FLOODPLAIN MAPPING IN DATA-SCARCE ENVIRONMENTS USING REGIONALIZATION TECHNIQUES

by

Keighobad Jafarzadegan

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Lyles School of Civil Engineering

West Lafayette, Indiana

May 2019

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Venkatesh Merwade, Chair
School of Civil Engineering
Dr. Rao Govindaraju
School of Civil Engineering
Dr. Dennis Lyn
School of Civil Engineering
Dr. David Johnson
School of Industrial Engineering

Approved by:

Dr. Dulcy Abraham
Head of the Graduate Program

To my Mother

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor, Dr. Venkatesh for his constant guidance and support throughout my graduate studies. I would also like to extend sincere gratitude to my PhD committee members, Dr. Rao Govindaraju, Dr. Dennis Lyn, and Dr. David Johnson, who have provided insights and helpful comments for improving the quality of my research. I specially thank the Purdue's Graduate School for awarding me the Ross fellowship, Dr. Jacques W. Delleur for the travel grant, and Lyles family for the Lyles Teaching Assistant support. I also like to thank all my colleagues and friends in hydro group: Zhu Liu, Siddharth Saksena, Sayan Dey, Jessica Eisma, Ganesh Mallya, David Canon, Nicholas Olsen, Abhinav Gupta, Abhishek Abhishek, Bruce Wang, Anzy Lee, Adnan Rajib, Liuying Du, Nikhil Sangwan, Kyungmin Sung, Becca Essig. I will never forget sharing my graduate life with all of you. Last but most crucially, I am truly grateful to my mother for her unlimited support during my PhD.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Author: Jafarzadegan, Keighobad. PhD
Institution: Purdue University
Degree Received: May 2019
Title: Floodplain Mapping in Data-Scarce Environments Using Regionalization Techniques
Committee Chair: Venkatesh Merwade

Flooding is one of the most devastating and frequently occurring natural phenomena in the world. Due to the adverse impacts of floods on the life and property of humans, it is crucial to investigate the best flood modeling approaches for delineation of floodplain areas. Conventionally, different hydrodynamic models are used to identify the floodplain areas. However, the high computational cost, and the dependency of these models on detailed input datasets limit their application for large scale floodplain mapping in data-scarce regions. Recently, a new floodplain mapping method based on a hydrogeomorphic feature, named Height Above Nearest Drainage (*HAND*), has been proposed as a successful alternative for fast and efficient floodplain mapping at the large scale. The overall goal of this study is to improve the performance of *HAND*-based method by overcoming its current limitations. The main focus will be on extending the application of the *HAND*-based method to data-scarce environments. To achieve this goal, regionalization techniques are integrated with the floodplain models at the regional and continental scales. Considering these facts, four research objective are established to (1) Develop a regression model to create 100-year floodplain maps at a regional scale (2) Develop a classification framework for creating 100-year floodplain maps for the Contiguous United States (3) Develop a new version of the *HAND*-based method for creating probabilistic 100-year floodplain maps, and (4) Propose a general regionalization framework for transferring information from data-rich basins to data-scarce environments.

In the first objective, the state of North Carolina is selected as the study area, and a regression model is developed to regionalize the available 100-year Flood Insurance Rate Maps (FIRMs) to the data-scarce regions. The regression model is an exponential equation with three independent variables including the average slope, the average elevation, and the main stream slope of the watershed. The results show that the estimated floodplains are within the expected range of

accuracy of C>0.6 and F>0.9 for majority of watersheds located in the mid-altitude regions, but it overpredicts and underpredicts in the flat and mountainous regions respectively.

The second objective of this research extends the spatial application of the *HAND*-based method to the entire United States by proposing a new classification framework. The proposed framework classifies the watersheds into three groups by using seven watershed characteristics related to the topography, climate and land use. The validation results show that the average error of floodplain maps is around 14% which demonstrate the reliability and robustness of the proposed framework for continental floodplain mapping. In addition to the acceptable accuracy, the proposed framework creates the floodplain maps for any watershed within the United States.

The *HAND*-based method is a deterministic modeling approach to floodplain mapping. In the third objective, the probabilistic version of this method is proposed. Using a probabilistic approach to floodplain mapping provides more informative maps. In this study, a flat watershed in the state of Kansas is selected as the case study, and the performance of four probabilistic functions for floodplain mapping is compared. The results show that a linear function with one parameter and a gamma function with two parameters are the best options for this study area. It is also shown that the proposed probabilistic approach can reduce the overpredictions and underpredictions made by the deterministic *HAND*-based approach.

In the fourth objective, a new regionalization framework for transferring the calibrated environmental models to data-scarce regions is proposed. This framework aims to improve the current similarity-based regionalization methods by reducing the subjectivity that exists in the selection of basin descriptors. Using this framework for the probabilistic *HAND*-based method in the third objective, the floodplains are regionalized for a large set of watersheds in the Central United States. The results show that “vertical component of centroid (or latitude)” is the dominant descriptor of spatial variabilities in the probabilistic floodplain maps. This is an interesting finding which shows how a systematic approach can help to explore the hidden descriptors for regionalization. It is demonstrated that using common methods, such as correlation coefficient calculation, or stepwise regression analysis, will not reveal the critical role of latitude on the spatial variability of floodplains.

CHAPTER 1. INTRODUCTION

1.1 Background and Motivation

Floods are the most frequent natural disasters in the world, leading to huge costs and damages annually. The estimated damages from this phenomenon between 2004 and 2013 exceeded 300 billion US dollars (Guha-Sapir et al., 2015), which places it next to earthquakes and windstorms as the three costliest catastrophes of the world. Figure 1-1 shows the frequency of different natural disasters[1] in different continents (University of Louvain Belgium 2014).

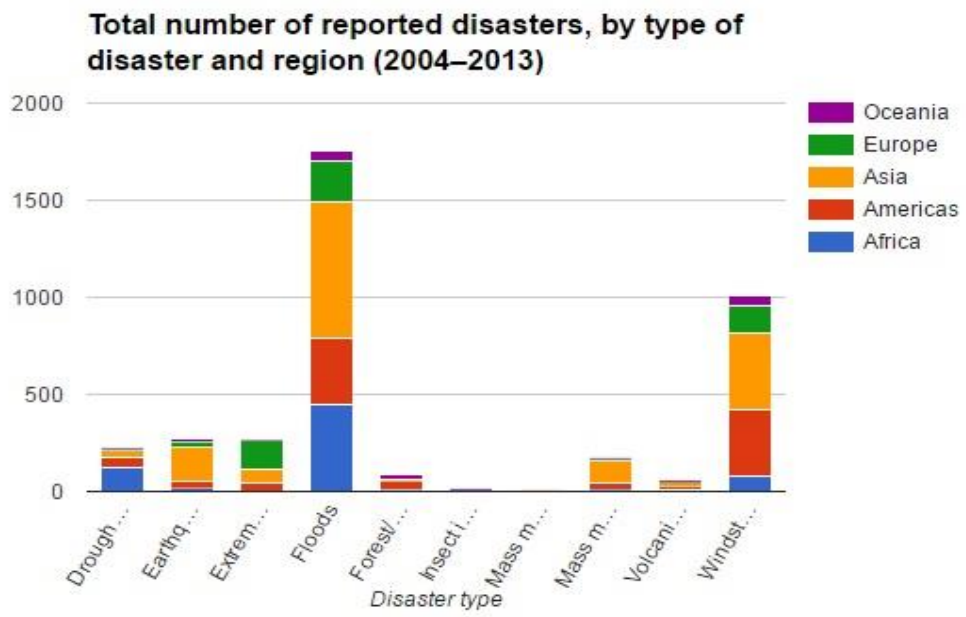


Figure 1-1World Disasters Report 2014 (Source: University of Louvain Belgium)

Considering the disastrous impacts of floods on human lives and property, there is a growing interest to perform flood risk management projects for individual streams as well as for entire stream networks in a small or large basin (Moel et al. 2009; Van Alphen et al. 2009). Land-use planners, flood risk managers, emergency response teams, utility companies, insurance companies and citizens have different stakes and objectives in a flood risk management project. However, one of the key steps in any flood risk management project is the identification of the floodplains.

[1] A flooding event is qualified as disaster if more than 10 people are killed, more than 100 people are affected, state of emergency is declared, or international assistance is requested (Guha-Sapir et al., 2015)

Delineation of floodplains is also vital for many ecological and environmental studies. Flooding plays a vital role in the growth and reproduction of the regional aquatic plants and animals (Walker et al., 1997). It keeps the lateral connection between the river and the floodplain and promotes the transport of nutrients, biota and organic carbon to the floodplains (Walling and He, 1998; Baldwin and Mitchell, 2000; Thoms and Sheldon, 2000; Thoms, 2003). The crucial ecological role of a floodplain as a productive environment is another reason for the increasing attention about the proper delineation of these areas in the last decades.

The United States Federal Emergency Management Agency (FEMA) has invested billions of dollars to create flood insurance rate maps (FIRMs) for the entire country (FEMA 2009). A similarly determined effort of flood inundation mapping exists in Europe where Directive 2007/60/EC required all member states to generate these maps. (Moel et al. 2009; Van Alphen et al. 2009). The majority of floodplain maps provided by FEMA correspond to the 100-year return period flood. The return period of a flood event, sometimes defined as the recurrence interval, is the inverse of the probability that a given flood event is equaled or exceeded in a year. For example, a 100-year flood is a flood event having an occurrence probability of 1% per year. The significant role of 100-year floodplain mapping as a primary step of any flood risk management problem is the major motivation for this study, which proposes a new alternative approach to 100-year floodplain mapping.

FEMA maps are the most reliable resources for obtaining the freely available 100-year floodplain maps across the United States. A huge investment has been made by FEMA to create accurate 100-year floodplain maps for around half of the US Rivers over the last decades. After selecting the study area, the FEMA floodplain mapping process is started by gathering information about hydrology, hydraulics, infrastructures, land use and existing flood maps. Then, statistical analysis and hydrologic models are used to estimate the 100-year steady and unsteady flow. The estimated flow is fed into a hydraulic model to create 100-year floodplain maps. Considering the broad availability and the reliability of these maps as a valuable source of information in the US, this research uses these maps as the primary input for development and validation of all the models proposed in the next chapters. However, FEMA maps are not the perfect reference maps that reflect the actual floodplain areas corresponding to a 100-year flood event. The uncertainties that exit in

the floodplain modeling process, such as the uncertainty in the model structure and the input data are one of the major drawbacks of using these maps as a reference. In addition, the effective FIRMs are sometimes adjusted from the original engineering assessment of 100-year depths through a political process of community-based appeals. This reduces the FEMA floodplain extents, specifically in the urban areas. Moreover, the models developed in this research ignore the impact of riverine structures (e.g. dams, levees and bridges) on the floodplains while FEMA considers them during the modeling process. Considering all these limitations, it is not expected to have a perfect overlap between the maps generated by the developed models and the FEMA maps. Specifically, it is highly recommended to exclude the urban watersheds in this research because of the differences that exist between the approach used by FEMA and the proposed models in these watersheds.

Conventionally, floodplains are delineated by using hydrodynamic models. These models use the basic physical laws in fluid mechanics, namely conservation of mass and momentum, to simulate the dynamics happening around a river in an extreme flood event. A vast literature exists on improving the performance of these hydrodynamic models for flood inundation mapping (Teng et al., 2017). Currently, 1D and 2D models can simulate the major components of a complex environment for estimating the floodplains with acceptable accuracy.

The main drawback of the hydrodynamic models is their inefficiency for delineating floodplains in large-scale problems and data-scarce regions. In data-scarce regions, the hydrologic and hydraulic models cannot be calibrated due to the lack of streamflow gauges for measuring the actual flow and water depth. In addition, the limited access to detailed information about the geometry of rivers reduces the accuracy of results generated by the hydrodynamic models in these regions. This issue creates a new research opportunity for developing another generation of floodplain mapping models where the focus is changed from accuracy to efficiency of inundation mapping.

We define the efficient floodplain mapping models as a series of methods that consist of the following four attributes:

Computational efficiency: The floodplain mapping models should be fast. This includes both modeling setup and running time. Currently, it takes days to setup a hydrodynamic model and additional hours to run these models for a few kilometers-long rivers in a watershed.

Cost effectiveness: The cost of these models should be reasonable, because floodplain mapping projects have limited budgets. The hydrodynamic models need high resolution data, such as bathymetry and topography, for accurate inundation mapping; collecting these datasets is expensive.

Transferability: The floodplain mapping models should be applicable to data-scarce regions and create floodplain maps at a large-scale. Because of the high computational cost and the long running time of hydrodynamic models, these models are not always suitable for simulating dense stream networks with more than one thousand kilometers of rivers. Moreover, the lack of detailed data (e.g. river geometry as well as streamflow and water depth for the model calibration) in data-scarce regions leads to a significant drop in the accuracy of floodplain maps generated by these models.

Accuracy: The floodplain mapping models should be accurate. This means the error of modeling, typically estimated from comparison of predicted maps with some reliable reference maps, should be negligible. This feature is the main advantage of complex hydrodynamic models. Usually there is trade-off between accuracy and the three aforementioned attributes. Therefore, the proposed modelling approach should maintain a balance among all four attributes.

Considering these four attributes as the basis of an efficient floodplain mapping approach, several methods have been proposed for preliminary delineation of floodplain areas using freely-available Digital Elevation Models (DEM) (Clubb et al., 2017; Dodov and Foufoula-Georgiou, 2005; Gallant and Dowling, 2003; Lhomme et al., 2008; McGlynn and Seibert, 2003; Nardi et al., 2013, 2006; Papaioannou et al., 2015; Teng et al., 2015; Williams et al., 2000). Currently, the research related to floodplain mapping using DEM-based methods is primarily focused on the first two attributes of efficiency. Due to the large number of simplifications and assumptions made for reducing the computational time and cost of these methods, the accuracy of these methods drops

significantly. Recently a series of DEM-based methods have been developed where a watershed is classified into flooded and non-flooded areas using a morphologic index. Among several indices, a hydrogeomorphic feature, named Height Above Nearest Drainage (*HAND)* has been shown as one of the best features for mapping floodplains. The acceptable accuracy of the *HAND*-based method, as well as its fast and cost-effective structure make this method an attractive option for efficient floodplain mapping. The main limitation of this method is its dependency on some reference floodplain maps, which limits its application to floodplain mapping in data-scarce regions.

Prediction in data-scarce environments is generally challenging due to the absence of sufficient input data for the modeling and the lack of reliable reference data for the model calibration in these regions. Regionalization techniques are a set of methods used commonly in the field of hydrology for transferring information from data-rich basins to data-scarce basins. There is a rich literature on the application of these techniques for streamflow prediction in ungauged basins. (Hrachowitz et al., 2013; Kay et al., 2007; Kim and Kaluarachchi, 2008; McIntyre et al., 2005; Merz and Blöschl, 2004; Parajka et al., 2005; Reed et al., 1999; Sefton and Howarth, 1998; Sivapalan, 2003; Vandewiele and Elias, 1995; Viviroli et al., 2009). However, regionalization is completely new in the field of floodplain mapping where there is a strong potential to integrate regionalization techniques with floodplain models and create floodplain maps in data-scarce environments. In this regard, there are important research questions such as whether the new generation of DEM-based methods can create floodplain maps in data-scarce regions within the expected range of accuracy, and how to improve the performance of the proposed DEM-based methods in these regions. Overall, the research on floodplain mapping using alternative DEM-based methods, which covers the four attributes of efficient modeling, is still young compared to advanced hydrodynamic models used for the accurate mapping of inundation areas. This implies a substantial need for developing novel approaches for efficient floodplain mapping.

1.2 Research Objectives

The overall goal of this research is to improve the performance of the *HAND*-based method and overcome its limitations for efficient floodplain mapping. The major focus will be on extending the application of the *HAND*-based method to data-scarce regions for large-scale floodplain

mapping. To achieve this goal, regionalization techniques are integrated with floodplain models at the state and continental scales. In addition, a new version of the *HAND*-based method for probabilistic floodplain mapping is developed and regionalized within a large landscape in the Central United States. It should be noted that, all the probabilistic floodplain maps created in this dissertation corresponds to a 100-year flood event. This means that when referring to the probability of inundation by a 1% Annual Exceedance Probability (AEP) event, we mean that there is a certain degree of confidence that the depths with 1% AEP are non-zero. Specifically, four research objectives are studied in this dissertation as follows:

1. Develop a regression model to create 100-year floodplain maps at a regional scale: This study focuses on regression-based regionalization techniques where the most significant watershed characteristics for transferring the available FEMA FIRMs to data-scarce watersheds are determined.

2. Develop a classification framework for creating 100-year floodplain maps for the Contiguous United States (CONUS): In this framework, the watersheds are classified based on their topographic, climatic and land use characteristics. Then, a probabilistic binary classifier uses the classification results and *HAND* as input to create the floodplain maps for any watershed within the United States. This research objective extends the spatial scale of the problem to the continental scale and proposes a novel classification framework to regionalize available FEMA FIRMs to all watersheds in the United States. The fast, cost-effective, acceptable accuracy and the broad application of this framework for floodplain mapping in any watershed across the United States, is one of the significant research accomplishments in this dissertation which aligns with the goal of efficient floodplain mapping.

3. Develop a new version of the *HAND*-based method for creating probabilistic 100-year floodplain maps: This objective focuses on rearranging the formulation of the traditional *HAND*-based method so that the threshold of *HAND*, selected for floodplain mapping in deterministic approach, is considered as a random variable with a probability density function in the probabilistic approach. The new model is able to create a grid of floodplain maps where each cell represents the probability of inundation by a 100-year flood event. Using the

probabilistic *HAND*-based method, this research objective compares the performance of four potential probabilistic functions for floodplain mapping in a flat watershed in the state of Kansas. The accuracy of the proposed method is evaluated by comparing the floodplain maps created by the proposed method with the FEMA FIRMS and the maps produced by the deterministic version of the *HAND*-based method.

4. Propose a general regionalization framework for transferring information from data-rich basins to data-scarce environments: Considering that high subjectivity exists for the selection of basin descriptors in a regionalization problem, this objective proposes a systematic approach where the most significant physical/climatic basin descriptors for regionalization of the basins are determined. The effectiveness of this framework is tested for the probabilistic *HAND*-based method developed in the third objective. The proposed framework uses the available FEMA FIRMs in the Arkansas-White-Red region in the U.S. to create probabilistic floodplain maps for all basins in this region.

1.3 Organization of this dissertation

This dissertation consists of six chapters. Chapters 2-5 describe the four objectives conducted during the PhD research. These chapters are presented in a self-contained manner, i.e., each chapter has an abstract, introduction, description of study area and data, methods, results, and conclusion sections. However, all four of these chapters are linked under the umbrella of the *HAND*-based method improvements for efficient floodplain mapping in data-scarce regions. In Chapter 6, the practical and theoretical contribution of this research in hydrology and the flood modeling community is discussed, and the primary findings are synthesized.

CHAPTER 2. A DEM-BASED APPROACH FOR LARGE-SCALE FLOODPLAIN MAPPING IN UNGAUGED WATERSHEDS

2.1 Abstract

Binary threshold classifiers are a simple form of supervised classification methods that can be used in floodplain mapping. In these methods, a given watershed is examined as a grid of multiple cells where each cell has a particular morphologic value. A reference map is a grid that all cells have already labeled as flood and non-flood from a precise hydraulic modeling or a remote sensing observation. By using the reference map, a threshold on morphologic feature is determined to label the unknown cells as flood and non-flood (binary classification). The main disadvantage of these methods is that a reference inundation map is required to train the classifier and find the threshold. These reference maps are not available in many regions including ungauged watersheds. In this chapter, regression modeling is used to predict the threshold by relating it to the watershed characteristics. Application of this approach for North Carolina shows that the threshold is related to main stream slope, average watershed elevation, and average watershed slope. By using the Fitness (F) and Correct (C) criteria of $C > 0.9$ and $F > 0.6$, results show the thereshold prediction and the corresponding floodplain for 100-year design flow are comparable to that from Federal Emergency Management Agency's (FEMA) Flood Insurance Rate Maps (FIRMs) in the region. However, the floodplains from the proposed model are underpredicted and overpredicted in the flat (average watershed slope < 1%) and mountainous regions (average watershed slope > 20%). Overall, the proposed approach provides an alternative way of mapping floodplain in data-scarce regions.

2.2 Introduction

Floodplain mapping is one of the required steps in the assessment process of flood risk management. Considering the disastrous impacts of floods on human lives and property, the United States Federal Emergency Management Agency (FEMA) has invested billions of dollars to create flood insurance rate maps (FIRMs) for the entire country (FEMA 2009). FIRMs provide inundation extent that corresponds to 100-year return period flood. A similarly determined effort of floodplain mapping exists in Europe where Directive 2007/60/EC required all member states to

generate these maps. (Moel et al. 2009; Van Alphen et al. 2009; "EXCIMAP, 2007). The conventional floodplain mapping approach involves both hydrologic and hydraulic modeling. A hydrologic model is used to generate design flow corresponding to a specific return period, which is generally 100-year. In gauged locations, flood frequency analysis can be performed using historical data to determine the design flow corresponding to a given return period. Once the design flow is known, it is fed to a 1D or 2D hydraulic model to generate water surface elevations and inundation extent for a river reach (Cobby et al. 2003; Hunter et al. 2007; Tayefi et al. 2007; Cook and Merwade 2009; Bates et al. 2010; Neal et al. 2012; Cantisani et al. 2014).

For ungauged sites, however, there are several arguments regarding the accuracy of the estimated design flow based on hydrologic modeling. In these problems, a Synthetic Unit hydrograph (SUH) related to a particular return period is created based on different techniques. Singh et al. (2014) categorized the available SUH models into four groups including traditional, conceptual, probabilistic and geomorphological. They reviewed the popular methods for each group and concluded that geomorphological models are the most useful approach for prediction in ungauged basins (Grimaldi et al. 2010; Grimaldi et al. 2012; Petroselli and Grimaldi 2015; Grimaldi and Petroselli 2015; Rigon et al. 2016). The uncertainties associated to SUH estimation, which is the main input of a hydraulic model, is a critical issue for flood mapping in ungauged basins. In order to overcome this issue, Grimaldi et al. (2013) proposed a fully continuous hydrologic–hydraulic modeling framework for flood mapping. In this method, instead of SUH estimation, a discharge time series is directly fed to a hydraulic model and the frequency analysis of the inundation area corresponding to a particular return period is implemented in the final step on the generated flood maps. Another fast and simple alternative approach for estimation of peak discharge in ungauged sites is the use of regression equations that relate streamflow statistics to watershed characteristics. For example, the StreamStats program developed by the United States Geological Survey (USGS) uses regionalized regression equations to estimate peak discharge at any location along a stream for a given return period (U.S. Geological Survey 2012).

The conventional hydrologic and hydraulic modeling approach requires resources to collect or gather the required data and run the models after proper calibration and validation. Some of the key data include digital elevation model (DEM), land use, soil, hydrologic data, river bathymetry,

and details of structures such as bridges and culverts along the reach. This approach is generally adopted for creating a flood map for individual river segments where such data either exist or can be acquired using available resources. In data-scarce regions, flood maps created through modeling can have high uncertainty (Merwade et al. 2008). The data and computational requirements increase significantly when flood maps for tens or hundreds of reaches need to be created for a region, thus making the conventional modeling approach unfeasible for large data-scarce regions. Absence of good datasets and computational resources has led to the development of alternative methods that process easily available public domain datasets over larger areas to create floodplain maps.

The free and widespread access to high resolution DEM for the entire globe (30 m or 90 m) in the recent years, has led to the generation of new geomorphologic Digital Terrain Model (DTM) floodplain delineation methods. The essence of these methods lies in the distinguishable geomorphic and hydrologic properties of floodplain from the neighboring hillslopes. Floodplain is the "concave depositional frequently saturated predominantly flat area" (Nardi et al. 2013) surrounding the streams. Therefore, the geomorphologic floodplain delineation methods make a preliminary estimation of potential flooding areas without considering the flood magnitudes. This is one of major differences of these methods with the conventional hydraulic modeling approaches. Although some recent geomorphic DTM-based methods are able to generate floodplain corresponding to a particular flood frequency, hydraulic models can create dynamic maps with varied inundation depth, which are event-based and are highly correlated to the flood magnitude. In one of the first geomorphic floodplain delineation studies conducted by Williams et al. (2000), the floodplain was estimated by comparing DEM and a constant water surface level for the entire drainage network. McGlynn and Seibert (2003) used a DTM-based algorithm and regional regression analysis to find the contribution of riparian area for stream networks (McGlynn and McDonnell 2003). In another study, Dodov and Foufoula-Georgiou (2006) proposed a fast algorithm based on regional geomorphologic analysis to estimate the floodplain morphometry. Nardi et al. (2006; 2013) used a hydrogeomorphic approach that obtains the flow discharge and depth at each stream node by using the flow at the watershed outlet in conjunction with a scaling relationship based on the Geomorphologic Instantaneous Unit Hydrograph (Rodríguez-Iturbe et al. 1979; Rodriguez-Iturbe 1993). Papaioannou et al. (2015) proposed a multi-criteria-analysis

framework incorporating geographic information systems (GIS), fuzzy logic and clustering techniques to map floodplain areas at the catchment scale.

Recently some new alternative methods based on supervised classification techiniques have been used for floodplain mapping. In these methods, parameters of classification are recognized by training the watershed on an available reference flood map. The trained model will be used to classify the watershed into flood and non-flood areas. De Risi et al. (2014) used topographic wetness index, derived from a DEM, in conjunction with a Bayesian updating framework to identify floodplains. Manfreda et al. (2008, 2011) used a binary threshold method in a supervised classification technique to identify flood and non-flood areas by using DEM based modified topographic index (TIm) as the classifier. Degiorgis et al. 2012, investigated the performance of binary threshold methods by creating several classifiers based on a single morphologic feature, including the distance from a DEM cell to the nearest stream (D), difference of elevation between a given cell and closest stream (H), surface curvature (ΔH), contributing area (A) and local slope (S). They demonstrated that the topographic feature, H, defined as the difference in elevation between a given cell and the nearest stream is the most significant morphologic feature for floodplain mapping using binary classifiers. Further studies on performance of single or a combination of multiple morphologic features also proved the effectiveness and applicability of feature H for flood mapping in supervised binary classification methods (Manfreda et al. 2014; Manfreda et al. 2015; Samela et al. 2016). It should be noted that Feature H firstly defined as an effective hydrologic descriptor by Rennó et al. (2008) and its application in the prediction of hydrologically relevant soil environments was investigated (Nobre et al. 2011).

Despite the advantages of the proposed geomorphic DTM-based methods for simple and preliminary large-scale flood mapping, their applicability and effectiveness are still controversial for data-scarce regions. For example, the supervised classification methods are all dependent on a reference map for training but these maps area not available in many regions. Moreover, the methods based on regional regressions analysis, which relate the floodplain geometry to contributing area, require large survey datasets, which are not available for many rivers. In one study Sangwan and Merwade (2015) used a simple GIS-based attribute query on the SSURGO soil database in the U.S. to map floodplains in Indiana, which was then expanded for the entire

U.S. (Merwade et al. 2015). Although this work and some other studies such as clustering methods and older low-valley detection approaches can be applied for any ungauged watershed, there are many assumptions and high uncertainties in the structure of such methods. Furthermore, they are not able to account for floodplain related to a particular flood frequency, which limits their applications for flood risk management purposes.

As mentioned before, one of the main drawbacks of supervised classification methods is that they require data for training the algorithm. In the case of flood mapping, it means that a portion of the watershed should have either some reference flood maps created by using models or observed historical flood extents as well as observed flood marks or reference maximum levels for floodplain modeling calibration and validation. Consequently, these methods are not entirely independent of hydrologic and hydraulic information, and they cannot be used for ungauged watersheds. The overall goal of this study is to overcome this limitation by proposing a new approach that can create floodplain maps for any gauged and ungauged watersheds without hydraulic/hydrologic data collection or modeling. Although much literature exists relating peak discharge with watershed characteristics such as morphometric, geomorphic and climatic features, the lack of regression models for direct prediction of floodplains from significant features of a given watershed is addressed in this study. Similarly, many studies have used the morphologic feature, H, to delineate floodplains, but this study relates this H with most significant watershed characteristics to develop a regression model with the broader goal of generalizing this regression based approach for application in ungaguged basins with different geophysical settings. As a first step towards this broader goal, this study develops the regression equation and tests its robustness and accuracy in mapping the 100-year floodplain in various geographic regions of North Carolina, USA. In general, this chapter introduces a fast, cost-effective and automated method for large-scale floodplain mapping in ungauged watersheds. The simplicity of regression models for direct prediction of floodplains, and the feasibility of this method for floodplain mapping in data-scarce regions are the factors that can be beneficial for decision makers and flood risk management agents.

This chapter is organized as follows: The study area and dataset is presented in section 2. In section 3, the proposed methodology is explained. First the TrH model is described, then details of regression model development and model validation are presented, respectively. In section 4, the

developed regression model with selected watershed characteristics for North Carolina is presented. Furthermore, the results including regression model predictions and the generated flood maps are validated by FEMA FIRMs and the outcomes are illustrated. In section 5, the limitations of the proposed approach are discussed and some solutions and alternativs are suggested for future studies. Finally, section 6 summarizes the proposed methodology by introducing an operative strategy for 100-year flood hazard mapping in ungauged watersheds. In addition, the main findings and conclusions are presented.

2.3 Study Area and Data

North Carolina is selected as a test bed to develop and test the proposed model due to data availability, physiographic diversity, and history of large flood events in the state in the 1990s, 2004 and 2015. In regards to data, FEMA FIRMs are available for all main river reaches in North Carolina, thus providing a rich resource for reference flood maps. Topographically, North Carolina is divided into three major regions, including the Atlantic Flat Plain, the Piedmont Plateau, and the Appalachian Mountains, thus covering a wide range of elevation from flat regions in the east to mountains in the west.

In this study, floodplain maps are created for HUC12 units (https://water.usgs.gov/GIS/huc.html), which are the smallest geographic units within the Watershed Boundary Dataset (WBD,"U.S. Geological Survey - National Hydrography Dataset" 2016). In order to create the model 185 HUC12 units were selected, and the performance of the model is validated on 105 additional HUC12 units (Figure 2-1). The GIS data used in this study include DEM, stream network, Flood Insurance Rate Maps (FIRMs), climate rasters and land use. One arc second (30 m) horizontal resolution DEMs are obtained from the United State Geological Survey's (USGS) National Elevation Dataset. The stream network for each HUC12 unit is derived from USGS's National Hydrography Dataset (NHD). Due to the large number of HUCs, both datasets are directly downloaded from a FTP site by writing a custom python script (ftp://rockyftp.cr.usgs.gov/vdelivery/Datasets/Staged/). FIRMs for all HUCs are obtained from FEMA service center. These maps are shape files with polygon feature classes that show the 100-year floodplain. Considering the key role of precipitation and temperature for simulating the streamflow in hydrologic models, and the importance of maximum precipitation in creating the

flood events, three climate datasets including average annual precipitation, average precipitation in the wettest month, and average annual temperature are obtained as raster files with 30 seconds resolution (around 1 km^2) for the entire state. The National Land Cover Database (NLCD) 2011 for the entire state is obtained from http://viewer.nationalmap.gov/basic/ to calculate the percentage of urban, water and forest areas in each watershed.

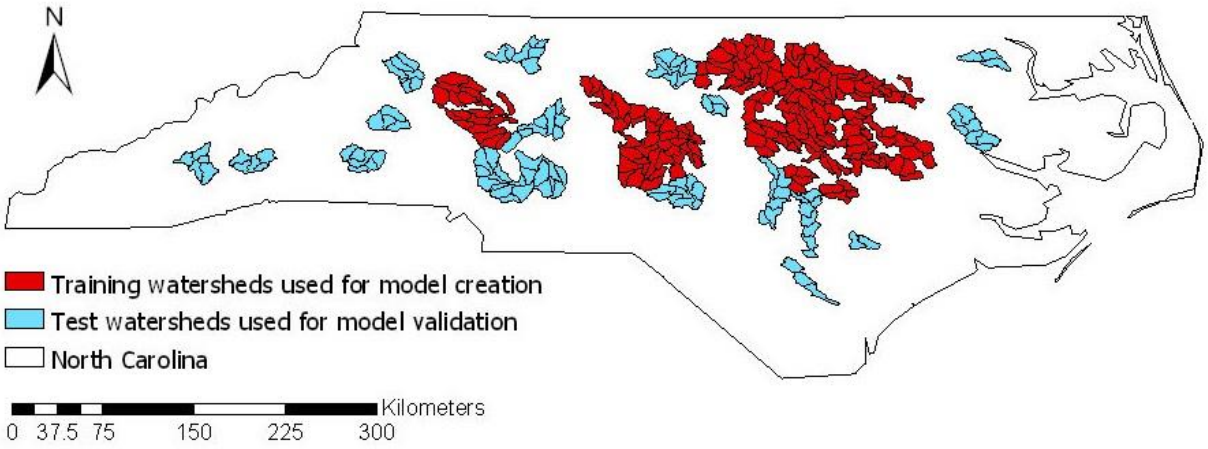


Figure 2-1 Geographical location of training and test watersheds used for the regression model creation and model validation respectively

2.4 Methodology

2.4.1 TrH Model Development

Finding the threshold of the morphologic feature H (TrH) forms the basis of the overall approach presented in this chapter. The TrH model needs a DEM, a stream network, and reference floodplain maps corresponding to a specific return period, which is 100-year in this study. This input/output structure is scripted as a core function and is run multiple times for the entire training watershed to get a TrH range for the watershed. In this study, FEMA maps are used as reference maps due to their easy availability for the study area. The stream network can be created from a DEM by using many terrain processing tools such as ArcGIS hydrology toolbox, ArcHydro or TAUDEM tools, among others. The resolution of the stream network created is dictated by the critical source area (CSA) threshold used in extracting the stream cells from the flow accumulation grid. In this study, however, the stream network is not generated through terrain processing. Instead, NHDPlus stream network is converted to a raster grid to create a stream network raster. While we understand that a processed stream network such as NHD may not be available easily outside the U.S., we chose

this route to avoid proper stream network generation issues in coastal flat areas, which is not the major focus of this analysis. However, it is important to note that generating a stream network in flat areas can be challenging using standard terrain processing tools. Next, the DEM is processed by using the following three steps: (i) fill the sinks; (ii) compute the flow direction grid; (iii) compute H grid. Figure 2-2 shows a hypothetical DEM where only two cells (1,3) and (5,6) are draining to a stream, and their H values are 5.5 and 3, respectively. Once the H grid is obtained, it is compared with a reference floodplain grid (FEMA polygon map converted to a raster). As with all terrain based processes, the H grid and the corresponding TrH is affected by the resolution of the DEM and its vertical accuracy (Gesch et al. 2002; Sanders 2007). An accurate higher resolution DEM exists for North Carolina, but the use of relatively less accurate 30m resolution DEM will make the finding from this study comparable with other studies in the literature.

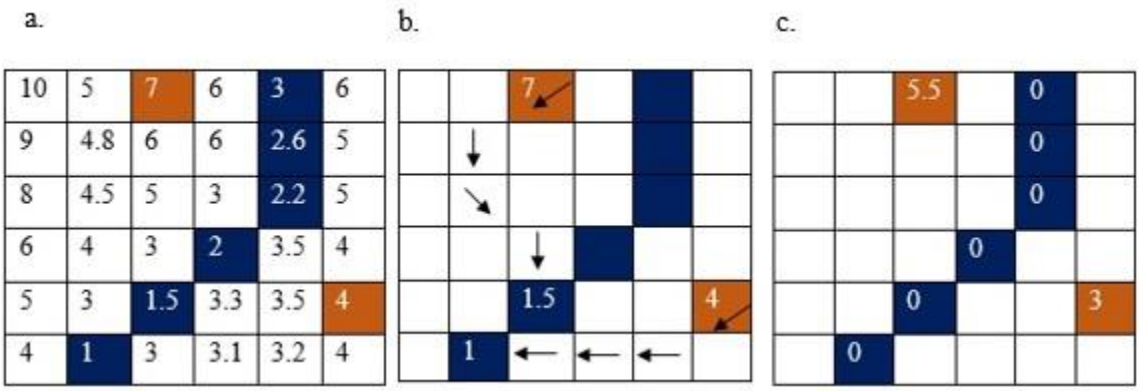


Figure 2-2 Feature H calculation process: Hypothetical DEM, Stream cells and two cells chosen for H calculation (a), flow direction and connection points on stream (b) and raster H (c)

The general technique for finding this threshold has root in the supervised learning techniques commonly used in remote sensing image classification problems (Nair and Bindhu 2016). In these methods, each image pixel (cell in this study) is defined as a pair consisting of feature input and target class. A reference map with an available set of target classes is used to train the model and calibrate the classifier parameters. The calibrated model will be used to find the target class of unknown cells. In this study, the feature input is raster H, the target classes are flood and non-flood labels (binary classification) and the reference map is the FEMA map. While the FEMA maps are generated through detailed hydrologic and hydraulic analysis by incorporating the influence of bridge, culvert and other in-stream structures, we do not expect the calibrated model to create 100

percent overlap with the reference map because the comparison is done over a number of reaches compared to a single reach.

Among many algorithms for supervised learning, finding a threshold on input feature H (TrH) for binary classification of unknown map into the flood and non-flood area has been of interest recently. In a binary classification using raster (Degiorgis et al. 2013), a positive instance can be labeled as true positive (tp) or false positive (fp) depending on whether the classified cell truly matches with the reference map cell or not. Similarly, a negative instance could be classified true negative (tn) or false negative (fn) as shown in Figure 2-3. Based on true/false positive/negative cells, the following equations can be defined:

$$rtp = \frac{True\ positive\ instances}{Total\ positives} \quad (2\text{-}1)$$

$$rfp = \frac{False\ positive\ instances}{Total\ negatives} \quad (2\text{-}2)$$

$$rtp + rfn = 1 \quad (2\text{-}3)$$

$$rfp + rtn = 1 \quad (2\text{-}4)$$

$$error = rfp + (1 - rtp) \quad (2\text{-}5)$$

where rtp, rfp, rfn and rtn denote the rate of true positive, false positive, false negative and true negative, respectively. Using Equation 2-5, the total error of a binary classification, the error between the classified raster and the reference FIRM, can be computed. Ideally, one would estimate a single TrH for a given stream network that gives the lowest error between the classified raster and reference map. Minimizing the total error of classification by finding the optimized threshold has been the typical way for flood mapping in the recent studies (Manfreda et al. 2011; Degiorgis et al. 2012). However, FEMA reference maps are not observed inundation maps and they can be deceiving due to the uncertain data and subjective models used in creating them. In addition, these flood maps are developed by using the conventional hydrologic and hydraulic modeling approach. In the conventional hydrologic-hydraulic modeling framework, the physics of a real flood event along with details of river geometry, impacts of structures such as dams and bridges and the urbanization effects are simulated. However, the proposed approach, which does not account for these small-scale details, is useful for large-scale application for preliminary estimation of floodplains. Taking all these considerations into account, relying on minimum error between two maps generated from completely different approaches is not reasonable. Therefore,

a new method for TrH estimation is proposed in which instead of using Equation 2-5 and selecting one optimized TrH for each watershed, a range of TrH that gives reasonable inundation maps is determined by using Equation 2-6.

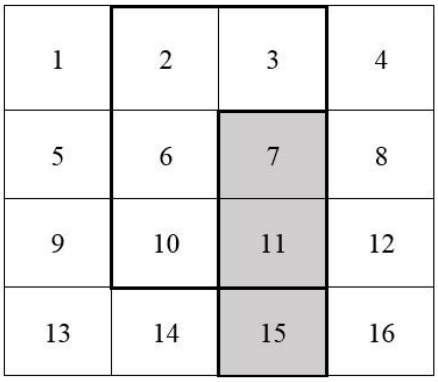
$$TrH_i \in TrH_{range} \ if \ C_{TrH_i} \geq \alpha and \ F_{TrH_i} \geq \beta \quad (2\text{-}6)$$

TrH_{range} is an interval of TrH values where any threshold inside this interval can generate an acceptable floodplain map. The range of TrH is determined based on the overlap between the classified map and the FEMA reference map as determined by two indices, namely the Correct (C) and Fit (F) index, as given by Equations 2-7 and 2-8 (Bates and De Roo 2000, Horritt and Bates 2002, Tayefi et al. 2007, Alfieri et al. 2014,Sangwan and Merwade 2015).

$$C = \frac{True\ positive\ instances}{Total\ positives} = \frac{flood\ cells\ \ predicted\ correctly}{flood\ cells} \quad (2\text{-}7)$$

$$F = \frac{True\ positive\ instances}{Total\ positives + False\ positives} = \frac{flood\ cells\ \ predicted\ correctly}{flood\ cells + nonflood\ cells\ predicted\ as\ flood} \quad (2\text{-}8)$$

The above equations show that the C index is the same as true positive (rtp) defined using Equation 2-1. This term is useful when the model is underpredicting. However, it cannot quantify the weakness of models in overprediction. On the other hand, F index considers both underprediction and overprediction together. For example, assume a biased model predicting the entire watershed as flood area. The C index of such model would be 1 because it predicts all the flood area correctly. However, the F index would be a small number because the term in the denominator, non-flooded cells predicted as flooded, is also a large value, which causes reduction in F index (Figure 2-3). A predicted flood map is considered acceptable if it can give C > α and F > β in relation to a reference map. The value of α and β can be determined based on the scale and expected accuracy of the problem. In this study α=0.9 and β=0.6 are chosen (Equation 2-6) which is a fairly high expectation for model accuracy. The C and F indices calculated for checking the performance of hydrodynamic flood inundation models in other studies vary from 0.6 to 0.95 for C, and 0.6 to 0.8 for F (e.g. Alfieri et al. 2014b; Bates and De Roo 2000). It should be noted that most studies in the literature calculate C and F for evaluating flood maps at reach scale. Additionally, the reference FIRMs used in this study are developed through detailed hydrologic and hydraulic modeling that include the influence of bridge, culvert and other structures on the flood inundation. Considering all these factors, the criteria of α = 0.9 and β = 0.6 seems quite stringent for comparing TrH based floodplain maps with the reference maps.



$Observed\ flood = \{7,11,15\}$
$Predicted\ flood = \{2,3,6,7,10,11\}$
$True\ positive = \{7,11\}$
$False\ positive = \{2,3,6,10\}$
$True\ negative = \{1,4,5,8,9,12,13,14,16\}$
$False\ negative = \{15\}$
$rtp = \frac{2}{3}$
$rfp = \frac{4}{13}$
$C = \frac{2}{3}$
$F = \frac{2}{3+4} = \frac{2}{7}$

Figure 2-3 A simple example for understanding the binary classification terms including rate of true positive (rtp) and rate of false positive (rfp) as well as two common indices, Correct (C) and Fit (F), used to validate flood mapping problems.

2.4.2 Regression Model Development

Once a TrH range is established, the next step is to relate this range with watershed characteristics. The watersheds in the study area are divided into training and test groups. The regression equation is developed based on information from training watersheds. Assume (X, Y) are the variables of the regression equation ($Y = f(x)$). If n and m are the total number of training watersheds and significant watershed characteristics respectively, X is the predictor array with n rows and m columns which contains watershed characteristics and Y is a one dimensional response variable array namely TrH_{range} with n rows. In order to find the response array, the TrH model is run for all the training watersheds. For each watershed, 19 different characteristics, presented in Table 2-1, associated with topography, shape, climate, and land use are calculated. A python script using ArcPy module is created to determine all these features simultaneously using the NHD, DEM,

climate, and land use data. Among the characteristics listed in Table 2-1, the most significant ones related to TrH are determined through the exploratory regression function in ArcGIS. This function, tests different combinations of watershed characteristics as independent variables against the mean of the TrH range determined from TrH model results, finds the highest adjusted R^2 associated with the best combinations and finally, summarizes the variable significance (Table 2-2). It should be noted that the highest adjusted R^2 values using four and five variables are 0.43 and 0.45, respectively. The negligible improvement in the value of adjusted R^2 for using more than three variables, and the tendency to have a simpler model that avoids overfitting are the reasons to select three features as the maximum number of model variables. In addition, the low value of adjusted R^2 (around 0.4 for the best combination of features in Table 2-2) is not critical at this preliminary stage because the exploratory regressionis function uses only the mean TrH value from a wider range of TrH values.

After recognizing the significant features, array X is made, which along with array Y (created from n times running of TrH model) are the main inputs of regression analysis. Four regression models, presented in Table 2-3, are defined, and cross-validation technique is performed to compare the performance of these models. The linear and exponential regression models are the functions commonly used in the regionalization of flood quantiles and stream characteristics. This study also uses these two model structures and defines four alternative regression models using either two or three features. The features are those that show the highest correlation with the mean of TrH in the previous step. More details on model selection and cross-validation are provided below.

2.4.2.1 Regression Model Selection

The performance of four proposed regression models is compared using k-fold cross-validation. In the k-fold cross-validation procedure, all 183 watersheds are divided into k number of groups. Usually, the value of k is selected based on trial and error, but k=10 is selected in this study based on similar past studies (Kohavi 1995; Dietterich 1998; Borra and Di Ciaccio 2010). Thus, 10 groups including approximately 18 watersheds are created randomly. Each regression model is trained on watersheds in k-1 groups (9 x 18 watersheds), and the root mean square error (RMSE) of the model is calculated on the remaining group (that includes 18 watersheds), “termed test group”. This process is repeated k times by changing the train and test data, and a RMSE value is

calculated for each trial. This results in k number of RMSE values, and their average is reported as the total error of each model. By comparing the error of these models, the model with minimum total error is selected for all study sites.

