal of Hydrologic Engineering, 23*(4), 05018003.

Hsu, K.-l., Gao, X., Sorooshian, S., & Gupta, H. V. (1997). Precipitation estimation from remotely sensed information using artificial neural networks. *Journal of Applied Meteorology, 36*(9), 1176-1190.

Hu, T., Lam, K., & Ng, S. (2001). River flow time series prediction with a range-dependent neural network. *Hydrological Sciences Journal, 46*(5), 729-745.

Humphrey, G. B., Gibbs, M. S., Dandy, G. C., & Maier, H. R. (2016). A hybrid approach to monthly streamflow forecasting: integrating hydrological model outputs into a Bayesian artificial neural network. *Journal of Hydrology, 540*, 623-640.

Isik, S., Kalin, L., Schoonover, J. E., Srivastava, P., & Lockaby, B. G. (2013). Modeling effects of changing land use/cover on daily streamflow: an artificial neural network and curve number based hybrid approach. *Journal of Hydrology, 485*, 103-112.

Islam, Z. (2011). A review on physically based hydrologic modeling. *University of Alberta: Edmonton, AB, Canada*.

Jain, A., & Srinivasulu, S. (2004). Development of effective and efficient rainfall-runoff models using integration of deterministic, real-coded genetic algorithms and artificial neural network techniques. *Water Resources Research, 40*(4).

Jakada, H., & Chen, Z. (2020). An approach to runoff modelling in small karst watersheds using the SWAT model. *Arabian Journal of Geosciences, 13*(8).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.

Jayakrishnan, R., Srinivasan, R., Santhi, C., & Arnold, J. (2005). Advances in the application of the SWAT model for water resources management. *Hydrological Processes: An International Journal, 19*(3), 749-762.

Jimeno-Sáez, P., Senent-Aparicio, J., Pérez-Sánchez, J., & Pulido-Velazquez, D. (2018). A Comparison of SWAT and ANN models for daily runoff simulation in different climatic zones of peninsular Spain. *Water, 10*(2), 192.

Joseph, J. F., Falcon, H. E., & Sharif, H. O. (2013). Hydrologic trends and correlations in south Texas River basins: 1950–2009. *Journal of Hydrologic Engineering, 18*(12), 1653-1662.

Kalin, L., Isik, S., Schoonover, J. E., & Lockaby, B. G. (2010). Predicting water quality in unmonitored watersheds using artificial neural networks. *Journal of environmental quality, 39*(4), 1429-1440.

Karunanithi, N., Grenney, W. J., Whitley, D., & Bovee, K. (1994). Neural networks for river flow prediction. *Journal of Computing in Civil Engineering, 8*(2), 201-220.

Kim, M., Baek, S., Ligaray, M., Pyo, J., Park, M., & Cho, K. (2015). Comparative studies of different imputation methods for recovering streamflow observation. *Water, 7*(12), 6847-6860.

Kişi, Ö. (2007). Streamflow forecasting using different artificial neural network algorithms. *Journal of Hydrologic Engineering, 12*(5), 532-539.

Koycegiz, C., & Buyukyildiz, M. (2019). Calibration of SWAT and two data-driven models for a data-scarce mountainous headwater in semi-arid Konya closed basin. *Water, 11*(1), 147.

Kreuter, U. P., Harris, H. G., Matlock, M. D., & Lacey, R. E. (2001). Change in ecosystem service values in the San Antonio area, Texas. *Ecological economics, 39*(3), 333-346.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26): Springer.

Li, D., Christakos, G., Ding, X., & Wu, J. (2018). Adequacy of TRMM satellite rainfall data in driving the SWAT modeling of Tiaoxi catchment (Taihu lake basin, China). *Journal of Hydrology, 556*, 1139-1152.

Licciardello, F., Rossi, C., Srinivasan, R., Zimbone, S., & Barbagallo, S. (2011). Hydrologic evaluation of a Mediterranean watershed using the SWAT model with multiple PET estimation methods. *Transactions of the ASABE, 54*(5), 1615-1625.