Table 2-1 Potential watershed characteristics for regression analysis

Features	Description
Main Stream Slope (MSS)	Slope of the stream in watershed with highest strahler's stream order (Strahler, 1957)
Main Stream Length (m) (MSL)	Length of the stream in watershed with highest strahler's stream order
Area (m^2) (A)	Area of watershed
Perimeter (m) (P)	Perimeter of watershed
Circulatory ratio (CR)	watershed area/area of a circle having a perimeter equal to that of watershed
Shape Factor (SF)	watershed area /(Stream Length)
Centroid_X (m) (CX)	the x component of the centroid of watershed
Centroid_y (m) (CY)	the y component of the centroid of watershed
Drainage Density (DD)	Total length of flowlines in watershed/Area of watershed
Average Elevation (m) (AE)	Average of elevation in watershed
Average Slope (%) (AS)	Average of slope in watershed
Relief (m) (R)	Maximum elevation of watershed-Minimum elevation of watershed
Drainage area (m^2) (DA)	Drainage area at outlet of watershed
Annual Precipitation (mm) (AP)	Average of annual precipitation in watershed
Wettest Precipitation (mm) (WP)	Average of precipitation at wettest month
Temperature (T)	Average of Annual Temperature
Urban (U)	Percentage of urban area in watershed
Water (W)	Percentage of water area in watershed
Forest (F)	Percentage of forest area in watershed

Table 2-2 Highest Adjusted R^2 of regression analysis for models with 1, 2 and 3 features

Feature Combination	1 of 19 Features			2 of 19 Features			3 of 19 Features		
	DA	MSS	CR	MSS,AS	MSS,DA	MSS,F	MSS, AE,AS	MSS, AS, DA	MSS, DA, F
Adjusted R^2	0.18	0.17	0.09	0.32	0.26	0.26	0.41	0.4	0.36

Table 2-3 Alternative regression models for TrH prediction in North Carolina

L1	$TrH = A(main\ stream\ slope) + B(average\ slope) + C$
P1	$TrH = e^{a}(main\ stream\ slope)^{b}(average\ slope)^{c}$
L2	$TrH = A(main\ stream\ slope) + B(average\ slope) + C(average\ elevation) + D$
P2	$TrH = e^{a}(main\ stream\ slope)^{b}(average\ slope)^{c}(average\ elevation)^{d}$

For each model shown in Table 2-3, the parameters are determined by running the regression model for 100000 times by randomly selecting a TrH value within the range for each of the 183 watersheds. The parameters that produce the minimum RMSE (Equations 2-9 and 2-10) are then chosen for a given model. Next, the model that gives the least RMSE (average of RMSE in cross-validation process) is selected as the best among the four models.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} e_i}{N}} \tag{2-9}$$

$$e_i = \begin{cases} (tr_i - trl_i)^2 & tr_i < trl_i \\ 0 & trl_i \le tr_i \le tru_i \\ (tr_i - tru_i)^2 & tr_i \ge tru_i \end{cases} \tag{2-10}$$

where e_i is the error of TrH prediction for watershed i and N is the total number of watersheds in test group. tr_i is the prediction from the regression equation, trl_i and tru_i are the lower and upper bounds of the TrH interval, respectively.

2.4.3 Floodplain Mapping

After predicting the TrH by regression model for a given watershed, this parameter should be converted to the 100-year floodplain. Binary Threshold Classifier (BTC) is a simple conditional function that is used for this conversion. In this method, the predicted TrH is compared with raster H calculated before as follows:

$$f(i,j) = \begin{cases} 1 & H_{i,j} \le TrH \\ 0 & H_{i,j} > \text{TrH} \end{cases} \tag{2-11}$$

where $f(i,j)$ is the value of floodplain raster for a given cell (i,j). $H_{i,j}$ is the value of raster H for a given cell (i,j) and TrH is the predicted value from proposed regression model. A value of 1 or 0 in Equation 2-11 is assigned based on whether the cell is within or outside the floodplain, respectively.

2.4.4 Model Validation

After determining the best regression model by using watershed characteristics from 183 watersheds, the model is validated by applying it to a separate set of 105 watersheds representing different geographic regions in North Carolina (Figure 2-1). For each watershed, independent variables namely average slope, average elevation, and main stream slope are calculated to estimate the TrH value by using the proposed regression equation found in the previous step. The model validation is performed in this study by two different approaches. First, the predicted TrH is compared with a desired TrH range of test watersheds to check the possibility of acceptable prediction, overprediction and underprediction for all test watersheds. In order to find the desired TrH, the TrH model described before, should be run for test watersheds as well. In the second validation approach, the predicted TrH is converted to the floodplain by using the BTC method and the generated flood maps are compared with the existing FIRMs by using the C and F indices.

2.5 Results

2.5.1 Selection of Dependent Variables for Regression Analysis

The 19 watershed characteristics shown in Table 2-1 are first filtered through an exploratory regression analysis to find the ones that are related to TrH.

Table 2-2 shows that a combination of main stream slope, average elevation, and average slope has the highest correlation with TrH for the study sites. Therefore, these three variables are chosen as the main independent variables to develop a regression model for estimating TrH and the corresponding floodplain. USGS has already conducted a scientific investigation to predict 100-year flood magnitude of ungauged sites based on watershed characteristics in North Carolina. USGS equations have drainage area as the only significant variable for flood magnitude prediction. The results of this analysis also highlight the role of drainage area because of its influence on stream and watershed average slope. The direct relation between drainage area and flow magnitude is explained by the fact that a larger drainage area will produce higher flows. In the case of

floodplains, for a given magnitude, the topography dictates the floodplain based on the channel and floodplain geometry. Specifically, the average watershed slope is a significant variable that affects the surface water movement towards the streams and the slope of the floodplain. The main stream slope controls the velocity of water in the channel and in the floodplain during a flood event. The average watershed elevation has correlation with feature H that is the main factor for mapping the floodplain. Overall, these variables show the fact that topography is the most significant criteria in explaining the variations in floodplains.

2.5.2 Selection of Regression Model

After determining the independent variables, k-fold cross-validation is used to select the best regression model structure. Results show that P2 has the least error (0.33) followed by L2 (0.35), P1 (0.36) and L1 (0.39) from the ten trials (Figure 2-4). Even though the error is not very different among the four models, P2 is selected for estimating the TrH in this study. Additionally, an exponential function is preferred over a linear function because the exponential function is the result of a linear regression on the log transformed data. Therefore, the independent variables can be directly used without further processing. However, for the linear functions (L1 and L2), an additional step of normalization is required where the mean and standard deviation of all independent variables should be estimated. The final form of P2 along with its parameters obtained through regression on all 183 watersheds is presented in Equation 2-12.

$$TrH = e^{-0.2}(main\ stream\ slope)^{-0.28}(average\ slope)^{0.5}(average\ elevation)^{-0.32} \quad (2\text{-}12)$$

Equation 2-12 reveals that TrH is inversely related to main stream slope and average elevation, and directly related to the average slope of the watershed. As mentioned earlier, both average slope and main stream slope affect the floodplain using TrH, and average elevation is a good regional indicator for separating mountainous areas from flat areas.

2.5.3 Model Validation

The proposed regression model, described by Equation 2-11, is validated by applying it to 105 watersheds in North Carolina. Figure 2-5 presents desired TrH intervals (box) as well as predicted TrH (red dot) for these watersheds. Desired TrH intervals represent a range of TrH that will produce an acceptable floodplain map for watersheds as defined by C> 0.9 and F > 0.6 in relation to FIRMs. Accordingly, this interval could be wide or narrow for a given watershed as shown by

the vertical boxes in Figure 2-5. For instance, watershed 32 in Figure 2-5 has a wide range of TrH from 2 to 12 meters, but watershed 4 has a narrow TrH range of 1.8 to 2.2 meters. The shape of the cross-section or the floodplain valley can explain the variability in TrH ranges. In mountainous areas, higher side slope in river cross-sections lead to the wider range while in flat areas; the same flood plain corresponds to the smaller TrH ranges. This figure illustrates that 55 watersheds were predicted well, meaning that the predicted TrH for these watersheds is within the desired range while 15 and 24 watersheds were underpredicted and overpredicted respectively. Among the 105 watersheds used for validation, 11 watersheds have empty TrH ranges. These watersheds are not included in validation and are explained further in the discussion section. The regression model overpredicts TrH for watersheds where the TrH range is narrow and the mean value is relatively small; whereas underprediction occurs in watersheds that have wider TrH range and the mean of the TrH is relatively higher.

The C and F indices (Equations 2-7 and 2-8) are used to estimate the accuracy of predicted flood maps of all 105 watersheds by comparing them with FIRMs. Figure 2-6 describes the accuracy of predicted flood maps with respect to average slope and mainstream slope. Usually, watersheds that have $C < 0.9$ underpredict the floodplain; whereas watersheds with $F > 0.6$ overpredict it. In Figure 2-6a, C index is mostly low for areas with high average slope. This results in underprediction of floodplain as seen earlier for mountainous regions. On the other hand, high values of C along with unacceptable values of F for areas with low average slope show that the model overpredicts the floodplain for flat regions. These regions have a mean average slope of less than 1% and mean average elevation of less than 20 meters. Figure 2-7b shows that for very small main stream slopes, F index is unacceptable, and C is high which proves the model is overpredicting the floodplain.

The method proposed here requires some training data to relate TrH with watershed characteristics. The training data in this study came from the FEMA maps, which are typically not available for ungauged watersheds. Thus, for this work to be useful for getting floodplain maps in ungauged watersheds, relationships that apply for specific geophysical and climate settings need to be developed. For example, it is possible that the regression equation for TrH developed in this chapter may apply to similar conditions, but that needs to be verified. Similarly, as found in this study, separate equations for flat and mountainous areas may give better results. Thus, this study

forms a foundation for more studies where more than one relationship needs to be developed to estimate TrH using watershed characteristics. As an extension of this study, future work is being carried out to develop relationships for TrH and watershed characteristics for the entire contiguous United States. The future study will actually accomplish the broader goal, where one could develop a floodplain map for an ungauged watershed by using the following procedure: (i) develop stream network by performing terrain analysis; (ii) get the H grid; (iii) get TrH for that particular region from a set of available equations based on the criteria related to factors such as topography and climate; and (iv) use the TrH to classify the H grid into floodplain (H < TrH) and non-floodplain (H>TrH) areas.

Overall, the proposed regression model provides satisfactory performance for most watersheds included in this study (Figure 2-7a), except for those located in mountain regions (Figure 2-7b) where the average watershed slope is more than 20% and those located in flat regions (Figure 2-7c) where the average slope is less than 1% (average elevation is less than 20 meters). Considering the socio-economic impacts, the maps that underpredict the floodplain can give a false sense of security and will lead to disastrous consequences compared to the ones that give either accurate or overpredicted floodplain. Thus, analysis of finding the TrH range and watershed characteristics is conducted for 50 additional mountainous watersheds in western parts of the state to gain a better understanding of the underperformance of TrH model and exponential regression in these areas. The results of TrH model show that, for 22 out of 50 watersheds, there is no particular TrH value that will generate acceptable flood maps to meet the overlap criteria (C>0.9 and F>0.6) specified in this study. The remaining 28 watersheds show high fluctuation of desired TrH interval as shown in Figure 2-8. Some of these watersheds have an extremely wide range of desired TrH intervals (more than 20 meters) while there are some watersheds with only one desired value. Additionally, the results of regression analysis to relate TrH to topography features do not yield any meaningful relationship. These results show just using a single feature H for the whole watershed may not be an appropriate approach for flood mapping in mountainous areas.

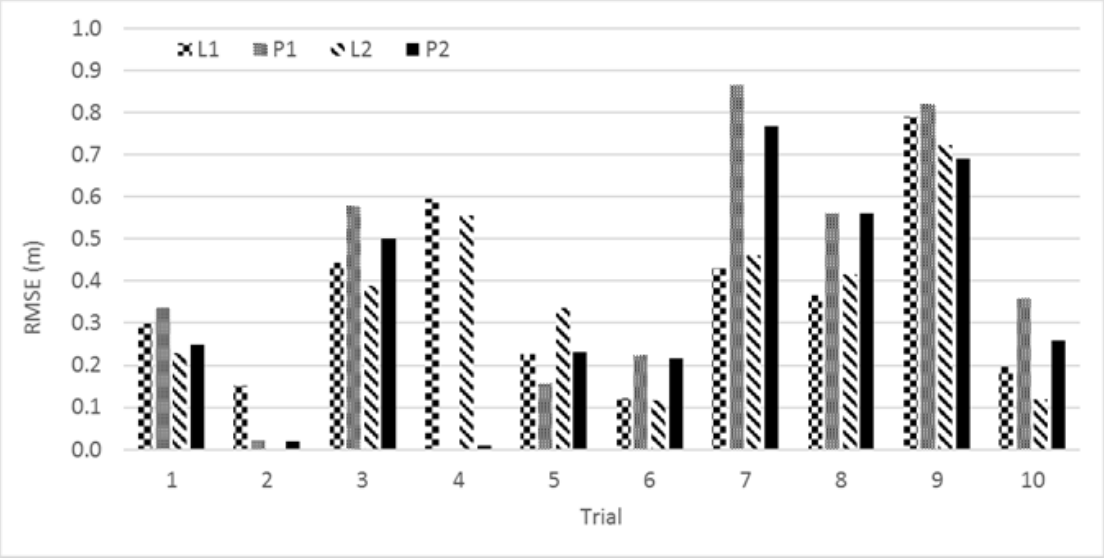


Figure 2-4 Comparison of four alternative models (L1, P1, L2, P2) using 10-fold Cross-Validation

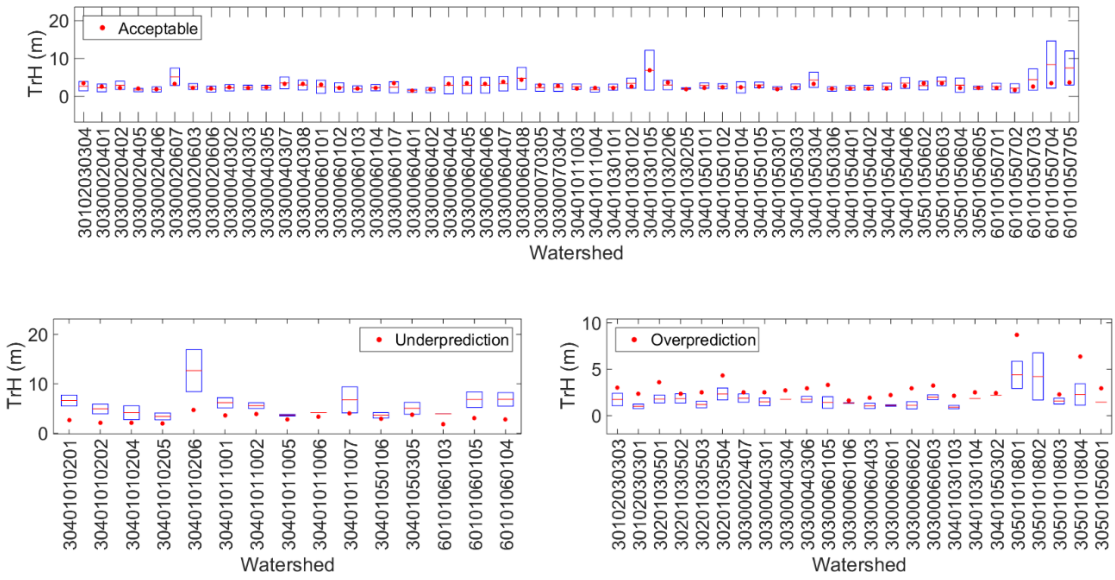


Figure 2-5 Desired TrH intervals (box plots) and predicted TrH (red dots) for test watersheds divided into three sub-plots for watersheds with acceptable prediction, underprediction and overprediction.

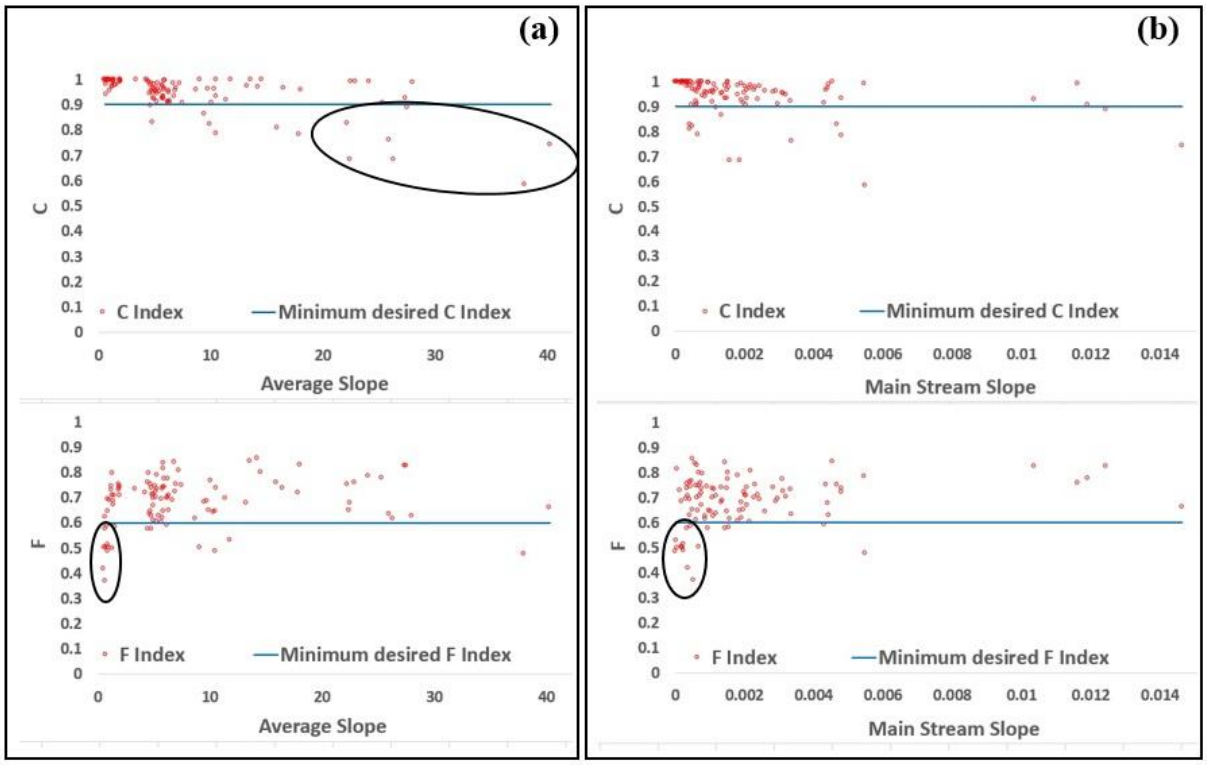


Figure 2-6 Variation of F and C index with respect to watershed characteristics: Average Slope (a) and Main Stream Slope (b). The red dots are watersheds and the blue lines (C=0.9 and F=0.6) are used to distinguish watersheds predicted well (above the line) from those predicted poorly (below the line). The ellipses highlight the critical areas where the majority of watersheds are predicted poorly.

Discussion and Limitation

The broader goal of the work proposed here is to develop a methodology for creating floodplain maps for ungauged watersheds. This study is based on previous studies that use geomorphic characteristics to delineate the floodplains and takes it a step further by relating the key geomorphic attribute (TrH) to watershed characteristics. A total of 19 characteristics related to topography, morphometry and climate were considered. While one would expect climate and/or hydrology to play a role in dictating the 100-year floodplain, topography related attributes emerged as the key independent variables, although it seems that climate factors would be absorbed into an intercept, or constant coefficient in the tested regression model. The fact that the dependent variable (TrH) is derived from topography and that floodplains are topographically controlled explains the emergence of slope and elevation in the regression equation. Additionally, climate variations in North Carolina are relatively smaller than the topographic variation

(http://climate.ncsu.edu/climate/ncclimate.html) As an extension of this study, we are expanding the scope to the contiguous United States, and there we see climate variables (e.g. annual average precipitation and temperature) related to TrH. Thus, it should be noted that the results found in this study are limited to North Carolina, and more work is needed to generalize these findings.

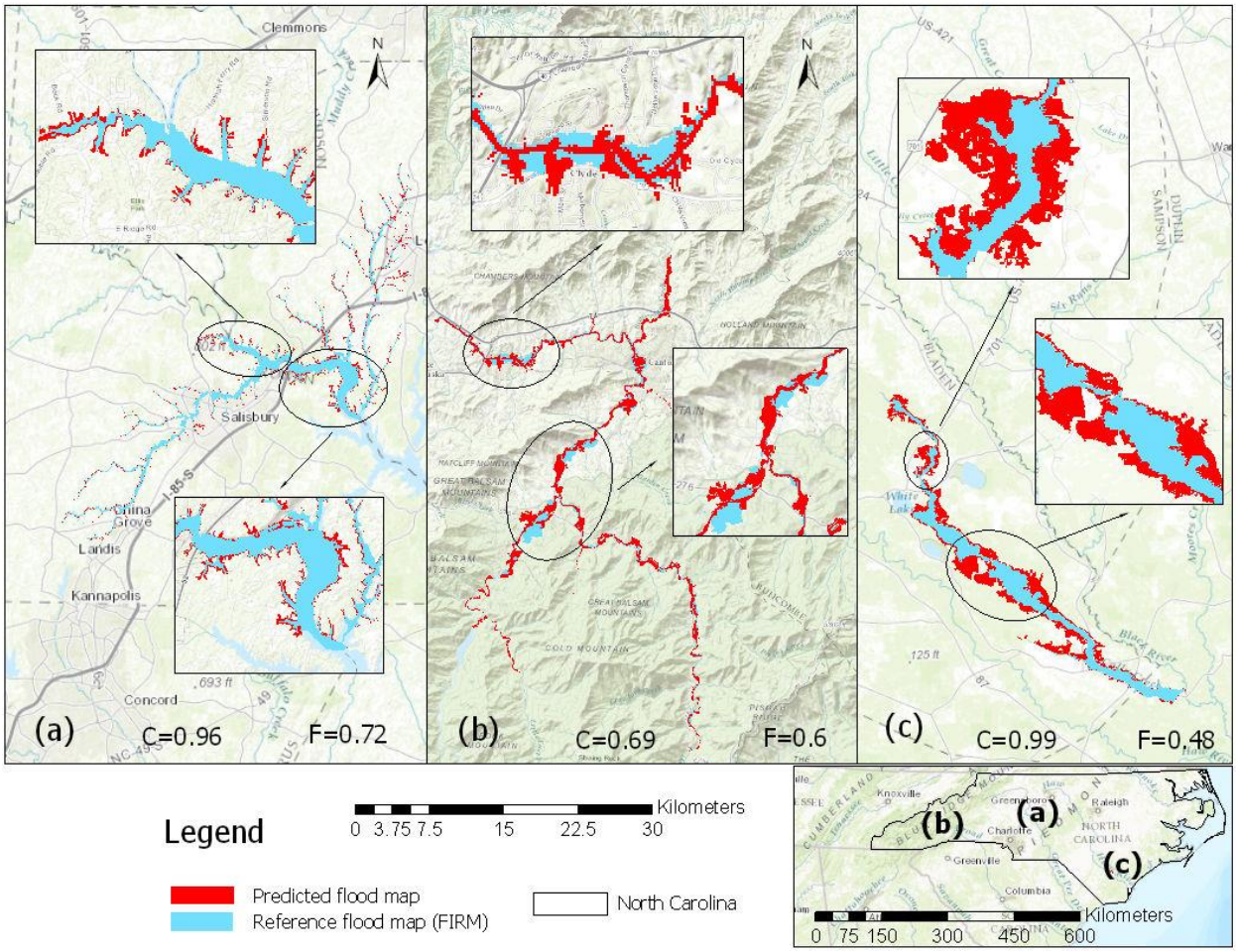


Figure 2-7 Comparison of predicted flood map with FIRM for three watersheds with different topography: Acceptable prediction for mid-altitude watershed (a), underprediction for flat watershed (b) and overprediction for mountainous watershed (c)

Due to a wide range of topographic variability from coasts in the east to mountains in the west, the results in the validation section demonstrated the drawback of the proposed method for mountainous and flat regions. A narrow TrH range in flat areas and high fluctuation of TrH range in mountainous areas are the issues that should be examined. The shape of a single cross section, as well as spatial variability of cross section shapes in a watershed, are two factors affecting the TrH range variabilities. For a single cross-section, higher side slope leads to wider range of TrH

and vice versa. This is the major reason for having wide and narrow TrH ranges in mountainous and flat watersheds respectively. Since the TrH range is representative of the entire watershed, a TrH value estimated for a single cross section should be converted to the TrH range that accommodates all single cross-sectional TrH values for the entire watershed. In some mountainous watersheds, catchments (drainage area for a particular reach) have high topographical variability, which results in more spatial variability of cross sections. Higher spatial variability of cross sections in a watershed reduce the chance of having a common TrH range for all cross sections which causes an empty TrH interval or a very narrow range. These two factors including high side slope of a single cross section and high spatial variability of cross sections, cause high fluctuations in TrH range in mountainous regions. In flat watersheds, single cross-sectional shapes lead to a narrow TrH range while the spatial variability is often low and it cannot change the final TrH range.

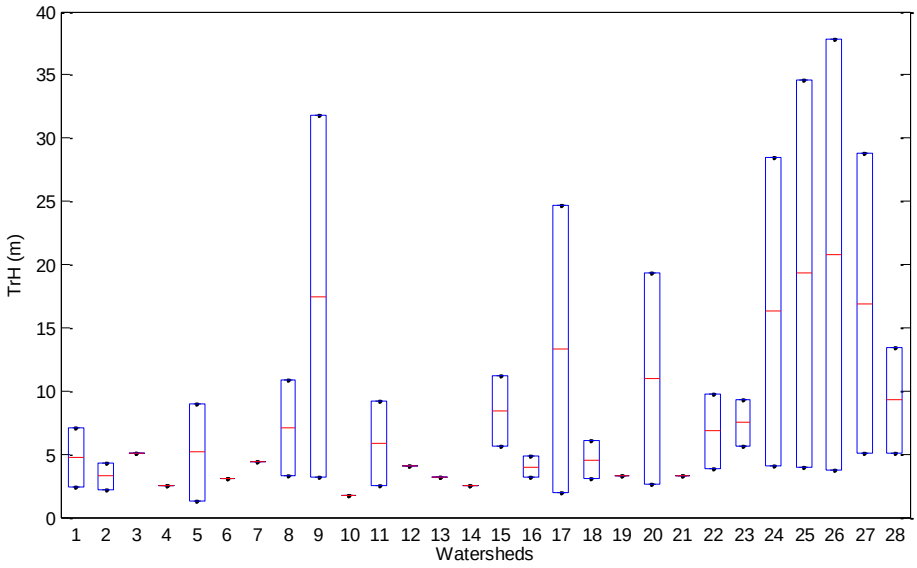


Figure 2-8 Desired TrH intervals (box plots) for mountainous watersheds

Based on these explanations, reducing the spatial unit from HUC12 to catchment unit can be an alternative approach to solve the issue related to the spatial variability of cross section and to decrease the number of empty and narrow TrH ranges particularly in mountainous areas. Another beneficial method to improve the proposed methodology, is regionalization. Using three different regression models for flat, mid-altitude and mountainous watersheds can be considered as a general framework for the entire state, which can also increase the accuracy of predictions.

In this study, all watersheds selected for training the algorithm, belong to central North Carolina. However, watersheds from the western (mountainous) and eastern (flat) zones were added to the test set to examine the transferability of the proposed model from central region to other regions in the state of North Carolina. The unsuccessful results in the western and eastern parts strengthens the idea that using one regression model for all watersheds within the state of North Carolina is not appropriate. A detailed look into the average slope of training watersheds reveal that the average slope for almost all of them varies between 1 and 20 percent which shows the lack of sufficient watersheds from flat and mountainous areas in the training step. This issue can be another reason for the weak performance of the model in these regions. Although selecting more watersheds from flat and mountainous regions in the training step can make some slight improvements for watersheds in these regions, it can decrease the accuracy of method for mid-altitude watersheds because of different behavior of TrH in the mid-altitude region compared to flat and mountainous areas. Therefore, separate modeling for three regions including flat, mid-altitude and mountainous areas is still the preferred approach for the future studies.

Besides the issues associated with the TrH approach and the regression modeling, it should be noted that proposed approach cannot account for human intervention in the system. For example, the FIRMs used for comparison in this study include the effects of structures like bridge, culverts, dams and levees on the flooding. Since the proposed model does not include the effect of such structures, there is high error between the predicted map and FEMA maps in some watersheds. Despite these limitations and drawbacks, the proposed model is a useful approach among alternative methods for fast and simple floodplain mapping. While the validation results prove the reliability and robustness of the proposed method for floodplain mapping in almost all watersheds in central North Carolina, this approach can be generalized for application at the continental scale. One of the practical advantages of this method is that it can generate floodplain maps for all the tributaries inside a watershed rapidly. This makes the algorithm a supplementary tool for extending the available floodplain maps (e.g., FIRMs) beyond their study areas. Figure 2-9 illustrates this advantage by presenting existing FIRMs for a sample watershed compared to the predicted floodplain generated by the proposed algorithm for all tributaries. The hydraulic models used to generate FIRMs need hours or days to create these floodplain maps for only the main streams. However, the proposed method generates floodplain maps for a dense stream network including

all tributaries in a few minutes using a computer desktop with Core i7, 3.6 GHz processor and 16 GB memory (RAM).

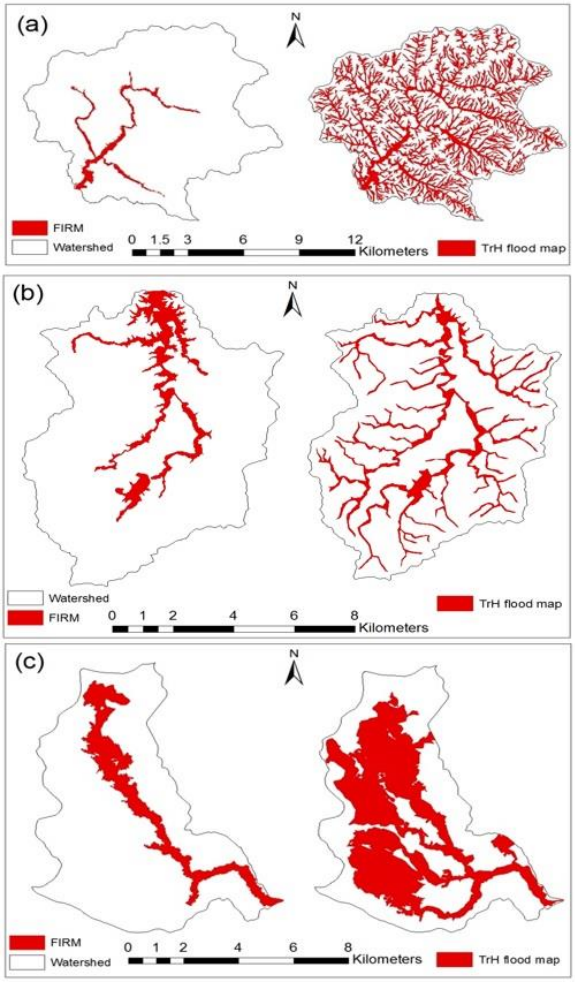


Figure 2-9 Existing floodplain maps (FIRMs) compared to the predicted floodplain maps extended to all tributaries using three different watersheds including a mid-altitude watershed with high fitness (a), a mountainous watershed with underestimation (b) and a coastal watershed with overestimation(c)

2.6 Conclusion and Future Work

The proposed DEM based approach provides an alternative method for mapping the 100-year floodplains in ungauged watersheds. Recent studies have shown that a threshold on topographic feature H (TrH) is useful in mapping the floodplain in an area. This study extends past works by developing regression equations for prediction of TrH based on watershed characteristics in North Carolina. Among the 19 different watershed features related to topography, shape, land use and climate, main stream slope, average watershed elevation and average watershed slope are selected

as the three main features that represent the highest correlation with inundation extent. The regression model also shows that average watershed slope and elevation have a directly proportional relationship with TrH value. Similarly, TrH is inversely proportional to the main stream slope. Thus, for two watersheds with comparable average elevation, the watershed with lower main stream slope will have a higher TrH. The results of model validation indicate the robustness and reliability of the proposed regression model for mapping the floodplain of all watersheds except for mountain and flat areas. Underprediction in mountainous regions and overprediction in the flat watersheds are the drawbacks of the proposed model. The main reason for the unsuccessful mapping results in these regions is related to the selection of training watersheds that mostly belong to the mid-altitude regions. Alghough incorporating more watersheds from flat and mountainous regions into the training set can solve the underprediction and overprediction issue, it will decrease the overall accuracy of the model in the mid-altitude region. Therefore, considering the fact that majority of watersheds belong to the mid-altitude region, using three different regresion models for these three regions is still the preferred approach. Additional analysis on mountainous areas shows that the morphologic feature H cannot provide an acceptable flood map for such a large-scale spatial unit. Considering more morphologic features for mountainous areas or decreasing the geographical unit from HUC12 to catchment are two alternatives to overcome the limitations of the proposed method in mountainous regions. Developing three separate regression models for mountainous, mid-altitude and flat regions is another suggestion that improves the performance of proposed method for future studies.

The method proposed here requires some training data to relate TrH with watershed characteristics. The training data in this study came from the FEMA maps, which are typically not available for ungauged watersheds. Thus, for this work to be useful for getting floodplain maps in ungauged watersheds, relationships that apply for specific geophysical and climate settings need to be developed. For example, it is possible that the regression equation for TrH developed in this chapter may apply to similar conditions, but that needs to be verified. Similarly, as found in this study, separate equations for flat and mountainous areas may give better results. Thus, this study forms a foundation for more studies where more than one relationship needs to be developed to estimate TrH using watershed characteristics. As an extension of this study, future work is being carried out to develop relationships for TrH and watershed characteristics for the entire contiguous

United States. The future study will actually accomplish the broader goal, where one could develop a floodplain map for an ungauged watershed by using the following procedure: (i) develop stream network by performing terrain analysis; (ii) get the H grid; (iii) get TrH for that particular region from a set of available equations based on the criteria related to factors such as topography and climate; and (iv) use the TrH to classify the H grid into floodplain (H < TrH) and non-floodplain (H>TrH) areas.

CHAPTER 4. A GEOMORPHIC APPROACH TO 100-YEAR FLOODPLAIN MAPPING FOR THE COTERMINIOUS UNITED STATES

4.1 Abstract

Floodplain mapping using hydrodynamic models is difficult in data-scarce regions. Additionally, using hydrodynamic models to map floodplain over large stream network can be computationally challenging. Some of these limitations of floodplain mapping using hydrodynamic modeling can be overcome by developing computationally efficient statistical methods to identify floodplains in large and ungauged watersheds using publicly available data. This chapter proposes a geomorphic model to generate probabilistic 100-year floodplain maps for the Coterminous United States (CONUS). The proposed model first categorizes the watersheds in the CONUS into three classes based on the height of the water surface corresponding to the 100-year flood from the streambed. Next, the probability that any watershed in the CONUS belongs to one of these three classes is computed through supervised classification using watershed characteristics related to topography, hydrography, land use and climate. The result of this classification is then fed into a probabilistic threshold binary classifier (PTBC) to generate the probabilistic 100-year floodplain maps. The supervised classification algorithm is trained by using the 100-year Flood Insurance Rate Maps (FIRM) from the U.S. Federal Emergency Management Agency (FEMA). FEMA FIRMs are also used to validate the performance of the proposed model in areas not included in the training. Additionally, HEC-RAS model generated flood inundation extents are used to validate the model performance at fifteen sites that lack FEMA maps. Validation results show that the probabilistic 100-year floodplain maps, generated by proposed model, match well with both FEMA and HEC-RAS generated maps. On average, the error of predicted flood areas is around 14 % across the CONUS. The high accuracy of the validation results shows the reliability of the geomorphic model as an alternative approach for fast and cost effective delineation of 100-year floodplains for the CONUS.

4.2 Introduction

Digital Elevation Models (DEMs) play a critical role in flood inundation mapping by providing floodplain topography as input to hydrodynamic models, and then enabling the mapping of the floodplain by using the resulting water surface elevations (Bates and De Roo, 2000a; Casas et al., 2006; Merwade et al., 2008; Noman et al., 2001; Tate et al., 2002; Townsend and Walsh, 1998). Most commonly, the hydrodynamic modeling approach is used to create flood hazard maps corresponding to a rare high flood frequency of 100-year return period or higher. Although this approach can provide very accurate floodplain maps, it is computationally demanding. As a result, the modeling approach to flood hazard mapping works well for individual streams, but its efficiency drops significantly when used to map floodplains over a large stream network (Cobby et al., 2003). Although there are ongoing efforts to using hydrodynamic models for large scale floodplain mapping (Sampson et al., 2015; Wing et al., 2017), the issue related to high computational demand still exists. In the recent years, geomorphic methods that use topography data in the form of digital elevation model (DEM) and its derivatives, such as wetness index and slope, have been used to map floodplains. Geomorphic methods are not only used to delineate the geomorphic floodplain, the riparian area just above the bank-full discharge corresponding to a 1.5-2 year flow, but they can be trained using 100-year hazard maps to provide 100-year flood inundation extent. While the accurate hydrodynamics resulting from river structures and complex geometry cannot be accounted by the geomorphic methods, they provide an efficient solution by providing the required accuracy in large scale floodplain mapping at a much lower cost (Bates, 2004; Bradley et al., 1996).

Considering the importance of flood hazard, it is important to understand the role of uncertainty and incorporate that information in flood hazard maps. The hydrodynamic modeling approach is suitable for accounting various uncertainties, and thus lends itself to creating probabilistic floodplain maps. Merwade et al., (2008) conducted a detailed analysis on the potential sources of uncertainty arising in floodplain mapping problems. They showed that uncertainty in design flow, terrain datasets and modeling approach are three major components affecting the inundation extents (Alfonso and Tefferi, 2015; Di Baldassarre et al., 2010; Yan et al., 2013). To generate a probabilistic floodplain map, a large number of hydrodynamic model configurations, corresponding to a distinct combination of uncertain data input and/or model parameters are

executed to generate an ensemble of flood inundation maps. This ensemble is then used to assign the probability of flooding to any given point within the floodplain to get a probabilistic floodplain map (Aronica et al., 2002; Domeneghetti et al., 2013; Neal et al., 2013; Purvis et al., 2008; Sarhadi et al., 2012). Besides providing a robust prediction for flood inundation, probabilistic presentation of floodplain areas is also beneficial for decision making and risk analysis (Alfonso et al., 2016). Again, this process is time consuming and computationally demanding. The objective of this chapter is to propose a method to avoid this computational burden in the hydrodynamic modeling approach by developing a geomorphic model based probabilistic floodplain mapping approach that relates the flood extent to watershed characteristics.

Wolman (1971) conducted one of the first studies to explore floodplain mapping using alternative approaches, in which flood mapping methods were compared by dividing them into several groups including physiographic, pedologic, vegetation, occasional floods, regional floods of selected frequency and flood profiles, and backwater curves. While this study did not focus on the details of any specific method, it provided a general insight on these alternative floodplain mapping methods. Williams et al. (2000) suggested a simple method to delineate floodplains by subtracting the DEM from an assumed constant water level for the entire stream network. The main limitation of this method was the assumption of constant water level and the lack of a reliable method to find the actual water depth in the rivers. Later, a series of methods to identify low-lying valleys based on DEM were developed (Dodov and Foufoula-Georgiou, 2005; Gallant and Dowling, 2003; McGlynn and Seibert, 2003). For example, Gallant and Dowling (2003) proposed a multiresolution index to estimate the valley bottoms. Although distinguishing valley bottoms from hillslopes is a valuable task for hydrologic purposes, these areas do not account for a particular flood magnitude or frequency. Nardi et al., (2013, 2006) proposed a hydrogeomorphic method for mapping floodplains in which the hydrologic characteristics of a flood event were also incorporated into the modeling process. Therefore, the method was able to map floodplains corresponding to specific flood frequencies. In addition to DEM, methods based on soil information have also been proposed for floodplain mapping (Sangwan and Merwade, 2015).

Some recently developed alternative methods for floodplain mapping are based on supervised and unsupervised classification (clustering) of data. Unsupervised methods attempt to group data

points into several clusters based on similarity of their attributes. A common form of clustering in hydrological problems is termed “regionalization”, where a large heterogeneous area is divided into smaller homogeneous regions based on multiple watershed characteristics (Chiang Shih-Min et al., 2002; Rao, 2004; Rao and Srinivas, 2006a, 2006b; Razavi, Tara and Coulibaly, Paulin, 2013; Ridolfi et al., 2016). Watershed characteristics have been widely used as reliable descriptors of hydrologic variables in ungauged basins (Berger and Entekhabi, 2001; Ganora et al., 2009; Patton and Baker, 1976; Sankarasubramanian and Vogel, 2002; Sefton and Howarth, 1998; Thomas and Benson, 1970). Specifically, several regional regression models have been developed in the past few decades to relate streamflow statistics (e.g.100-year flood, mean annual flow, 7-day low flow frequencies) with watershed characteristics (Acreman, 1985; Crippen and Bue, 1977; Ries, 2007; Sauer et al., 1983; Thomas and Benson, 1970; Turnipseed and Ries III, 2007). Besides regionalization, clustering methods can also be used to map flood risk areas. In one study, Papaioannou et al. (2014) employed a clustering method to classify a raster into different levels of flood risk areas. They used multi-criteria evaluation methods to select and find the weights of the most significant factors for clustering. Selection of proper factors and weights can add huge uncertainties in the results of unsupervised classification methods, largely due to the high sensitivity of clustering results to the unknown weight of factors.

Spatial supervised classification methods attempt to find a pattern in the attributes of some labeled data (training data), and utilize that pattern to classify the unkown data (test data). These methods have been successfully used in floorplain mapping by dividing a watershed into a grid of cells where each cell can be classified as “flood” or “non-flood” (binary classification). An observed or reliable floodplain map is required as a reference to train the classifer, and then the trained classifier predicts the class labels of unknown cells (De Risi et al., 2014; Degiorgis et al., 2013; Manfreda et al., 2015, 2014; Samela et al., 2016). In order to find the best classifier, different morphological features have been proposed, including the modified topographic index (Manfreda et al., 2011, 2008), and topographic wetness index (De Risi et al. 2014). Degiorgis et al. (2012) compared the performance of several single morphologic features, and suggested that feature *H*, defined as the difference in elevation between a given cell and the lowest elevation in the nearest stream as represented in a DEM (Nobre et al., 2011; Rennó et al., 2008), plays the most significant role in these methods. Their results showed that using several features and/or more complicated

classification methods, such as support vector machine, is not necessary. In these past studies, a binary threshold classification is used so that a threshold on the morphologic feature is chosen based on minimizing the error between the reference and predicted flood extents.

The supervised binary classification based on finding a threshold on morphologic feature *H* (*TrH*) is a reliable approach for floodplain mapping over large areas because it is simple, fast, and accurate. In addition, it can identify the floodplains associated with a particular flood frequency such as 100-year floods. However, this method, like any supervised classification problem, needs some training based on a reference map. The reference maps are usually provided by collecting detailed survey data from field measurements and running hydrodynamic models. The dependency of this method on the reference map and hydraulic data limits its application for ungauged basins where no reference maps are available. Samela et al. (2017) used supervised classification methods to identify floodplains for both gauged and ungauged basins by assuming that the threshold determined from training watersheds can be used for other ungauged watersheds in a large region. This threshold transferability assumption considers the entire study area as a homogenous region where hydrological and morphological factors in the training and test areas are considered identical.

Watershed characteristics have been widely used in hydrology to convert flood magnitudes from gauged sites to ungauged sites. Using this concept, Jafarzadegan and Merwade (2017) developed a regression model which used watershed characteristics to predict *TrH* corresponding to the 100-year floodplain for North Carolina. The predicted *TrH* from the regression model was then used to identify floodplains. Although the method worked, it was not able to satisfactorily predict *TrH* for flat and mountainous watersheds because the regression model was site dependent, thus limiting its application in areas with different topographic, climatic and land use settings. In order to overcome this limitation, this study proposes a geomorphic model in which the classification method is used to classify watersheds based on watershed characteristics and then a range of *TrH* values are used to map probabilistic floodplains for a given watershed. The proposed method also overcomes the threshold transferability assumption of Samela et al (2017) by acknowledging spatial heterogeneity in the landscape to relate the spatial *TrH* variability to watershed characteristics. Considering the generality of the proposed model, it is developed and applied for stream networks across the CONUS.

4.3 Dataset and Study Area

When a single *TrH* is used for an entire stream network in a watershed, it is assumed that all rivers and tributaries in the watershed have the same hydrological and morphological characteristics. The assumption of hydrological and morphological homogeneity and unique *TrH* can generate unreal results with high uncertainties for a large watershed, but working with smaller watersheds can provide relatively accurate results. In this study, a Hydrologic Unit Code 12 (HUC12) is used as the computational unit for floodplain mapping. The United States Geological Survey (USGS) has divided the U.S. using six levels of hydrologic unit codes (HUC, watershed boundaries) from the largest HUC2, called regions, to the smallest HUC12, called subwatersheds. According to the HUC classification, the U.S. is divided into 22 regions, and each region is subdivided into around 7600 subwatersheds. ("U.S. Geological Survey - National Hydrography Dataset," 2014.) A total of 216 HUC12 units, referred to hereafter as just watersheds, across the CONUS are selected (Figure 4-1) in this study. The watersheds are selected to capture the variability in topography, climate, land use and geography to develop and test the proposed model (Table 4-1)

Table 4-1 Summary statistics of samples compared to the population

Statistics	Topography (Average Elevation) (m)		Climate (Annual Precipitation) (mm)	
	Population	Samples	Population	Samples
Mean	578	428	974	956
STD	588	402	446	433
Min	45	63	124	166
Max	2420	1962	2815	2123

Additionally, training of the classification algorithm requires reference floodplain maps, which are available from FEMA for these watersheds. In addition to the 216 training watersheds, 145 watersheds are chosen to validate the performance of the proposed model in the first phase. The same criteria used for selection of training watersheds is considered for choosing the validation watersheds. The second phase of validation is performed by generating probabilistic 100-year floodplain maps for 15 more watersheds that do not have any reference FEMA maps. Considering the lack of FEMA reference maps for these 15 sites, the predicted flood extents are validated against the results obtained from HEC-RAS modeling at these sites. Figure 4-1 depicts the spatial distribution of the watersheds selected for training and validation. Figure 4-1 clearly shows uneven distribution of areas between the eastern and western part of the U.S. due to the absence of reliable

reference maps (FEMA) for some states such as Washington, Utah, Idaho, and Wyoming in the west.

Other datasets including stream networks, DEMs, Land use and climatic rasters are also used in this study to compute watershed characteristics. The sources for these data include the USGS 30m horizontal resolution National Elevation Dataset DEM, USGS's National Hydrography Dataset (NHD) for the stream networks, the National Land Cover Database (NLCD) 2011 for Land use and WorldClim-Global Climate Data for the average annual precipitation and temperature. Flood Insurance Rate Maps (FIRMs) provided by FEMA are used as reference for training and validation of the proposed methodology.

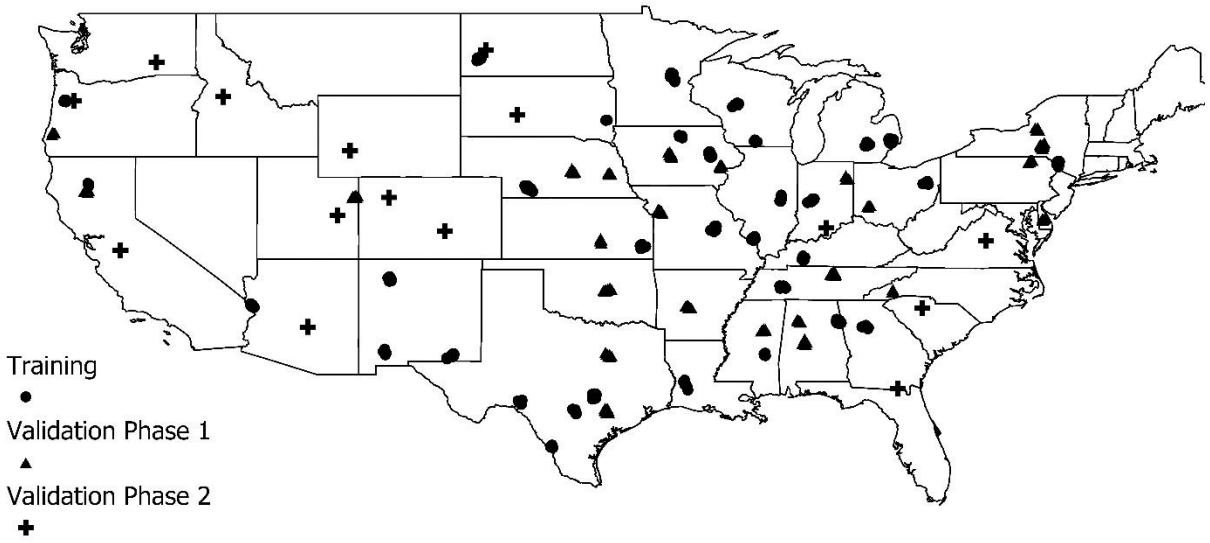


Figure 4-1 Map of the United States with geographic location of watersheds for training and validation phases

4.4 Methodology

In this section, the geomorphic model for probabilistic mapping of 100-year floodplains for CONUS is introduced. The proposed model consists of two classification modules. In the first module, all study watersheds within the CONUS are classified into three different classes of TrH range based on multiple watershed characteristics using supervised classification. In the second module, each study watershed, represented using a raster grid, is classified into flooded and non-flooded cells using the probabilistic threshold binary classification (PTBC). It should be noted that

both classifications are applied in the probabilistic mode. In the first module, the probability that a given watershed belongs to one of the three TrH classes is determined by using watershed characteristics derived from DEM, land use and climate data. The second module uses H raster, a lookup table and the class probabilities derived from the first module as input to determine the probability of each grid cell within the watershed getting inundated from a 100-year flood event. A flowchart of the proposed model is presented in Figure 4-2, and more specific details are provided below.

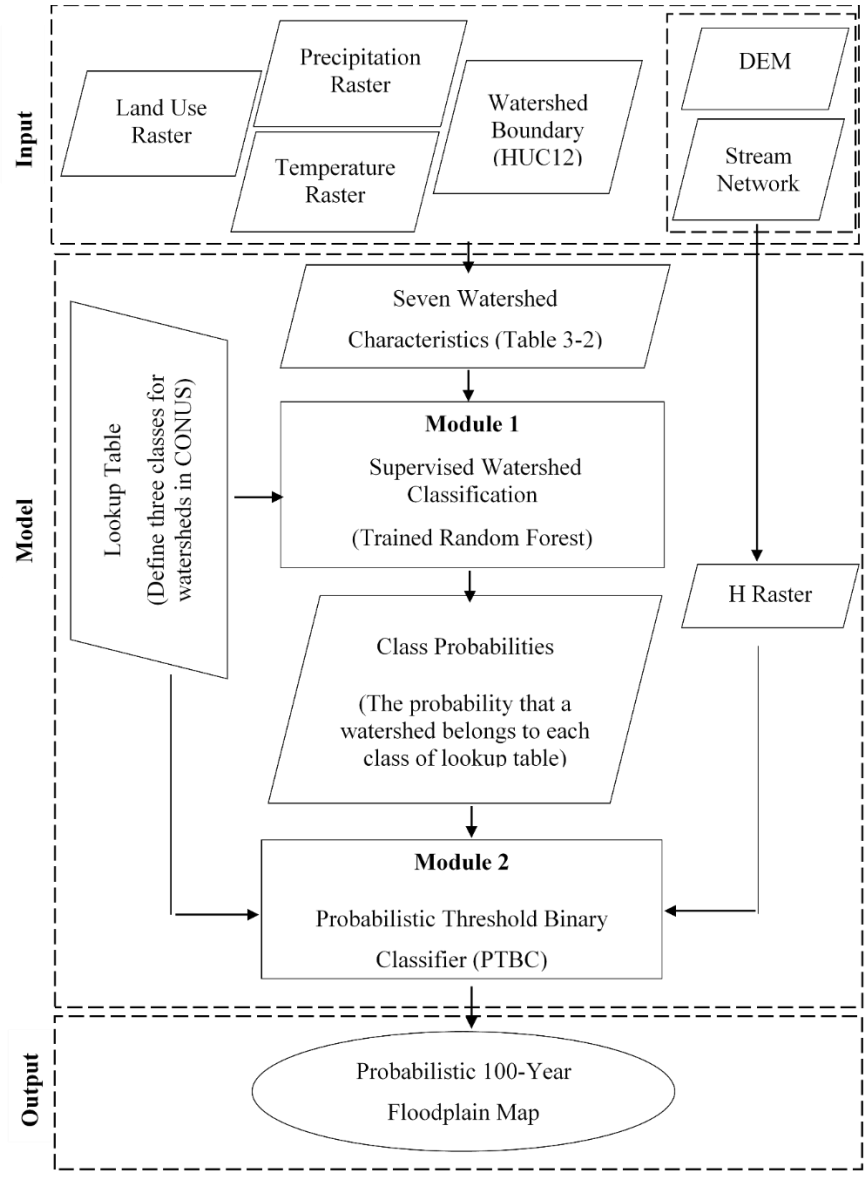


Figure 4-2 Flowchart of the proposed model for probabilistic 100-year floodplain mapping

4.4.1 Supervised Watershed Calssification

In a supervised classification problem, each data point is defined as a pair consisting of attributes and target class. The first component, attribute, is a vector of input features describing the status of data point while the second component, target class, is a discrete label assigned to a data point as the output of classification. The objective is to establish a rule and find a relation between these two components (train the classifier) to predict the class label of unknown data points. In order to find this relation, data points with known target classes are chosen as training dataset. In this study, each watershed represents a data point, which has watersheds characteristics as attributes, and the target classes include discrete labels associated with three *TrH* ranges from a lookup table.

To compile the attributes of training dataset, multiple watersheds from different geographical locations are selected, and several watershed characteristics based on hydrography, topography, climate, and land use are calculated (Table 4-2). The main stream slope, drainage density, and drainage area of a watershed are three features related to the hydrography because the stream network and flow path in the watershed are required. To calculate the main stream slope, the stream with the highest Strahler's stream order is selected as the main stream, and then the DEM is used to find the slope of this river. Two climatic features including precipitation and temperature are also calculated to include climate variability. Topography features including average slope and average elevation of the watershed are determined from the DEM. Another possible effective variable in floodplain mapping is the surface roughness in the form of Manning's coefficient[2]. The surface roughness is computed by taking the mean of the Manning's coefficients for the watershed as found from different land use types in the watershed (Kalyanapu et al., 2010). Percent urban cover and water are calculated from the land use raster.

The list of potential watershed characteristics in Table 4-2 includes some features that may not be strongly correlated with *TrH*. Thus, to increase the efficiency of the classifier, watershed characteristics that are poorly correlated with *TrH* are removed from the analysis. Two commonly used correlation coefficients, namely Pearsons's r and Kendall's tau, are used in this study to test the correlation of watershed characteristics with the *TrH* ranges (Kendall, 1949; Pearson, 1904). While the Pearson's r tests the linear relation, Kendall's tau is a rank based coefficient that tests

[2] The Manning's coefficient represents the roughness or friction applied to the flow by the channel

non-linear monotonic correlations. In order to find the correlation coefficients between a vector of distinct values (a given watershed characteristic such as average slope) and a vector of interval numbers (*TrH* range), a uniformly distributed random number is generated over the *TrH* range for all watersheds each time. Then the correlation coefficient between the random *TrH* vector and the watershed characteristic vector is determined. This process is repeated 10000 times, and the maximum correlation coefficient is reported as the correlation coefficient between the given watershed characteristic and the *TrH* range.

Table 4-2 List of potential watershed characteristics

Factors	Watershed Characteristics	Description
Hydrography	Main Stream Slope (MSS)	Slope of the stream in watershed with highest strahler's stream order
	Drainage Density (DD)	Total length of flowlines in watershed/area of watershed
	Drainage Area (DA) (km^2)	Total area directing water toward the outlet of watershed
Topography	Average Elevation (AE) (m)	Average of elevation in watershed
	Average Slope (AS)	Average of slope in watershed
Climate	Annual Precipitation (AP) (mm)	Average of annual precipitation in watershed
	Annual Temperature (AT)	Average of annual temperature in watershed
Land Use	Urban (PU)	Percentage of urban area in watershed
	Water (PW)	Percentage of water area in watershed
	Roughness Coefficient (RC)	Average of Manning's roughness coefficients in watershed

To assign the target classes to the training watersheds, two important variables namely optimum *TrH* and *TrH* range are calculated for each watershed. A FEMA map is required as a reference map to find these values for a given watershed. If one assumes raster *H* for a watershed, all cells with *H* less than *TrH* are labeled as flood and others will be non-flood cells. This simple "if and else rule" is used for floodplain mapping based on *TrH*. In general, each instance of a binary classification problem is positive or negative which can be renamed with flood and non-flood cells in a flood mapping problem. The optimum *TrH* is determined by minimizing the total error between predicted and reference maps where the total error is the summation of all misclassified cells (Flood predicted as non-flood and vice versa). In order to find the *TrH* ranges, two indices namely C and F are used (Equations 3-1 and 3-2). These indices have been widely used in the literature to estimate the performance of a predicted flood inundation extents (Alfieri et al., 2014;

Bates and De Roo, 2000b; Horritt and Bates, 2002; Sangwan and Merwade, 2015). While C index only recognizes the underpredictions in a model, F gives more information about both underpredictions and overpredictions. In this study, *TrH* range is defined as an interval of the *TrH* values where any threshold inside this interval can generate an acceptable flood map with $C > \alpha$ and $F > \beta$ (Equation 3-3).

$$C = \frac{flood\ cells\ \ predicted\ correctly}{flood\ cells} \quad (4\text{-}1)$$

$$F = \frac{flood\ cells\ \ predicted\ correctly}{flood\ cells + nonflood\ cells\ predicted\ as\ flood} \quad (4\text{-}2)$$

$$TrH_i \in TrH_{range}\ \ if\ C_{TrH_i} \geq \alpha\ and\ F_{TrH_i} \geq \beta \quad (4\text{-}3)$$

In this study, α and β for TrH range calculation are 0.8 and 0.5 respectively. Jafarzadegan and Merwade (2017) used α = 0.9 and β = 0.6 for TrH range calculation in North Carolina, but considering the broader applicability of the proposed work, the criteria for α and β is slightly relaxed in this study by using lower values for α and β. The lookup table is created by looking into the variability of *TrH* range, and optimum *TrH* for the training watersheds. This table defines three classes of *TrH* ranges and assumes that any watershed in CONUS belongs to one of these three classes. Based on this table, the calculated *TrH* range and optimum *TrH*, a target class label is assigned to each training watershed.

The significant watershed characteristics and the assigned class labels of training watersheds are the major inputs for developing a supervised classifier. In this study, four common classifiers, namely, logistic regression, support vector machine, decision tree and random forest, are fit to the data. The performance of these classifiers is compared using K-fold cross validation and the Root Mean Square Error (RMSE). The best classifier is selected to perform the supervised classification for the proposed model (Module1). The selected classifier creates the probability that a watershed belongs to each class, as defined by a *TrH* range presented in the lookup table (Table 4-3). These class probabilities, as well as their corresponding *TrH* range from the lookup table and the *H raster*, are used in PTBC to generate the probabilistic 100-year floodplain maps.

Table 4-3 Lookup table including TrH ranges and their corresponding class labels for CONUS

Class	TrH Range (m)
1	0.5-2.5
2	2-5
3	4-8

4.4.2 Probabilistic Threshold Binary Classifier (PTBC)

PTBC is the second classification module used in the proposed model to generate the 100-year floodplains. The essence of this classifier is similar to the threshold binary classifiers used in the literature for floodplain mapping (Degiorgis et al., 2013, 2012). Those simple threshold classifiers use raster *H* as input and generate deterministic floodplain maps based on a threshold (*TrH*). The PTBC, proposed in this study, uses additional information, including the class probabilities (from Module 1), and a set of *TrH* ranges (lookup table) instead of a single *TrH* to generate probabilistic floodplain maps. In order to employ PTBC and generate the probabilistic 100-year floodplain maps, first the *TrH* ranges from lookup tables are discretized into eleven *TrH* values (Table 4-4). Considering the *TrH* range as a set of *TrH* values between two endpoints as $a \leq TrH \leq b$, ten equal increments are defined to discretize the *TrH* range as follows:

$$\Delta = \frac{b-a}{10} \tag{4-4}$$

$$TrH\ range \approx \{a, a+\Delta, a+2\Delta, \dots, a+9\Delta, b\} \tag{4-5}$$

For each discretized *TrH* value, the raster is classified into flood and non-flood areas using a simple conditional function (Equation 3-6). In order to use this function, raster *H* for a given watershed should be computed and all cells with corresponding *H* values less than *TrH* are labeled as flood and others are labeled as non-flood cells. This process is repeated for all eleven discretized *TrH* values and the mean of flood and non-flood cells are calculated (Equation 3-7). A weighted average of probabilistic flood maps for each class is calculated to find the final floodplain maps (Equation 3-8). The weight of each class, defined as the probability of watershed belonging to a given class (P_m) from Module 1, is used as input to PTBC.

$$\mathrm{f_{k,s}(i,j)} = \begin{cases} 1 & H_{i,j} \leq TrH_{k,s} \\ 0 & H_{i,j} > TrH_{k,s} \end{cases} \tag{4-6}$$

$$Pr_s(i,j) = \frac{\sum_{k=1}^{K} f_{k,s}(i,j)}{K} \tag{4-7}$$

$$Pr(i,j) = \sum_{s=1}^{S} \left(P_s \times Pr_s(i,j)\right) \tag{4-8}$$

Table 4-4 Discretized version of Lookup table by using ten increments for each range

K_{th} discretization	Class 1 [0.5 2.5]	Class 2 [2 5]	Class 3 [4 8]
1	0.5	2	4
2	0.7	2.3	4.4
3	0.9	2.6	4.8
4	1.1	2.9	5.2
5	1.3	3.2	5.6
6	1.5	3.5	6
7	1.7	3.8	6.4
8	1.9	4.1	6.8
9	2.1	4.4	7.2
10	2.3	4.7	7.6
11	2.5	5	8

In these equations, *K* refers to the total eleven discretized *TrH* values inside a *TrH* range, where index *k* is the counter of these eleven numbers k=(1,2,…,11) , *S* is total number of classes where index *s* is the counter of classes (*s*=1,2,3), $TrH_{k,s}$ is the k^{th} discretized *TrH* in class s, $Pr(i,j)$ is the probability of 100-year flood for a given cell *(i,j)*, P_s is the probability that watershed belongs to class s, $Pr_s(i,j)$ is the probability of 100-year flood for given cell (i,j) if the watershed belongs to class s, $f_{k,s}(i,j)$ is a conditional function for kth discretized *TrH* in class s for a given cell *(i,j)*, $H_{i,j}$ is morphologic feature *H* for a given cell *(i,j)*.

To understand the approach, consider a hypothetical example where the probability of flooding for two cells a and b within a watershed needs to be determined. Cell a is near a stream with $H = 2$ and b is away from the stream with $H = 4.2$. First, seven watershed characteristics for the watershed are calculated using the DEM, land use, and climate data. The watershed characteristics are used as input to the classifier (module 1) that has already been trained for the CONUS. Assume that Module 1 classifier predicts the class probabilities as $P_s = [0.1,0.7,0.2]$, which means the given watershed most likely belongs to the second class (probability = 0.7) of lookup table (Table 4-3). The discretized *TrH* range for this class is available in Table 4-4. Using Equation 3.6, the conditional function ($f_{k,s}$) is calculated for both points "a" and "b" (Table 4-5). The probability of flooding for each class (Pr_s) is determined by taking the average of conditional functions at each

column (Eq. 7) (Table 4-6). Finally, the numbers from Table 4-6 together with output of module one ($P_s = [0.1,0.7,0.2]$) are used in Equation 3-8 to find the probability of 100-year floodplain for cell "a" and "b" as follows:

$pr_a = 0.1 \times 0.273 + 0.7 \times 1 + 0.2 \times 1 = 0.93$

$pr_b = 0.1 \times 0 + 0.7 \times 0.273 + 0.2 \times 0.909 = 0.37$

These values show that point "a" with a probability of 0.93 is very likely to get inundated while point "b" with a 0.37 chance of flooding is less likely. Point "b" and other points with a probability of flooding around 0.5 refer to areas with highest uncertainty near the floodplain boundary that need further evaluation to decide whether they will get inundated or not.

Table 4-5 Conditional function values ($f_{k,s}$) for all 33 discretized *TrH* values at point (a) and (b)

Point	*K* th discretization	Class 1 [0.5 2.5]	Class 2 [2 5]	Class 3 [4 8]
(a)	1	0	0	0
	2	0	0	1
	3	0	0	1
	4	0	0	1
	5	0	0	1
	6	0	0	1
	7	0	0	1
	8	0	0	1
	9	0	1	1
	10	0	1	1
	11	0	1	1
(b)	1	0	1	1
	2	0	1	1
	3	0	1	1
	4	0	1	1
	5	0	1	1
	6	0	1	1
	7	0	1	1
	8	0	1	1
	9	1	1	1
	10	1	1	1
	11	1	1	1

Table 4-6 probability of flooding for each class of lookup table at points a and b

	Class 1 [0.5 2.5]	Class 2 [2 5]	Class 3 [4 8]
a	0	0.273	0.909
b	0.273	1	1

4.4.3 Validation Phase 1: Comparison with FEMA

In order to validate the effectiveness and reliability of the geomorphic model, probabilistic 100-year floodplain maps are generated for multiple watersheds across the CONUS and their overlap with FEMA maps is examined. To compare a deterministic map (reference map) with a probabilistic map (predicted map), two methods are used. In the first method, the Overestimation Flood Index (OFI) and the Underestimation Flood Index (UFI) are defined using Equations 3-9 and 3-10 respectively.

$$\mathrm{UFI} = \frac{\sum_{i=1}^{N}(1-P_i)}{N} \times 100 \qquad i \in F \qquad (4\text{-}9)$$

$$\mathrm{OFI} = \frac{\sum_{j=1}^{M}(P_j)}{M} \times 100 \qquad j \in NF \qquad (4\text{-}10)$$

In these equations, F and NF refer to the flood and non-flood areas of reference map respectively. P_i and P_j are the probability of flooding for cell i and j obtained from the predicted probabilistic map. Cell i represents a cell inside the FEMA floodplains (F); whereas cell j represents a cell outside of FEMA floodplains (inside the non-flood areas (NF)). N and M are the total number of cells inside the F and NF respectively. After finding these two indices for each validating watershed, the performance of watershed is presented as a point in the OFI-UFI space.

Performance of the geomorphic model for estimating the extent of floodplains is also evaluated using the Receiver Operating Characteristic (ROC) graphs, which are one of the most commonly used methods for validation of probabilistic classifiers. For a given threshold between 0 and 1, the probabilistic map is converted to deterministic one and the rate of true positive (rtp) and rate of false positive (rfp) are calculated (Fawcett, 2006):

$$rtp = \frac{True\ positive\ instances}{Total\ positives} \qquad (4\text{-}11)$$

$$rfp = \frac{False\ positive\ instances}{Total\ negatives} \qquad (4\text{-}12)$$

Here, positive and negative refer to the flood and non-flood cells, respectively. ROC graph is a curve showing the relation of rtp and rfp for different thresholds. In order to quantify the performance of such a graph, the area under the curve (AUC) is calculated (Figure 4-3). For a

random classification, AUC value is 0.5, but in this study, watersheds with flood maps having AUC more than 0.9 are considered good and flood maps with AUC less than 0.8 are considered poor. The AUC values calculated in similar geomorphic floodplain modeling studies vary from 0.55 to 0.95 (Manfreda et al. 2014; Samela et al., 2016). Therefore, regarding the continental extent of this study, the selected constraints for AUC are considered reasonable.

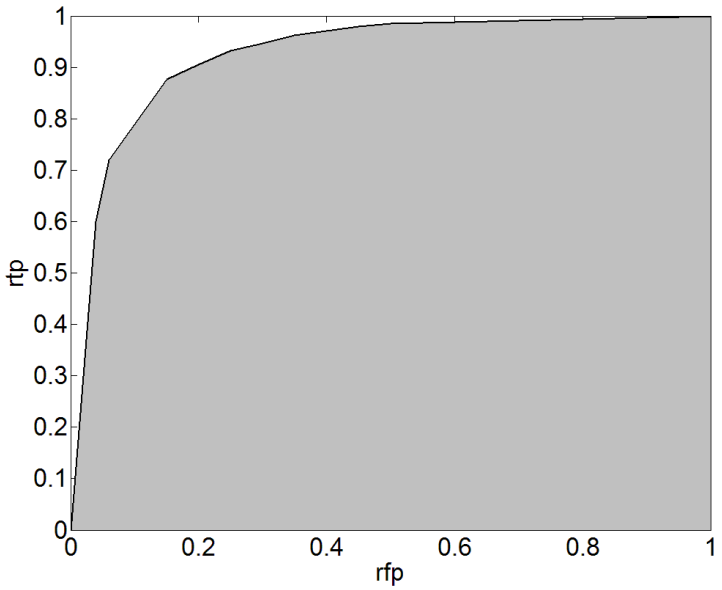


Figure 4-3 Template of Receiver Operating Curve (ROC) and Area Under the Curve (AUC) used for evaluation of a probabilistic floodplain map compared to a deterministic reference map

4.4.4 Validation Phase 2: Comparison with HEC-RAS Results

Many areas in the CONUS do not have FEMA maps, and creating these maps would require hydrologic and hydraulic modeling. To evaluate the reliability of the geomorphic model, the probabilistic 100-year floodplain maps are also compared with HEC-RAS generated inundation extents in areas (Figure 4-1) where FEMA maps do not exist. The 100-year flow magnitude for some gauged streams are found using flood frequency analysis by fitting the Log-Pearson Type 3 distribution to the annual maximum series (Griffis and Stedinger, 2007). For ungauged reaches, the 100-year flow magnitude is estimated using StreamStats, which is a web interface developed by United States Geological Survey (USGS) to estimate 100-year flood magnitudes at any location along the ungauged reaches (U.S. Geological Survey, 2012). StreamStats uses watershed characteristics (e.g. drainage area, stream slope, basin length, average precipitation, fraction of urban area) in a regression model to estimate the target flood magnitudes. The estimated 100-year

flow rate as well as the geometry data, generated using HEC-GeoRAS (Ackerman, 2005) are used to create the HEC-RAS model and the inundation extent. The inundation from HEC-RAS model and the predicted flood extents by the proposed model are compared the same way as FEMA maps. Considering that HEC-RAS modeling was performed only on the main reach, the comparison is conducted on a single reach instead of the entire network.

4.5 Results

4.5.1 Geomorphic Model Setup

A total of 216 watersheds with various climate, land use and topography from 43 different geographical regions are selected (Figure 4-1). To perform supervised classification, significant watershed characteristics, from a set of ten, are selected by using the correlation coefficients between these characteristics and the *TrH* range. As presented in Table 4-7, two land use characteristics, namely PU and PW, as well as AE show low correlation with *TrH* range, and thus are removed from further analysis. In order to assign the class labels to these watersheds, a lookup table including three TrH ranges is created (Table 4-3).

Table 4-7 Correlation between *TrH* range and watershed characteristics

Features	Pearson's r	Kendall tau
MSS	-0.16	-0.26
AS	0.4	0.41
AE	-0.14	-0.2
DA	0.54	0.32
DD	0.32	0.28
AP	0.19	0.22
AT	0.24	0.31
RC	0.24	0.28
PU	-0.09	-0.1
PW	-0.07	-0.03

Supervised classification is then performed using four methods, namely logistic regression, support vector machine, decision tree and random forest. Comparison of these methods using K-fold cross-validation with k=10 demonstrates that random forest classifier with an accuracy of 0.776 performs the best for the study data, followed by logistic regression, decision tree, and support vector machine with an accuracy of 0.736, 0.735 and 0.529, respectively. Thus, random

forest classifier is used to classify the watersheds in this study. Random forest classifier is an ensemble of multiple tree classifiers which combine the decisions of all tree classifiers by weighted or unweighted voting to classify the unknown examples (Pal, 2005). Each tree casts a unit vote for the most popular class to classify an input vector (Breiman, 1999). In this study ten sub-samples of the training dataset are generated by replacement (Bootstrapping) where the sub-sample size is the same as the original dataset. Then ten tree classifiers are fitted to sub-samples. Finally, for each given vector of watershed characteristics, the decisions of these ten trees is averaged to find the probability of all three class labels (P_s).

Some additional feature analysis on the developed random forest indicates that average slope (AS) with the weight of 0.33 is the most significant factor for the classification. Annual temperature (TR) and roughness coefficient (RC) have weights of 0.15 and 0.14 respectively, and main stream slope (MSS) and drainage area (DA) have weights of 0.11 and 0.1. Annual precipitation (AP) and drainage density (DD) have the lowest weights, with values of 0.09 and 0.07 in this classifier. All these low weight variables have a relatively equal significance in the classification of watersheds. After determining the most significant watershed characteristics and the best classifier for watershed classification, the geomorphic model is then used for floodplain mapping using PTBC. To generate a probabilistic 100-year floodplain map for a given watershed, *H* raster is calculated from a DEM and stream network. Furthermore, seven watershed characteristics (Table 4-7) are calculated and used as input to the trained random forest classifier. The random forest classifier estimates three class probabilities for three class labels. The three *TrH* ranges corresponding to the class labels in lookup table (Table 4-3), the three probabilities from the random forest result as well as *H* raster are used as the main inputs to PTBC to generate the final floodplain maps (Figure 4-2).

4.5.2 Validation Phase 1: Comparison with FEMA Maps

In order to validate the effectiveness of the geomorphic model for floodplain mapping, the floodplain of 145 watersheds from various geographical regions is mapped by using the proposed model. Figure 4-4 illustrates the position of the validating watersheds in the OFI-UFI space. The performance of predicted flood extents for each watershed can be evaluated by using the distance of the watershed position in the OFI-UFI from the origin. The high density of points near the origin

in Figure 4-4 shows that the predicted flood extents by the proposed model is satisfactory compared to the FEMA reference maps. In order to quantify the validation results, the average OFI and UFI with 95 percent confidence interval is determined. The results show that the average of overprediction and underprediction for watersheds in CONUS vary from 12.6% to 16%, and 12.2% to 15.2% respectively.

In addition to the OFI-UFI plot, the high frequency of AUC values around 1 for all watersheds as shown in Figure 4-5 demonstrates the ability of the geomorphic model to reliably create 100-year floodplain maps. Based on the results, 81% of predicted maps have an AUC>0.9, and 14% fall in the range of 0.8-0.9. Only 5% of watersheds with AUC less than 0.8 have poor estimation of flood extent. In order to check the overall fit between the probabilistic maps and the FEMA maps, the flood probability values for all cells in 145 validating watersheds are rounded to one decimal digit numbers (0, 0.1,0.2,…,0.9,1), and their occurrence inside the FEMA floodplains and FEMA non-flood areas are presented in Figure 4-6. This figure shows that 75% of reference non-flood areas include cells with zero probability of flooding. Moreover, around 75% of reference floodplain area includes cells with probability of 0.9 or 1. This proves that almost 75% of entire validating watersheds has a complete fit with FEMA map. The advantage of probabilistic map can be explained by looking at the 25% remaining cells. In a deterministic map, if 75% of cells predict truly, the remaining 25% are definitely the errors. However, these probabilistic maps show that less than 5% of cells have been predicted incorrectly (cells with probability of zero inside the flood area or cells with probability of one inside the non-flood area), and more than 20% of cells show probability of flooding between zero and one. These 20% of cells are areas with some level of uncertainty that need further investigation before deciding their flooding status. The uncertainty for making a decision will increase as the probabilities move to the middle of the range (0.5). On the contrary, recognizing the flood and non-flood areas for small or large probabilities would be easier. Therefore, a probabilistic presentation of flood extent helps decision makers to recognize that areas near the boundary of floodplains need further evaluation to decide their flooding status.

Spatial distribution of poorly predicted areas among validating watersheds in CONUS in Figure 4-7 shows that results are not affected by the location. Some of the poorly predicted watershed lie next to a well-predicted watershed as illustrated by examples for New York, Tennessee, and Texas

in Figure 4-7. The seven watershed characteristics of these poorly predicted watersheds are also compared with those of the training watersheds to examine any pattern in their characteristics. Figure 4-8 shows that the watershed characteristics for poorly predicted areas lie randomly with a wide variability without any peculiar pattern. Further investigation of these poorly predicted watersheds reveals that: (i) their topography is heterogeneous (e.g. two cases in Texas and one case in Tennessee); (ii) they are located in coastal areas with nested stream networks (e.g. two cases in California); or they are located in urban areas with artificial channels and many riverine structures (e.g. one case in Indiana).

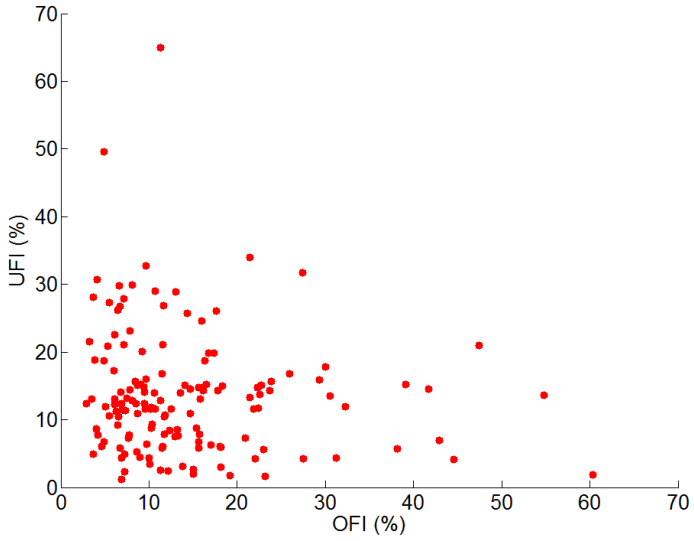


Figure 4-4 Performance of 145 validating watersheds (red dots) in OFI-UFI space after comparing with FEMA floodplain maps (Validation Phase 1)

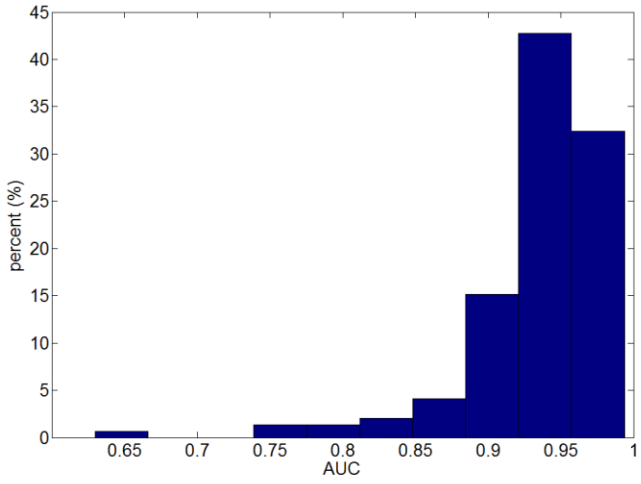


Figure 4-5 Histogram of AUC for 145 validating watersheds

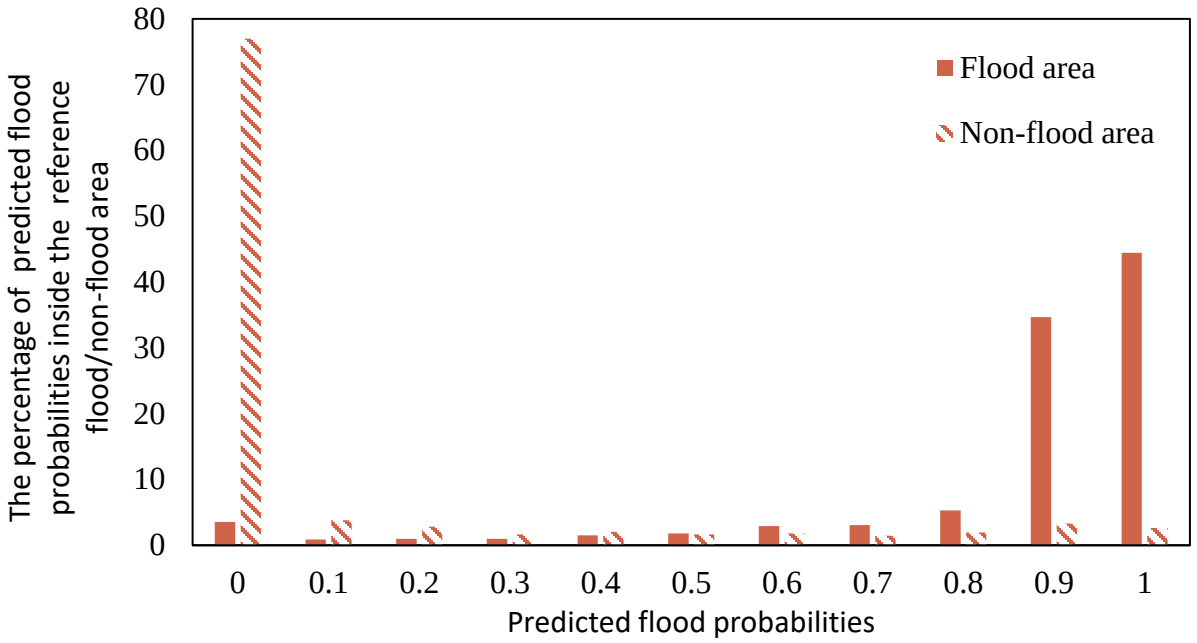


Figure 4-6 Distribution of predicted flood probabilities inside the flood and Non-flood area of reference map

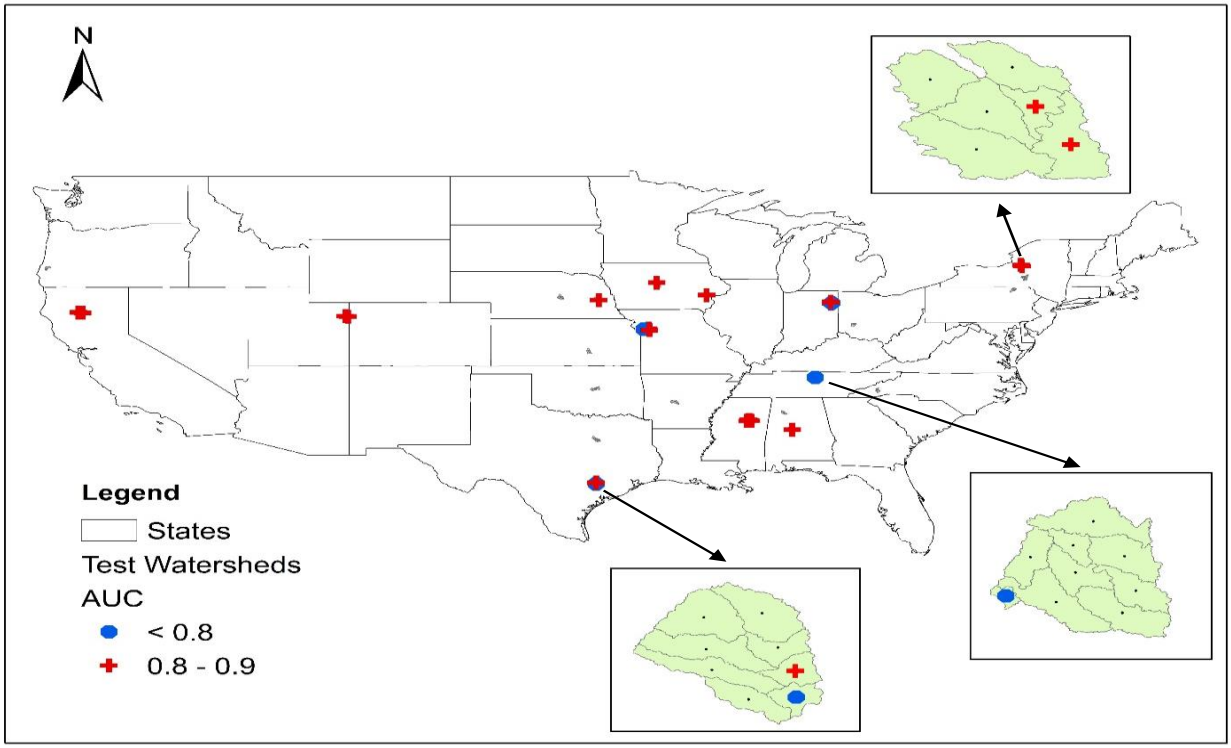


Figure 4-7 Distribution of poorly predicted watersheds in CONUS

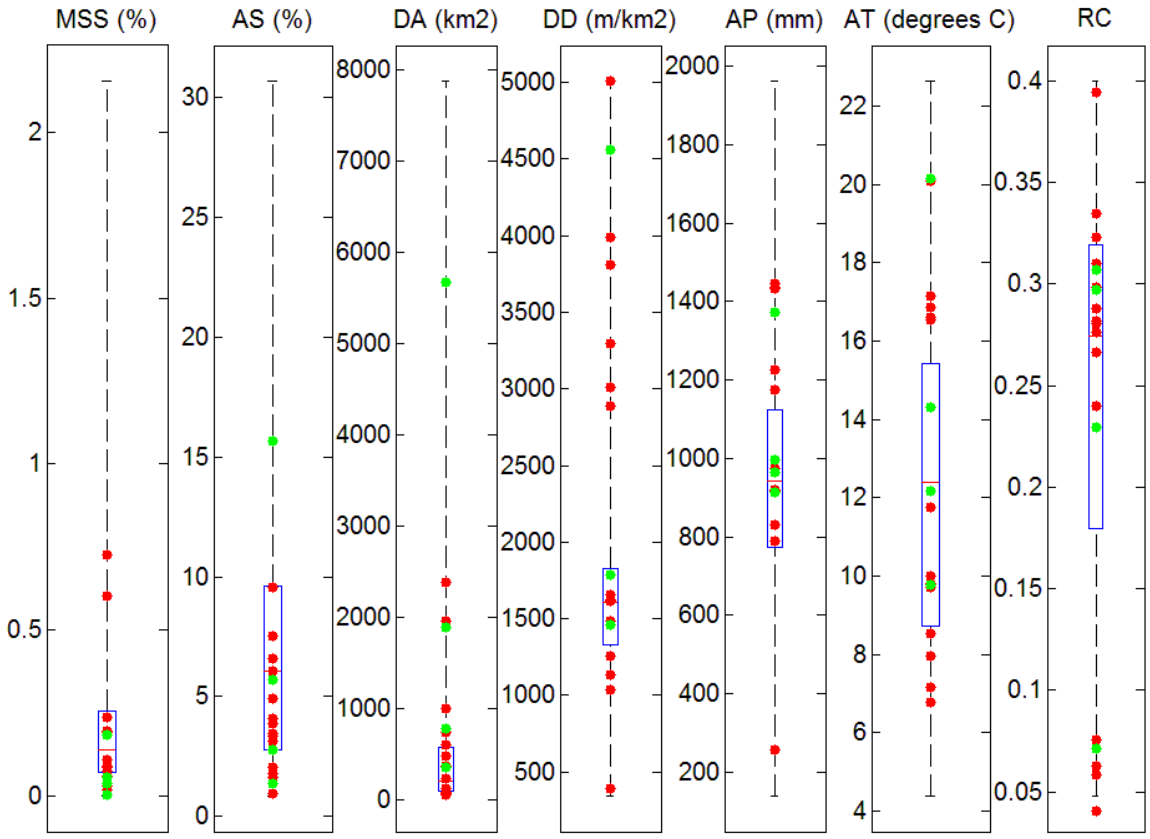


Figure 4-8 Distribution of watershed characteristics used for all training watersheds showing the minimum, 25 percentile, median, 75 percentile and maximum values together with values of poorly predicted watersheds presented by dots. The red dots refer to predicted watersheds with 0.8 < AUC < 0.9 and the green dots corresponds to poorest watersheds with AUC<0.8.

4.5.3 Validation Phase 2: Comparison with HEC-RAS

The floodplain maps are also compared with HEC-RAS generated inundation maps in 15 data-scarce regions that do not have FEMA maps. According to the watershed characteristics for these 15 watersheds, as presented in Table 4-8, seven watersheds fall in the mid-latitude regions, and the remaining eight watersheds fall in the flat and mountainous areas with respect to the average slope. The corresponding class probabilities generated from Random Forest classifier for these watersheds are listed in Table 4-8. The OFI and UFI for these areas (presented in Table 4-10 and Figure 4-9), calculated by comparing the predicted flood maps with HEC-RAS generated inundation for a single river reach show that ten out of 15 areas have good prediction. Two reaches show underprediction (UFI > 30%) and three show overprediction (OFI >30%). The average of OFI and UFI for these fifteen watersheds considering the 95 percent confidence interval are in the

range of 7.7 to 24.8 % and 2.9 to 20.3 % respectively. The larger confidence interval of results at validation phase 2 compared to phase 1 can be explained by the smaller sample size in phase 2 (15 watersheds compared to 145 watersheds used in phase 1).

In Figure 4-10 to Figure 4-11, the probabilistic 100-year floodplain maps for three watersheds in Wyoming, South Dakota, and Idaho are presented. The results for Wyoming and South Dakota are the examples of well-predicted watersheds. Fifty percent of flood extent is underpredicted for the Idaho reach (Figure 4-12) because the estimated *TrH* from the geomorphic model is lower than what it should be. The Idaho reach should belong to Class 1 (with average slope, AS = 40.42%) due to its hilly terrain, but the random forest classifier puts the Idaho watershed belonging in both class 1 and 2 with probability of 0.6 and 0.4, respectively (Table 4-9). This example demonstrates the limitation of the random forest in correct classification of this watershed. The performance of random forest can be improved by adding more training data to capture the variation of watershed characteristics and generate a better model fit to data. Also, there are other factors in addition to the seven selected watershed characteristics affecting the *TrH* which have been neglected in classification. These factors can be more dominant in areas such as Idaho.