Loáiciga, H., Maidment, D., & Valdes, J. B. (2000). Climate-change impacts in a regional karst aquifer, Texas, USA. *Journal of Hydrology, 227*(1-4), 173-194.

Maier, H. R., & Dandy, G. C. (1996). The use of artificial neural networks for the prediction of water quality parameters. *Water Resources Research, 32*(4), 1013-1022.

Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental modelling & software, 15*(1), 101-124.

Malagò, A., Efstathiou, D., Bouraoui, F., Nikolaidis, N. P., Franchini, M., Bidoglio, G., & Kritsotakis, M. (2016). Regional scale hydrologic modeling of a karst-dominant geomorphology: The case study of the Island of Crete. *Journal of Hydrology, 540*, 64-81.

Mararakanye, N., Le Roux, J., & Franke, A. (2020). Using satellite-based weather data as input to SWAT in a data poor catchment. *Physics and Chemistry of the Earth, Parts A/B/C, 117*, 102871.

Maraun, D., Wetterhall, F., Ireson, A., Chandler, R., Kendon, E., Widmann, M., . . . Themeßl, M. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of geophysics, 48*(3).

Meresa, H. (2019). Modelling of river flow in ungauged catchment using remote sensing data: application of the empirical (SCS-CN), Artificial Neural Network (ANN) and Hydrological Model (HEC-HMS). *Modeling Earth Systems, 5*(1), 257-273.

Milly, P. C., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., & Stouffer, R. J. (2008). Stationarity is dead: Whither water management? *Science, 319*(5863), 573-574.

Minns, A., & Hall, M. (1996). Artificial neural networks as rainfall-runoff models. *Hydrological Sciences Journal, 41*(3), 399-417.

Moradkhani, H., & Sorooshian, S. (2009). General review of rainfall-runoff modeling: model calibration, data assimilation, and uncertainty analysis. In *Hydrological modelling and the water cycle* (pp. 1-24): Springer.

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.

Moriasi, D. N., Gitau, M. W., Pai, N., & Daggupati, P. (2015). Hydrologic and water quality models: Performance measures and evaluation criteria. *Transactions of the ASABE, 58*(6), 1763-1785.

Muche, M. E., Sinnathamby, S., Parmar, R., Knightes, C. D., Johnston, J. M., Wolfe, K., . . . Smith, D. (2019). Comparison and Evaluation of Gridded Precipitation Datasets in a Kansas Agricultural Watershed Using SWAT. *Journal of the American Water Resources Association*.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology, 10*(3), 282-290.

Neitsch, S. L., Arnold, J. G., Kiniry, J. R., & Williams, J. R. (2011). *Soil and water assessment tool theoretical documentation version 2009*. Retrieved from

Noori, N., & Kalin, L. (2016). Coupling SWAT and ANN models for enhanced daily streamflow prediction. *Journal of Hydrology, 533*, 141-151.

Ochoa, A., Pineda, L., Crespo, P., & Willems, P. (2014). Evaluation of TRMM 3B42 precipitation estimates and WRF retrospective precipitation simulation over the Pacific–Andean region of Ecuador and Peru. *Hydrology and Earth System Sciences, 18*(8), 3179-3193.

Olivera, F., Valenzuela, M., Srinivasan, R., Choi, J., Cho, H., Koka, S., & Agrawal, A. (2006). ARCGIS-SWAT: A GEODATA MODEL AND GIS INTERFACE FOR SWAT 1. *Journal of the American Water Resources Association, 42*(2), 295-309.

Qi, M., & Zhang, G. P. (2001). An investigation of model selection criteria for neural network time series forecasting. *European Journal of Operational Research, 132*(3), 666-680.

Qi, Z., Kang, G., Chu, C., Qiu, Y., Xu, Z., & Wang, Y. (2017). Comparison of SWAT and GWLF model simulation performance in humid south and semi-arid north of China. *Water, 9*(8), 567.