Table 4-8 Watershed characteristics for 15 validating watersheds

No	HUC12	State	MSS	AS (%)	DA (m^2)	DD	AP (mm)	AT ©	RC
1	150602030605	Arizona	0.0062	15	500	0.003	387	20	0.397
2	110200020704	Colorado	0.0056	3.51	327	0.0014	350	10.7	0.35
3	140500050108	Colorado	0.0069	24.14	1143	0.0031	500	3.6	0.348
4	031102010505	Georgia	0.0001	0.47	3238	0.001	1338	19.5	0.222
5	051401040706	Indiana	0.0007	5.95	338	0.0047	1152	12.1	0.277
6	030501060305	South Carolina	0.0003	9.3	2025	0.0018	1205	15.9	0.308
7	140600080205	Utah	0.0009	35.33	5070	0.0011	212	10.1	0.343
8	020802040501	Virginia	0.0004	8.24	255	0.0017	1082	13.4	0.292
9	170200160505	Washington	0.0019	4.48	1645	0.001	210	10.8	0.215
10	140401040110	Wyoming	0.0007	2.7	997	0.0018	225	2.6	0.353
11	101401021103	South Dakota	0.0007	7.47	3013	0.0018	428	8.7	0.316
12	170602080412	Idaho	0.0032	40.42	937	0.0019	640	3.5	0.346
13	170900050604	Oregon	0.0007	4.67	4018	0.0027	1118	11.3	0.162
14	180400080803	California	0.0003	1.46	5203.6	0.0012	296	16.4	0.103
15	101302010107	North Dakota	0.0007	6	271	0.0016	417	5.4	0.287

Table 4-9 Class probabilities generated by random forest for 15 validating watersheds

No	HUC12	State	Class 1	Class 2	Class 3
1	150602030605	Arizona	0.5	0.5	0
2	110200020704	Colorado	0.1	0.4	0.5
3	140500050108	Colorado	0.5	0.3	0.2
4	031102010505	Georgia	0.4	0.5	0.1
5	051401040706	Indiana	0.1	0.8	0.1
6	030501060305	South Carolina	1	0	0
7	140600080205	Utah	0.9	0.1	0
8	020802040501	Virginia	0.3	0.7	0
9	170200160505	Washington	0.1	0.8	0.1
10	140401040110	Wyoming	0	0.5	0.5
11	101401021103	South Dakota	0.2	0.8	0
12	170602080412	Idaho	0.6	0.4	0
13	170900050604	Oregon	0.3	0.7	0
14	180400080803	California	0	0.3	0.7
15	101302010107	North Dakota	0	0.9	0.1

Table 4-10 Performance of predicted flood extents by proposed model for 15 validating rivers compared with floodplain maps generated by HEC-RAS

Rivers	HUC12	State	UFI (%)	OFI (%)
1	150602030605	Arizona	2.4	24.5
2	110200020704	Colorado	4.7	11.6
3	140500050108	Colorado	2.6	26.1
4	031102010505	Georgia	2.4	31.7
5	051401040706	Indiana	15.7	38.4
6	030501060305	South Carolina	3.9	11.2
7	140600080205	Utah	15.5	3
8	020802040501	Virginia	15.8	0.4
9	170200160505	Washington	7.3	3.5
10	140401040110	Wyoming	1	8
11	101401021103	South Dakota	6.1	6.7
12	170602080412	Idaho	57.5	0.8
13	170900050604	Oregon	7.8	50.4
14	180400080803	California	31.5	11.2
15	101302010107	North Dakota	0.2	15.9

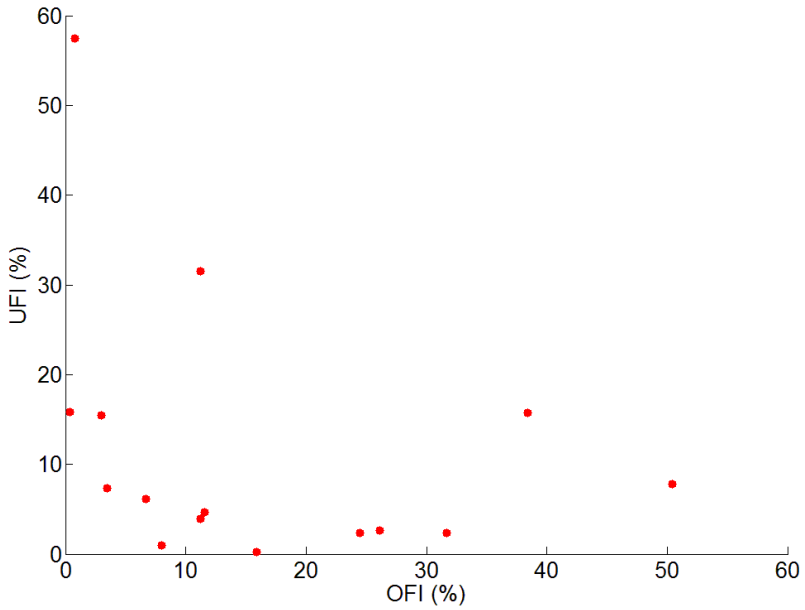


Figure 4-9 Performance of 15 validating rivers (red dots) in OFI-UFI space after comparing with floodplain maps generated by HEC-RAS (Validation Phase 2)

4.6 Discussion and Conclusions

In this study, a geomorphic model for probabilistic mapping of 100-year floodplains in CONUS is proposed by using attributes derived from freely available topography, land use and climate data. Overall results, computed in terms of AUC and UFI-OFI, show that the proposed model provides a relatively reliable and robust method to generate probabilistic 100-year floodplain maps for an entire stream network in a HUC12 unit. The proposed model is scalable to identify floodplains for all stream reaches in the CONUS by delineating floodplains for each HUC12 unit. The proposed model is a fast and cost-effective method for primary estimation of floodplain areas for an entire stream network in any gauged or ungauged watershed. For example, for a HUC 12 unit used in this study with the combined stream lengths in the range of 50-150 km, the proposed approach created the probabilistic floodplain map in 5-10 minutes using a computer desktop with Core i7, 3.6 GHz processor and 16 GB memory (RAM). The computing time also included time of downloading DEM and NHD stream network for the unit (around 2-3 minutes). Creating a probabilistic flood inundation map for the same length of stream network using a hydrodynamic model would take hours or days (including both model setup and running time) depending on the model used. In addition to the actual computing time of a conventional hydrodynamic model, the

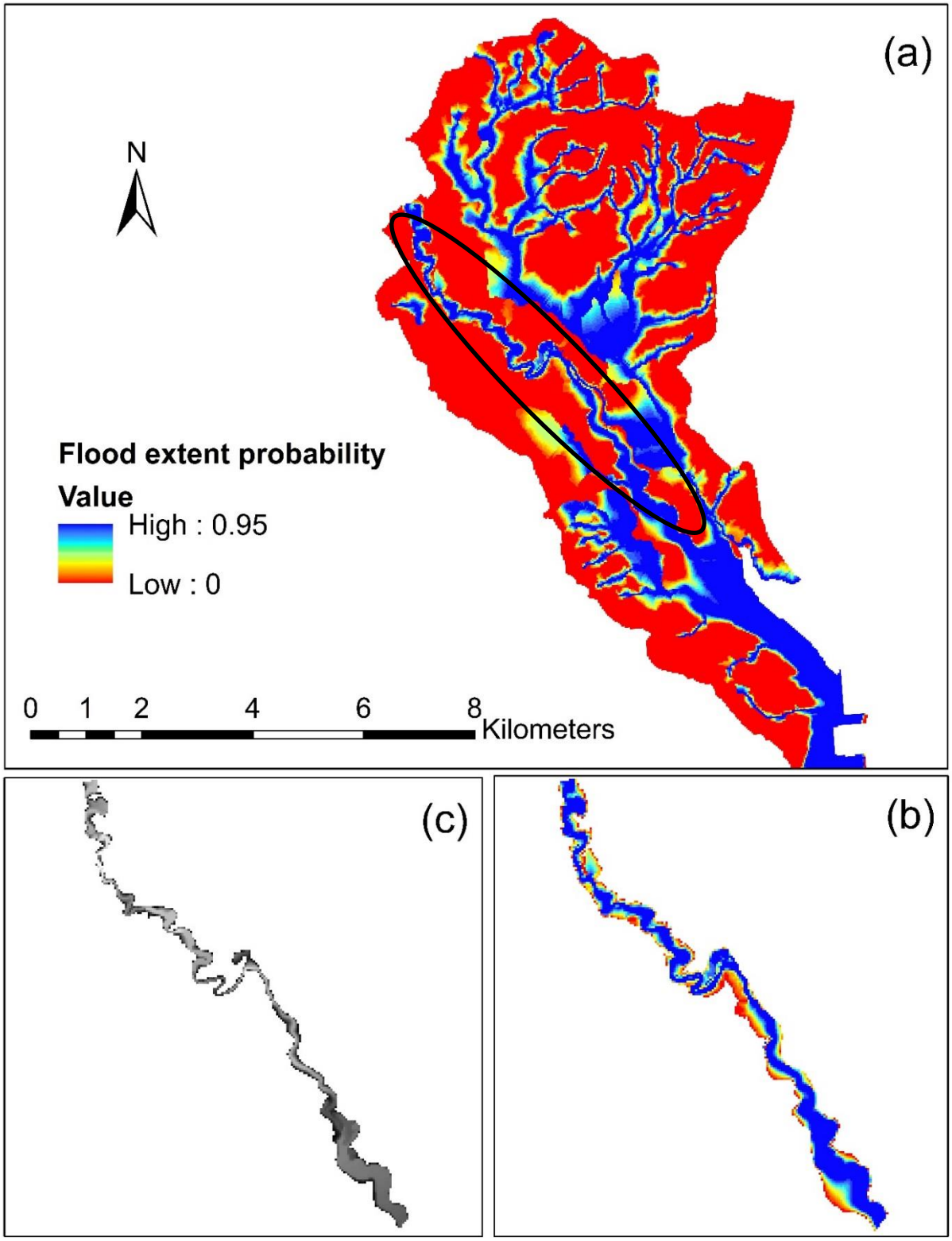


Figure 4-10 Probabilistic 100-year floodplain map generated by proposed model for entire watershed in Wyoming (a): The ellipse highlights the portion of watersheds used for comparison of model prediction (b) with HEC-RAS results (c).

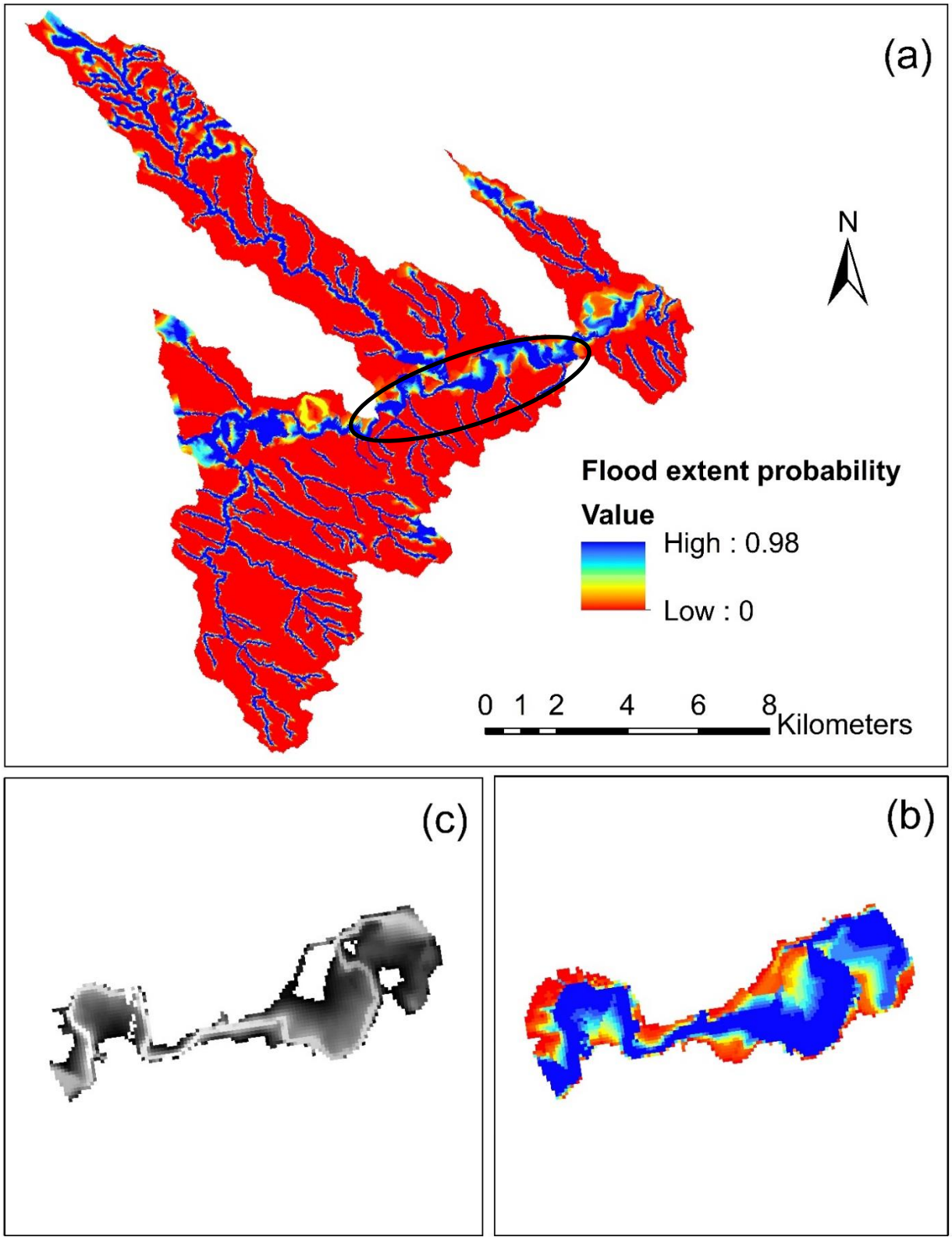


Figure 4-11 Probabilistic 100-year floodplain map generated by proposed model for entire watersheds in South Dakota (a): The ellipse highlights the portion of watersheds used for comparison of model prediction (b) with HEC-RAS results (c).

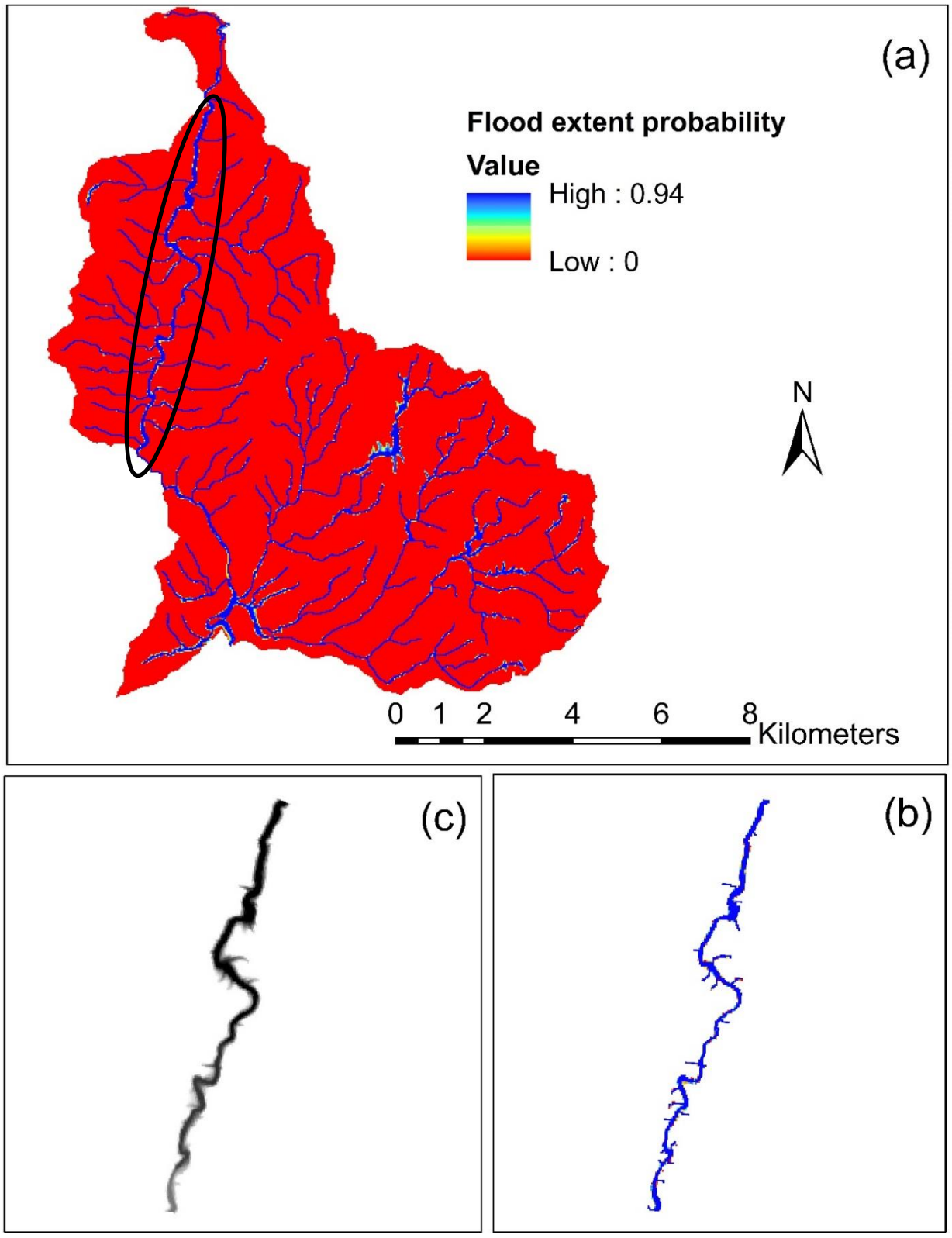


Figure 4-12 Probabilistic 100-year floodplain map generated by proposed model for entire watersheds in Idaho (a): The ellipse highlights the portion of watersheds used for comparison of model prediction (b) with HEC-RAS results (c).

lack of reliable data for all tributaries including 100-year flow and the bathymetry data, and the high cost of field measurement to derive accurate data pose additional challenges in applying conventional probabilistic floodplain modeling approaches for large areas.

The validation results illustrated that around 80% of watersheds are predicted well by the proposed approach in comparison to the FEMA maps. It should be noted that FEMA maps are used as reference only to train and validate the proposed geomorphic approach. It is unrealistic to expect an exact overlap between the FEMA maps and the geomorphic model predicted maps because FEMA maps are generated by hydrologic and hydraulic models that account for accurate hydrodynamics and geometric details. It is also true that the modeling approach used in FEMA mapping has uncertainties related to 100-year flow estimation, model structure and assumptions (Merwade et al., 2008; Saksena and Merwade, 2017, 2015) so some of the discrepancies between flood extent predicted by the proposed approach and FEMA maps could be related to these uncertainties. Similar arguments can be made about the comparison between the proposed approach and HEC-RAS predicted outputs that show around 67% of satisfactory prediction, 20% overprediction and 13% underprediction. The higher rate of overprediction compared to underprediction and the lower fitness with HEC-RAS maps compared to FEMA maps can be explained by two arguments. First, the geomorphic approach makes prediction for the entire stream network including the tributaries, but HEC-RAS maps are created only for a single reach of a river. This scale difference is the major reason of overprediction in most of the selected watersheds. Second, the bathymetry data used in HEC-RAS has more uncertainties than ones used by FEMA. In our HEC-RAS modeling approach, the bathymetry data is generated using DEM and digitizing cross sections on the river. However, most FEMA maps include accurate bathymetry using field measurements. Furthermore, the poor results for the extreme case of Idaho demonstrates the limitation of random forest classifier in true classification of all watersheds. The performance of random forest classifier can be improved by increasing the number of training data to capture more variability in the watershed characteristics.

While floodplain mapping is traditionally being done using computational models, the data and resources need to undertake modeling studies make the task of floodplain mapping difficult in data-scarce low income rural areas. This study is motivated by the desire to make floodplain maps

more accessible in such regions using machine learning techniques. However, the quantity and quality of training data is critical in developing any methods using machine learning. In the proposed approach, the use of FEMA maps, which are not true observations, and somewhat uncertain products, can significantly affect the model performance. Therefore, one of the major limitations of the geomorphic model developed in this study is its dependence on the FEMA maps accuracy. The lack of detailed information in urban areas and exclusion of riverine structures in mapping the floodplains are some limitations of the geomorphic model, these details are not easy to incorporate at river network scale. Detailed hydrodynamic models are more useful for local regions of importance, but large-scale methods such as the geomorphic method proposed here can be more effective for estimating the flood extents in data-scarce regions and rural areas.

The probabilistic watershed classification by random forest classifier, the range of *TrH* values used in the lookup table instead of one certain value, and the PTBC module used to convert these uncertain data to a probabilistic map demonstrate that the model structure is the only source of uncertainty considered in the proposed approach. The other potential sources of uncertainty in the proposed model are associated with two major inputs, topography data (DEM) and the reference maps. As a suggestion for future studies, the uncertainty in topography data can be incorporated into the proposed model by using several DEMs. The quantification of uncertainty in reference maps is a challenging task because the modeling approach used in creating a flood inundation map has several uncertainties, including data sources, model structure and its parameters. However, if the uncertainty in the reference maps is known, it can be incorporated by rearranging the optimum *TrH* and the *TrH* range values determined based on reference maps for training watersheds. The new values could affect the class labels of training watersheds which will produce new watershed classification results in the first module of the proposed model.

It should be stressed that, the continental scale floodplain mapping for the CONUS has been performed by different studies recently, including (Sangwan and Merwade 2015; Wing et al. 2017; Samela et al. 2017). The soil-based approach by Sangwan and Merwade (2015) can generate 100-year floodplain maps for CONUS, but the soil-based approach ignores topographic attributes, which play an important role in forming floodplains. The recent proposed hydrodynamic approach by Wing et al. (2017) is a significant contribution in continental scale floodplain mapping as it

relies on freely available open sourced data for numerical hydrodynamic modeling in such a large scale domain. However, the potential source of uncertainties in simplifying the channel geometries obtained from a DEM without any detailed field measurement, and the errors in estimating the flow rate from regional regression equations significantly reduce the model accuracy. A detailed comparison of accuracy between some of these related studies is not easy as some of them use different performance measures such as the C and F indices compared to UFI, OFI and AUC used in this study. Additionally, the computational units or domains for these studies also vary. For example, this study uses HUC12 as one computational unit, but other studies use county or climate regions or the whole CONUS for creating floodplain maps. However, a simple comparison can be made by considering the fact that our proposed model is developed based on the criteria of C>0.8 and F>0.5 and 80% of watersheds have been predicted well. Considering the validation results from other studies that have an average C and F of around 0.8 and 0.5, respectively, the results from this study are reasonable. The major advantage of the proposed model compared to the conventional hydrodynamic approach is the fact that one can get an acceptable floodplain map in a data-scarce region without investing considerable amount of computational and monetary resources.

Samela et al. (2017) proposed a continental scale geomorphic approach, similar to the one proposed here for the CONUS. Their approach yielded an average AUC of 83.3% for the CONUS, with majority of the areas having an AUC ranging between 80-90 %. In this study, the average AUC is 93.3% for the validating watersheds, with most areas giving an AUC ranging between 90-100%. The increase in accuracy in our approach is attributed to the consideration of heterogeneity in the topography by using HUC 12 for computations compared to HUC 2 by Samela et al who also assumed constant GFI for the entire HUC2. Finally, the proposed geomorphic method is able to create probabilistic presentation of floodplains which is not possible at such a scale from other related studies. The probabilistic presentation of floodplains is more realistic because of the stochastic nature of flood events and the huge uncertainties associated with their predictions. While the probabilistic maps do not account for uncertainties related to rainfall and flow predictions, they consider the uncertainties in the model structure by assigning a range of *TrH* instead and finding the probability of a watershed belonging to different *TrH* classes. In addition, the probabilistic

maps would be more useful for generating flood risk maps and decision making (Alfonso et al., 2016).

The proposed model, like any geomorphic method, considers topography as the key factor in defining the floodplains. In addition, the higher impact of average slope, derived from DEM, on *TrH* variability in CONUS, confirms the dominant role of topography in the utilization and success of the proposed model. Consequently, it is expected that the quality of DEM, including its horizontal resolution and vertical accuracy can highly affect the model results (Manfreda et al., 2011; Rexer and Hirt, 2014; Saksena and Merwade, 2015; Sanders, 2007; Yamazaki et al., 2012). In this study, the USGS NED was used to generate the floodplain maps for CONUS because of its higher quality compared to DEMs provided by Shuttle Radar Topography Mission (SRTM) (Gesch et al., 2002; Sanders, 2007). The availability of higher quality DEMs, such as NED with 1/9 arc second resolution or LIDAR data, in the future will certainly improve the proposed model performance significantly.

Looking into the characteristics of the poorly predicted watersheds, it is found that proposed model is not influenced by any particular topographic, climatic or land use setting. A uniform distribution of poorly predicted watersheds across the CONUS also shows that the proposed model is not affected by the geographic loactions of the watersheds. However, a closer look into poorly predicted watersheds revealed that watersheds with extreme topographic heterogeneity performed relatively poorly. For future studies, defining and adding a new morphologic index, which explains the level of topographic heterogeneity, to the current seven watershed characteristics may improve the performance of the proposed model in such regions. In addition, mapping the floodplain in coastal and urban areas needs additional considerations because of different parameters affecting the floodplain in these areas. Therefore, it is highly recommended to exclude coastal and urban watershed from the proposed model application and use separate models developed exclusively for these watersheds.

Overall, the findings from this study suggest that the approach may be extended to floodplain mapping at the global scale because of the strong dependence of *TrH* on topography and its attributes. While good topography data is available in developed nations, developing nations rely on globally available dataset such as SRTM and ASTER DEM. It is known that the accuracy of

globally available DEMs is not as good as the DEM used in this study so the proposed approach will require some modifications to account for the lower accuracy of data at the global scale. Additionally, data-scarce regions will also not have access to 100-year hazard maps for training, and in such cases other resources including the global flood map repositories (e.g. the floodplain maps created in 19 European countries and Japan (Van Alphen and Passchier, 2007)) and satellite derived flood inundation information may be used to train and validate the geomorphic model.

CHAPTER 6. PROBABILISTIC FLOODPLAIN MAPPING USING *HAND*-BASED STATISTICAL APPROACH

6.1 Abstract

Detection of 100-year floodplains is one of the major tasks in flood risk management. In recent years, a variety of DEM-based methods have been developed for preliminary estimations of 100-year floodplains over large regions. The higher efficiency of these methods for large-scale problems and data-scarce regions compared to the conventional hydrodynamic methods is a big advantage. However, unlike considerable advances in the field of probabilistic mapping by hydrodynamic models, these methods are mostly deterministic and cannot provide a probabilistic presentation of the floodplains. In this study, a new method is proposed to combine both advantages of probabilistic mapping compared to deterministic ones and DEM-based methods against conventional models. This method includes a probabilistic function, which uses a morphologic feature, Height Above Nearest Drainage (*HAND*), as the independent variable. *HAND* is defined as the difference in elevation between a given point and the nearest stream based on the flow direction and can be calculated from a Digital Elevation Model (DEM). The parameters of the probabilistic function are determined by using a heuristic optimization algorithm named Particle Swarm Optimization (PSO) by minimizing the error of a predicted 100-year floodplain map compared to a reference map. The results illustrate that a linear function with one parameter is an appropriate function for the study site. In addition, a comparison of the proposed method with its deterministic version demonstrates the higher effectiveness and reliability of the proposed probabilistic method for a flat watershed where the overpredictions and underpredictions generated by a deterministic threshold method are reduced.

6.2 Introduction

Floods are one of the most frequent natural disasters in the world, leading to huge economic and human losses annually (Baker, 1994). Considering the disastrous impacts of floods on human lives and property, there is a growing interest to perform flood risk management projects for individual streams as well as for entire stream networks in a small watershed or large basin (Moel et al. 2009; Van Alphen et al. 2009). Land-use planners, flood risk managers, emergency response teams,

utility companies, insurance companies and citizens have different stakes and objectives in a flood risk management project. However, one of the key steps in any flood risk management project is the identification of the floodplains. Delineation of floodplains is also vital for many ecological and environmental studies. Flooding plays a vital role in the growth and reproduction of the regional aquatic plants and animals (Walker et al., 1997). It keeps the lateral connection between the river and the floodplain and promotes the transport of nutrients, biota and organic carbon to the floodplains (Walling and He, 1998; Baldwin and Mitchell, 2000; Thoms and Sheldon, 2000; Thoms, 2003). The crucial ecological role of a floodplain as a productive environment is another reason for the increasing attention about the proper delineation of these areas in the last decades.
Many approaches or models exist in the literature for floodplain mapping. The area of the landscape, desired accuracy of the floodplain maps, computational and monetary cost of the modelling and the type of the required maps (deterministic or probabilistic) are some of the factors that dictate the selection of an appropriate model for floodplain mapping. Floodplains can be delineated by using either the conventional hydrodynamic models, or the new generation low-complexity methods (Afshari et al., 2018). Conventional hydrodynamic models delineate the inundation extent by simulating the physics of the stream and using detailed information related to the channel geometry (planform and cross-sectional shape), surface roughness and the riverine structures (Bates and De Roo, 2000; Musser and Dyar, 2007; Tayefi et al., 2007; Kim et al., 2011; Liu et al., 2018). On the other hand, low-complexity methods use easily available data such as a digital elevation model (DEM) or soil maps for preliminary estimation of floodplains over a larger area (Sangwan and Merwade, 2015; Samela et al., 2017; Jafarzadegan et al., 2018).

Floodplain modeling is primarily driven by the need to map flood inundation extent by coupling well-calibrated streamflow forecast models with hydrodynamic models (Wright et al., 2008; Patro et al., 2009; Nguyen et al., 2015; Nguyen et al. 2016). In order to generate flood inundation extents at a regional or continental scale, hydrodynamic models are replaced with the low-complexity methods such as AutoRoute (Follum, 2013; Follum et al., 2017), which relies just on the Manning's equation to get the flood width and the inundation extent. Recently Height Above Nearest Drainage (*HAND*) has been used to map inundation extent by using streamflow forecasts in conjunction with Manning's equation based hydraulic parameters (Maidment et al., 2016). *HAND* is a morphologic raster-based feature (Rennó et al., 2008; Nobre et al., 2011) which is

defined as the elevation difference between a given raster cell and the nearest stream cell into which it drains. In addition to the flood inundation mapping corresponding to a given flow value, the flood inundation maps can be generated for design flows corresponding to different return periods (e.g. 50-year, 100-year, 500-year). These maps are widely produced and utilized in Europe and the United States as the primary component of flood risk management projects (Martini and Loat, 2007; Alphen et al., 2009; Moel et al., 2009; FEMA, 2015). The focus of this chapter is on the mapping of the flood extent corresponding to the 100-year design flow (100-year floodplain). The 100-year floodplain is the area adjacent to a river that will be inundated due to a flood event with 1% chance of annual exceedance**.** The 100-year flood is suggested as a medium frequency flood event and is widely used as a common standard for flood mapping and risk analysis globally (Hazen, 1914; Watt, 2000; Martini and Loat, 2007; Merz et al., 2007; Lóczy et al., 2012). The term floodplain used hereafter in this article refers to 100-year floodplain.

The mapping of floodplains is complicated due to the large uncertainties involved in the overall procedure including the models, their parameterization and data inputs. In addition to the stochastic nature of a flood event, mapping the corresponding flood inundation includes several uncertain components related to precipitation and streamflow data, topographic representation, model structures and geospatial operations (Merwade et al., 2008). Considering all these uncertainties, the results from a deterministic approach, which only accounts for a single system configuration, could be spuriously precise (Beven and Freer, 2001; Bates et al., 2004; Beven, 2006; Di Baldassarre et al., 2010**).** In a probabilistic floodplain mapping approach, Monte-Carlo methods are used to generate an ensemble of results from different combinations of uncertain inputs and model structures. The weighted average of the results from all model configurations are used to derive a probabilistic floodplain map (Aronica et al., 2002; Romanowicz and Beven, 2003; Beven, 2006; Verbunt et al., 2007; Purvis et al., 2008; Sarhadi et al., 2012; Domeneghetti et al., 2013; Neal et al., 2013; Pedrozo-Acuña et al., 2015).

Deterministic 100-year floodplain maps define a rigid boundary where any property just inside the boundary is 100% prone to a 100-year flood event; whereas any property just outside the boundary is 100% safe from 100-year flooding. Because of the uncertainties involved in determining the floodplain boundary, a deterministic floodplain map creates a false sense of flood safety just

outside its boundary, and vice versa. In a probabilistic floodplain map, the chance of flooding is described in terms of probability that decreases as the distance increases from the stream. Considering all the uncertainties involved in predicting the 100-year flood event and mapping the inundation extent, probabilistic presentation of flood extent is sensible compared to a deterministic map representing only one of many "behavioral" model realizations (Bates et al., 2004; Di Baldassarre et al., 2010). These maps are also more reliable for decision making and risk analysis (Alfonso et al., 2016). Despite the benefits of probabilistic maps, the methods used to generate these maps require a lot of computational power for running hundreds to thousands of hydrodynamic simulations. Probabilistic floodplain mapping becomes even more computationally challenging when more streams within a network need to be included. Additionally, absence of detailed bathymetry and hydraulic data for an entire stream network pose challenges in creating accurate hydrodynamic models.

Availability of Geographic Information System (GIS) tools and data, such as DEM, in the last few decades has provided an unique opportunity for developing the low-complexity DEM-based methods for floodplain delineation (Williams et al., 2000; Gallant and Dowling, 2003; McGlynn and Seibert, 2003; Dodov and Foufoula-Georgiou, 2005; Nardi et al., 2006; Grimaldi et al., 2013; Nardi et al. 2013; Papaioannou et al., 2015; Teng et al., 2015; Clubb et al., 2017; Jafarzadegan and Merwade, 2017; Jafarzadegan et al., 2018). Among these DEM-based methods, the ones that are based on a morphologic classifier to separate the floodplain from non-flooded areas have shown to be more accurate and computationally efficient (Manfreda et al., 2011; Degiorgis et al., 2012; Degiorgis et al., 2013; Manfreda et al., 2014; Manfreda et al., 2015; Samela et al., 2015; Samela et al., 2017). These methods use a DEM to compute the *HAND* raster, which is then classified in two steps. First, a reference map is used to find a threshold on *HAND* (calibration stage), and then all cells that have *HAND* values less than the threshold are marked as belonging to the floodplain (prediction stage).

The DEM-based floodplain mapping methods discussed earlier, are good alternatives for preliminary estimation of floodplains over a larger spatial domain, but they are mostly used to generate deterministic maps which is a big drawback compared to the conventional hydrodynamic models that can also produce probabilistic floodplain maps. The objective of this chapter is to

overcome this drawback by proposing a *HAND*-based statistical approach to probabilistic floodplain mapping. The proposed method uses the main idea of classification based on the morphologic feature *HAND*, but instead of finding a threshold for *HAND* in the calibration stage, a probabilistic function is estimated for the given watershed. This function is then used to predict the probabilistic map in the prediction stage. We are aware that probabilistic floodplain mapping is not new using hydraulic models, but creating probabilistic floodplain maps over large regions is not feasible using hydraulic models due to data and computational requirements (Aggett and Wilson, 2009). The proposed approach is novel in the sense that a well accepted *HAND*-based method is modified to create a probabilistic version for mapping floodplains over large regions. The proposed method requires some reliable floodplain maps, which are usually generated by existing hydraulic models, for calibration. However, this method has the advantage of being calibrated on a small portion of watershed to generate floodplain maps for the entire stream network. The probabilistic map generated by this method does not directly account for common uncertain variables in hydrologic or hydraulic simulations, but it provides greater confidence by assigning probabilities to the areas that are completely predicted wrongly by deterministic methods. In addition, the probability of flooding a given point can be used to find the risk measures directly.

6.3 Dataset and Study Area

This study uses the Middle Neosho watershed, a relatively flat watershed in south Kansas, USA, for developing and testing the proposed probabilistic floodplain mapping approach (Figure 6-1). This watershed is selected for the following two reasons: (i) floodplain mapping using the morphologic feature *HAND* has shown relatively poor performance in flat watersheds (Manfreda et al., 2015; Jafarzadegan and Merwade, 2017; Samela et al., 2017). Thus, developing and testing the proposed approach in a flat watershed will highlight the improvements over the deterministic approach; (ii) Streamflow records at one of the stream gauges (United States Geological Survey (USGS) # 07183500) in this watershed show that this region is prone to frequent floods (Flood stage is 6.4m at this gauge station). The flood stage defined by USGS is the stage at which overflow of the natural banks of a stream begins to cause damage in the local area from inundation (flooding). Considering the number of days that gauge height is higher than flood stage, this gauge has experienced flooding for more than 120 days since 2007. The high intensity rainfall events occurring in the late Spring and summer (Monthly average precipitation higher than 120mm) as

well as the soil type in the area which is predominantly categorized as hydrologic group D (soil with least infiltration rate and highest potential for runoff) are the major reasons that contribute to flooding in this area. The highest peak discharge recorded at this gauge is 11610 m^3/s in 1951. The Neosho River, which is the major river in this watershed, and its tributaries drain into Grand Lake in Oklahoma. Grand Lake is a major economic resource for Oklahoma, and it supplies the surface water to many communities in this region. The average annual precipitation in this watershed is 1100 mm (Kansas State Research and Extension, 2011). The land use in this watershed is dominated by grassland and cropland, which cover 48 and 33 percent of the total area, respectively. The Middle Neosho watershed is categorized as a HUC8 watershed by USGS (https://water.usgs.gov/GIS/huc.html) which consists of 33 sub-basins (HUC12). The performance of the proposed probabilistic approach is validated by using three different samples, and for each sample, around half of the sub-basins (16 or 17 sub-basins) are selected for training and the rest are used for testing/validation (Figure 6-2).

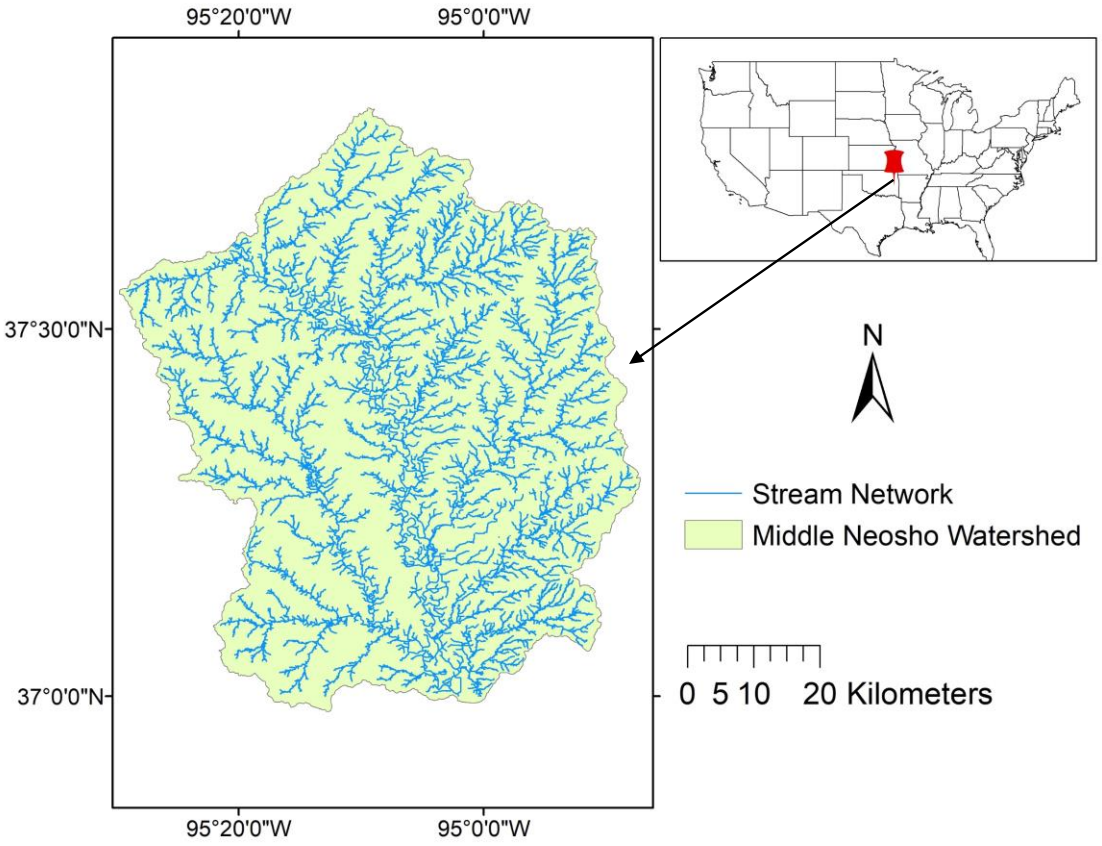


Figure 6-1 The geographic location of Middle Neosho Watershed and its stream network

The main datasets used in this study include the one arc second (approximately 30 m) horizontal resolution DEM and the stream network obtained from the USGS National Elevation Dataset and National Hydrography Dataset, respectively. The probabilistic floodplain maps are developed and validated by comparing against the Flood Insurance Rate Maps (FIRMs) provided by the Flood Emergency Management Agency (FEMA). FEMA has invested billions of dollars in the last decades to create the most updated floodplain maps within an appropriate range of accuracy (Maidment, 2009). These maps are produced using field measurements related to detailed river bathymetry data and riverine structures. FEMA uses calibrated HEC-RAS, a well-known hydraulic model, to delineate floodplain maps for most of the areas in the U.S. Depending on the modeling approach and the accuracy of input data used for generating these maps, accuracy of the FEMA FIRMs is variable across the U.S. It is true that these maps are not observed floodplain maps and have a certain amount of uncertainties. However, FEMA FIRMs are the only well-documented, and free source of floodplain maps in the U.S. Therefore, these maps are used as the reference maps in this study. In sample 1, a combination of main rivers and tributaries are randomly selected for training and testing. In sample 2, the model is trained mostly on the main rivers and tested on the tributaries. On the contrary, the training and testing data in sample 3 are the tributaries and the main rivers, respectively.

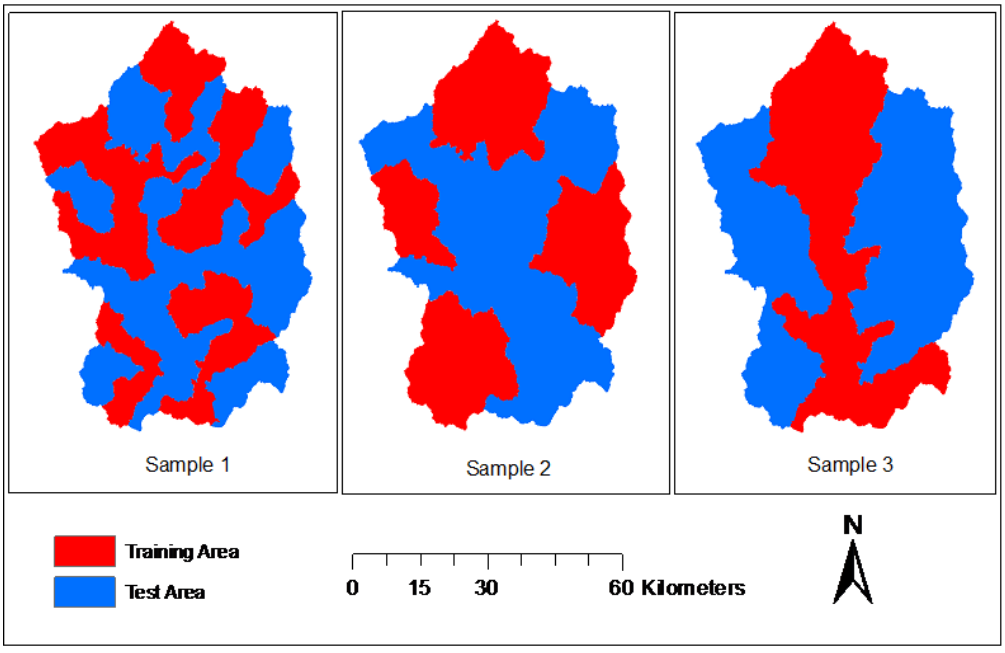


Figure 6-2 Distribution of training and test areas for three different samples used in this study

6.4 Methodology

This study defines a probabilistic floodplain map using a floating point raster in which each cell has a value ranging from zero to one that corresponds to the probability of being inundated from a 100-year flood event. A deterministic floodplain map in this study is also described using a raster in which the cells contain a value of either zero or one to represent non-flooded and flooded areas, respectively. In order to generate the probabilistic floodplain raster, a method based on finding a probabilistic function using *HAND* is proposed to estimate the probability of flooding.

6.4.1 Computing the HAND Raster and its Probabilistic Function

The first step in generating the probabilistic floodplain map is to compute the *HAND* raster using the DEM and the stream network for the study watershed. The *HAND* raster is generated through the following three steps: First, a flow direction raster is computed from the DEM (Figure 6-3a) using the D8 method (Greenlee, 1987; Jenson and Domingue, 1988) to determine the flow from each cell to one of its neighboring cells (Figure 6-3b). Next, the flow direction raster is used to find the nearest stream cell for each cell (Figure 6-3c). It should be noted that the term 'nearest stream cell' in the definition of *HAND* refers to the first stream cell into which the cell flows and it is not the nearest based on the Euclidean distance. The raster in Figure 6-3c shows the coordinates (row, column) of stream cells drained by each cell. For example, the DEM cell with the value of 10 drains to a stream cell with the value of 1, located in the third row and the second column (3,2). Finally, the elevation of the nearest stream cell is deducted from the elevation of each cell to get its corresponding *HAND* value (Figure 6-3d).

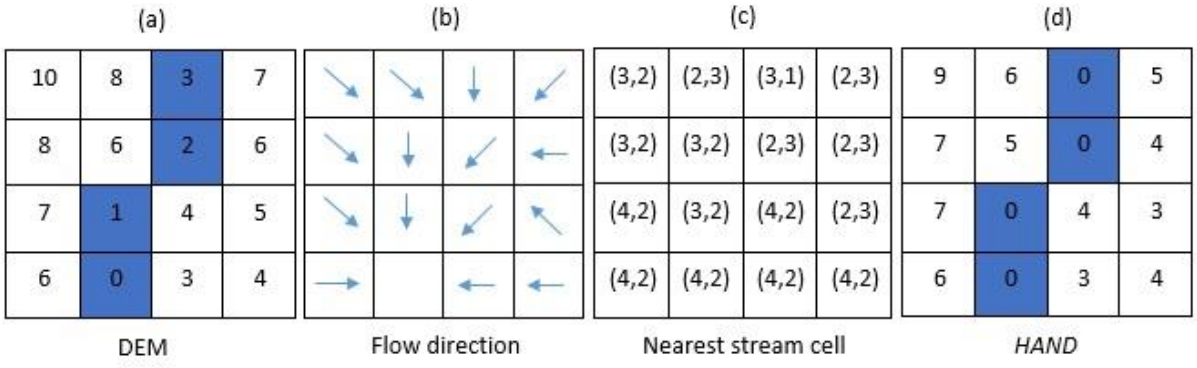


Figure 6-3 Calculation of HAND raster for a hypothetical DEM and stream (blue cells): A DEM (a) is used to generate flow direction (b). Using flow direction, the coordinates (row, column) of the nearest stream cell, drained by each cell, are determined (c). The final output is HAND raster (d) created by deducting the elevation of nearest stream cell from DEM

After the *HAND* raster is computed, a probabilistic function (φ) is determined that uses *HAND* as the independent variable to find the probability of flooding for a given cell. For a deterministic mapping approach, instead of finding a probabilistic function (φ), the best threshold on *HAND* (*TrH*) is determined. This threshold is found by minimizing the error of the predicted floodplain map compared to a reference floodplain map. The deterministic approach assumes that a single *TrH* is applicable for the entire stream network. Considering the heterogeneity of stream morphology within a watershed, the *TrH* can be different for each stream. That is why *TrH* is used as a random variable with a probability density function for probabilistic mapping. The deterministic threshold method assigns binary numbers (flooded as one and non-flooded as zero) to all cells by comparing the *HAND* values of cells with *TrH*. This rule can be reformatted as follows:

$$pr(TrH > HAND) = \{0,1\} \tag{6-1}$$

In Equation 4-1, $pr(TrH > HAND)$ presents the probability of flooding for a given cell when TrH is greater than the *HAND* value for that cell. For the probabilistic approach, assuming *TrH* as a random variable, the cumulative distribution function (*F*) and the probability function (φ) are defined by Equations 4-2 and 4-3, respectively

$$F_{TrH}(HAND) = pr(TrH < HAND) = \int_0^{HAND} f_{TrH}(u)du \tag{6-2}$$

$$\varphi(HAND) = pr(TrH > HAND) = 1 - pr(TrH < HAND) = 1 - F_{TrH}(HAND) \tag{6-3}$$

where $f_{TrH}(u)$ is the probability density function (PDF) of the random variable (*TrH*) and can be calculated by taking the derivative of the cumulative distribution function (CDF). Therefore, $F_{TrH}(HAND)$ is the CDF for random variable *TrH* that is less than a given value of *HAND*. $\varphi(HAND)$ is the probabilistic function used to convert the *HAND* into a probabilistic map directly and refers to the probability of flooding using the *HAND* value for any particular cell. In this chapter, the parameters of probabilistic function φ are determined by minimizing the error between predicted probabilistic floodplain extents and the available FEMA FIRMs. The estimated probabilistic function can be converted to the CDF and PDF (Equations 4-2 and 4-3).

In Figure 6-4, the function φ, CDF and PDF for a deterministic floodplain mapping approach is presented. The PDF shows that when a single *TrH* value is used, φ takes the form of a step function for a given cell i such that when $HAND < TrH$, probability of flooding is 1, and zero otherwise. In the probabilistic approach, the φ function takes the form of a descending curve because the points with higher *HAND* values should be less prone to flooding. Based on this logic, five potential probabilistic functions for φ are defined. In Figure 6-5a simple linear probabilistic function is used with a single parameter H1 that defines the point when the probability of function becomes zero (L1). This PDF uses a uniform distribution for *TrH*. Figure 6-5b uses a two parameter (H1 and H2) linear function for φ where the parameters H1 and H2 define when the probability will be less than one and greater than zero, respectively (L2). The PDF for *TrH* is still a uniform distribution, but it is shifted to the right compared to the first function. It should be noted that L1 is a specific case of L2 ($H_1 = 0$). The third function is a combination of two linear lines with different slopes. The PDF consists of two different uniform distributions, and the φ function is estimated by finding three parameters (H_1, H_2, α) as shown in Figure 6-5c (L3).

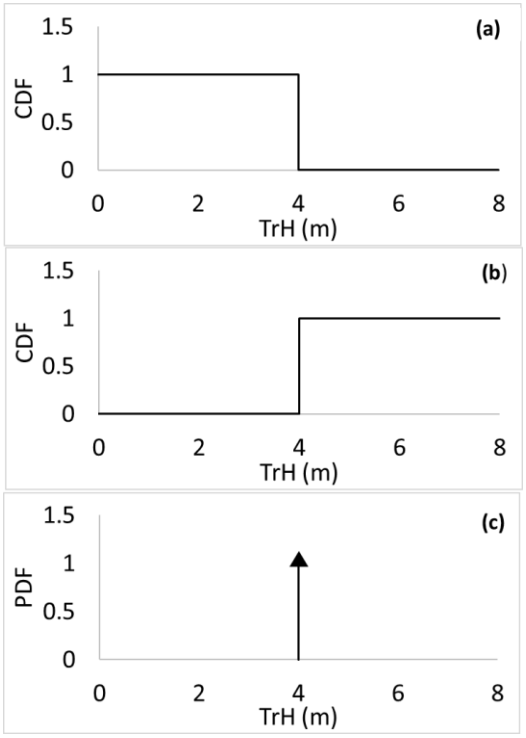


Figure 6-4 Template of φ function (a), CDF (b) and PDF (c) for deterministic floodplain mapping approach

Two other φ functions corresponding to the standard distributions namely, lognormal and gamma, are also considered. Given the application of lognormal distribution in fitting hydrological variables such as flood volume, flood peak discharge and rainfall (Hazen, 1914; Chow, 1954), and the fact that *TrH* values are always positive, lognormal is considered as an appropriate candidate. Similarly, considering the exponential and descending shape of the φ function (1-CDF), Gamma distribution is another good option for probabilistic mapping using *TrH*. Using Equations 4-2 and 4-3, the φ function for these two distributions are determined as follows:

Lognormal:

$$\varphi(HAND) = 0.5 - 0.5\, erf\left(\frac{\ln(HAND)-\mu}{\sqrt{2}\sigma}\right) \quad (6\text{-}4)$$

$$\text{erf}(a) = \frac{2}{\sqrt{\pi}}\int_0^a e^{-x^2} dx \quad (6\text{-}5)$$

Gamma:

$$\varphi(HAND) = 1 - \frac{1}{\Gamma(k)}\gamma(k, \frac{HAND}{\theta}) \quad (6\text{-}6)$$

$$\Gamma(a) = \int_0^\infty x^{a-1}\, e^{-x} dx \quad (6\text{-}7)$$

$$\gamma(a,b) = \int_0^b x^{a-1}\, e^{-x} dx \quad (6\text{-}8)$$

To use φ function for lognormal distribution (Equation 4-4), the Gauss error function (erf) should be calculated (Equation 4-5). The φ function corresponding to the Gamma distribution (Equation 4-6 is determined by calculating the complete and the lower incomplete gamma functions ($\Gamma(\text{a})\,, \gamma(a,b)$) respectively (Equations 4-7 and 4-8). A template of these two functions as well as their CDF and PDF are presented in Figure 6-6 (LN, G). The mean and variance of lognormal distribution (μ and σ), and the k and θ parameters for gamma distribution must be determined to define the φ function. Table 6-1 provides a summary of the five alternative φ functions and their parameters for probabilistic floodplain mapping.

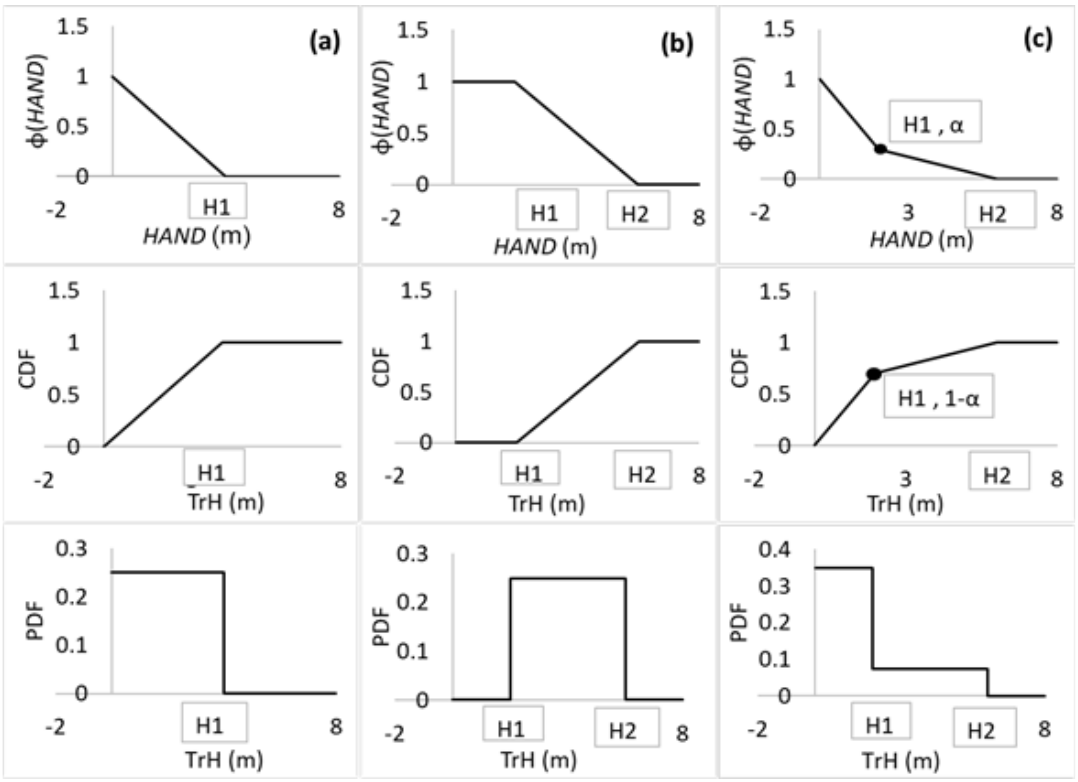


Figure 6-5 Template of three different φ functions, L1 (a), L2 (b) and L3 (c), with their corresponding CDF and PDF

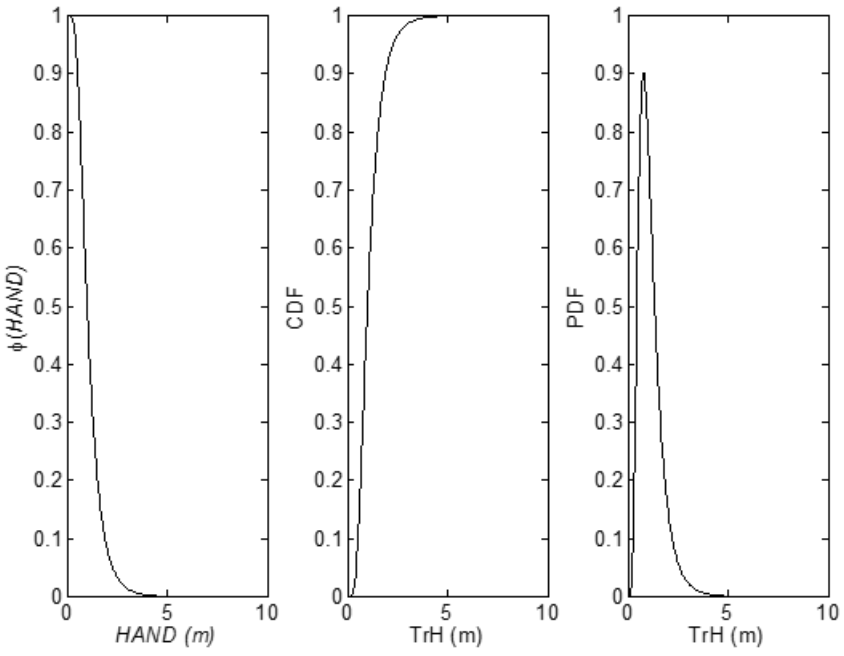


Figure 6-6 Template of φ function, CDF and PDF for Lognormal or Gamma distributions

Table 6-1 Five potential φ functions for probabilistic floodplain mapping

Name	φ functions	PDF	Parameters
L1	Linear with one parameter	Uniform	H_1
L2	Combination of step and linear with 2 parameters	shifted Uniform	H_1, H_2
L3	Two linear line with three parameters	Two uniforms	H_1, H_2, α
LN	Conversion of log-normal	Log-normal	μ, σ
G	Conversion of Gamma	Gamma	k, θ

6.4.2 Parameter Estimation for φ Function (Calibration)

The unknown parameters of the five φ functions presented in Table 1 should be estimated by solving an optimization problem, which minimizes the total error of predicted flood extent compared to the reference map. This nonlinear optimization problem is defined below

$$arg \min_{\beta_i}(error) \tag{6-9}$$

$$error = \frac{\sum_{i=1}^{N}(1-P_i)^2+\sum_{j=1}^{M}P_j^2}{N+M} \times 100 \qquad i \in F,\ j \in NF \tag{6-10}$$

where F and NF refer to the flood extent of the reference flooded and non-flooded areas, respectively from the FEMA maps. N and M are the total number of cells inside the F and NF, and P_i, P_i are the probability of flooding for cell i inside F, and cell j inside NF obtained from the predicted probabilistic map. The unknown parameters of φ function are denoted by β_i in this optimization, subject to the constraints presented in Table 6-2. For the first two functions, decision parameters are in order and should be between 0 to 40. For the last two functions, the parameters of distributions are estimated by searching in a range of 0 to 5.

Table 6-2 Constraints of optimization problem for five φ functions

Name	φ functions	Constraints
L1	Linear with one parameter (L1)	$H_1 < 40\ (m)$
L2	Combination of step and linear with 2 parameters (L2)	$H_1, H_2 < 40\ (m)$ and $H_1 < H_2$
L3	Two linear lines with three parameters (L3)	$H_1, H_2 < 40\ (m)$ and $H_1 < H_2$ and $0 < \alpha < 1$
LN	Conversion of log-normal (LN)	$0 < \mu, \sigma < 5$
G	Conversion of Gamma (G)	$0 < k, \theta < 5$

The selection of upper bound (40) basically means the flooding depth cannot be greater than 40 m. Previous work by (Jafarzadegan et al., 2018) shows that the *TrH* for many watersheds across the

U.S. is less than 4 m and 20 m in flat and mountainous watersheds, respectively. Thus, the use of 40 as a bound on H1 ensures that no flood cells are ignored during the optimization. The LN and G with parameters higher than 5 also act as L1 with *H1* > 40. In other words, if a φ function with parameters higher than 5 is used for floodplain mapping, all areas with *HAND* less than 40 m are prone to flooding.

The objective function (Equation 4-10) can only be calculated if the probabilistic map is generated for the entire watershed (P_i and P_j). In a typical optimization problem, the objective function is a function of parameters, so for any set of parameters it can be calculated directly. However, in this problem, first, the probabilistic map is generated from the probabilistic function and then the objective function is calculated using the calculated probabilities. Due to the importance of the objective function calculation as the basis of any optimization problem (the objective function should be calculated many times for a different combination of parameters) and the complexity of the objective function calculation in this problem as a multi-step process, a hypothetical example is provided below to elaborate these steps.

Figure 6-7 illustrates *HAND* raster calculated from a small DEM with only 6 cells as well as a reference floodplain map for this area converted to a binary map. (flood=1 and non-flood=0). Assume a random solution $[\mu, \sigma] = [1.2,0.5]$ has been generated and the objective function of this solution based on LN function should be calculated.

(a)

8	6
9	0.5
12	1.5

(b)

0	0
0	1
0	1

Figure 6-7 The hypothetical *HAND* raster (a) and the reference floodplain map (b)

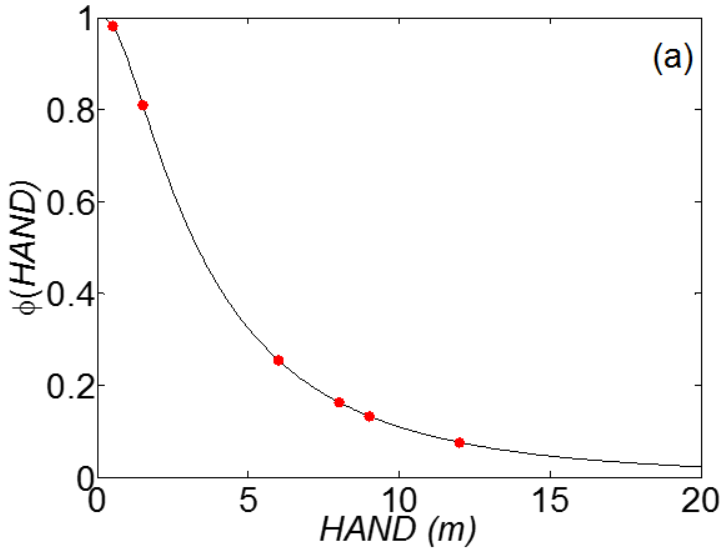
As the first step, the LN function related to the solution $[1.2,0.5]$ is determined using Equations 4-4 and 4-5. Then by using the *HAND raster* values and LN function, the probability values for probabilistic floodplain map are determined. Figure 6-8 presents the created LN function and the

floodplain map probabilities. The points on the function show the *HAND* values on the curve, which result in their corresponding probabilities.

When the probabilistic map is created, Equation 4-10 is used to compare this map with the reference map and find the total error as follows:

$$error = \left(\frac{(1-0.811)^2 + (1-0.982)^2 + 0.164^2 + 0.134^2 + 0.077^2 + 0.255}{2+4} \right) \times 100 = \ 2.5\ \%$$

By solving this optimization problem, the best probabilistic function that provides the highest fitness with the reference map is estimated. Due to the complex format of the objective function that is computed in several steps involving high computational efforts, using a derivative-based optimization method is not feasible for this problem. Among several available optimization methods, heuristic algorithms are known for their efficiency in solving the optimization problems with complicated objective functions. In this study, one of the common heuristic algorithms named Particle Swarm Optimization (PSO) is used to find the optimum parameters of the probabilistic function.



(b)

0.23	0.3
0.2	0.9
0	0.73

Figure 6-8 Function LN created based on the solution [1.2,0.5] with dots presenting the position of HAND raster values on the curve (a) and probabilistic floodplain map provided from position of dots on the LN function (b)

In heuristic methods, one or a set of solutions are generated in the parameter space. A solution is a vector of parameters which can be presented as a point in the parameter space. Based on the structure of optimization algorithm, the solution(s) are updated and move toward the optimum

point in the parameter space. In order to update the positions, the objective function should be calculated for each solution at any iteration.

6.4.3 PSO Algorithm

In the PSO algorithms (Marini and Walczak, 2015), a set of N candidate solutions are generated and the value of the objective function for each solution is calculated. Each candidate solution is called a particle in a D-dimensional space where D is the number of decision parameters. N is the total number of solutions that makes a swarm of particles. After generating a random swarm in the first iteration, the particles update their position in space (Equation 4-11) based on trajectories calculated from Equation 4-12 as follows:

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{6-11}$$

$$v_i(t+1) = w(t+1)v_i(t) + c1\big(p_i - x_i(t)\big)R1 + c2\big(g - x_i(t)\big)R2 \tag{6-12}$$

where $x_i(t)$ and $x_i(t+1)$ show the position of particle i in iteration t and t+1 respectively. $v_i(t+1)$ is the velocity vector of particle i which is updated based on three terms including: velocity in the previous iteration $v_i(t)$, direction of particle toward the "personal best" (p_i) and the "global best" (g). For "personal best", the best position of particle i among the entire positions it has already experienced from the first iteration to the current time is recorded. This position can be updated after a new iteration if the particle moves to a better position. For "global best", the best position among all particles of swarm for a particular iteration is recorded. Other parameters, namely c1 and c2 are named cognitive and social coefficient and usually are selected in range of [0, 4]. These coefficients are used to modulate the effects of second and third term in Equation 4-12. In addition, two random numbers R1 and R2 are generated from a uniform distribution in [0, 1] to consider the stochastic nature of the problem. $w(t+1)$ is called inertia weight and is used to control the impact of previous velocity compared to the other two terms. Based on the findings from several past studies (e.g., Arumugam and Rao, 2006; Bansal et al., 2011; Eberhart and Shi, 2001; Feng et al., 2007; Xin et al., 2009), this study uses linearly decreasing inertia weight as follows:

$$w(t) = w_{max} - \frac{(w_{max} - w_{min})}{t_{max}}t \tag{6-13}$$

In this equation, w_{max} and w_{min} are 0.9 and 0.4 respectively and the t_{max} is the maximum number of iterations in the algorithm.

6.4.4 Floodplain Mapping

In this section, three floodplain maps for the Middle Neosho watershed are generated and compared. After finding the best probabilistic function, the probabilistic floodplain map based on the selected function is generated for the test area. In addition to this map, a deterministic floodplain map is developed. The deterministic mapping by *HAND* is performed by finding the best threshold which minimizes the error of prediction (Figure 6-4). Finally, a reference map obtained from FEMA is used as the third floodplain map.

It should be noted that the floodplain maps provided by FEMA are not for the whole stream networks in the watershed. Therefore, the areas falling outside of FEMA floodplains are a combination of non-flooded and unstudied areas. Degiorgis et al. (2013) defined non-flooded areas as a collection of DEM cells that are: (1) Directly drained by the studied streams; (2) Not located inside the floodplain polygons and (3) Not flowing through the unstudied streams. By using this method, the non-flooded areas are detected and are merged with FEMA polygons to form the entire reference area. The reference area is used to clip all predicted maps to the same extent of flooded and non-flooded cells. In Figure 6-9, the reference area detection for one of the sub-basins in the test area is presented. This figure shows that reference area consists of flooded and non-flooded areas and exclude the non-study area from the sub-basin.

6.5 Results and Discussion

6.5.1 Calibration of φ Functions

The PSO algorithm is run in this study to find the best probabilistic function for floodplain mapping. A swarm of 20 particles is initialized and the particles update their location in the decision space using a maximum of 40 iterations. The performance of PSO for all five functions for sample 3 (Figure 6-2) is presented in Figure 6-10. It illustrates the gradual decrease in the error of predicted flood extents by updating the particle locations (probabilistic function) in the next iterations. In the last iterations, all particles are converged toward the same solution with minimum error (dark blue). Among these five plots, LN and G need less effort to reach the optimal compared to the first three functions with uniform PDFs. LN parameters vary less than G and converge sooner than

other functions. On the contrary, the L3 with more parameters (3 parameter) shows the highest fluctuation with the slowest convergence. These results suggest LN is a relatively robust function with little effort for optimization. However, the relatively less sensitivity of LN to its parameters causes some difficulties for optimization algorithm to find the global optima of LN. For example, for the second sample (Figure 6-2), the PSO is trapped in local optima while finding the LN parameters which results in highest error among all functions. This issue is addressed by increasing the initial number of particles from 20 to 50.

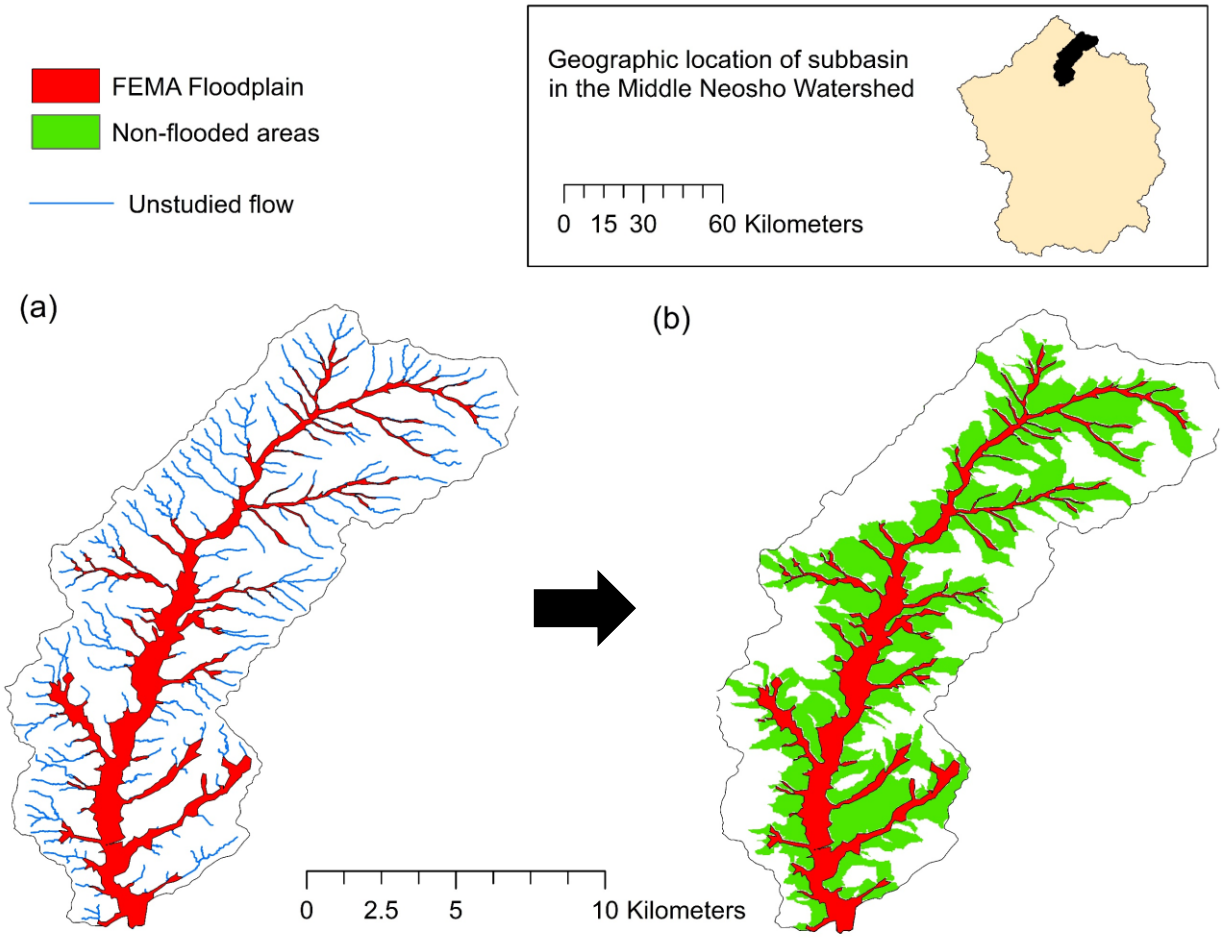


Figure 6-9 Reference area detection by finding the non-flooded areas (b) from studied and unstudied flow and available FEMA floodplains (a)

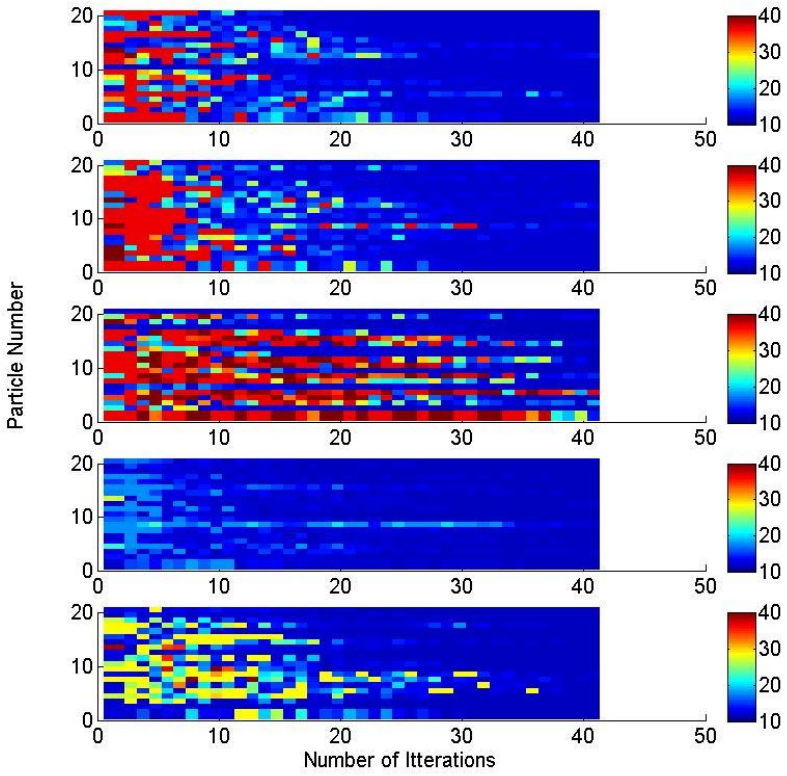


Figure 6-10 Performance of PSO in finding the best probabilistic function for L1 (a), L2 (b), L3 (c), LN (d) and G (e). The color bar shows the value of objective function (error of prediction).

Table 6-3 shows the optimal parameters of all five φ functions for three samples. The optimization results reveal that the first parameter of the L2 function (H1) is zero which means L2 and L1 are the same. Therefore, the L2 function is removed and the four remaining functions of L1, L3, LN and G are used for further analysis.

Table 6-3 The optimal parameters of φ functions for three samples

Functions	L1	L2		L3		LN			G	
Parameters	H1	H1	H2	H1	H2	α	μ	σ	k	Θ
Sample 1	3.91	0	3.91	3.19	9.1	0.14	0.44	0.96	1.2	1.84
Sample 2	3.92	0	3.92	2.34	6.93	0.29	0.45	0.97	1.2	1.85
Sample 3	3.91	0	3.91	2.48	7.19	0.26	0.44	0.96	1.2	1.84

6.5.2 Comparison of φ Functions

In order to compare the performance of these four functions, the probabilistic functions determined from optimization algorithms are applied on test areas, and the error of predicted flood extents

with respect to FEMA is calculated. This process is repeated for three different samples where the distribution of training and test area is changed. The errors corresponding to each training-test Sample and each function is shown in Figure 6-11. This figure demonstrates that the three functions L3, LN and G generate an identical error for floodplain mapping in this study area while the error of predicted flood extents by L1 is slightly higher.

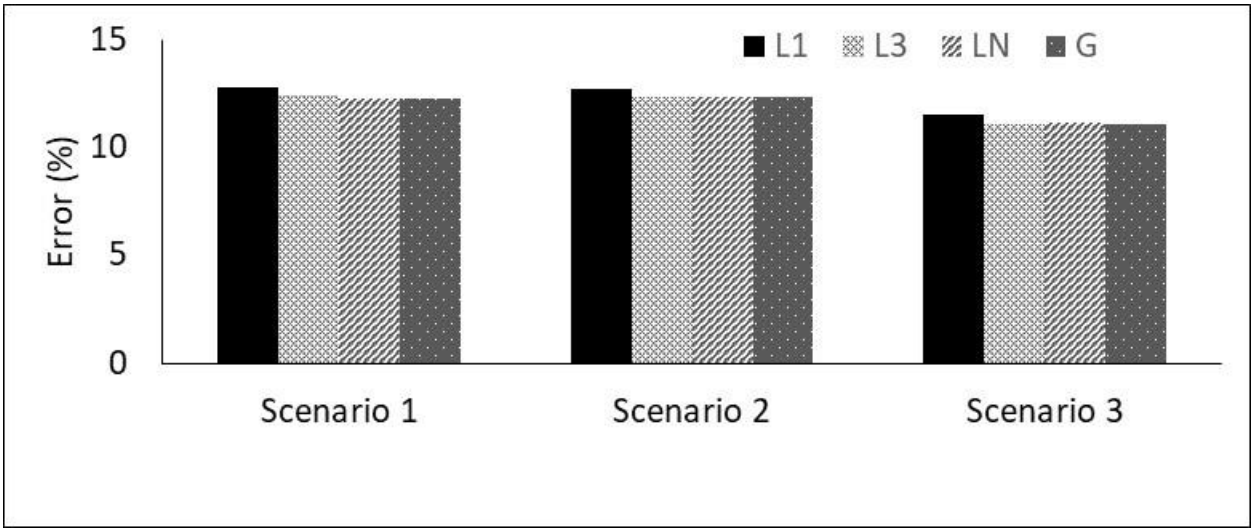


Figure 6-11 Performance of four different probabilistic functions for floodplain mapping using three different training-test samples

In Figure 6-12, the optimum probabilistic functions as well as their corresponding PDF for sample 3 are presented. Figure 6-12a shows that all four functions perform equally for floodplain mapping. It is true that the uniform distribution of TrH has a completely different shape compared to the log normal and gamma distributions. However, Figure 6-12 demonstrates how discrepancies in the PDFs of these three distributions for TrH leads to almost the same CDFs (and φ) after a mathematical integration. Since the floodplain maps are the output of φ functions, the distinction between PDFs do not affect the final outcomes. The main difference of L1 compared to the other three functions is seen at the tail end where it provides zero probability of flooding for values greater than 3.91m but other functions predict small probabilities at this range. The discrepancies in the tails do not add significant errors due to the small values of probability in these regions. Besides, these additional low probabilities at tails of Gamma and Log normal increase the number of probabilistic values in the floodplain map, which causes a more difficult decision making, given the higher number of uncertain values in the map. For example, the L1 function assigns zero

probability to some areas very far away from the main river and considers them totally safe against flooding. However, the other three functions assign small probabilities of 0.05 or 0.1 to these areas. The PDF for these four functions in Figure 6-12b shows that, although a bell shaped distribution of *TrH* (LN or G) is more realistic and intuitive considering the physics of the problem assuming a uniform distribution for *TrH* also provides equally reasonable results for floodplain mapping. The robustness of these four functions for three different samples show that except for L3 (Figure 6-13) the other three functions are identical for all three samples. A relatively lower performance and robustness of L3 as well as the need for three parameters versus two for other functions makes it the least desirable for use in probabilistic floodplain mapping.

Consequently, the simplicity of the linear function L1 with only one parameter and its similar performance compared to the other three functions, make it optimal for use in this study area. The use of LN and G with a slightly higher accuracy and one more parameter are both good choices for floodplain mapping as well. It should be noted that the conclusions related to comparisons of these four functions are only limited to this study area and drawing a general conclusion for all areas in the US will need further research.

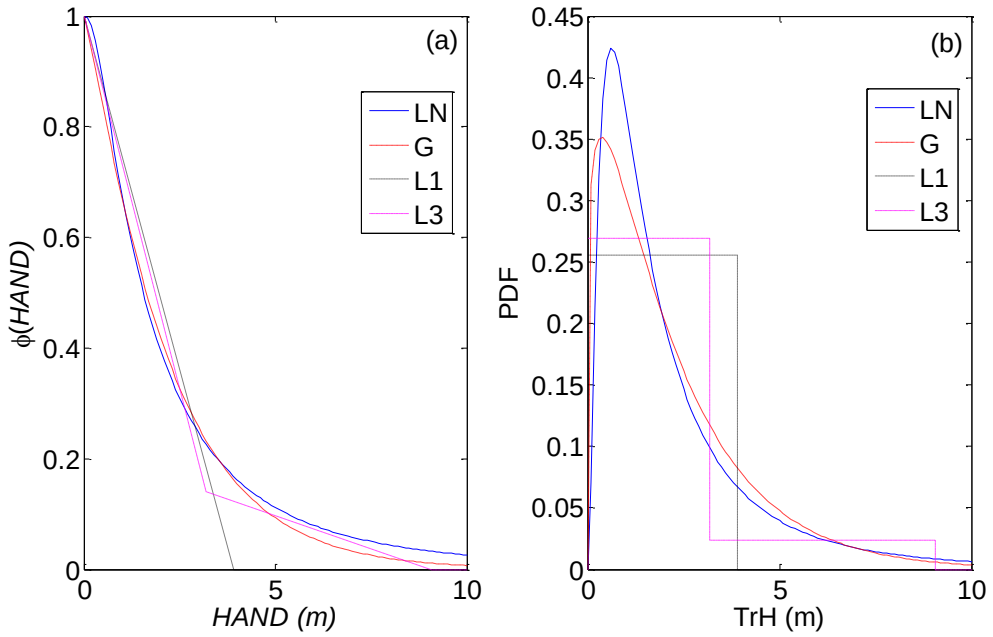


Figure 6-12 Four different probabilistic functions for direct estimation of floodplains (a), and their corresponding PDF (b)

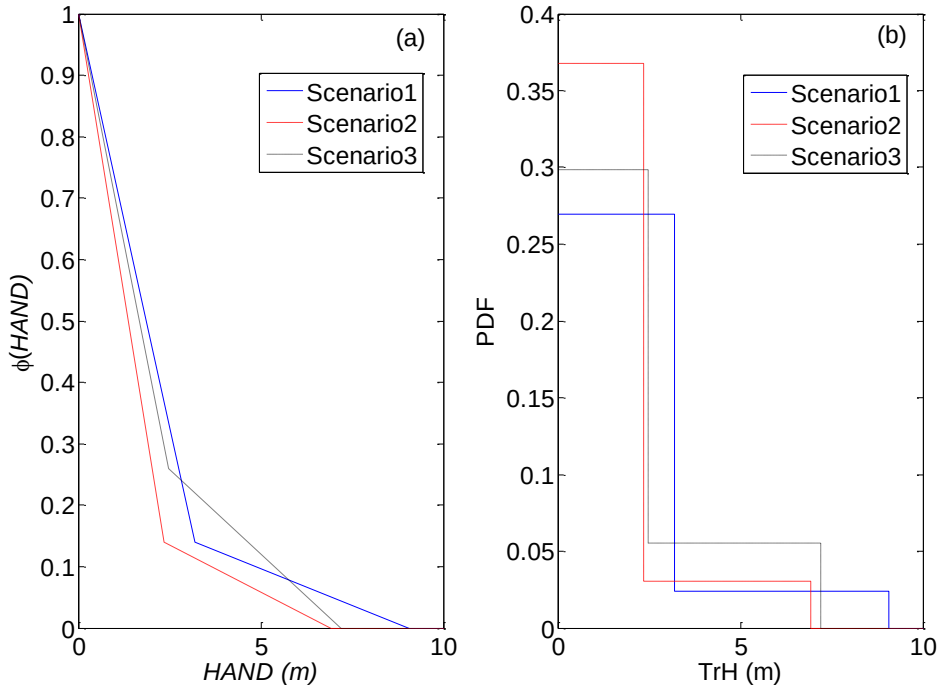


Figure 6-13 Change in probabilistic function (a), and the PDF (b) of L3 using three different training-test samples

6.5.3 Probabilistic Floodplain Mapping Compared to Deterministic Maps

In this study, L1 is used to generate the probabilistic floodplain maps. In Figure 6-14 and Figure 6-15, three floodplain maps including the reference map derived from FEMA FIRMs, and two predicted flood extent maps based on deterministic and probabilistic methods are displayed for two different regions in the test area (Figure 6-14a and Figure 6-15a). Figure 6-14 is from a flat region near the outlet of a sub-basin. It highlights the effectiveness of probabilistic mapping (Figure 6-14d) in reducing the underpredictions from the deterministic method (Figure 6-14c). The dark blue color area in the probabilistic map has 50% probability of flooding, but the deterministic method shows no inundation in this area when the reference map shows complete inundation. Figure 6-15 shows another region where overprediction from the deterministic map (Figure 6-15c) is presented with low probability of flooding by the probabilistic map (dark blue color in Figure 6-15d).

The areas with close to 50 percent probability of flooding are critical regions that need additional consideration because it is hard to classify them as either flooded or non-flooded points. However,

it is still better than a deterministic map which may classify these areas in a wrong category. Basically, a probabilistic floodplain map can be used for preliminary estimation of risk areas, so researchers and decision makers can devote more resources and efforts on these critical uncertain areas. The flood managers and decision makers can also derive more information from a probabilistic map and have the flexibility to convert a probabilistic map into a binary deterministic map based on their own considerations. For example, the importance of the region of interest in terms of land-use and the cost of flooding are some of the factors to use different and variable thresholds (less or more than 50 percent) on a probabilistic map. They also have the option to categorize a probabilistic map into more than two classes of flood and non-flood by using multiple thresholds. For example, an insurance company can convert a probabilistic map into a map with four classes including areas prone to flood by more than 90%, 50-90%, 10-50% and less than 10%. These maps are more informative compared to a single binary map for customers of the insurance company as well.

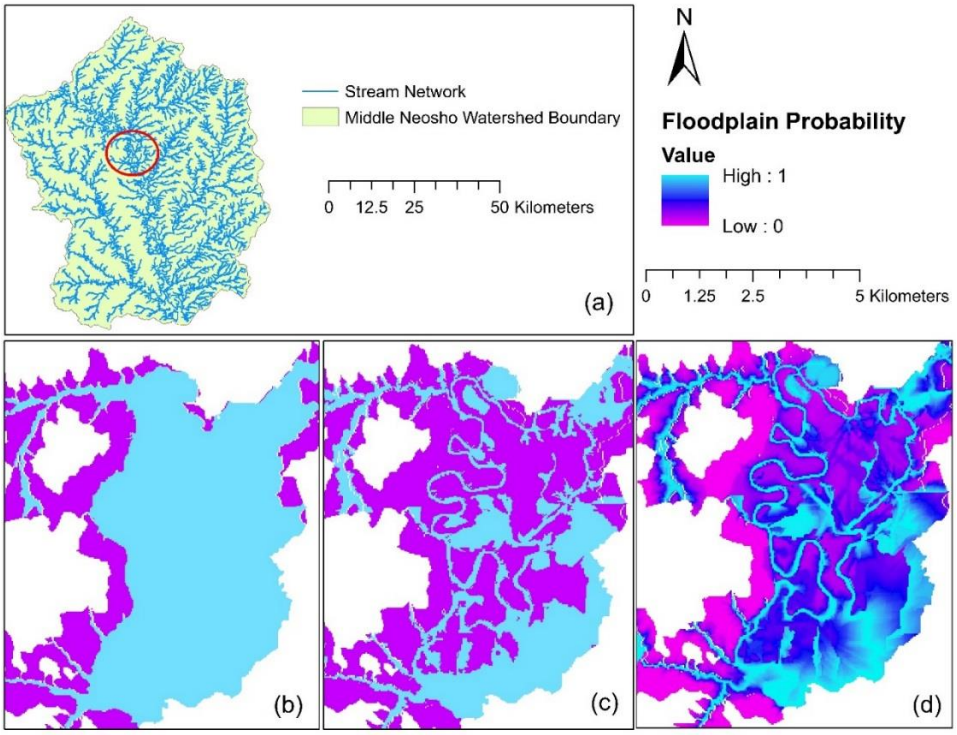


Figure 6-14 A visual comparison of three floodplain maps for a flat region in the center of Middle Neosho watershed highlighted by a red circle (a); Reference floodplain map developed by FEMA (b), predicted flood extents by deterministic (c) and probabilistic (d) methods: The colorbar shows the probability of flooding starting from zero as non-flooded (purple) to one as flooded (cyan) areas. The probabilistic method is reducing the underpredictions where purple areas in the deterministic map change to the dark blue areas in the probabilistic map.

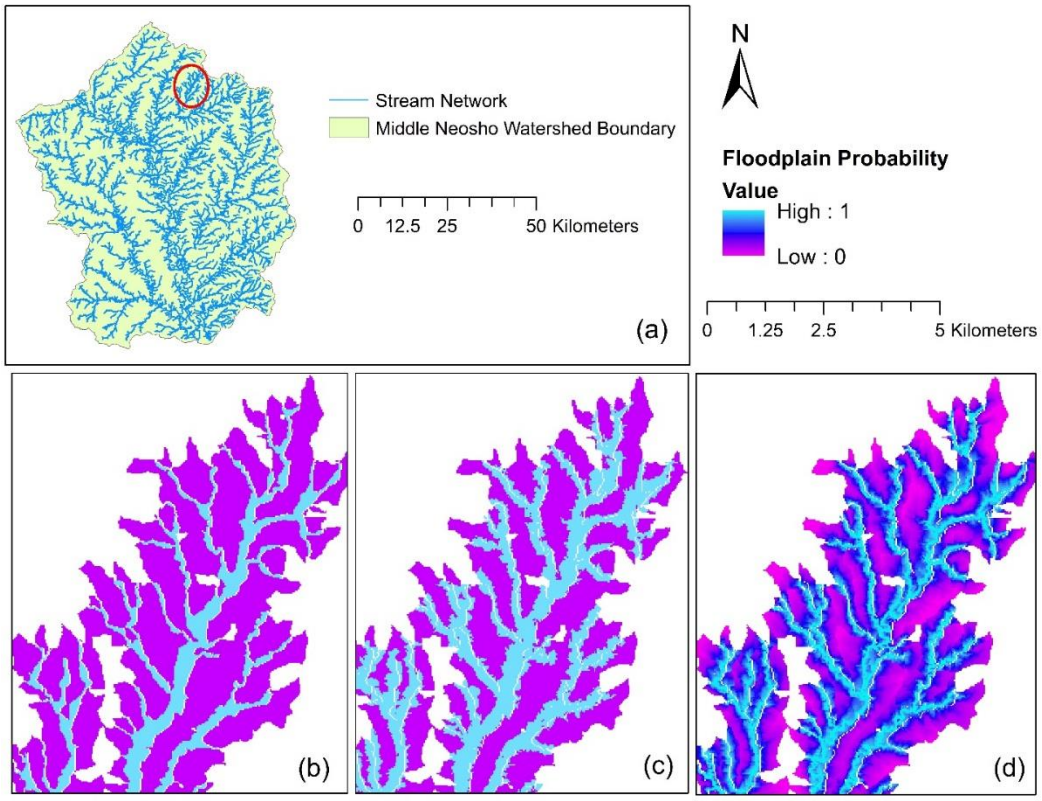


Figure 6-15 A visual comparison of three floodplain maps for the upstream of a region in the Middle Neosho watershed highlighted by a red circle (a); Reference floodplain map developed by FEMA (b), predicted flood extents by deterministic (c) and probabilistic (d) methods: The colorbar shows the probability of flooding starting from zero as non-flooded (purple) to one as flooded (cyan) areas. The probabilistic method is reducing the overpredictions where cyan areas in the deterministic map change to the dark blue areas in the probabilistic map.

In Figure 6-16, the distribution of predicted cell values inside the whole test area for both deterministic and probabilistic methods are provided. The results from the deterministic method (Figure 6-16a) illustrates that 23% of inundated cells are underpredicted and 15% of inundation cells are overpredicted. The probabilistic method (Figure 6-16b) changes the absolute result of overpredicted flooded and underpredicted flooded areas into probable areas prone to flooding. Comparison of probabilistic floodplain maps and deterministic ones generated by φ (*HAND*) and *TrH*, respectively shows the advantage of the probabilistic method in reducing the overprediction and underprediction. Considering FEMA maps as “truth”, the probabilistic approach replaces the errors from deterministic method by assigning flood probabilities to the areas that are completely predicted wrongly by the deterministic method. For example, flood regions in the FEMA maps are predicted as non-flooded cells by the deterministic method, but the proposed method assigns some

probability of flooding to these cells. These results agree with the Keynesian view that being approximately right is better than precisely wrong (Alfonso et al., 2016; Dottori et al., 2013). Overall, this new presentation of floodplain in the format of probabilistic values, which avoids the under and over prediction errors, is more reliable for decision makers and provides better information about the risk areas.

The probabilistic presentation of floodplain proposed in this study is different compared to other recent *HAND*-based methods because the proposed method is the probabilistic version of the "threshold binary classifier based on morphologic feature "*HAND*" introduced by Degiorgis et al., 2012. The threshold binary classifier approach has proved to be a useful geomorphologic method for many studies in steeper terrain watersheds, but has provided relatively poorer results in flat watersheds (Manfreda et al., 2015; Jafarzadegan and Merwade, 2017; Samela et al., 2017). The floodplain width in the flat watersheds is more sensitive to the change of TrH because of the low lateral slope across the stream lines. In other words, a small error in the estimation of TrH in these areas (e.g. less than one meter) can result in significant overprediction and underprediction of floodplains. Therefore, representing the uncertainty associated with TrH through a probability distribution is a more reliable approach compared to the deterministic methods which rely on one distinct TrH value for all streams.

In a recent paper, (Jafarzadegan et al., 2018), another *HAND*-based method for probabilistic floodplain mapping in the conterminous U.S. was proposed. This approach extended the application of *HAND*-based methods to the entire U.S. and used a range of TrH values to create a probabilistic floodplain map. The range of TrH used was quite broad due to its application at the continental scale and the lack of sufficient reference maps. This study, however, aims to determine the best distribution of TrH which fit the available reference maps. The approach presented in this chapter uses data at the watershed scale to create the floodplain map, which is found to be more accurate compared to what was found in Jafarzadegan et al (2018).