R Core Team. (2019). R: A language and environment for statistical computing: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Radcliffe, D., & Mukundan, R. (2017). PRISM vs. CFSR precipitation data effects on calibration and validation of SWAT models. *JAWRA Journal of the American Water Resources Association, 53*(1), 89-100.

Rezaeianzadeh, M., Stein, A., Tabari, H., Abghari, H., Jalalkamali, N., Hosseinipour, E., & Singh, V. (2013). Assessment of a conceptual hydrological model and artificial neural networks for daily outflows forecasting. *International journal of environmental science and technology, 10*(6), 1181-1192.

Rogers, L. L. (1992). Optimal groundwater remediation using artificial neural networks and the genetic algorithm.

Roth, V., & Lemann, T. (2016). Comparing CFSR and conventional weather data for discharge and soil loss modelling with SWAT in small catchments in the Ethiopian Highlands. *Hydrology and Earth System Sciences, 20*(2), 921-934.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., . . . Iredell, M. (2014). The NCEP climate forecast system version 2. *Journal of climate*, *27*(6), 2185-2208.

Sahoo, G., & Ray, C. (2006). Flow forecasting for a Hawaii stream using rating curves and neural networks. *Journal of Hydrology*, *317*(1-2), 63-80.

Sahoo, G., Ray, C., & De Carlo, E. (2006). Use of neural network to predict flash flood and attendant water qualities of a mountainous stream on Oahu, Hawaii. *Journal of Hydrology, 327*(3-4), 525-538.

Schwarz, G. E., & Alexander, R. (1995). *State soil geographic (STATSGO) data base for the conterminous United States* (2331-1258). Retrieved from

Shao, G., Zhang, D., Guan, Y., Xie, Y., & Huang, F. (2019). Application of SWAT model with a modified groundwater module to the semi-arid Hailiutu River Catchment, Northwest China. *Sustainability, 11*(7), 2031.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica sinica*, 221-242.

Sheather, S. (2009). *A modern approach to regression with R*: Springer Science & Business Media.

Singh, K. P., Basant, A., Malik, A., & Jain, G. (2009). Artificial neural network modeling of the river water quality—a case study. *Ecological Modelling, 220*(6), 888-895.

Spruill, C. A., Workman, S. R., & Taraba, J. L. (2000). Simulation of daily and monthly stream discharge from small watersheds using the SWAT model. *Transactions of the ASAE, 43*(6), 1431.

Srinivasan, R., & Arnold, J. G. (1994). INTEGRATION OF A BASIN-SCALE WATER QUALITY MODEL WITH GIS 1. *Journal of the American Water Resources Association, 30*(3), 453-462.

Srivastava, P., McNair, J. N., & Johnson, T. E. (2006). Comparison of process-based and artificial neural network approaches for streamflow modeling in an agricultural watershed. *Journal of the American Water Resources Association, 42*(3), 545-563.

Stampoulis, D., & Anagnostou, E. N. (2012). Evaluation of global satellite rainfall products over continental Europe. *Journal of hydrometeorology, 13*(2), 588-603.

Starrett, S. K., Starrett, S. K., Heier, T., Su, Y., Tuan, D., Bandurraga, M. J. A. J. o. E., & Sciences, A. (2010). Filling in missing peak flow data using artificial neural networks. *ARPN Journal of Engineering and Applied Sciences, 5*(1), 49-55.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting, 16*(4), 437-450.

Tobin, K. J., & Bennett, M. E. (2013). Temporal analysis of Soil and Water Assessment Tool (SWAT) performance based on remotely sensed precipitation products. *Hydrological Processes, 27*(4), 505-514.

Tuppad, P., Douglas-Mankin, K., Lee, T., Srinivasan, R., & Arnold, J. (2011). Soil and Water Assessment Tool (SWAT) hydrologic/water quality model: Extended capability and wider adoption. *Transactions of the ASABE, 54*(5), 1677-1684.