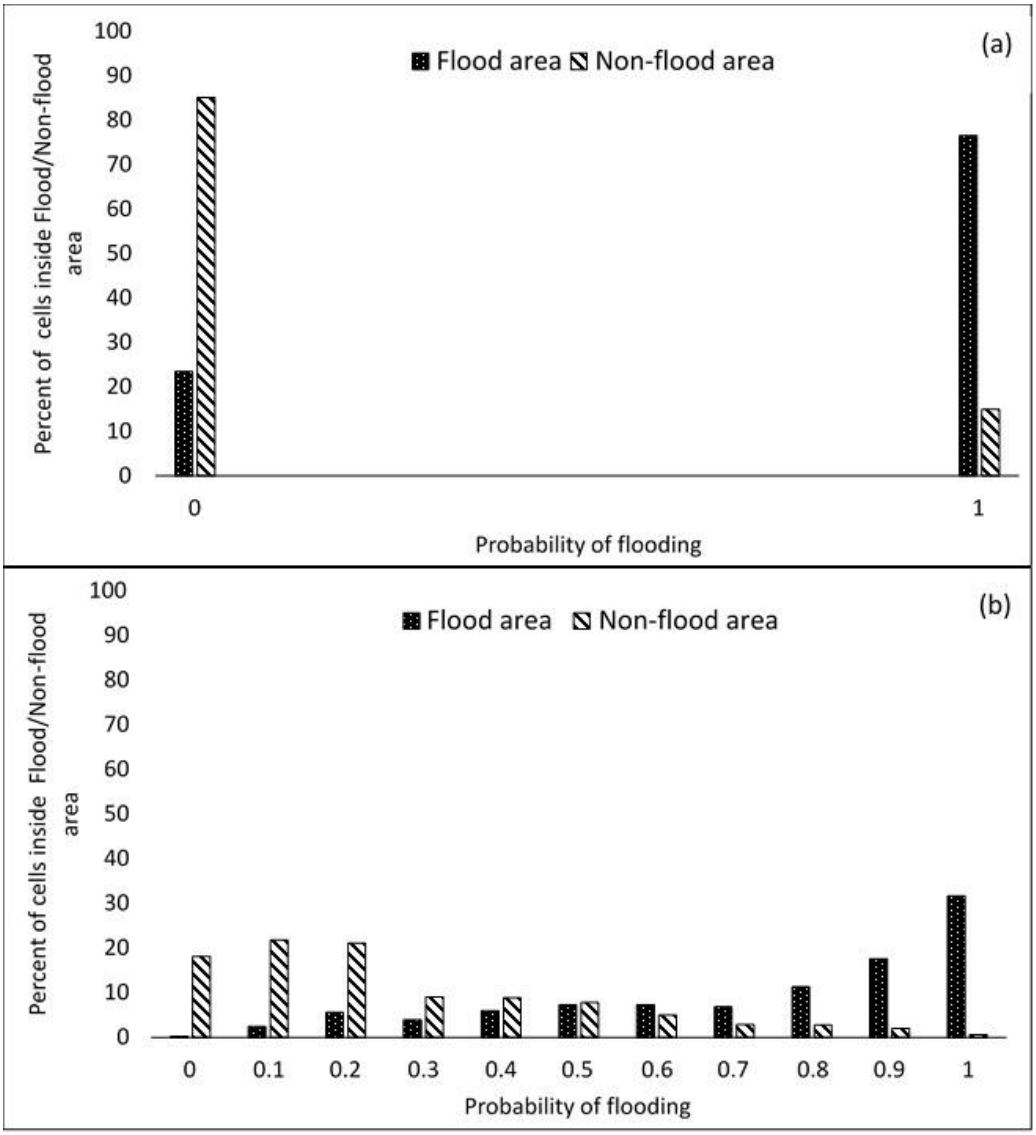


Figure 6-16 Distribution of predicted cell values for deterministic (a) and probabilistic (b) methods: the solid and hashed bars show the distribution of predicted cell probabilities inside the flooded and non-flooded areas of reference map respectively.

In floodplain mapping using a *HAND* based statistical approach, two major sources of uncertainty, including model parameters and input variables exist. The proposed approach focuses on the uncertainty of model parameter, *TrH*, by estimating its distribution for a watershed to generate a probabilistic map. The input variables of a proposed model are an important source of uncertainty, which affect the accuracy of floodplain maps as well. In the proposed approach, the topography data, DEM, used for *HAND* raster calculation, and the available FEMA maps used for generating probabilistic functions are two major inputs with their own level of uncertainties. The spatial resolution of DEM and its vertical accuracy (Gesch et al., 2002; Sanders, 2007) can affect the

resulting floodplain maps. In addition, the FEMA maps, which are the reference of this study, are floodplains delineated by hydraulic modeling (most of them by HEC-RAS models). Therefore, the reliability and accuracy of these maps as a reference, considering that they are not real observed maps, is a matter of controversy. As a result, in order to cover the majority of uncertain sources and improve the performance of the probabilistic approach, the uncertainties of input from DEM and the available FEMA maps should be considered in the structure of the proposed model in future studies.

6.6 Conclusion

In this study, the performance of five probabilistic functions for probabilistic floodplain mapping, including three linear (L1, L2, L3), one log-normal (LN), and one gamma (G), are compared. The results demonstrate that non-linear functions including log-normal and gamma offer very little advantage over simple linear functions in estimating *TrH* variability for floodplain mapping. Thus, a probabilistic linear function with only one parameter (L1), or a uniform distribution of *TrH*, is able to provide probabilistic floodplain maps for the study area with acceptable accuracy. The optimization results show that the linear function with two parameter (L2) is identical to the linear function with one parameter (L1). In addition to accuracy, the robustness of the four probabilistic functions is also compared by defining three samples including different combinations of training and test areas. Result from this analysis shows the three-parameter linear function (L3) is less robust compared to other functions, and that one parameter linear function (L1), log-normal (LN) and gamma (G) are equally robust for all samples. While the suitability of a one parameter linear function for mapping floodplains using the *HAND* approach is encouraging, more studies involving areas from different topography, climate and land at multiple spatial scales must be conducted to draw general conclusions about the performance of the proposed probabilistic functions. The comparison of probabilistic and deterministic floodplain maps generated by φ (*HAND*) and *TrH*, respectively shows the advantage of the probabilistic method in reducing the overprediction and underprediction. Overall, the probabilistic floodplain maps are more reliable and more informative as they incorporate the uncertainty in the floodplain mapping process. Moreover, the *HAND*-based statistical approach used for floodplain mapping has an advantage over conventional probabilistic hydrodynamic models because of its higher computational efficiency for fast and cost effective mapping of floodplains in data-scarce regions.

CHAPTER 7. A SYSTEMATIC APPROACH TO SIMILARITY-BASED REGIONALIZATION TECHNIQUES IN ENVIRONMENTAL MODELS

7.1 Abstract

Prediction in data-scarce regions is one of the challenging issues in environmental and water resources problems. The lack of reliable benchmarks in data-scarce regions limits the possibility of a straightforward prediction because the models cannot be calibrated in these regions. In hydrology, this issue is referred to as prediction in ungauged basins (PUB) and is accomplished by developing regionalization techniques that transfer information from gauged basins to ungauged ones by using either regression or similarity-based methods. This study proposes a novel regionalization similarity-based framework that can be used to transfer any calibrated model, including hydrologic, hydraulic, geomorphic and statistical, to data-scarce regions. The core of similarity-based regionalization methods is a physical/climatic similarity metric. This metric uses the basin descriptors to determine those basins that are similar to a target basin. Typically, the physical/climatic similarity metric is predetermined from historical experiences about the physics of the problem and study area. The main focus of this study is to reduce the subjectivity that exists in this process by establishing a systematic approach to obtaining an appropriate physical/climatic similarity metric. In the proposed framework, first a hierarchical clustering algorithm classifies the data-rich basins, and then a supervised classifier uses the clustering results as input and provides an appropriate physical/climatic similarity metric for a target basin in a data-scarce region. The effectiveness of the proposed regionalization framework is tested using a statistical gamma-based model to create the probabilistic floodplain maps. The statistical model is calibrated for 30 basins in the central region of the United States where benchmarks are available for calibration. Using the proposed framework, a simple hierarchical structure with three basin descriptors is defined as an appropriate physical/climatic similarity metric. The vertical component of geographical location (latitude) of basin is found as the most significant basin descriptor along with two other basin descriptors, namely the standard deviation of elevation and the average of precipitation in the wettest month. The proposed metric is applied to estimate the parameters of the gamma model in data-scarce regions. The validation results demonstrate the successful regionalization of 7 out of 8 basins using the proposed framework. For all these 7 basins, the regional errors of prediction

using the proposed regionalization framework compared to the errors using locally calibrated models is less than 0.5 percent.

7.2 Introduction

Simulation models are simplified representation of real world systems, and are widely used to understand the behavior of a complex system (Devia et al., 2015; Silvert, 2001). One of the key tasks in modeling is its calibration that involves the proper estimation of model parameters by adjusting their values to generate outputs that match closely with some benchmarks. Benchmarks refer to any reliable data (e.g. observed data) which can be used as a reference for calibration. The existence of benchmarks at some points inside the domain of the model is crucial for a successful calibration. In environmental and water resources problems, different types of the models (e.g. climate, hydrologic, hydraulic, geomorphic and statistical models) are built and calibrated based on users' needs and the problem in hand. Additionally, these models are implemented at different spatial scales ranging from few square meters of area to several thousand square kilometers. As the spatial scale increases, the heterogeneity in representing the physical characteristics of the area as well as the actual physical processes increases. Thus, model calibration becomes challenging for large scale environmental or hydrologic models. This is especially true in data-scarce regions due to the absence of benchmarks for the model calibration. In hydrology, this issue, termed "Prediction in Ungauged Basins (PUB)", has gained a great deal of attention for many years (Hrachowitz et al., 2013; Sivapalan, 2003). Ungauged basins generally refer to basins with no available data, specifically streamflow. Thus, hydrologic models cannot be calibrated in ungauged basins, and flow prediction without reliable calibration will involve a lot of uncertainty.

To create predictions in ungauged basins, a large number of methods, named regionalization techniques, have been proposed in the literature. Regionalization refers to all methods used to transfer information from gauged basins to ungauged ones by relating hydrologic phenomena to basin descriptors (Blöschl and Sivapalan, 1995; Oudin et al., 2010; Young, 2006). These techniques can be categorized into two groups: regression and similarity-based methods. Regression based methods are the most common tools used for regionalizing hydrologic models. In these methods, a hydrologic model is calibrated on a large number of gauged basins, and the parameters of the hydrologic model are related to some basin descriptors by establishing multiple

regression relationships. These regression relationships can then be used to estimate the model parameters in an ungauged basin (Sefton and Howarth, 1998; Tung et al., 1997). Despite their popularity in the last two decades, the regression-based methods suffer from certain limitations (Fernandez et al., 2000; Hundecha and Bárdossy, 2004; Kim and Kaluarachchi, 2008; Lee et al., 2006; Merz and Blöschl, 2004; Seibert, 1999). Specifically, regression based methods assume that the model parameters are independent, and that the error residuals are normally distributed, but these assumptions are not true in many cases (McIntyre et al., 2005). Furthermore, there is considerable uncertainty in the model parameters, due to the calibration method, model structure, and uncertainty of the model inputs. Therefore, applying the optimum parameter set as dependent variable in regression equations is not the best choice approach when multiple sets of parameters produce almost similar model performance (Anderson et al., 2001; Beven and Freer, 2001).

Similarity-based regionalization methods use the concept of similarity-based on certain basin descriptors and transfer the entire model parameter set of these basins to a similar ungauged basin. Transfer of all model parameters as a set in similarity-based methods is useful compared to traditional regression-based methods, which neglect the interdependencies of parameter sets (McIntyre et al., 2005; Parajka et al., 2005). Kokkonen et al. (2003) also concluded that "when there is a reason to believe that, in the sense of hydrological behavior, a gauged catchment resembles the ungauged catchment to a sufficient extent, it is worthwhile to adapt the entire calibrated parameters from the gauged basin". Typically, a similarity-based method consists of two steps including 1) selecting similar basins, named donor basins, and 2) transferring the information from donor basins to an ungauged basin, named target basin. Most studies found in the literature focus on the second step (Holmes et al., 2002; Kay et al., 2007; Masih et al., 2010; McIntyre et al., 2005). For example, McIntyre et al. (2005) introduced an extension of the generalized likelihood uncertainty estimation (GLUE) framework (Beven and Binley, 1992) for the regionalization problems by proposing the weighting average of donor basin simulation results. The weights were defined by the product of a prior likelihood of a model, and the relative likelihood of that model being applicable to the target basin. Kay et al. (2007) proposed another weighted averaging method where the uncertainty of model calibration was taken into account during the parameter transposition. Despite these promising advances in transferring information from donor basins to target basins, the methods proposed in these studies do not have any strong

do not have any strong grounds for selecting the donor basins. A proper selection of donor basins in the first step increases the chance of reliable and robust prediction for target basins regardless of using a simple or sophisticated method in the second step. To properly determine the donor basins in a similarity-based regionalization method, two key questions should be answered:

1) Which basin descriptors should be selected to reflect the functional similarity of basins?
2) Which similarity metric should be used to find the most appropriate donor basins?

In the general context, “functional behavior” of a basin is the series of physical processes that take place to create the target outputs in the basin. Therefore, functional similarity is the similarity in the functional behavior of two basins and is used as a generalization of the term "hydrologic similarity". The hydrologic behavior of a basin should be distinguished from hydraulic, geomorphic, or any other behavior because the type and distribution of processes involved vary. The primary goal of this study is to extend the concept of regionalization beyond hydrology for a broad range of environmental models. To achieve this goal, we propose a hybrid classification framework which outlines the overall steps to successfully regionalize calibrated environmental models. It should be noted that the term "environmental model" is used without specifying its type to make the proposed framework applicable for different purposes and modeling types (e.g. hydrologic, hydraulic, geomorphic, climate, groundwater, and statistical models). While the water basins are the common computational unit of hydrologic models, the proposed framework can be used for other computational units as well. For example, a river and an aquifer can be used as a computational unit for sediment and groundwater modeling respectively. For the latter, the proposed framework can be applied for regionalizing groundwater levels using a groundwater model, such as MODFLOW. In this case, the parameters of MODFLOW are estimated for aquifers in data-scarce regions where field data is not available for the model calibration. Despite the generality of the proposed framework, its application and effectiveness is demonstrated in this study by focusing on a probabilistic floodplain mapping problem where a statistical model already calibrated on several basins is transferred to data-scarce basins.

7.3 Related Work

A wide variety of physical and climatic attributes have been used in the past studies as basin descriptors in similarity-based regionalization methods. Some studies have also used spatial proximity and advanced geo-statistical methods (e.g. kriging) to find the donor basins. Literature shows that the role of spatial proximity in the regionalization techniques can be mixed in the sense that it can provide both good (Merz and Blöschl, 2004; Parajka et al., 2005; Sawicz et al., 2011; Vandewiele and Elias, 1995; Viviroli et al., 2009), and poor (Ouarda et al., 2001; Reed et al., 1999; Shu and Burn, 2003) results. A review of similarity-based studies in the literature shows that the basin descriptors are usually determined based on available information about the history of the study area, e.g., physical characteristics and the hydrological response of the basin. In other words, the selection of basin descriptors is a subjective process. The role of subjectivity in the selection of basin descriptors is a major concern when reliable information about the history of the study area does not exist, or when the physical process and the structure of the model is not fully understandable (e.g. black box statistical models). The probabilistic floodplain mapping model used in this study is a good example of such cases. The physical processes are not concrete to the modeler, and it is difficult to pre-determine significant basin descriptors. These issues indicate the importance and the necessity of using a more systematic approach for defining the most significant basin descriptors.

Selection of appropriate donor basins has been the topic of significant interest in the field of Regional Flood Frequency Analysis (RFFA) for many years (Burn, 1990; Burn and Goel, 2000; Castellarin et al., 2001; Laaha and Blöschl, 2006; Ramachandra Rao and Srinivas, 2006) . In RFFA, the flood quantiles for an ungauged site are estimated by using the information from a group of donor basins that are hydrologically similar to the target site. The flood quantiles in the gauged sites are usually estimated by fitting a statistical density function (e.g. Log Pearson 3, Log normal or Gumbel distributions) on the historical observed flows. Among several proposed techniques for identifying homogeneous regions (or selecting the donor basins), the region-of-influence (ROI) method (Burn, 1990; Zrinji and Burn, 1996) has been widely used in regionalization problems. In this method, the donor basins are selected based on a similarity metric which is defined by Euclidean distance in the basin descriptor space (De Coursey, 1973). A threshold is usually applied on the similarity metric to decide the number of donor basins for a given target basin. Considering

the Euclidean distance and the linear combination of descriptors in the formation of this similarity metric is a major limitation of ROI. Additionally, the basin descriptors used as inputs to ROI are usually predetermined attributes, which may not be the best indicators of basin functionality. In a promising study, Oudin et al. (2010) focused on the concept of hydrological similarity, as a type of functional similarity, and its relation to the physical similarity of basins. They used the ROI method for selecting the donor basins and demonstrated that the basins which are hydrologically similar to a given basin are not always the same with those that are physically similar. These results illuminate the weakness of the physical/climatic similarity metric and the predetermined basin descriptors in reflecting the hydrological response of the basins.

Basin classification (or catchment classification (McDonnell and Woods, 2004; Wagener et al., 2007)) is another approach commonly used in RFFA to identify the homogeneous regions (Castellarin et al., 2001; Laaha and Blöschl, 2006; Ramachandra Rao and Srinivas, 2006). Classification techniques are typically categorized into supervised and unsupervised methods. A common input for both methods is a set of attribute vectors, which could be the basin descriptors in a basin classification problem. In unsupervised methods, referred to as clustering, the attribute vectors are classified into groups where the dissimilarity within each group, and between different groups, are minimized and maximized, respectively. In supervised learning, however, more information about the class labels of a portion of attribute vectors is needed. This additional piece of information provides a significant advantage for the classifier to find appropriate attributes and meaningful patterns. In other words, known class labels shrink the range of the potential attributes for a classifier and reduce the chance of finding irrelevant patterns. A supervised algorithm detects the relationship between the attribute vectors and corresponding class labels (referred to as training stage) and applies this relationship on other attribute vectors with unknown class labels. Considering this fact, supervised learning should be the dominant method used for basin classification. However, finding these class labels is a critical issue in the absence of any observed information. The lack of reliable information about the class labels of attribute vectors reduces the applicability of supervised methods for many problems including basin classification in RFFA.

Unlike supervised classification, two clustering algorithms, namely Agglomerative hierarchical clustering (Burn et al., 1997; Nathan and McMahon, 1990; Tasker, 1982) and K-means (Bhaskar

and O'Connor, 1989; Burn, 1989; Burn and Goel, 2000; Wiltshire, 1986), have been widely used in the literature of RFFA. The core of hierarchical clustering methods is the dissimilarity measure used within the algorithm. In the basin classification problems, the dissimilarity measure is a distance-based function of basin descriptors. Basins that have more similar basin descriptors are grouped in a hierarchical structure. The possibility of applying various distance-based dissimilarity measures within the hierarchical clustering algorithm can provide some degree of flexibility and improvements in the selection of donor basins. For example, using measures other than Euclidean distance may help to capture non-linear similarities between basins, but the issue related to choosing appropriate basin descriptors still exists. The hybrid classification framework, described in the next section, addresses the two key issues related to the selection of basin descriptors and physical/climatic similarity metric for regionalization problems. The proposed framework reduces the current subjectivity in the selection of basin descriptors by proposing a systematic approach. Additionally, the supervised classifier embedded in the hybrid classification framework has the flexibility to fit any linear or non-linear function on data and estimate the best physical/climatic similarity metric for recognizing the donor basins.

7.4 Hybrid Classification Framework

This section describes a framework to estimate the model parameters for basins located in data-scarce regions using the calibrated models in data-rich regions. A model, denoted by f in Equation 5-1, is a function of inputs and parameters to produce output(s) related to a specific domain of application.

$$Y = f(X \mid \theta) \tag{7-1}$$

where X and Y denotes the model inputs and outputs respectively, and θ represents a vector with t model parameters. As mentioned earlier, model f is a general term which can be used for a broad range of problems. For example, for streamflow prediction f refers to a hydrologic model where Y is the streamflow time series, and X is a set of climate, topographic and land use data. For a hydraulic model, Y denotes the water depth in a river while the geometry of a river, and streamflow information may be used as input X. In data-rich basins, benchmarks are used to calibrate model f and determine its parameter set (θ). Considering a problem where benchmarks are available for n basins, model f can be calibrated for all n basins which results in n known vector of θ. The

hybrid classification framework in this section proposes a general systematic procedure for estimating the model parameter set (θ) in data-scarce basins using the available model parameters in data-rich basins.

As a similarity-based approach, the goal is to find the best donor basins (data-rich basins that are similar to a given data-scarce basin) using an appropriate physical/climatic metric. The training step in the hybrid classification framework, illustrated in Figure 7-1a, explores this metric by using two classification algorithms. In the first classification algorithm, hierarchical clustering is used with a novel dissimilarity measure to group similar data-rich basins into clusters. The proposed dissimilarity measure guarantees that the basins included in a cluster have the maximum functional similarity. The class labels of data-rich basins determined by the hierarchical clustering algorithm are then used to perform a second classification using a supervised classifier. The supervised classifier screens the most significant physical/climatic basin descriptors among a large set of potential basin descriptors and relates them to the class labels. The classification pattern detected by the classifier will be the final physical/climatic similarity metric which can be used for a data-scarce basin to find the donor basins. In addition to the similarity metric generated in the training step, new models should be created as the representative of each class of basins using the aggregation method. In the second step, namely testing (Figure 7-1b), data-scarce basins are used as inputs, and two estimated outputs generated in the training step are used as processors to estimate the calibrated models for data-scarce basins.

This framework can also be described in the context of a similarity-based method where the donor basins for a data-scarce basin are those data-rich basins that their class labels match with that of the data-scarce basin. To transfer the information from donor basins to the target data-scarce basin, the calibrated models of these donor basins are aggregated during the aggregation process and are utilized as calibrated models of the data-scarce basin. The techniques used to transfer the calibration parameter sets from donor basins to a target basin is defined as aggregation process in this study. Aggregation process is the second step for all similarity-based regionalization methods and is highly dependent on the structure and the purpose of problem. Although aggregation process is not the main focus of this study, an aggregation technique which is exclusively used for this case study is also explained.

7.4.1 Hierarchical Clustering Using a New Dissimilarity Measure to Classify Data-Rich Basins

Basin clustering is aimed to convert a *n* dimensional problem into *m* dimensions where *n* is the total number of data-rich basins (*n* basins with *n* calibrated models), *m* refers to the number of classes (*m* group of basins with *m* calibrated models), and $m << n$. Agglomerative hierarchical clustering, a well-known algorithm for grouping the data points, can be applied to union all similar basins into one class. In this algorithm, pairwise comparisons are applied among all data-rich basins, and a multi-level hierarchy, named dendrogram, is created.

The essential component of this clustering algorithm is the pairwise comparison step where the similarity of two basins is evaluated and two basins of a pair with the highest similarity are joined. This process is repeated until the dendrogram is formed. Various dissimilarity measures have been introduced to decide the priority for joining the most similar data points in the hierarchy. The "similarity" between two given basins can vary depending on the problem in hand. For example, two basins that show a similar hydrologic response to a rainfall event can present a different behavior in converting the streamflow to flood inundation areas. To define the dissimilarity measure, first the attributes of classification should be determined.

In regionalization problems, typically two types of attributes exist for each basin: The basin descriptors (e.g. topographic, climate and land use characteristics), and the calibrated model parameters. At this stage, finding the significant attributes from a long list of potential basin descriptors is difficult, and a poor selection of dominant basin descriptors can result in erroneous classification. The second type of attributes, calibrated model parameters, may not also reflect the similarity of two basins if they are used directly. For example, assume basins 1, 2 and 3 and their calibrated model parameters as ($\alpha_1 = 0.2, \beta_1 = 0.5, \gamma_1 = 0.2$), ($\alpha_1 = 0.5, \beta_1 = 0.5, \gamma_1 = 0.3$), ($\alpha_1 = 0.3, \beta_1 = 1, \gamma_1 = 0.8$), respectively. A distance-based dissimilarity measure will show that basin 1 and 2 are more similar than basin 1, 3 (Table 7-1). However, if the model is not sensitive to parameters β and γ, basins 1 and 3 should be considered more similar, or there could be other attributes corresponding to the physics of the problem which have not been considered in the model parameters. Other information that can be used to create the dissimilarity measure includes the model outputs and the reference data for comparing the model results. Since clustering is applied

on the basins with calibrated models (data-rich basins), the reference data is available for these basins. Therefore, the error of calibrated models can be calculated by comparing the model output with reference data for all basins. Thus, basin *i* can be considered similar to basin *j*, if the additional errors generated by running the calibrated model of basin *i* on the basin *j* and running the calibrated model of basin *j* on the basin *i* are negligible. Hence, the dissimilarity measure between basins *i* and *j* are defined using Equations 5-2 to 5-4 below.

$$\Delta_{ij} = e_{ij} - e_{ii} \tag{7-2}$$

$$\Delta_{ji} = e_{ji} - e_{jj} \tag{7-3}$$

$$d_{ij} = d_{ji} = \frac{\Delta_{ij}+\Delta_{ji}}{2} \tag{7-4}$$

where e_{ii} (or e_{jj}), referred to as existing error, is the error in basin i when the model is calibrated on the same basin i (j). This is the error that exists in all models when the calibration does not result in a perfect match between model results and reference data. e_{ij} (or e_{ji}), referred to as cross-modeling error, is the error when the model calibrated from basin j (i) is applied to basin *i* (*j*). Δ_{ij} (or Δ_{ji}), referred to as net error, is the difference between the cross-modeling and existing error in basin i(j). d_{ij} (or d_{ji}) is the dissimilarity measure between basin *i* and *j*. Using these equations, two important criteria, including the functional behavior of the basin and the role of the model structure (model parameters) on the basins, are incorporated in the calculation of the dissimilarity measure.

Table 7-1 The values of several distance-based measures between basins 1,2 and 1,3 for a simple example

Distance measure	Basins 1, 2	Basins 1, 3
Euclidean	0.32	0.79
Sueclidean	1.99	2.63
Citiblock	0.4	1.2
Minkowski (p=3)	0.3	0.7
Chebychev	0.3	0.6
Hamming	0.67	1

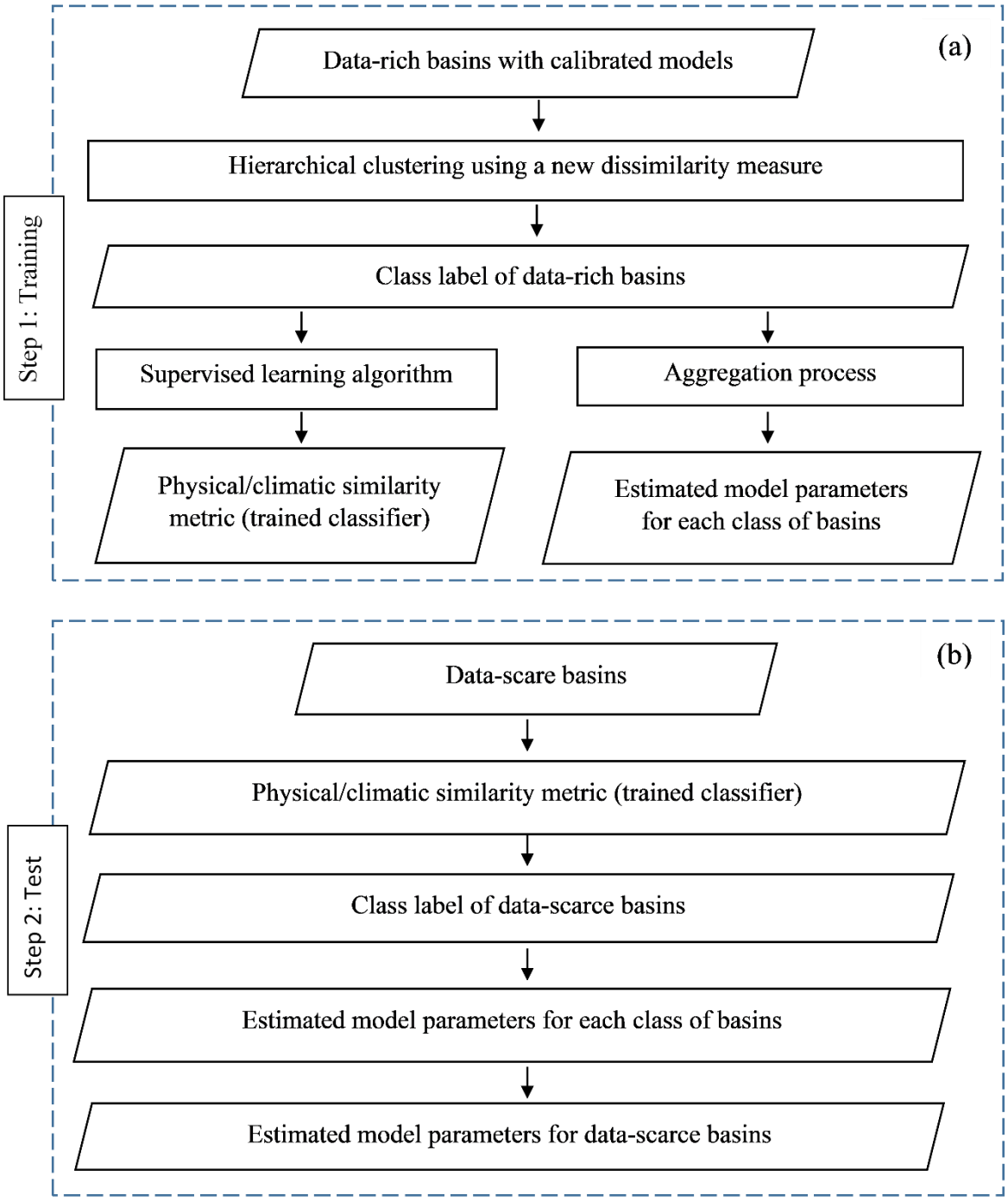


Figure 7-1 Proposed hybrid classification framework for transferring the calibrated models to data-scarce environments

The dissimilarity measure should be calculated for all possible pairs of *n* data-rich basins, resulting in $n \times (n-1)/2$ dissimilarity values. Considering d_{ii}=0 and d_{ij}=d_{ji}, a n by n dissimilarity matrix is created which is the basis of a hierarchical algorithm for deciding the basins that should

be joined at each stage of the hierarchy. After calculating the dissimilarity matrix, the linkage method should be determined. The linkage method is used to define the dissimilarity values between the groups of basins. For example, "single", "complete" and "average" are three popular linkage methods used in the agglomerative hierarchical algorithms. The first two methods use the minimum, or the maximum dissimilarity values between the basins of two groups, and the third method take the average of dissimilarity values between the two groups. A dendrogram is created based on the dissimilarity matrix and the assigned linkage methods. A Dendrogram is a tree of basins which show how similar basins are joined at each level of the hierarchy. At this stage, depending on the number of clusters decided by the modeler, a cutoff is applied on dendrogram and the class labels for each cluster are determined. The main limitation of the proposed dissimilarity measure is that it can only estimate dissimilarity between two data-rich basins. However, to recognize the donor basins for a target basin, the dissimilarity (or similarity) between the target basin and data-rich basins should be determined. To overcome this issue, a second classification is linked to the hierarchical clustering algorithm where the class labels of data-rich basins generated by the hierarchical clustering algorithm is fed into a supervised classifier to estimate a physical/climatic similarity metric. This metric can be later used to estimate the similarity of data-rich basins with the target basin.

7.4.2 Training a Supervised Classifier to Find the Significant Basin Descriptors, and the Similarity Metric

In supervised learning techniques, first a classifier is trained on data with available class labels, referred to as training step, and then it will be used for predicting the class labels of unknown data, referred to as test step. A schematic diagram of supervised learning algorithms used in the hybrid classification framework is illustrated in Figure 7-2. Let X and Y represent the attribute and class label matrices for n data-rich basins where x_{ij} and $y_i$represent the value of j^{th} attribute for the i^{th} basin, and the class label of i^{th} basin respectively (i=1,2,…,n and j=1,2,…,k). In the training step, a supervised classifier finds the best fit function for relating attributes (X) to class labels (Y). Another feature of supervised classifiers is their capability to estimate the significance of attributes. For example, assume five-dimensional attribute vectors, denoted by [a, b, c, d, e], are used as input to a supervised classifier and the significance of these attributes are estimated as 2%, 3%, 50%, 45%, 0% respectively. This shows that attributes c and d are the significant basin descriptors with 95% contribution in the classification process, and the other three attributes can be removed from

the analysis. Using this feature of supervised classifiers, the potential basin descriptors in the attribute matrices are filtered and the most significant ones are selected for prediction. The fit function (or trained classifier) in the proposed framework is a physical/climatic similarity metric used for estimating the class label of basins in data-scarce environments. The performance of a supervised classifier is highly dependent on the reliability of training class labels (Y). In the hybrid classification framework, these class labels are produced by the hierarchical clustering algorithm. This shows the high impact of hierarchical clustering algorithm and its dissimilarity measure on the overall performance of the supervised classifier.

7.4.3 Aggregation Process

The hierarchical clustering algorithm reduces an *n* dimensional problem into *m* dimensions by classifying *n* basins into *m* groups. To complete the process of dimension reduction, the *n* calibrated models corresponding to the *n* basins should also be reduced to *m* calibrated models.

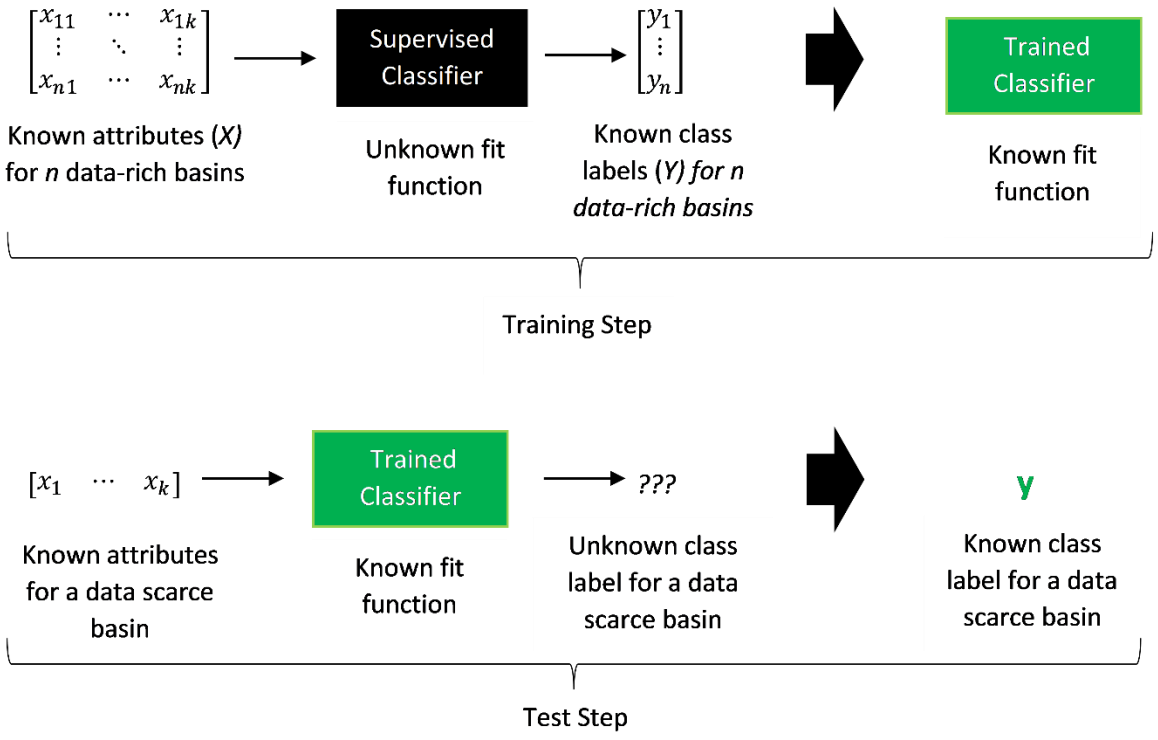


Figure 7-2 Schematic diagram of supervised learning algorithms used for basin classification in the hybrid classification framework

This step is named aggregation process because the calibrated models inside each group should be aggregated and form a new model where the new model is the representative of all basins included in that group. The method used to aggregate the models is highly dependent on the case study and the model structure. A pairwise averaging aggregation based on the dendrogram can be a simple approach for aggregation. For example, Figure 7-3 is a hypothetical dendogram created for 5 basins using the proposed dissimilarity measure. This dendogram shows that basins 1 and 4, as well as, basins 2 and 5 have the highest functional similarities in the first level of dendogram. It means the model calibrated in basin 1 (f_1) can create acceptable results in basin 4. Similarly, the model calibrated in basin 4 (f_4) will be a good choice for creating results in basin 1. Therefore, these two models can be replaced with a new aggregated model (f_{14}) where the parameters of this model are the average of f_1 and f_4. Using this concept, f_2 and f_5 are also aggregated to create f_{25}. In the second level of dendogram f_{14} should be aggregated with f_3 to create f_{143}. The final aggregated model that is the representative of all five models is created by aggregating f_{143} with f_{25}. In this study, the overall aggregation process is based on dendogram similar to the procedure explained in Figure 7-3. However, since the models are statistical gamma distributions, instead of simple averaging between the parameter of two models in the pairwise aggregation, the aggregated models are estimated by generating samples from the initial distributions. More details about this approach will be provided in the case study section.

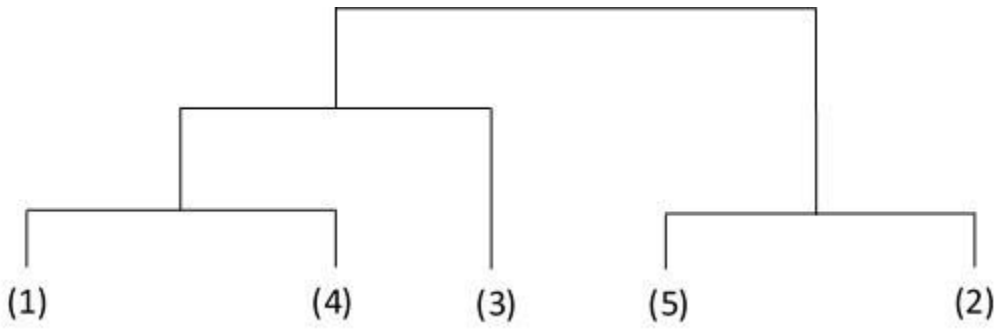


Figure 7-3 A hypothetical dendogram created for clustering five basins using a hierarchical clustering algorithm

7.5 Framework Application for Probabilistic Floodplain Mapping in Data-Scarce Environments

Jafarzadegan and Merwade (2018) proposed a statistical function named $\varphi(HAND)$ that can be used for a given basin to generate a probabilistic floodplain map. The independent variable, *HAND*, in the $\varphi(HAND)$ is a hydrogeomorphic feature defined as Height Above Nearest Drainage (Nobre et al., 2011; Rennó et al., 2008). To create $HAND$, Digital Elevation Model (DEM) and stream network of the basin are needed (Jafarzadegan and Merwade, 2017). The φ function is derived from the gamma Cumulative Density Function (CDF_{gamma}) using Equation 5-5, and can be directly calculated by Equation 5-6:

$$\varphi = 1 - CDF_{gamma} \tag{7-5}$$

$$\varphi(HAND) = 1 - \frac{1}{\tau(k)}\gamma\left(k, \frac{HAND}{\theta}\right) \tag{7-6}$$

where k and θ are the shape and scale parameters of the φ function which should be estimated for each basin. The $\gamma(a)$ and $\gamma(a,b)$ are the complete and the lower incomplete gamma functions (Equations 5-7 and 5-8) calculated as:

$$\tau(a) = \int_0^\infty x^{a-1}\, e^{-x} dx \tag{7-7}$$

$$\tau(a,b) = \int_0^b x^{a-1}\, e^{-x} dx \tag{7-8}$$

The optimum parameters of φ function are determined by minimizing the error of predicted flood extent compared to a reference floodplain map. Readers are referred to Jafarzadegan and Merwade (2019) for more details related to estimating the error function. In this study, the Flood Insurance Rate Maps (FIRMs) provided by the U.S. Federal Emergency Management Agency (FEMA) are used as reference maps. FEMA FRIMS were created using detailed field measurements and modeling for many areas in the U.S., and thus form a good basis for comparing results from other floodplain mapping efforts.

The $\varphi(HAND)$ function provides a simple and computationally efficient probabilistic floodplain mapping approach over large areas compared to the conventional modeling approach, which is limited to small reaches due to data and computational demands. The development of $\varphi(HAND)$, through calibration of its parameters, depends on the availability of reference data. Thus,

estimating its parameters for areas with limited or no reference data poses the classical challenge of prediction in ungauged basins. In other words, how to transfer the calibrated φ functions to data-scarce environments. This question is addressed by applying the proposed regionalization framework to the Arkansas-White-Red region in the U.S. with highly variable topography and climate. Specifically, the western part of this area is mountainous while the eastern and the southern parts are flat. To utilize the proposed framework and define a physical/climatic similarity metric, 30 basins that have FEMA FIRMs are selected as training basins. The framework is then validated by applying it to eight basins as shown in Figure 7-4. All the basins are selected to ensure variability in geographic, topographic and climatic conditions. Using the FEMA FIRMs as reference floodplain maps, the φ function is calibrated for all training basins. Figure 7-5 shows the calibrated parameters of φ function for 30 basins in a two-dimensional parameter space.

7.5.1 Hierarchical Clustering Using a New Dissimilarity Measure to Classify the Data-Rich Basins

All the data-rich basins are clustered using an agglomerative hierarchical clustering algorithm with the proposed dissimilarity measure. To cluster these basins, the proposed dissimilarity measure is calculated for all possible pairs of basins and the dissimilarity matrix is generated. Using this matrix and the 'average' linkage method, the agglomerative hierarchical clustering algorithm creates a dendogram (Figure 7-6). It is decided to cluster the basins into two groups. Therefore, a cutoff is selected near the top of the dendrogram in which the red dash line is separated from the blue solid line.

The geographical location of basins that belong to each class are displayed in Figure 7-7. The map of clustered basins shows that each class consists of basins from both western and eastern parts of the study area. This is interesting because the climatic and topographic condition in the western area is completely different from the eastern regions. Although the significant basin descriptors that affect the spatial variation of the calibrated models are investigated later after supervised classification, this map strengthens the hypothesis that climatic or topographic basin characteristics are not the significant attributes.

In Figure 7-8, the results of clustering using two common distance-based measures, Euclidean and Seuclidean, on model parameters are compared with the proposed measure. In the first row, the

basins are clustered into two classes. The distance based measures (Figure 7-8a and Figure 7-8c) separate one basin with the high shape parameter from the rest because of its far distance from the other points in the parameter space. However, the proposed measure is able to find 11 more basins similar to this basin. This means the overall performance of φ function for floodplain mapping cannot be estimated by just looking at the values of its parameters, and the basin response to the model structure is a more important factor that should be taken into account. The proposed dissimilarity measure considers this factor by running the φ function on other basins. The second row of Figure 7-8 displays the difference of three measures when basins are clustered into three groups. The completely different clustering results generated by each of these measures explain the importance of selecting an appropriate dissimilarity measure for a clustering problem.

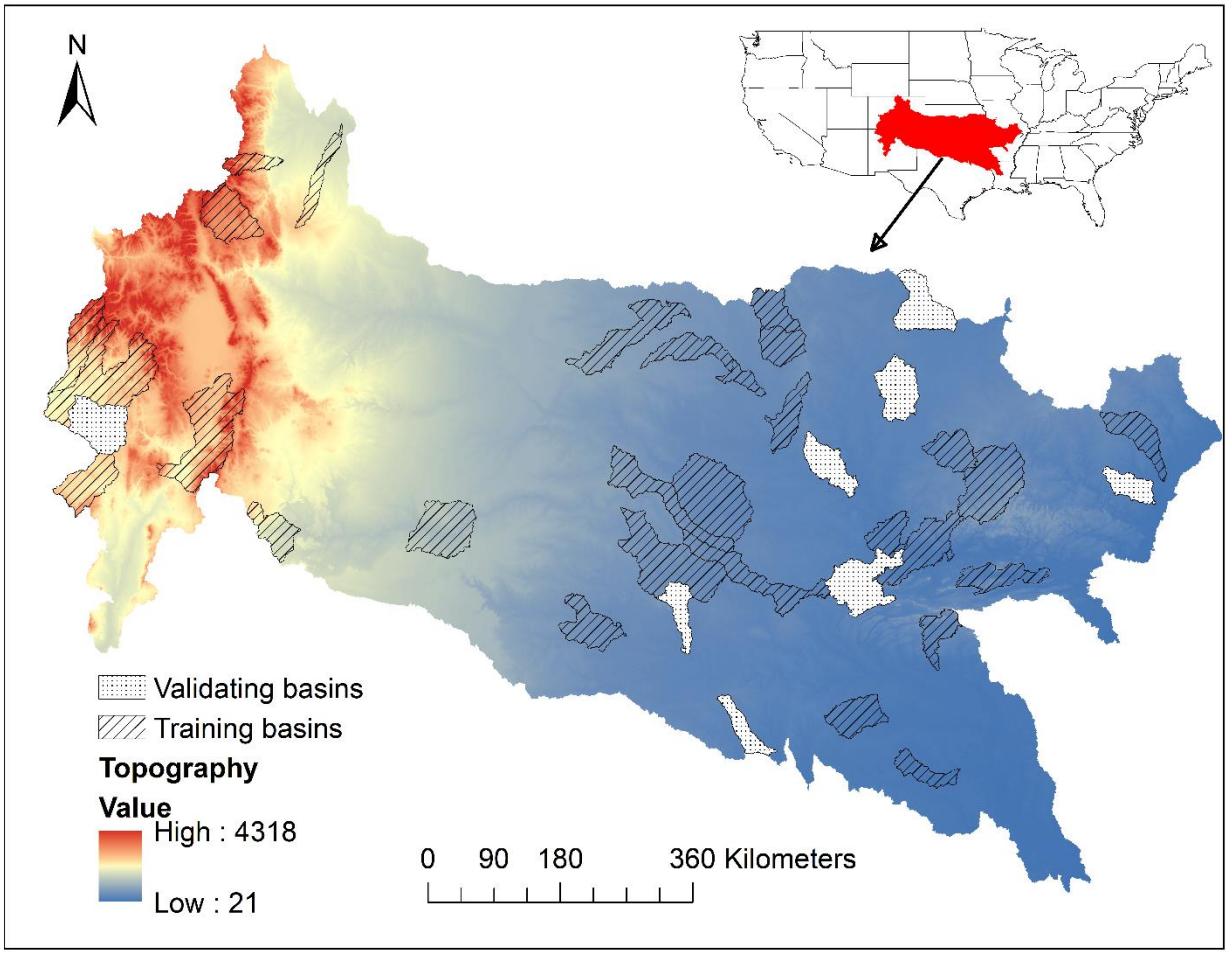


Figure 7-4 Location of study area inside the United States as well as location of training and validating basins inside the study area. The color bar shows the topographic change across the study area.