U.S. Geological Survey. (2016). National Water Information System data available on the World Wide Web (USGS Water Data for the Nation). Retrieved from https://waterdata.usgs.gov/nwis

USDA-NRCS. (2014). Geospatial data gateway.

Wang, Y., Shao, J., Su, C., Cui, Y., & Zhang, Q. (2019). The Application of Improved SWAT Model to Hydrological Cycle Study in Karst Area of South China. *Sustainability, 11*(18), 5024.

Worland, S. C., Steinschneider, S., Asquith, W., Knight, R., & Wieczorek, M. (2019). Prediction and Inference of Flow Duration Curves Using Multioutput Neural Networks. *Water Resources Research, 55*(8), 6850-6868.

Worqlul, A. W., Maathuis, B., Adem, A. A., Demissie, S. S., Langan, S., & Steenhuis, T. S. (2014). Comparison of rainfall estimations by TRMM 3B42, MPEG and CFSR with ground-observed data for the Lake Tana basin in Ethiopia. *Hydrology and Earth System Sciences, 18*(12), 4871-4881.

Wu, J. S., Han, J., Annambhotla, S., & Bryant, S. (2005). Artificial neural networks for forecasting watershed runoff and stream flows. *Journal of Hydrologic Engineering, 10*(3), 216-222.

Wurbs, R. A., & James, W. P. (2002). *Water resources engineering*. Upper Saddle River, United States: Prentice Hall.

Yang, L., Jin, S., Danielson, P., Homer, C., Gass, L., Bender, S. M., . . . Fry, J. (2018). A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies. *ISPRS journal of photogrammetry and remote sensing, 146*, 108-123.

Yaseen, Z. M., El-Shafie, A., Jaafar, O., Afan, H. A., & Sayl, K. N. (2015). Artificial intelligence based models for stream-flow forecasting: 2000–2015. *Journal of Hydrology, 530*, 829-844.

Yaseen, Z. M., Sulaiman, S. O., Deo, R. C., & Chau, K.-W. (2018). An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. *Journal of Hydrology*.

Yen, H., Daggupati, P., White, M. J., Srinivasan, R., Gossel, A., Wells, D., & Arnold, J. G. (2016). Application of large-scale, multi-resolution watershed modeling framework using the hydrologic and water quality system (HAWQS). *Water, 8*(4), 164.

Zakizadeh, H., Ahmadi, H., Zehtabian, G., Moeini, A., & Moghaddamnia, A. (2020). A novel study of SWAT and ANN models for runoff simulation with application on dataset of metrological stations. *Physics and Chemistry of the Earth, Parts A/B/C*, 102899.

Zhang, B., & Govindaraju, R. S. (2000). Prediction of watershed runoff using Bayesian concepts and modular neural networks. *Water Resources Research, 36*(3), 753-762.

Zhang, G. P., & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research, 160*(2), 501-514.

Zhang, Q., & Stanley, S. J. (1997). Forecasting raw-water quality parameters for the North Saskatchewan River by neural network modeling. *Water Research, 31*(9), 2340-2350.

Zhang, X., Srinivasan, R., & Liew, M. V. (2010). On the use of multi-algorithm, genetically adaptive multi-objective method for multi-site calibration of the SWAT model. *Hydrological processes, 24*(8), 955-969.

Zhang, X., Srinivasan, R., Zhao, K., & Liew, M. V. (2009). Evaluation of global optimization algorithms for parameter calibration of a computationally intensive hydrologic model. *Hydrological processes, 23*(3), 430-441.

Zhang, Y., & Wurbs, R. (2018). Long-term changes in river system hydrology in Texas. *Proceedings of the International Association of Hydrological Sciences, 379*, 255-261.

Zhao, G., Gao, H., & Cuo, L. (2016). Effects of urbanization and climate change on peak flows over the San Antonio River Basin, Texas. *Journal of hydrometeorology, 17*(9), 2371-2389.