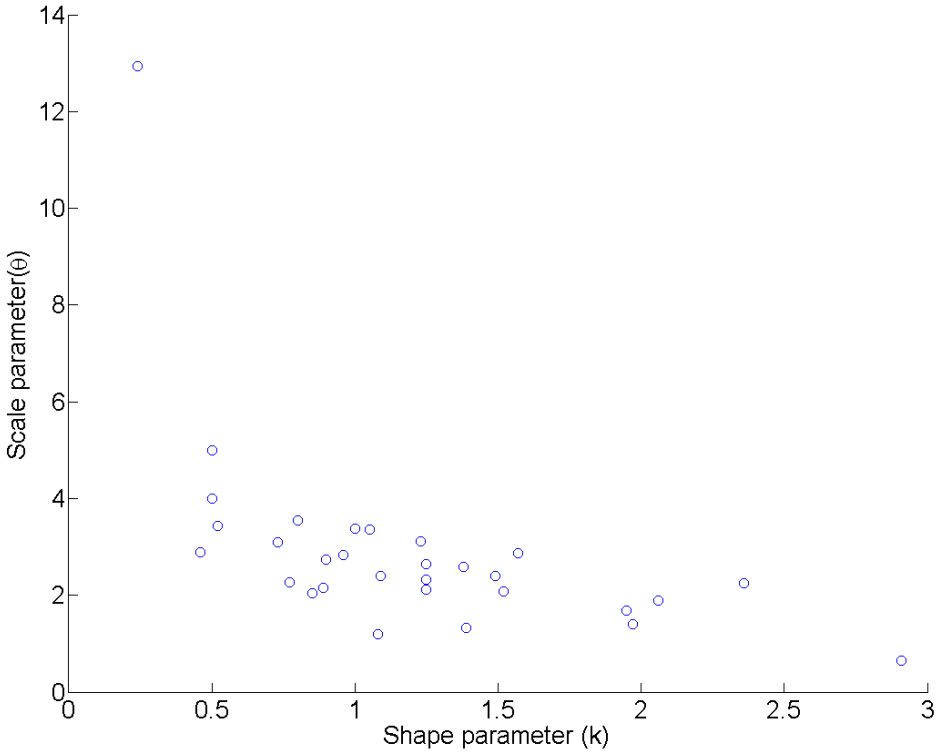


Figure 7-5 Position of data-rich basins in the parameter space. Each point refers to the parameters of a calibrated function for a given basin

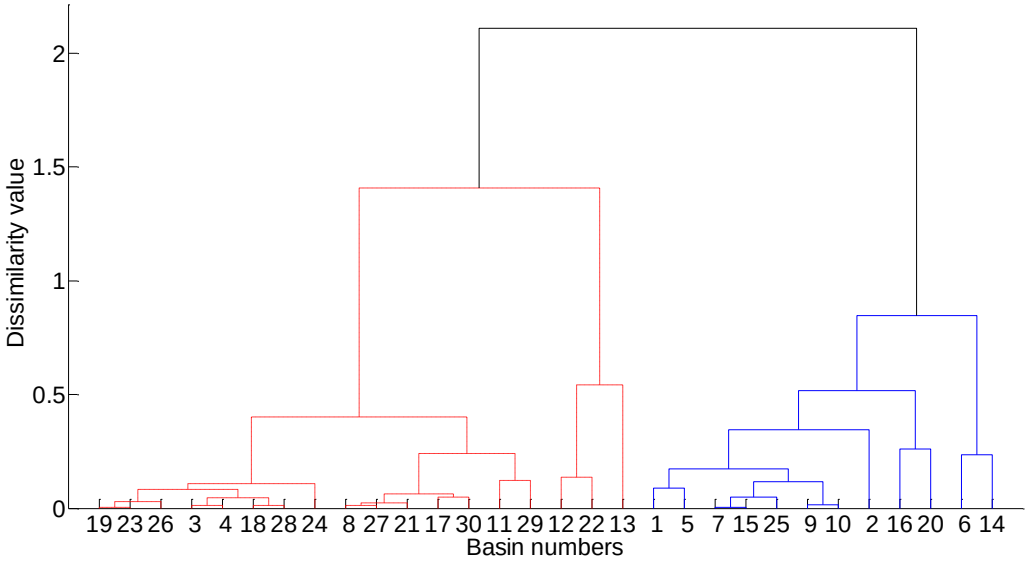


Figure 7-6 Dendrogram shows how the basins are joined based on their similarity at different levels of the tree cluster. The red and blue colors are used to separate the final two clusters.

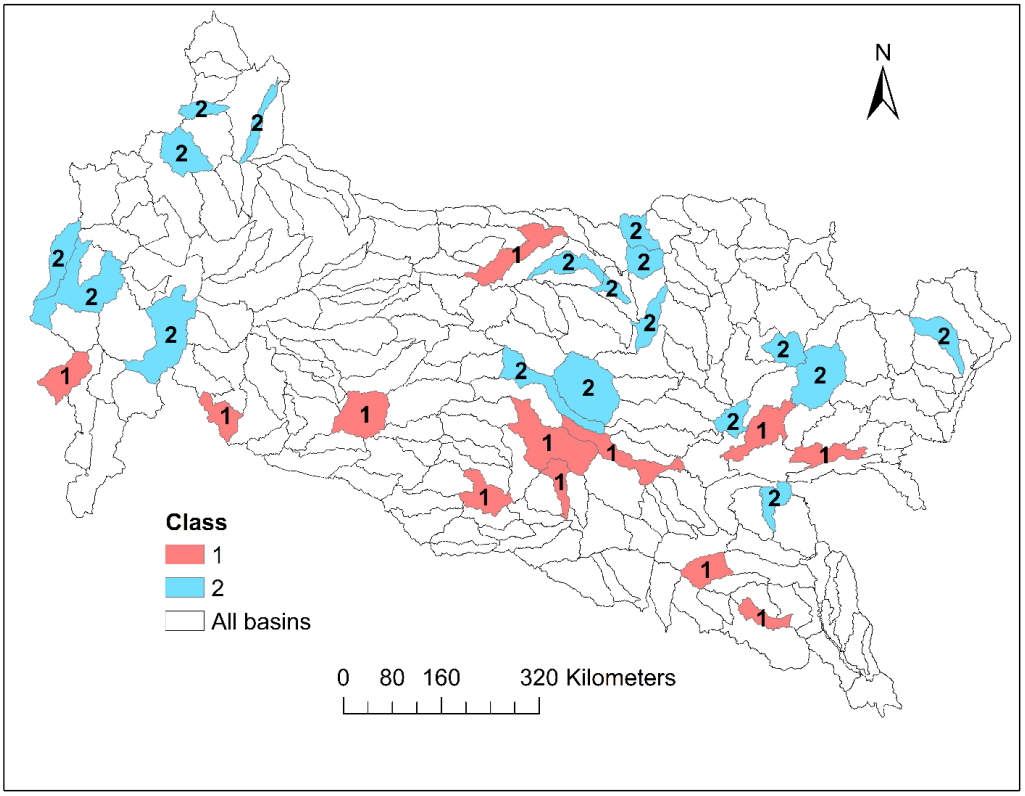


Figure 7-7 Map of clustered training basins

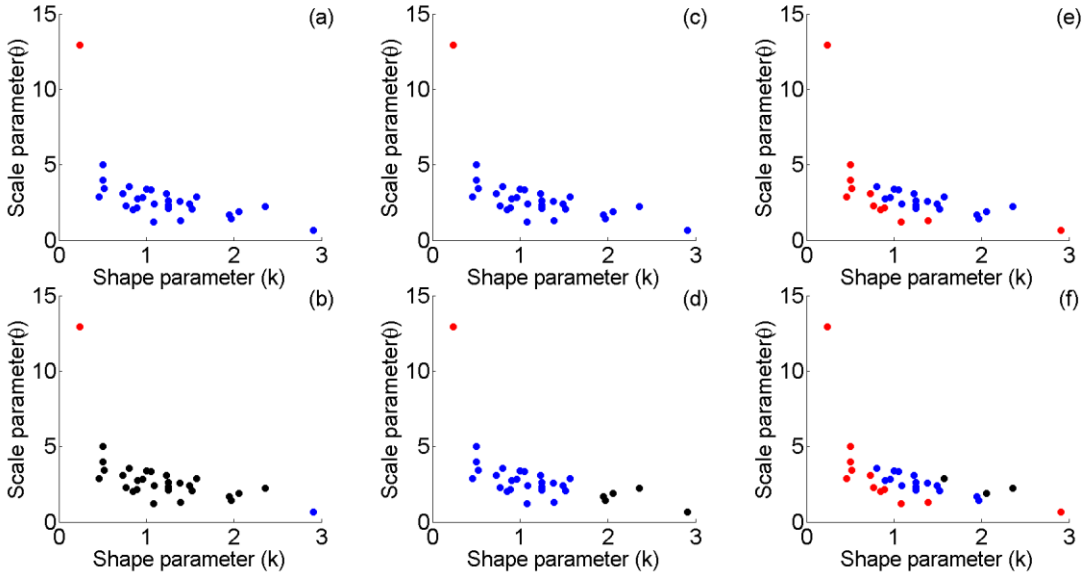


Figure 7-8 Clustered basins in the parameters space using: Euclidean measure for two (a) and three (b) classes, Seuclidean measure for two (c) and three (d) classes, and the new proposed measure for two (e) and three classes (f). The blue and red colors are used to distinguish two different classes and, the black color is added when three classes are generated

7.5.2 Training a Supervised Classifier to Find the Significant Basin Descriptors, and the Similarity Metric

After establishing the class labels for all the data-rich basins, a supervised classifier, named decision tree algorithm (Jafarzadegan et al., 2018), is performed using 25 basin descriptors related to shape, topography, climate, land use, and hydrography of the basins as presented in Table 7-2. Considering the important role of spatial proximity in many of the past regionalization studies, two additional attributes related to the geographical location of basins (latitude and longitude) are also considered in the list of potential basin descriptors. These basin descriptors are derived using a 30m-horizontal resolution DEM from the National Elevation Dataset (NED), a set of climate rasters (average temperature, average precipitation, wettest month precipitation) from WorldClim-Global data, the 2011 National Land Cover Dataset, the National Hydrography Dataset (NHD) flowlines and the NHD basin boundaries. The output from this supervised learning is a list of significant basin descriptors and the classification pattern.

Figure 7-9 presents the trained decision tree algorithm for classifying the basins. The tree classifier recognizes the Centroid Y (CY), the Average Highest Precipitation (AHP), and the STD Elevation (SE) as the only 3 attributes out of 25 potential ones which are required to estimate the class labels of target basins. In other words, this trained tree is proposed as an appropriate physical/climatic similarity metric for selecting the donor basins. The weights of CY, AHP and SE are also determined as 0.52, 0.23 and 0.25 respectively which shows the dominant role of CY for classification of basins in this problem. Basically, it shows that most of the basins above the line CY=1424294 m are similar and should be distinguished from those located below this line. In other words, the φ function developed for class 1 can be used for almost all basins above this line while the other φ function developed for class 2 can be used for remaining basins below the line. Using only CY for classifying the basins can put a few basins in the wrong class. Therefore, two additional attributes related to the climate and topography, namely the precipitation in the wettest month of the year, and the standard deviation of the elevation in basin, are used inside the tree classifier to improve the classification results.

Table 7-2 Potential basin descriptors related to the shape, location, hydrography, climate, topography and land use of a basin

Factors	Basin descriptors	Description
Shape and location	Area (A) (km^2)	Area of basin
	Perimeter (P) (km)	Perimeter of basin
	Circulatory Factor (CF)	Basin area/area of a circle having a perimeter equal of that basin
	Centroid_X (CX) (m)	Horizontal component of centroid of basin
	Centroid_Y (CY) (m)	Vertical component of centroid of basin
Hydrography	Main Stream Length (MSL) (m)	Length of stream with the highest Strahler's stream order in basin
	Main Stream Slope (MSS) (m)	Slope of stream with the highest Strahler's stream order in basin
	Drainage Density (DD) (1/km)	Total length of flowlines in basin/area of basin
	Drainage Area (DA) (km^2)	Total area directing water toward outlet of basin
	Stream order range (SOR)	Difference between maximum and minimum stream's order in basin
Climate	Average Annual Temperature (AAT) ©	Average of annual temperature in basin
	Average Annual Precipitation (AAP) (mm)	Average of annual precipitation in basin
	Average Highest Precipitation (AHP) (mm)	Average of annual precipitation in wettest month
	STD Annual Temperature (SAT) ©	Standard deviation of annual temperature in basin
	STD Annual Precipitation (SAP) (mm)	Standard deviation of annual precipitation in basin
	STD Highest Precipitation (SHP) (mm)	Standard deviation of annual precipitation in wettest month
Topography	Average Elevation (AE) (m)	Average of elevation in basin
	Average Slope (AS) (%)	Average of slope in basin
	STD Elevation (SE) (m)	Standard deviation of elevation in basin
	STD Slope (SS) (%)	Standard deviation of slope in basin
	Relief (R) (m)	Difference between elevation of highest and lowest points in basin
Land Use	Water Percentage (WP) (%)	Percentage of water area in basin
	Urban Percentage (UP) (%)	Percentage of urban area in basin
	Average Roughness Coefficient (ARC)	Average of manning's roughness coefficient in basin
	STD Roughness Coefficient (SRC)	Standard deviation of manning's roughness coefficient in basin

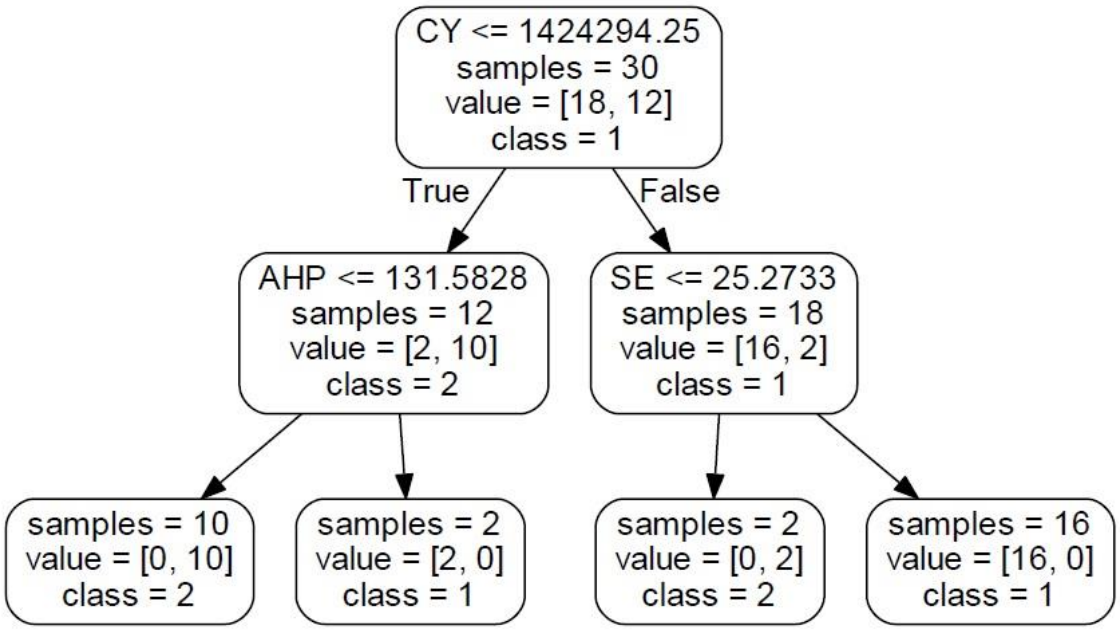


Figure 7-9 Trained Decision tree algorithm includes three significant basin descriptors (CY,AHP, SE) used as final physical/climatic similarity metric to select the donor basins for a target basin

7.5.3 Aggregation Process

In addition to the clustering and classification step, the 30 calibrated φ functions are aggregated to form m new φ functions denoted by $\hat{\varphi}_t$ (t=1,2,...,m) using a pairwise aggregation based on dendrogram. Because φ function can be easily derived from CDF, the aggregation process between two φ functions can be considered as aggregation of CDFs. To make a pairwise aggregation between two calibrated models of basin i and j (CDF_{gamma} (k_i, θ_i) and CDF_{gamma} (k_j, θ_j)), n_i and n_j samples generated from two CDF_{gamma} are combined, and a new CDF_{gamma} is fit on the new sample. The numbers of generated samples are defined using Equations 5-9 and 5-10.

$$n_i = \frac{\Delta_{ij}}{\Delta_{ij}+\Delta_{ji}} \times 10000 \tag{7-9}$$

$$n_j = \frac{\Delta_{ji}}{\Delta_{ij}+\Delta_{ji}} \times 10000 \tag{7-10}$$

where n_i and n_j are the number of samples generated from the CDF_{gamma} (k_i, θ_i) and CDF_{gamma} (k_j, θ_j) respectively. By using these two equations, the higher proportion of the total samples

belong to the CDF that generates less error on the other basin. This assures that the CDF with higher stability has more weight in generating the aggregated CDF. Using Eq. 5 the aggregated CDF is converted to $\hat{\varphi}$, which is representative of a particular class of basins.

Figure 7-10 illustrates $\hat{\varphi}_1$ and $\hat{\varphi}_2$ functions aggregated from the original calibrated φ functions. Looking into this figure, for example, the probability of flooding for a given cell inside a basin with $HAND = 2m$ is estimated to be around 65% if the basin belongs to class 1, and the probability of flooding for this cell drops to a value around 35% if the basin belongs to class 2. $\hat{\varphi}_1$ and $\hat{\varphi}_2$ are the results of converting a 30 dimensional problem into 2 dimensions. Therefore, for any basin inside the study area, one of these two functions can be used to generate the probabilistic floodplain map.

7.5.4 Framework Validation

To validate the performance of the proposed framework, first the efficacy of the aggregation process is explored by applying a regionalization test on 30 training basins. Then, the same regionalization test is applied on 8 validating basins to investigate the performance of the hybrid classification framework in selecting the proper donor basins. For applying the regionalization test, *m* floodplain maps are generated for each basin using $\hat{\varphi}_t$ (*t*=1,2,...,*m*). Considering the available reference floodplain maps for these basins, the error of predictions corresponding to $\hat{\varphi}$ functions are calculated. The net regionalization error is calculated for each basin using Equation 5-11.

$$\Delta_{r,it} = e_{r,it} - e_{ii} \tag{7-11}$$

where $e_{r,it}$ denotes the regional error in basin *i* using aggregated function $\hat{\varphi}_t$. e_{ii} is the local error which is calculated by using locally calibrated φ on the same basin *i*. Deducting the local error from the regional error gives $\Delta_{r,it}$ which is the net regional error on basin *i* using aggregated function $\hat{\varphi}_t$. Assuming a given basin *i* classified as *c*, the regionalization test is successful if these two conditions are met:

$$min\left(\Delta_{r,it}, t = 1,2,\dots,m\right) = \Delta_{r,ic} \tag{7-12}$$

$$\Delta_{r,ic} \leq \varepsilon \tag{7-13}$$

where ε denotes the maximum error, which is still acceptable for the prediction. This number is decided by the modeler based on the structure, requirements and limitations of the problem. A successful regionalization reflects the success of both basin classification and aggregation process. To test the aggregation process, the regionalization test is applied on 30 training basins. Using the same basins which are already used for training the classifier removes the impact of classification step. Therefore, class *c* in condition (a) of test, refers to the result of hierarchical clustering. To test the success of aggregation process, the value of $\Delta_{r,i1}$ and $\Delta_{r,i2}$ are calculated and compared for all 30 basins ($i = 1,2,\dots,30$) in Figure 7-11. For each basin, the lowest error should be the same with the class of that basin to meet the first condition of the test. Basins 17, 27 and 30 are the only ones that fail this condition, but the difference between $\Delta_{r,i1}$ and $\Delta_{r,i2}$ is negligible for these three basins (specifically for basin 17 and 27). The second condition is also met very well for all basins where the minimum error is always a small value between 0 and 1 percent.

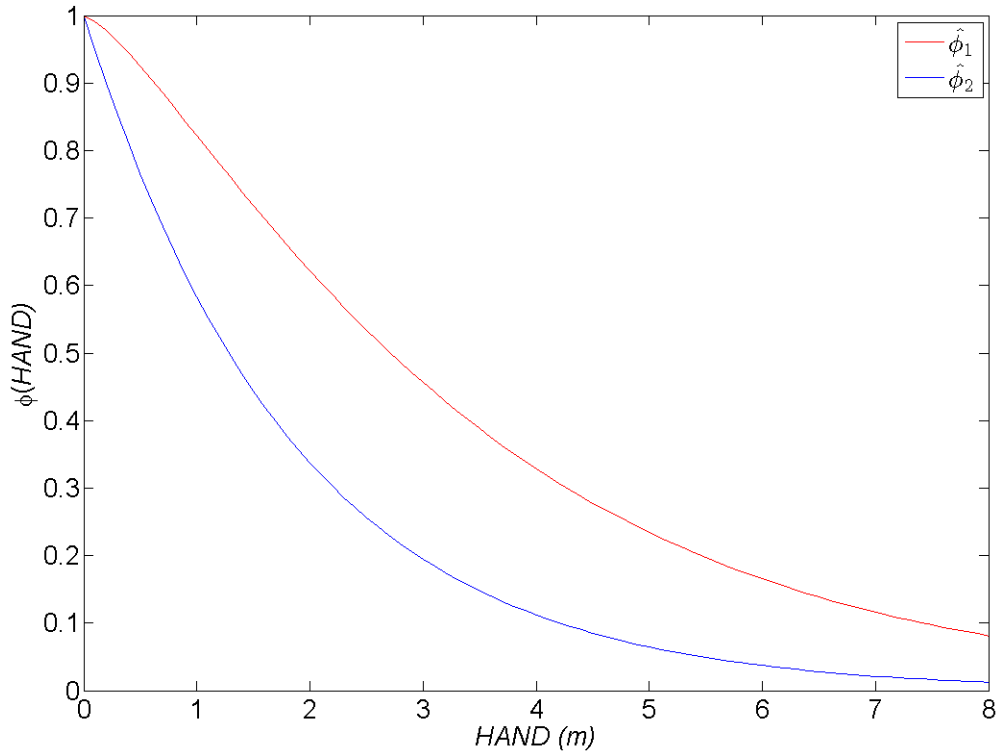


Figure 7-10 Two aggregated functions developed for probabilistic floodplain mapping in the study area: The red and blue curves are used for the basins that belong to class 1 and 2 respectively.

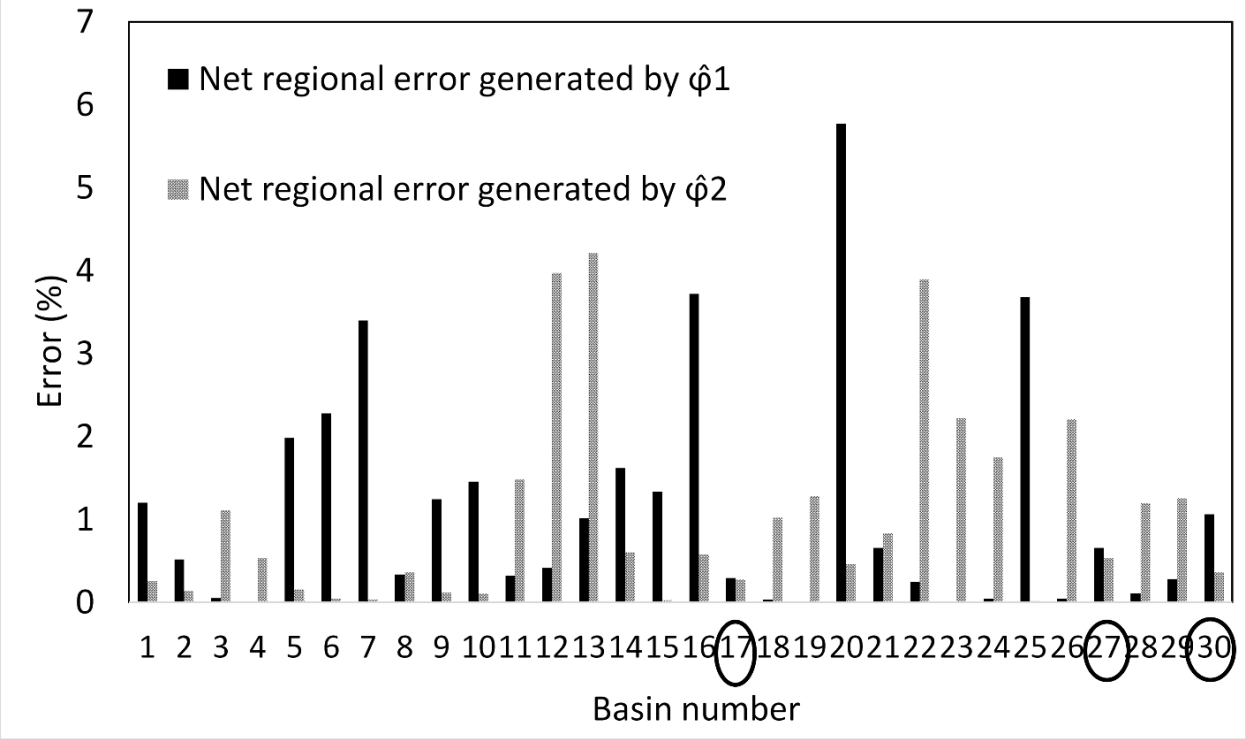


Figure 7-11 Net regional errors ($\Delta_{r,i1}$ and $\Delta_{r,i2}$) generated for each training basin by $\hat{\varphi}_1$ and $\hat{\varphi}_2$. The basin numbers, highlighted by circle, are those basins which failed the first condition of regionalization test because min ($\Delta_{r,i1}$ and $\Delta_{r,i2}$) $\neq \Delta_{r,ic}$

In the second test, the regionalization test is applied on 8 validating basins from the study area. This time the outcome of hybrid classification framework, namely the trained classifier, is used to determine class *c*. Assuming the success of aggregation process on training basins from the first test, the regionalization test highlights the effectiveness of the proposed framework in proper classification of basins. In other words, the regionalization test on 30 training basins investigates the success of aggregation process while the regionalization test on 8 validating basins explores the success of developed physical/climatic similarity metric (or trained classifier) in basin classification. Three significant basin descriptors, CY, AHP and SE, are calculated, and the trained tree classifier (Figure 7-9) is used to identify the class labels of these basins (Table 7-3). Figure 7-12 presents the net regional errors corresponding to both aggregated functions ($\hat{\varphi}_1$ and $\hat{\varphi}_2$). To evaluate the first condition of regionalization test, the identified class labels for these basins (the last column of Table 7-3) is compared with the minimum net regional errors in Figure 7-12. The comparison shows that except basin 5, all basins are classified correctly. The minimum net regional errors are also smaller than 0.5 percent which shows the efficacy of the proposed

framework. Figure 7-13 illustrates the location of classified validating basins. Basin 5, which is the only incorrectly classified basin among the eight validating basins, is located in the western part of the study area. It shows that western basins located in the mountainous regions of the study area need additional investigation.

Table 7-3 The value of three significant basin descriptors and the identified class labels for validating basins

Basin	HUC8	AHP (mm)	SE (m)	CY (m)	Class label
1	11130202	128	75	1292632	2
2	12030104	126	50	1131129	2
3	11090204	151	79	1330058	1
4	11070107	126	41	1490105	1
5	14080103	53	119	1545022	1
6	11070205	135	18	1592690	2
7	10290102	144	29	1704667	1
8	11010012	125	52	1461296	1

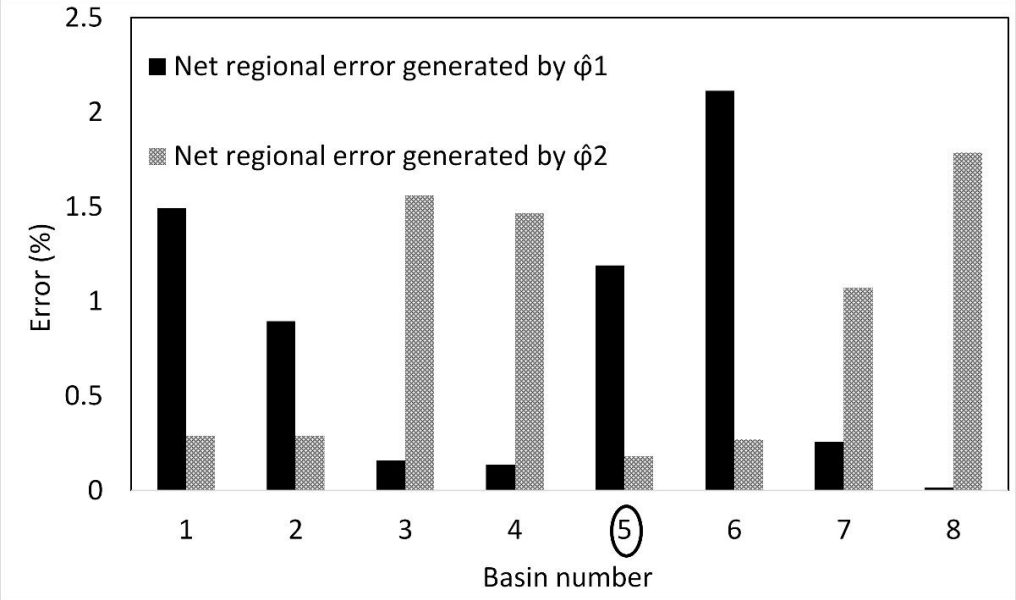


Figure 7-12 Net regional errors ($\Delta_{r,i1}$ and $\Delta_{r,i2}$) generated for each validating basin by $\hat{\varphi}_1$ and $\hat{\varphi}_2$. The basin numbers, highlighted by circle, are those basins which failed the first condition of regionalization test because min ($\Delta_{r,i1}$ and $\Delta_{r,i2}$) $\neq \Delta_{r,ic}$

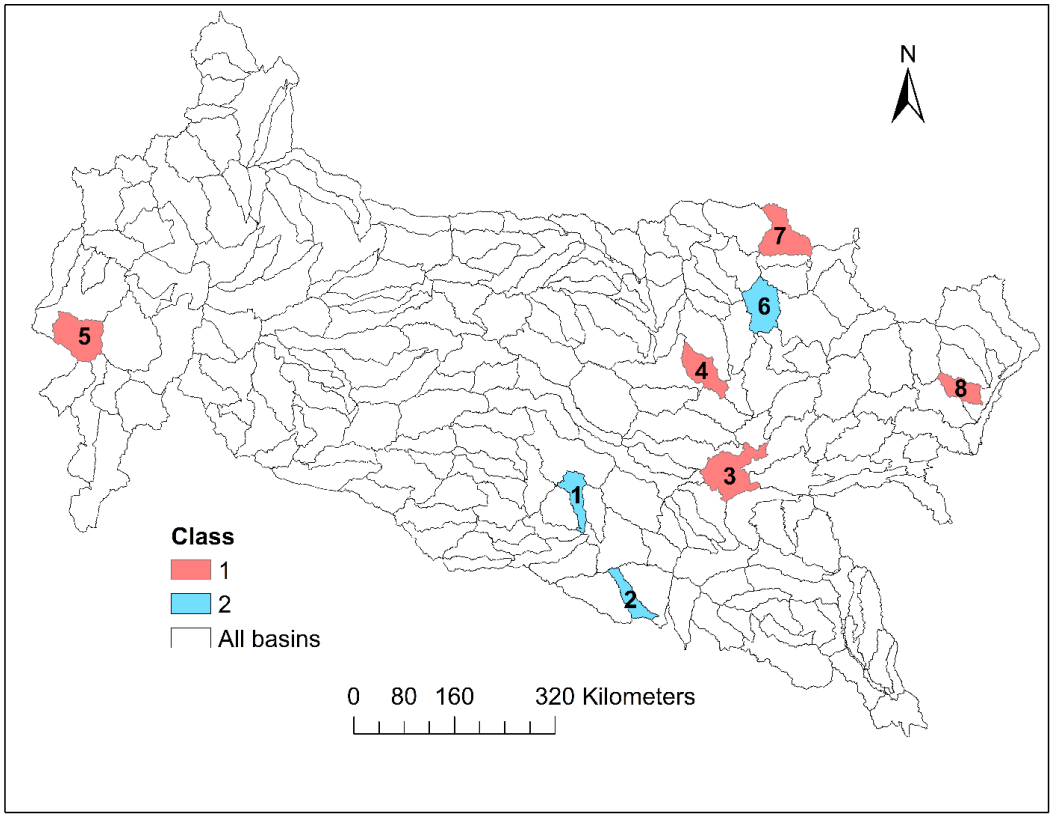


Figure 7-13 Map of classified test basins

7.6 Discussion

The fundamental role of donor basin selection in the success of a similarity-based regionalization method is the main motivation for proposing the hybrid classification framework in this study. Overall, the systematic approach used for the selection of donor basins in a similarity-based regionalization, and the generality of the proposed framework being applicable for different modeling purpose are two major novelties of this chapter which are discussed further below.

Applying a systematic approach for finding an appropriate similarity metric is of paramount importance when the modeling process and the physics of the problem are not completely understood. It is difficult to predetermine the basin descriptors and define a similarity metric in these cases. The case study used here is a good example where a statistical gamma-based function uses a hydrogeomorphic feature to generate probabilistic floodplain maps.

Finding the significant basin descriptors and the structure of similarity metric is not possible without using a systematic approach. Figure 7-14 and Figure 7-15 present the scatter plots of

different basin descriptors versus the gamma-based function parameters. The linear trend line and the R^2generated for each subplot show the linear relationship between the basin descriptors and φ function parameters. The shape parameter shows some correlations with three climatic descriptors, namely AAP, AHP and AAT as well as AE. For scale parameter, MSS shows high correlation compared to other basin descriptors. Besides, AAT, SAT, AS, AE, R and SE are also correlated to scale parameter. Comparing these basin descriptors with three significant attributes selected by the proposed hybrid classification framework, namely CY, AHP and SE, reveals large differences. CY, the most dominant attribute for regionalizing the basins, is not among these linearly correlated attributes. AHP and SE show small correlation with shape and scale parameters, respectively. Furthermore, considering the factors affecting flood extent in a basin, and the results of previous studies on flood inundation mapping, one would likely select a set of topographic, climatic and land use attributes for defining the physical/climatic similarity metric which is different from the solution provided by the proposed framework. The significant differences between basin descriptors selected by the proposed framework and those selected by either regression-based analysis or subjective decisions demonstrates the importance of this systematic approach.

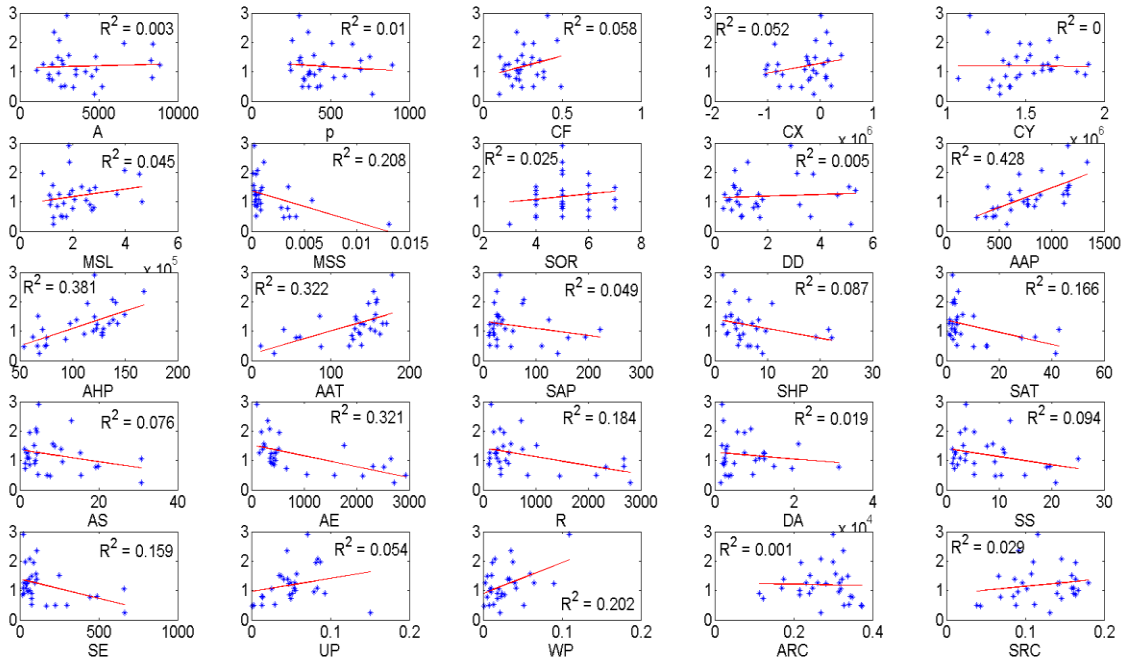


Figure 7-14 Linear correlation between basin descriptors (x axis) and shape parameter of φ function (k) (y axis)

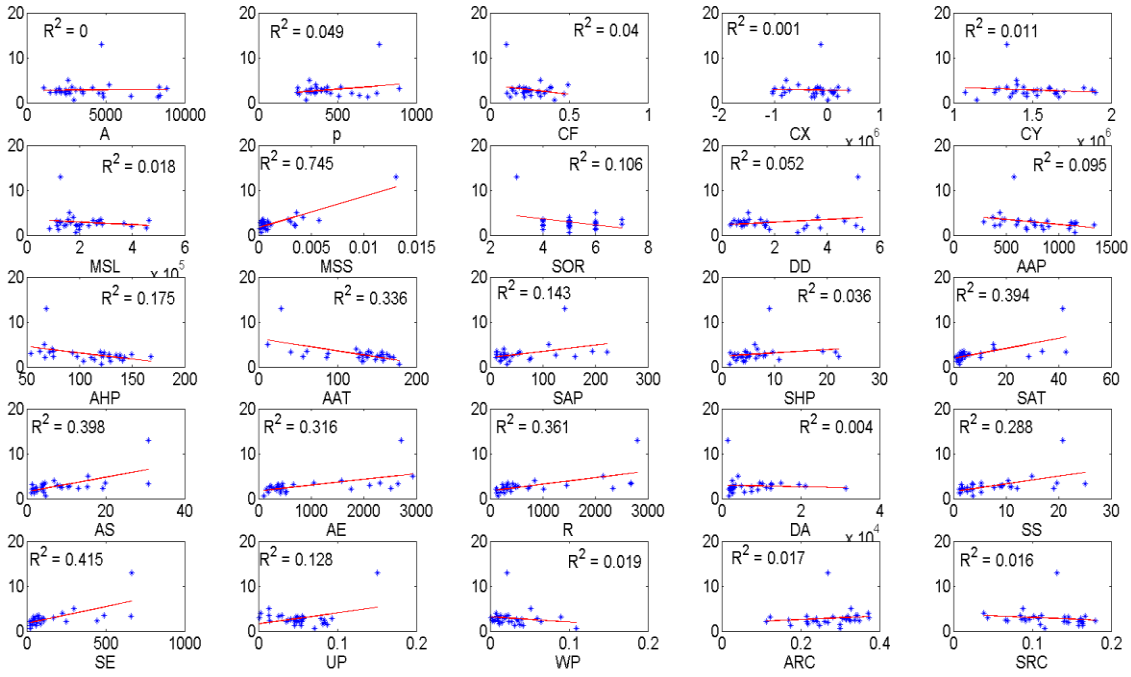


Figure 7-15 Linear correlation between basin descriptors (x axis) and scale parameter of φ function (θ) (y axis)

One of the most interesting findings of this study is the recognition of CY as the most important factor for basin classification in floodplain mapping. This means basins above a given latitude responds differently to a given flood event compared to those located below this line. The impact of latitude on floodplain mapping is a new finding that was unknown using the past linear regression-based analyses. In addition, considering that the aggregated φ for class 1 is above the aggregated φ for class 2 (Figure 7-10), northern basins inside the study area are generally more prone to larger inundation extents for a given flood magnitude.

Regionalization can be utilized for a broad range of problems and it shouldn't be considered only for transferring the hydrologic models or flood quantiles in RFFA. To demonstrate this fact, the hybrid classification framework proposed in this study is used for regionalizing a statistical model used for probabilistic floodplain mapping and its efficacy is evaluated. The validation results show that only one basin out of eight is classified wrongly, giving a high success rate of 87.5%. It should be noted that like any machine learning algorithms, the classification methods are prone to some degree of errors, and no single algorithm will fit perfectly to the available data to provide 100% success rate. The possible sources of errors in the proposed framework arises from uncertainty in

two classification algorithms, uncertainty in calibration procedure for finding the locally calibrated parameters of the model, and uncertainty in basin descriptor calculations. The linkage method used to union the groups of data points in the hierarchical clustering algorithm is one of the possible sources of classification errors. While this study uses the "average" of dissimilarity measure values, using other linkage methods may provide slight changes in the clustering results. A different set of class labels generated by hierarchical clustering algorithm can alter the trained classifier which means the significant descriptors and the structure of decision tree can be changed. The type of supervised classifier is another important component which can affect the accuracy of the final results. In this study, the decision tree algorithm is used because of its broad applications in the similar past works. To consider the uncertainty of classification framework and reduce the subjectivity involved in the proposed framework, various linkage methods in the hierarchical clustering algorithm, and an ensemble of supervised learning algorithms for the second classification can be applied and tested in future studies.

Uncertainty in the model calibration is related to the optimization procedure used to find the model parameters. In addition, the equifinality concept, defined as different sets of parameters providing similar results, is another issue which can pose some errors. In this study, the uncertainty in model parameters is neglected to rely on the most optimum parameter set for each basin. Using a range of calibrated parameter sets instead of a single optimum set may reduce the parameter uncertainty. Lastly, the uncertainty in the basin descriptor calculations originates from the uncertainty of input data and the averaging. The DEM, climate and land use rasters used to find the basin descriptors are all resultant of remote sensing imagery which has their own level of uncertainties. Also, we take the spatial average of all pixels inside a basin to calculate the basin descriptor. Using one single value, which is representative of a large heterogeneous basin, cannot reflect the basin characteristics properly and causes some additional errors in the final results.

7.7 Conclusion

The proposed hybrid classification framework provides a systematic approach for selecting the appropriate donor basins in a similarity-based regionalization. Typically, a physical/climatic similarity metric is used to identify the donor basins. The main assumption is that the basins, determined as physically similar by this metric, are functionally similar as well which is not always

true. By focusing on this assumption, the proposed framework produces a physical/climatic similarity metric which tries to classify basins based on functional similarity.

This framework uses a large number of data-rich basins as input and provides a trained classifier as output referred to as a physical/climatic similarity metric. The trained classifier can identify the class label of a data-scarce basin. The data-rich basins which have class label similar to data-scarce target basin are recognized as donor basins. The physical/climatic similarity metric explored by this framework is the result of a supervised learning algorithm. In supervised learning algorithms, the best mapping function for relating the attributes to class labels are determined. Therefore, if the class labels used for training the classifier are selected properly, the final similarity metric would be an appropriate metric as well. To assure this, we propose a novel dissimilarity measure within a hierarchical clustering algorithm which generates the class labels of data-rich basins as input to the supervised classifier. This dissimilarity measure considers the model structure and the functional behavior of a basin by running the locally calibrated models on other basins.

The efficacy of the proposed framework is tested for regionalizing a statistical gamma-based function for probabilistic floodplain mapping. Results show that the vertical component of geographical location of a basin (latitude) is the dominant attribute for basin classification in response to floodplain mapping. These results show that the northern basins mostly belong to class 1 and are prone to a larger inundation extent for a given flood event. The standard deviation of basin elevation and the average of precipitation in the wettest month are two other important basin descriptors which are selected for defining the physical/climatic similarity metric. The developed metric is tested for 8 validating basins, and the errors of produced floodplain maps, based on the proposed regionalization framework, is compared with the errors of floodplain maps generated by locally calibrated models. Results show an 87.5% success rate in which the errors of regionalization for 7 out of 8 basins are very similar to the errors of local calibration. In future studies, the proposed framework can be applied for other regionalization purposes. Specifically, this framework can be applied for hydrologic similarity-based regionalization problems to determine a proper physical/climatic similarity metric. The capability of this framework for detecting some hidden basin descriptors which are not easily found, makes this approach an attractive solution to different regionalization problems.

CHAPTER 8. SYNTHESIS

The practical contribution of this research to efficient floodplain mapping is presented by developing statistical models which create 100-year floodplain maps in data-scarce regions within the state of North Carolina and the Contiguous United States (CONUS). In Chapter 2, a regression model is developed for 100-year floodplain mapping in North Carolina. The regression model works well in mid-altitude regions, but it underestimates and overestimates in the flat and mountainous areas respectively. The three different behaviors of the *HAND*-based method shown by the regression model creates the idea that watersheds should be classified into three groups for floodplain mapping. In Chapter 3, a geomorphic framework including a supervised random forest classifier and a Probabilistic Threshold Binary Classifier (PTBC) are coupled to create 100-year probabilistic floodplain maps for any watershed in the United States. The average error of the predicted flood extent maps is around 14% which demonstrates the reliability and robustness of the proposed framework for large-scale floodplain mapping across the United States. Overall, the fast and cost-effective structure of these models, as well as their reasonable accuracy for preliminary estimation of floodplains demonstrate the practical application of these models in data-scarce regions.

In addition to the practical role of the developed models for efficient floodplain mapping, the theoretical contribution and the major findings of this dissertation are summarized as follows:

- **The role of basin descriptors in floodplain mapping**

This research evaluates the significance of different physical/climatic basin descriptors on spatial variability of floodplains across the US. The results of two studies explained in Chapter 2 and 3, indicate that depending on the scale and geographical location of the study area, the impact of basin descriptors can be different. In the study conducted in North Carolina, three topographic characteristics, namely average slope, average elevation and main stream slope are the major drivers that reflect the spatial variability of floodplains. In the continental scale study for the entire United States, however, seven basin descriptors corresponding to climate, land use and topography of the watersheds are the main reflectors of the spatial variability of floodplains. Comparing the

role of selected basin descriptors at both scales also reveals that the average slope of watershed is the most significant characteristic for estimating the floodplains in data-scarce regions.

- **Probabilistic version of the *HAND*-based method**

In Chapter 4, it is demonstrated that the simple thresholding method used in the deterministic *HAND*-based approach can be replaced by a probabilistic function of *HAND*, derived from the Cumulative Distribution Function (CDF) of threshold. The results of this study show that aside from the inherent benefits of probabilistic maps compared to deterministic ones (e.g. giving more information for decision making and risk analysis), the probabilistic approach can improve the accuracy of floodplain mapping by reducing the overprediction and underpredictions generated by the deterministic approach.

- **Advancement in regionalization techniques**

One of the most notable achievements of this research is presented in Chapter 5, where a general regionalization framework is proposed. First a novel dissimilarity measure is introduced inside the hierarchical clustering algorithms. This measure improves the performance of clustering algorithm for deciding the best donor basins. Furthermore, the regionalization framework suggested in Chapter 5 proposes a systematic approach for selecting the most significant basin descriptors and an appropriate physical/climatic similarity metric. The proposed framework reduces the high subjectivity that exists in the selection of donor basins. This framework finds "vertical component of centroid (or latitude)" as a dominant descriptor of spatial variabilities in the probabilistic floodplain maps. This is an interesting finding which shows how a proper selection of dissimilarity measure and using a systematic approach can help to explore the hidden descriptors. It is demonstrated that using common methods, such as correlation coefficient calculation, or stepwise regression analysis, will not reveal the critical role of latitude on the spatial variability of floodplains.

8.1 Limitation and future work

The focus of this study is on the identification of 100-year floodplains. Although the 100-year return period is the most common recurrence interval used for flood risk management tasks, the integration of these maps with floodplain areas corresponding to other return periods (e.g. 50, 200

and 500) provides much more information for decision makers. The FEMA FIRMs used as the main input of the proposed models are mostly available for 100-year flood events. The lack of reliable floodplain maps corresponding to other return periods limits the application of the proposed models to 100-year floodplain mapping problems. A potential future research objective is to create floodplain maps corresponding to other return periods by well-calibrated hydrodynamic models at different locations and use them as the input to the proposed models. Relying on FEMA FIRMs as the reference maps for training the proposed models is another limitation of this study. These maps have variable levels of uncertainties at different locations which affect the performance of the proposed models. For future work, using observed floodplains for training rather than FEMA FIRMs, or incorporating the uncertainty of FEMA FIRMs into the modeling task will provide a more rational approach to floodplain mapping.

Making a proper selection of training watersheds and increasing their total numbers in the regionalization techniques are other important factors which can be considered for improving the model performances in future work. The proposed models in this study may be extended to floodplain mapping at the global scale. This needs a thorough collection of training reference floodplain maps from different location in the globe. Lastly the regionalization framework proposed in Chapter 5 can be used for other environmental problems where the main drivers defining the similar basins with respect to the purpose of the problem are not well understood, and a systematic approach for finding the most significant basin descriptors is required.

LIST OF REFERENCES

Ackerman, C.T., 2005. HEC-GeoRAS; GIS Tools for support of HEC-RAS using ArcGIS. U. S. Army Corps Eng. Davis.

Acreman, M.C., 1985. Predicting the mean annual flood from basin characteristics in Scotland. Hydrol. Sci. J. 30, 37–49. https://doi.org/10.1080/02626668509490970

Afshari, S., Tavakoly, A.A., Rajib, M.A., Zheng, X., Follum, M.L., Omranian, E., Fekete, B.M., 2018. Comparison of new generation low-complexity flood inundation mapping tools with a hydrodynamic model. J. Hydrol. 556, 539–556. https://doi.org/10.1016/j.jhydrol.2017.11.036

Aggett, G.R., Wilson, J.P., 2009. Creating and coupling a high-resolution DTM with a 1-D hydraulic model in a GIS for scenario-based assessment of avulsion hazard in a gravel-bed river. Geomorphology 113, 21–34.

Alfieri, L., Salamon, P., Bianchi, A., Neal, J., Bates, P., Feyen, L., 2014. "Advances in Pan-European Flood Hazard Mapping." Hydrological Processes 28 (13): 4067–4077.

Alfonso, L., Tefferi, M., 2015. Effects of uncertain control in transport of water in a river-wetland system of the Low Magdalena River, Colombia, in: Transport of Water versus Transport over Water, Operations Research/Computer Science Interfaces Series. Springer, Cham, pp. 131–144. https://doi.org/10.1007/978-3-319-16133-4_8

Alfonso, L., Mukolwe, M.M., Di Baldassarre, G., 2016. Probabilistic Flood Maps to support decision-making: Mapping the Value of Information. Water Resour. Res. 52, 1026–1043. https://doi.org/10.1002/2015WR017378

Alphen, J.V., Martini, F., Loat, R., Slomp, R., Passchier, R., 2009. Flood risk mapping in Europe, experiences and best practices. J. Flood Risk Manag. 2, 285–292. https://doi.org/10.1111/j.1753-318X.2009.01045.x

Lhomme, J., Sayers, P., Gouldby, B., Samuels, P., Wills, M., Mulet-Marti, J., 2008. Inundation modelling Recent development and application of a rapid flood spreading method, in: Flood Risk Management: Research and Practice. CRC Press, pp. 30–39.

Teng, J., Jakeman, A.J., Vaze, J., Croke, B.F., Dutta, D., Kim, S., 2017. Flood inundation modelling: A review of methods, recent advances and uncertainty analysis. Environ. Model. Softw. 90, 201–216.

Aronica, G., Bates, P.D., Horritt, M.S., 2002. Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE. Hydrol. Process. 16, 2001–2016.

Arumugam, M.S., Rao, M.V.C., 2006. On the performance of the particle swarm optimization algorithm with various inertia weight variants for computing optimal control of a class of hybrid systems. Discrete Dyn. Nat. Soc. 2006.

Baker, V.R., 1994. Geomorphological understanding of floods, in: Geomorphology and Natural Hazards. Elsevier, pp. 139–156.

Baldwin, D.S., Mitchell, A.M., 2000. The effects of drying and re-flooding on the sediment and soil nutrient dynamics of lowland river–floodplain systems: a synthesis. Regul. Rivers Res. Manag. 16, 457–467. https://doi.org/10.1002/1099-1646(200009/10)16:5<457::AID-RRR597>3.0.CO;2-B

Bansal, J.C., Singh, P.K., Saraswat, M., Verma, A., Jadon, S.S., Abraham, A., 2011. Inertia Weight strategies in Particle Swarm Optimization, in: 2011 Third World Congress on Nature and Biologically Inspired Computing. Presented at the 2011 Third World Congress on Nature and Biologically Inspired Computing, pp. 633–640. https://doi.org/10.1109/NaBIC.2011.6089659

Bates, P.D., De Roo, A.P.J., 2000. A simple raster-based model for flood inundation simulation. J. Hydrol. 236, 54–77. https://doi.org/10.1016/S0022-1694(00)00278-X

Bates, P.D., Horritt, M.S., Aronica, G., Beven, K., 2004. Bayesian updating of flood inundation likelihoods conditioned on flood extent data. Hydrol. Process. 18, 3347–3370. https://doi.org/10.1002/hyp.1499

Bates, P.D., 2004. Remote sensing and flood inundation modelling. Hydrol. Process. 18, 2593–2597. https://doi.org/10.1002/hyp.5649

Bates, Paul D., Matthew S. Horritt, and Timothy J. Fewtrell. 2010. "A Simple Inertial Formulation of the Shallow Water Equations for Efficient Two-Dimensional Flood Inundation Modelling." Journal of Hydrology 387 (1–2): 33–45. doi:10.1016/j.jhydrol.2010.03.027

Beven, K., Binley, A., 1992. The Future of Distributed Models: Model Calibration and Uncertainty Prediction. Hydrol. Process. 6, 279–298.

Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol. 249, 11–29. https://doi.org/10.1016/S0022-1694(01)00421-8

Beven, K., 2006. A manifesto for the equifinality thesis. J. Hydrol., The model parameter estimation 320, 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Berger, K.P., Entekhabi, D., 2001. Basin hydrologic response relations to distributed physiographic descriptors and climate. J. Hydrol. 247, 169–182.

Bhaskar, N.R., O'Connor, C.A., 1989. Comparison of Method of Residuals and Cluster Analysis for Flood Regionalization. J. Water Resour. Plan. Manag. 115, 793–808.

Blöschl, G., Sivapalan, M., 1995. Scale Issues in Hydrological Modelling: A Review. Hydrol. Process. 9, 251–290.

Borra, S., Di Ciaccio, A., 2010. "Measuring the Prediction Error. A Comparison of Cross-Validation, Bootstrap and Covariance Penalty Methods." Computational Statistics & Data Analysis 54 (12): 2976–89. doi:10.1016/j.csda.2010.03.004.

Bradley, A.A., Cooper, P.J., Potter, K.W., Price, T., 1996. Floodplain mapping using continuous hydrologic and hydraulic simulation models. J. Hydrol. Eng. 1, 63–68. https://doi.org/10.1061/(ASCE)1084-0699(1996)1:2(63)

Breiman, L., 1999. Random forests-random features Technical Report 576. Stat. Dep. UC Berkeley USA.

Burn, D.H., 1990. Evaluation of Regional Flood Frequency Analysis with a Region of Influence Approach. Water Resour. Res. 26, 2257–2265.

Burn, D.H., 1989. Cluster Analysis as Applied to Regional Flood Frequency. J. Water Resour. Plan. Manag. 115, 567–582.

Burn, D.H., Goel, N.K., 2000. The Formation of Groups for Regional Flood Frequency Analysis. Hydrol. Sci. J. 45, 97–112.

Burn, D.H., Zrinji, Z., Kowalchuk, M., 1997. Regionalization of Catchments for Regional Flood Frequency Analysis. J. Hydrol. Eng. 2, 76–82.

Cantisani, A., Giosa, L., Mancusi, L., Sole, A., 2014. "FLORA-2D: A New Model to Simulate the Inundation in Areas Covered by Flexible and Rigid Vegetation." Int J Eng Innov Technol 3 (8): 179–186.

Casas, A., Benito, G., Thorndycraft, V., Rico, M., 2006. The topographic data source of digital terrain models as a key element in the accuracy of hydraulic flood modelling. Earth Surf. Process. Landf. 31, 444–456. https://doi.org/10.1002/esp.1278

Castellarin, A., Burn, D.H., Brath, A., 2001. Assessing the Effectiveness of Hydrological Similarity Measures for Flood Frequency Analysis. J. Hydrol. 241, 270–285. https://doi.org/10.1016/S0022-1694(00)00383-8

Chiang Shih-Min, Tsay Ting-Kuei, Nix Stephan J., 2002. Hydrologic regionalization of watersheds. I: methodology development. J. Water Resour. Plan. Manag. 128, 3–11. https://doi.org/10.1061/(ASCE)0733-9496(2002)128:1(3)

Chow, V.T., 1954. The log probability law and its engineering applications Proc. Am Soc Civ. Engrs Sep 80.

Clubb, F.J., Mudd, S.M., Milodowski, D.T., Valters, D.A., Slater, L.J., Hurst, M.D., Limaye, A.B., 2017. Geomorphometric delineation of floodplains and terraces from objectively defined topographic thresholds. Earth Surf. Dyn. Gottingen 5, 369–385. http://dx.doi.org/10.5194/esurf-5-369-2017

Cobby, D.M., Mason, D.C., Horritt, M.S., Bates, P.D., 2003. Two-dimensional hydraulic flood modelling using a finite-element mesh decomposed according to vegetation and topographic features derived from airborne scanning laser altimetry. Hydrol. Process. 17, 1979–2000. https://doi.org/10.1002/hyp.1201

Committee on FEMA Flood Maps; Mapping Science Committee; Board on Earth Sciences and Resources; Water Science and Technology Board; Division on Earth and Life Studies; National Research Council. 2009. Mapping the Zone: Improving Flood Map Accuracy. Washington, D.C.: National Academies Press. http://www.nap.edu/catalog/12573.

Cook, A., Merwade, V., 2009. "Effect of Topographic Data, Geometric Configuration and Modeling Approach on Flood Inundation Mapping." Journal of Hydrology 377 (1): 131–142.

Crippen, J.R., Bue, C.D., 1977. Maximum floodflows in the conterminous United States.

De Coursey, D.G., 1973. Objective Regionalization of Peak Flow Rates. Floods Droughts Proc. Second Int. Symp. Hydrol.

De Risi, R., Jalayer, F., De Paola, F., Giugni, M., 2014. Probabilistic delineation of flood-prone areas based on a digital elevation model and the extent of historical flooding: The case of Ouagadougou. Bol. Geológico Min. 125, 329–340.

Degiorgis, M., Gnecco, G., Gorni, S., Roth, G., Sanguineti, M., Taramasso, A.C., 2013. Flood hazard assessment via threshold binary classifiers: case study of the Tanaro River basin. Irrig. Drain. 62, 1–10. https://doi.org/10.1002/ird.1806

Degiorgis, M., Gnecco, G., Gorni, S., Roth, G., Sanguineti, M., Taramasso, A.C., 2012. Classifiers for the detection of flood-prone areas using remote sensed elevation data. J. Hydrol. 470–471, 302–315. https://doi.org/10.1016/j.jhydrol.2012.09.006

Devia, G.K., Ganasri, B.P., Dwarakish, G.S., 2015. A Review on Hydrological Models. Aquat. Procedia 4, 1001–1007

Di Baldassarre, G., Schumann, G., Bates, P.D., Freer, J.E., Beven, K.J., 2010. Flood-plain mapping: a critical discussion of deterministic and probabilistic approaches. Hydrol. Sci. Journal–Journal Sci. Hydrol. 55, 364–376.

Dietterich, Thomas G. 1998. “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms.” Neural Computation 10: 1895–1923.

Dodov, B., Foufoula-Georgiou, E., 2005. Fluvial processes and streamflow variability: Interplay in the scale-frequency continuum and implications for scaling. Water Resour. Res. 41, W05005. https://doi.org/10.1029/2004WR003408

Dodov, B. A., and E. Foufoula-Georgiou. 2006. “Floodplain Morphometry Extraction from a High-Resolution Digital Elevation Model: A Simple Algorithm for Regional Analysis Studies.” IEEE Geoscience and Remote Sensing Letters 3 (3): 410–13. doi:10.1109/LGRS.2006.874161.

Domeneghetti, A., Vorogushyn, S., Castellarin, A., Merz, B., Brath, A., 2013. Probabilistic flood hazard mapping: effects of uncertain boundary conditions. Hydrol. Earth Syst. Sci. 17, 3127.

Dottori, F., Di Baldassarre, G., Todini, E., 2013. Detailed data is welcome, but with a pinch of salt: Accuracy, precision, and uncertainty in flood inundation modeling. Water Resour. Res. 49, 6079–6085.

Eberhart, R.C., Shi, Y., 2001. Tracking and optimizing dynamic systems with particle swarms, in: Evolutionary Computation, 2001. Proceedings of the 2001 Congress On. IEEE, pp. 94–100.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett., ROC Analysis in Pattern Recognition 27, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Feng, Y., Teng, G.-F., Wang, A.-X., Yao, Y.-M., 2007. Chaotic inertia weight in particle swarm optimization, in: Innovative Computing, Information and Control, 2007. ICICIC’07. Second International Conference On. IEEE, pp. 475–475.

Fernandez, W., Vogel, R.M., Sankarasubramanian, A., 2000. Regional Calibration of a Watershed Model. Hydrol. Sci. J. 45, 689–707. https://doi.org/10.1080/02626660009492371

Follum, M.L., 2013. AutoRoute Rapid Flood Inundation Model. Engineer Research and Development Center Vicksburg MS Coastal and Hydraulics Lab.

Follum, M.L., Tavakoly, A.A., Niemann, J.D., Snow, A.D., 2017. AutoRAPID: a model for prompt streamflow estimation and flood inundation mapping over regional to continental extents. JAWRA J. Am. Water Resour. Assoc. 53, 280–299.

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39, 1347. https://doi.org/10.1029/2002WR001426

Ganora, D., Claps, P., Laio, F., Viglione, A., 2009. An approach to estimate nonparametric flow duration curves in ungauged basins. Water Resour. Res. 45.

Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., Tyler, D., 2002. The national elevation dataset. Photogramm. Eng. Remote Sens. 68, 5–32.

Greenlee, D.D., 1987. Raster and vector processing for scanned linework. Photogramm. Eng. Remote Sens. 53, 1383–1387.

Griffis, V.W., Stedinger, J.R., 2007. Log-Pearson Type 3 distribution and Its application in flood frequency analysis. I: distribution characteristics. J. Hydrol. Eng. 12, 482–491. https://doi.org/10.1061/(ASCE)1084-0699(2007)12:5(482)

Grimaldi, S., Petroselli, A., 2015. “Do We Still Need the Rational Formula? An Alternative Empirical Procedure for Peak Discharge Estimation in Small and Ungauged Basins.” Hydrological Sciences Journal 60 (1): 67–77.

Grimaldi, S., Petroselli, A., Alonso, G., Nardi, F., 2010. “Flow Time Estimation with Spatially Variable Hillslope Velocity in Ungauged Basins.” Advances in Water Resources 33 (10): 1216–1223.

Grimaldi, S., Petroselli, A., Arcangeletti, E., Nardi., 2013. “Flood Mapping in Ungauged Basins Using Fully Continuous Hydrologic–hydraulic Modeling.” Journal of Hydrology 487: 39–47.

Grimaldi, S., Petroselli, A., Nardi, F., 2012. “A Parsimonious Geomorphological Unit Hydrograph for Rainfall–runoff Modelling in Small Ungauged Basins.” Hydrological Sciences Journal 57 (1): 73–83.

Guha-Sapit, D., Hoyois, P., Below, R., 2015. Annual Disaster Statistical Review: The numbers and trends. Centre for Research on the Epidemiology of Disasters ((Belgium) Brussels)

Hazen, A., 1914. Discussion on ‘Flood flows’ by WE Fuller. Trans ASCE 77, 526–563.

Holmes, M.G.R., Young, A.R., Gustard, A., Grew, R., 2002. A Region of Influence Approach to Predicting Flow Duration Curves within Ungauged Catchments. Hydrol Earth Syst Sci 6, 721–731. https://doi.org/10.5194/hess-6-721-2002

Horritt, M. S., Bates, P., 2002. “Evaluation of 1D and 2D Numerical Models for Predicting River Flood Inundation.” Journal of Hydrology 268 (1): 87–99.

Hrachowitz, M., Savenije, H.H.G., Blöschl, G., McDonnell, J.J., Sivapalan, M., Pomeroy, J.W., Arheimer, B., Blume, T., Clark, M.P., Ehret, U., Fenicia, F., Freer, J.E., Gelfan, A., Gupta, H.V., Hughes, D.A., Hut, R.W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P.A., Uhlenbrook, S., Wagener, T., Winsemius, H.C., Woods, R.A., Zehe, E., Cudennec, C., 2013. A Decade of Predictions in Ungauged Basins (PUB)—a Review. Hydrol. Sci. J. 58, 1198–1255. https://doi.org/10.1080/02626667.2013.803183

Hundecha, Y., Bárdossy, A., 2004. Modeling of the Effect of Land Use Changes on the Runoff Generation of a River Basin through Parameter Regionalization of a Watershed Model. J. Hydrol. 292, 281–295. https://doi.org/10.1016/j.jhydrol.2004.01.002

Hunter, N.M., Bates, P.D., Horritt, M.S., Wilson, M.D., 2007. “Simple Spatially-Distributed Models for Predicting Flood Inundation: A Review.” Geomorphology, Reduced-Complexity Geomorphological Modelling for River and Catchment Management, 90 (3–4): 208–25. doi:10.1016/j.geomorph.2006.10.021.

Jafarzadegan, K., Merwade, V., 2019. Probabilistic floodplain mapping using HAND-based statistical approach. Geomorphology 324, 48–61. https://doi.org/10.1016/j.geomorph.2018.09.024

Jafarzadegan, K., Merwade, V., 2017. A DEM-based approach for large-scale floodplain mapping in ungauged watersheds. J. Hydrol. 550, 650–662. https://doi.org/10.1016/j.jhydrol.2017.04.053

Jafarzadegan, K., Merwade, V., Saksena, S., 2018. A Geomorphic Approach to 100-Year Floodplain Mapping for the Conterminous United States. J. Hydrol. 561, 43–58. https://doi.org/10.1016/j.jhydrol.2018.03.061

Jenson, S.K., Domingue, J.O., 1988. Extracting topographic structure from digital elevation data for geographic information system analysis. Photogramm. Eng. Remote Sens. 54, 1593–1600.

Kalyanapu, A.J., Burian, S.J., McPherson, T.N., 2010. Effect of land use-based surface roughness on hydrologic model output. J. Spat. Hydrol. 9.

Kansas State Research and Extension, 2011. Middle Neosho Watershed: Watershed Restoration and Protection Strategy. Retrieved from http://www.kswraps.org/files/attachments/middleneosho_plansummary.pdf

Kay, A.L., Jones, D.A., Crooks, S.M., Kjeldsen, T.R., Fung, C.F., 2007. An Investigation of Site-Similarity Approaches to Generalisation of a Rainfall? Runoff Model. Hydrol. Earth Syst. Sci. Discuss. 11, 500–515.

Kendall, M.G., 1949. Rank and product-moment correlation. Biometrika 177–193.

Kim, U., Kaluarachchi, J.J., 2008. Application of Parameter Estimation and Regionalization Methodologies to Ungauged Basins of the Upper Blue Nile River Basin, Ethiopia. J. Hydrol. 362, 39–56. https://doi.org/10.1016/j.jhydrol.2008.08.016

Kim, M.H., Morlock, S.E., Arihood, L.D., Kiesler, J.L., 2011. Observed and forecast flood-inundation mapping application-A pilot study of an eleven-mile reach of the White River, Indianapolis, Indiana. US Geological Survey.

Kohavi, Ron. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." In IJCAI, 14:1137–1145.

Kokkonen, T.S., Jakeman, A.J., Young, P.C., Koivusalo, H.J., 2003. Predicting Daily Flows in Ungauged Catchments: Model Regionalization from Catchment Descriptors at the Coweeta Hydrologic Laboratory, North Carolina. Hydrol. Process. 17, 2219–2238.

Laaha, G., Blöschl, G., 2006. A Comparison of Low Flow Regionalisation Methods—Catchment Grouping. J. Hydrol. 323, 193–214. https://doi.org/10.1016/j.jhydrol.2005.09.001

Lee, H., McIntyre, N.R., Wheater, H.S., Young, A.R., 2006. Predicting Runoff in Ungauged UK Catchments, in: Proceedings of the Institution of Civil Engineers-Water Management. THOMAS TELFORD PUBLISHING, pp. 129–138.

Lhomme, J., Sayers, P., Gouldby, B., Samuels, P., Wills, M., Mulet-Marti, J., 2008. Inundation modelling Recent development and application of a rapid flood spreading method, in: Flood Risk Management: Research and Practice. CRC Press, pp. 30–39.

Liu, Z., Merwade, V., Jafarzadegan, K.,2018. Investigating the role of model structure and surface roughness in generating flood inundation extents using one- and two-dimensional hydraulic models. J. Flood Risk Manag. 0, e12347. https://doi.org/10.1111/jfr3.12347

Lóczy, D., Pirkhoffer, E., Gyenizse, P., 2012. Geomorphometric floodplain classification in a hill region of Hungary. Geomorphology 147, 61–72.

Maidment David R., 2009. FEMA Flood Map Accuracy. World Environ. Water Resour. Congr., Proceedings. https://doi.org/10.1061/41036(342)492

Maidment, D.R., Rajib, A., Lin, P., Clark, E.P., 2016. National Water Center Innovators Program Summer Institute Report 2016. Res. Summ. 4.

Manfreda, S., Leo, M.D., Sole, A., 2011. Detection of flood-prone areas using digital elevation models. J. Hydrol. Eng. 16, 781–790. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000367

Manfreda, S., Nardi, F., Samela, C., Grimaldi, S., Taramasso, A.C., Roth, G., Sole, A., 2014. Investigation on the use of geomorphic approaches for the delineation of flood prone areas. J. Hydrol. 517, 863–876. https://doi.org/10.1016/j.jhydrol.2014.06.009

Manfreda, S., Samela, C., Gioia, A., Consoli, G.G., Iacobellis, V., Giuzio, L., Cantisani, A., Sole, A., 2015. Flood-prone areas assessment using linear binary classifiers based on flood maps obtained from 1D and 2D hydraulic models. Nat. Hazards 79, 735–754. https://doi.org/10.1007/s11069-015-1869-5

Manfreda, S., Sole, A., Fiorentino, M., 2008. Can the basin morphology alone provide an insight into floodplain delineation? WIT Press, pp. 47–56. https://doi.org/10.2495/FRIAR080051

Marini, F., Walczak, B., 2015. Particle swarm optimization (PSO). A tutorial. Chemom. Intell. Lab. Syst. 149, 153–165.

Martini, F., Loat, R., 2007. Handbook on good practices for flood mapping in Europe.

Masih, I., Uhlenbrook, S., Maskey, S., Ahmad, M.D., 2010. Regionalization of a Conceptual Rainfall–Runoff Model Based on Similarity of the Flow Duration Curve: A Case Study from the Semi-Arid Karkheh Basin, Iran. J. Hydrol. 391, 188–201. https://doi.org/10.1016/j.jhydrol.2010.07.018

McDonnell, J.J., Woods, R., 2004. On the Need for Catchment Classification. J. Hydrol. 299, 2–3.

McGlynn, B.L., Seibert, J., 2003. Distributed assessment of contributing area and riparian buffering along stream networks. Water Resour. Res. 39, 1082. https://doi.org/10.1029/2002WR001521

McGlynn, Brian L., and Jeffrey J. McDonnell. 2003. "Quantifying the Relative Contributions of Riparian and Hillslope Zones to Catchment Runoff." Water Resources Research 39 (11): 1310. doi:10.1029/2003WR002091.

McIntyre, N., Lee, H., Wheater, H., Young, A., Wagener, T., 2005. Ensemble Predictions of Runoff in Ungauged Catchments. Water Resour. Res. 41.

Merwade, V., Cook, A., Coonrod, J., 2008. GIS techniques for creating river terrain models for hydrodynamic modeling and flood inundation mapping. Environ. Model. Softw. 23, 1300–1311. https://doi.org/10.1016/j.envsoft.2008.03.005

Merwade, V., Olivera, F., Arabi, M., Edleman, S., 2008. Uncertainty in flood inundation mapping: current issues and future directions. J. Hydrol. Eng. 13, 608–620. https://doi.org/10.1061/(ASCE)1084-0699(2008)13:7(608)

Merwade, V., Du, L., Sangwan, N., 2015. "Creating a National Scale Floodplain Map for the United States Using Soil Information." presented at the 2015 Fall Meeting, AGU, San Francisco, Calif.

Merz, R., Blöschl, G., 2004. Regionalisation of Catchment Model Parameters. J. Hydrol. 287, 95–123.

Merz, B., Thieken, A.H., Gocht, M., 2007. Flood risk mapping at the local scale: concepts and challenges. Flood Risk Manag. Eur. 231–251.

Moel, H. de, Alphen, J. van, Aerts, J., 2009. Flood maps in Europe-methods, availability and use. Nat. Hazards Earth Syst. Sci. 9, 289–301. https://doi.org/10.5194/nhess-9-289-2009

Musser, J.W., Dyar, T.R., 2007. Two-dimensional flood inundation model of the Flint River at Albany. Ga. US Geol. Surv. Sci. Investig. Rep. 5107, 49.

Nair, Minu, and J. S. Bindhu. 2016. "Supervised Techniques and Approaches for Satellite Image Classification." International Journal of Computer Applications 134 (16). http://search.proquest.com/openview/93c2350270de3c57a58cc03c4606fb58/1?pq-origsite=gscholar&cbl=136216.

Nardi, F., Biscarini, C., Di Francesco, S., Manciola, P., Ubertini, L., 2013. Comparing a large-scale DEM-based floodplain delineation algorithm with standard flood maps: the TIBER river basin case study. Irrig. Drain. 62, 11–19.

Nardi, F., Vivoni, E.R., Grimaldi, S., 2006. Investigating a floodplain scaling relation using a hydrogeomorphic delineation method. Water Resour. Res. 42.

Nathan, R.J., McMahon, T.A., 1990. Identification of Homogeneous Regions for the Purposes of Regionalisation. J. Hydrol. 121, 217–238.

Neal, J., Schumann, G., Bates, P., 2012. "A Subgrid Channel Model for Simulating River Hydraulics and Floodplain Inundation over Large and Data Sparse Areas." Water Resources Research 48 (11): W11506. doi:10.1029/2012WR012514.

Neal, J., Keef, C., Bates, P., Beven, K., Leedal, D., 2013. Probabilistic flood risk mapping including spatial dependence. Hydrol. Process. 27, 1349–1363.

Nguyen, P., Thorstensen, A., Sorooshian, S., Hsu, K., AghaKouchak, A., 2015. Flood Forecasting and Inundation Mapping Using HiResFlood-UCI and Near-Real-Time Satellite Precipitation Data: The 2008 Iowa Flood. J. Hydrometeorol. 16, 1171–1183. https://doi.org/10.1175/JHM-D-14-0212.1

Nguyen, P., Thorstensen, A., Sorooshian, S., Hsu, K., AghaKouchak, A., Sanders, B., Koren, V., Cui, Z., Smith, M., 2016. A high resolution coupled hydrologic–hydraulic model (HiResFlood-UCI) for flash flood modeling. J. Hydrol., Flash floods, hydro-geomorphic response and risk management 541, 401–420. https://doi.org/10.1016/j.jhydrol.2015.10.047

Nobre, A.D., Cuartas, L.A., Hodnett, M., Rennó, C.D., Rodrigues, G., Silveira, A., Waterloo, M., Saleska, S., 2011. Height Above the Nearest Drainage – a hydrologically relevant new terrain model. J. Hydrol. 404, 13–29. https://doi.org/10.1016/j.jhydrol.2011.03.051

Noman, N.S., Nelson, E.J., Zundel, A.K., 2001. Review of automated floodplain delineation from digital terrain models. J. Water Resour. Plan. Manag. 127, 394–402. https://doi.org/10.1061/(ASCE)0733-9496(2001)127:6(394)

Ouarda, T.B., Girard, C., Cavadias, G.S., Bobée, B., 2001. Regional Flood Frequency Estimation with Canonical Correlation Analysis. J. Hydrol. 254, 157–173.

Oudin, L., Kay, A., Andréassian, V., Perrin, C., 2010. Are Seemingly Physically Similar Catchments Truly Hydrologically Similar? Water Resour. Res. 46.

Parajka, J., Merz, R., Blöschl, G., 2005. A Comparison of Regionalisation Methods for Catchment Model Parameters. Hydrol. Earth Syst. Sci. Discuss. 9, 157–171.

Pal, M., 2005. Random forest classifier for remote sensing classification. Int. J. Remote Sens. 26, 217–222.

Papaioannou, G., Vasiliades, L., Loukas, A., 2014. Multi-criteria analysis framework for potential flood prone areas mapping. Water Resour. Manag. 29, 399–418. https://doi.org/10.1007/s11269-014-0817-6

Patro, S., Chatterjee, C., Mohanty, S., Singh, R., Raghuwanshi, N.S., 2009. Flood inundation modeling using MIKE FLOOD and remote sensing data. J. Indian Soc. Remote Sens. 37, 107–118. https://doi.org/10.1007/s12524-009-0002-1

Patton, P.C., Baker, V.R., 1976. Morphometry and floods in small drainage basins subject to diverse hydrogeomorphic controls. Water Resour. Res. 12, 941–952.

Pearson, K., 1904. Mathematical contributions to the theory of evolution.—XII. On a generalised Theory of alternative Inheritance, with special reference to Mendel's laws. Philos. Trans. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character 203, 53–86. https://doi.org/10.1098/rsta.1904.0015

Pedrozo-Acuña, A., Rodríguez-Rincón, J.P., Arganis-Juárez, M., Domínguez-Mora, R., González Villareal, F.J., 2015. Estimation of probabilistic flood inundation maps for an extreme event: Pánuco River, México. J. Flood Risk Manag. 8, 177–192.

Petroselli, A., and S. Grimaldi. 2015. “Design Hydrograph Estimation in Small and Fully Ungauged Basins: A Preliminary Assessment of the EBA4SUB Framework.” Journal of Flood Risk Management. http://onlinelibrary.wiley.com/doi/10.1111/jfr3.12193/pdf.

Policy for Use of Hydrologic Engineering Center-River Analysis System in the National Flood Insurance Program, 2015. URL https://www.fema.gov/policy-use-hydrologic-engineering-center-river-analysis-system-national-flood-insurance-program (accessed 4.27.18).

Purvis, M.J., Bates, P.D., Hayes, C.M., 2008. A probabilistic methodology to estimate future coastal flood risk due to sea level rise. Coast. Eng. 55, 1062–1073.

Ramachandra Rao, A., Srinivas, V.V., 2006. Regionalization of Watersheds by Hybrid-Cluster Analysis. J. Hydrol. 318, 37–56. https://doi.org/10.1016/j.jhydrol.2005.06.003

Romanowicz, R., Beven, K., 2003. Estimation of flood inundation probabilities as conditioned on event inundation maps. Water Resour. Res. 39.

Rao, A., 2004. Regionalization of Indiana watersheds for flood flow predictions phase I: Studies in regionalization of Indiana watersheds. Jt. Transp. Res. Program 180.

Rao, A., Srinivas, V.V., 2006a. Regionalization of watersheds by fuzzy cluster analysis. J. Hydrol. 318, 57–79.

Rao, A., Srinivas, V.V., 2006b. Regionalization of watersheds by hybrid-cluster analysis. J. Hydrol. 318, 37–56. https://doi.org/10.1016/j.jhydrol.2005.06.003

Razavi, Tara, Coulibaly, Paulin, 2013. Streamflow prediction in ungauged basins: review of regionalization methods. J. Hydrol. Eng. 18, 958–975. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690

Reed, D.W., Jakob, D., Robson, A.J., Faulkner, D.S., Stewart, E.J., 1999. Regional Frequency Analysis: A New Vocabulary. IAHS-AISH Publ. 237–243.

Rennó, C.D., Nobre, A.D., Cuartas, L.A., Soares, J.V., Hodnett, M.G., Tomasella, J., Waterloo, M.J., 2008. HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia. Remote Sens. Environ. 112, 3469–3481. https://doi.org/10.1016/j.rse.2008.03.018

Rexer, M., Hirt, C., 2014. Comparison of free high resolution digital elevation data sets (ASTER GDEM2, SRTM v2.1/v4.1) and validation against accurate heights from the Australian National Gravity Database. Aust. J. Earth Sci. 61, 213–226. https://doi.org/10.1080/08120099.2014.884983

Ridolfi, E., Rianna, M., Trani, G., Alfonso, L., Di Baldassarre, G., Napolitano, F., Russo, F., 2016. A new methodology to define homogeneous regions through an entropy based clustering method. Adv. Water Resour. 96, 237–250. https://doi.org/10.1016/j.advwatres.2016.07.007

Ries, K.G., 2007. The national streamflow statistics program: A computer program for estimating streamflow statistics for ungaged sites. DIANE Publishing.

Rigon, Riccardo, Marialaura Bancheri, Giuseppe Formetta, and Alban de Lavenne. 2016. “The Geomorphological Unit Hydrograph from a Historical-Critical Perspective.” Earth Surface Processes and Landforms 41 (1): 27–37.

Rodriguez-Iturbe, I. 1993. “The Geomorphological Unit Hydrograph.” Channel Network Hydrology, 43–68.

Rodríguez-Iturbe, Ignacio, Gustavo Devoto, and Juan B. Valdés. 1979. “Discharge Response Analysis and Hydrologic Similarity: The Interrelation between the Geomorphologic IUH and the Storm Characteristics.” Water Resources Research 15 (6): 1435–1444.

Saksena, S., Merwade, V., 2017. Integrated modeling of surface-subsurface processes to understand river-floodplain hydrodynamics in the upper Wabash river basin, in: World Environmental and Water Resources Congress. ASCE, Sacramento, CA, pp. 60–68.

Saksena, S., Merwade, V., 2015. Incorporating the effect of DEM resolution and accuracy for improved flood inundation mapping. J. Hydrol. 530, 180–194.

Samela, C., Manfreda, S., Paola, F.D., Giugni, M., Sole, A., Fiorentino, M., 2016. DEM-based approaches for the delineation of flood-prone areas in an ungauged basin in Africa. J. Hydrol. Eng. 21, 06015010. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001272

Samela, C., Troy, T.J., Manfreda, S., 2017. Geomorphic classifiers for flood-prone areas delineation for data-scarce environments. Adv. Water Resour. 102, 13–28.

Sampson, C.C., Smith, A.M., Bates, P.D., Neal, J.C., Alfieri, L., Freer, J.E., 2015. A high-resolution global flood hazard model. Water Resour. Res. 51, 7358–7381.

Sanders, B.F., 2007. Evaluation of on-line DEMs for flood inundation modeling. Adv. Water Resour. 30, 1831–1843. https://doi.org/10.1016/j.advwatres.2007.02.005

Sangwan, N., Merwade, V., 2015. A faster and economical approach to floodplain mapping using soil information. JAWRA J. Am. Water Resour. Assoc. 51, 1286–1304. https://doi.org/10.1111/1752-1688.12306

Sankarasubramanian, A., Vogel, R.M., 2002. Comment on the paper:“Basin hydrologic response relations to distributed physiographic descriptors and climate” by Karen Plaut Berger, Dara Entekhabi, 2001. Journal of Hydrology 247, 169–182. J. Hydrol. 263, 257–261.

Sarhadi, A., Soltani, S., Modarres, R., 2012. Probabilistic flood inundation mapping of ungauged rivers: Linking GIS techniques and frequency analysis. J. Hydrol. 458, 68–86.

Sauer, V.B., Thomas Jr, W.O., Stricker, V.A., Wilson, K.V., 1983. Flood characteristics of urban watersheds in the United States. USGPO,.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P.A., Carrillo, G., 2011. Catchment Classification: Empirical Analysis of Hydrologic Similarity Based on Catchment Function in the Eastern USA. Hydrol Earth Syst Sci 15, 2895–2911. https://doi.org/10.5194/hess-15-2895-2011

Sefton, C.E.M., Howarth, S.M., 1998. Relationships between dynamic response characteristics and physical descriptors of catchments in England and Wales. J. Hydrol. 211, 1–16.

Seibert, J., 1999. Regionalisation of Parameters for a Conceptual Rainfall-Runoff Model. Agric. For. Meteorol. 98, 279–293.

Shu, C., Burn, D.H., 2003. Spatial Patterns of Homogeneous Pooling Groups for Flood Frequency Analysis. Hydrol. Sci. J. 48, 601–618.

Silvert, W., 2001. Modelling as a Discipline. Int. J. Gen. Syst. 30, 261–282.

Singh, P. K., S. K. Mishra, and M. K. Jain. 2014. “A Review of the Synthetic Unit Hydrograph: From the Empirical UH to Advanced Geomorphological Methods.” Hydrological Sciences Journal 59 (2): 239–261.

Sivapalan, M., 2003. Prediction in Ungauged Basins: A Grand Challenge for Theoretical Hydrology. Hydrol. Process. 17, 3163–3170. https://doi.org/10.1002/hyp.5155

Strahler, A.N., 1957. Quantitative analysis of watershed geomorphology. Trans. Am. Geophys. Union 38, 913–920. https://doi.org/10.1029/TR038i006p00913

Tasker, G.D., 1982. Comparing Methods of Hydrologic Regionalization. JAWRA J. Am. Water Resour. Assoc. 18, 965–970.

Tate, E.C., Maidment, D.R., Olivera, F., Anderson, D.J., 2002. Creating a terrain model for floodplain mapping. J. Hydrol. Eng. 7, 100–108. https://doi.org/10.1061/(ASCE)1084-0699(2002)7:2(100)

Tayefi, V., Lane, S.N., Hardy, R.J., Yu, D., 2007. A comparison of one- and two-dimensional approaches to modelling flood inundation over complex upland floodplains. Hydrol. Process. 21, 3190–3202. https://doi.org/10.1002/hyp.6523

Teng, J., Vaze, J., Dutta, D., Marvanek, S., 2015. Rapid Inundation Modelling in Large Floodplains Using LiDAR DEM. Water Resour. Manag. 29, 2619–2636. https://doi.org/10.1007/s11269-015-0960-8

Teng, J., Jakeman, A.J., Vaze, J., Croke, B.F., Dutta, D., Kim, S., 2017. Flood inundation modelling: A review of methods, recent advances and uncertainty analysis. Environ. Model. Softw. 90, 201–216.

Thomas, D.M., Benson, M.A., 1970. Generalization of streamflow characteristics from drainage-basin characteristics. US Government Printing Office Washington, DC.

Thoms, M.C., 2003. Floodplain–river ecosystems: lateral connections and the implications of human interference. Geomorphology, Floodplains: environment and process 56, 335–349. https://doi.org/10.1016/S0169-555X(03)00160-0

Thoms, M.C., Sheldon, F., 2000. Lowland rivers: an Australian introduction. Regul. Rivers Res. Manag. 16, 375–383.

Townsend, P.A., Walsh, S.J., 1998. Modeling floodplain inundation using an integrated GIS with radar and optical remote sensing. Geomorphology, Application of remote sensing and GIS in geomorphology 21, 295–312. https://doi.org/10.1016/S0169-555X(97)00069-X

Tung, Y.-K., Yeh, K.-C., Yang, J.-C., 1997. Regionalization of Unit Hydrograph Parameters: 1. Comparison of Regression Analysis Techniques. Stoch. Hydrol. Hydraul. 11, 145–171.

Turnipseed, D.P., Ries III, K.G., 2007. The national streamflow statistics program: Estimating high and low streamflow statistics for ungaged sites. Geological Survey (US).

U.S. Geological Survey. 2012. "The StreamStats Program." http://streamstats.usgs.gov.

U.S. Geological Survey, 2014 - National Hydrography Dataset. URL http://nhd.usgs.gov/wbd.html

Van Alphen, J., Passchier, R., 2007. Atlas of Flood Maps, examples from 19 European countries, USA and Japan, Ministry of Transport. Public Works Water Manag. Hague Neth. Available Httpec Eur. Euenvironmentwaterflood Riskfloodatlasindex Htm Last Access 12 March 2013.

Vandewiele, G.L., Elias, A., 1995. Monthly Water Balance of Ungauged Catchments Obtained by Geographical Regionalization. J. Hydrol. 170, 277–291. https://doi.org/10.1016/0022-1694(95)02681-E

Verbunt, M., Walser, A., Gurtz, J., Montani, A., Schär, C., 2007. Probabilistic flood forecasting with a limited-area ensemble prediction system: selected case studies. J. Hydrometeorol. 8, 897–909.

Viviroli, D., Mittelbach, H., Gurtz, J., Weingartner, R., 2009. Continuous Simulation for Flood Estimation in Ungauged Mesoscale Catchments of Switzerland – Part II: Parameter Regionalisation and Flood Estimation Results. J. Hydrol. 377, 208–225. https://doi.org/10.1016/j.jhydrol.2009.08.022

Wagener, T., Sivapalan, M., Troch, P., Woods, R., 2007. Catchment Classification and Hydrologic Similarity. Geogr. Compass 1, 901–931.

Walker, K.F., Puckridge, J.T., Blanch, S.J., 1997. Irrigation development on Cooper Creek, central Australia—prospects for a regulated economy in a boom-and-bust ecology. Aquat. Conserv. Mar. Freshw. Ecosyst. 7, 63–73. https://doi.org/10.1002/(SICI)1099-0755(199703)7:1<63::AID-AQC218>3.0.CO;2-5

Walling, D.E., He, Q., 1998. The spatial variability of overbank sedimentation on river floodplains. Geomorphology 24, 209–223. https://doi.org/10.1016/S0169-555X(98)00017-8

Watt, W.E., 2000. Twenty years of flood risk mapping under the Canadian national flood damage reduction program, in: Flood Issues in Contemporary Water Management. Springer, pp. 155–165.

Williams, W.A., Jensen, M.E., Winne, J.C., Redmond, R.L., 2000. An Automated Technique for Delineating and Characterizing Valley-Bottom Settings. Environ. Monit. Assess. 64, 105–114. http://dx.doi.org/10.1023/A:1006471427421

Wiltshire, S.E., 1986. Regional Flood Frequency Analysis II: Multivariate Classification of Drainage Basins in Britain. Hydrol. Sci. J. 31, 335–346.

Wing, O.E.J., Bates, P.D., Sampson, C.C., Smith, A.M., Johnson, K.A., Erickson, T.A., 2017. Validation of a 30 m resolution flood hazard model of the conterminous United States. Water Resour. Res. 53. https://doi.org/10.1002/2017WR020917

Wolman, M.G., 1971. Evaluating alternative techniques of floodplain mapping. Water Resour. Res. 7, 1383–1392.

Wright N. G., Villanueva I., Bates P. D., Mason D. C., Wilson M. D., Pender G., Neelz S., 2008. Case Study of the Use of Remotely Sensed Data for Modeling Flood Inundation on the River Severn, U.K. J. Hydraul. Eng. 134, 533–540. https://doi.org/10.1061/(ASCE)0733-9429(2008)134:5(533)

Xin, J., Chen, G., Hai, Y., 2009. A particle swarm optimizer with multi-stage linearly-decreasing inertia weight, in: Computational Sciences and Optimization, 2009. CSO 2009. International Joint Conference On. IEEE, pp. 505–508.

Yamazaki, D., Baugh, C.A., Bates, P.D., Kanae, S., Alsdorf, D.E., Oki, T., 2012. Adjustment of a spaceborne DEM for use in floodplain hydrodynamic modeling. J. Hydrol. 436–437, 81–91. https://doi.org/10.1016/j.jhydrol.2012.02.045

Yan, K., Baldassarre, G.D., Solomatine, D.P., 2013. Exploring the potential of SRTM topographic data for flood inundation modelling under uncertainty. J. Hydroinformatics Lond. 15, 849–861. http://dx.doi.org/10.2166/hydro.2013.137

Young, A.R., 2006. Stream Flow Simulation within UK Ungauged Catchments Using a Daily Rainfall-Runoff Model. J. Hydrol. 320, 155–172.

Zrinji, Z., Burn, D.H., 1996. Regional Flood Frequency with Hierarchical Region of Influence. J. Water Resour. Plan. Manag. 122, 245–252.

VITA

Keighobad Jafarzadegan was born in Isfahan, Iran. He graduated with a B.S. in Civil Engineering (major in Structural Engineering) from Isfahan University of Technology in 2009. He received his M.S. degree in Civil Engineering (major in Water Resources) from University of Tehran in 2012. He joined the graduate program in Civil Engineering at Purdue University in Fall 2014 and received Doctor of Philosophy degree in May 2019.