A BAYESIAN BASED GRAPHICAL MODEL FRAMEWORK FOR ESTIMATION AND FORECAST OF STREAM FLOW

By

CAROLYN R. KREKELER

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2012

To Dr. K. Clint Slatton

ACKNOWLEDGMENTS

I thank the late Dr. K. Clint Slatton for his years of guidance as my advisor, and his constant encouragement and confidence in my ability to finish. I also thank my current advisor, Dr. Jose Principe, for volunteering to take me on as his student and all that involves. Finally, I thank my parents, family, friends, and lab mates for their continual patience with me throughout my graduate studies.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

A BAYESIAN BASED GRAPHICAL MODEL FRAMEWORK FOR ESTIMATION AND FORECAST OF STREAM FLOW

By

Carolyn R. Krekeler

May 2012

Chair: Jose Principe
Major: Electrical and Computer Engineering

For watershed monitoring and management, it is important to be able to predict with measurable accuracy the flow rates of major streams as a function of time. Traditionally this has been done through either time series analysis, based largely on regression and second order statistics, or through complex physical models that incorporate many contributing hydrologic and geologic features surrounding the area of interest. Time series models exploit time structure at a location, are often limited to second order statistics, and require data to be stationary in nature. Physical models can produce accurate forecasts, but rely on a large amount of input features (i.e. ground water, rainfall, topography, soil moisture content) and necessitate the specification of boundary conditions. This causes them to be very site specific and to have long calibration times.

In this work, I explore the possibility of using graphical models (in the form of a Bayesian network) as an alternate methodology that exploits full statistical information of the stream flow data over a geographic area. Graphical models have the ability to both forecast flow temporally and to predict within stream flow at unknown locations, even though conventionally they do not have an explicit representation for time

dependencies. Being probabilistic methods, estimates are based on the probability distribution functions formed from training data, and so only trends that are present in the training data can accurately be modeled. Specifically, I investigate the advantages and disadvantages of a variety of models in three domains: temporal (forecasting), spatial (within stream prediction), and spatial-temporal. The results show that the conventional graphical model framework requires modifications to fully describe the temporal dependencies found in natural stream flow over large geographic areas.

Additionally, a method of feature ranking and measuring information content is developed for determining which governing features are most useful in estimating stream flow. Features are ranked according to both the conditional entropy of feature combinations as well as the quadratic mutual information using the Cauchy-Schwartz divergence definition. It is seen that each information metric is useful under certain circumstances.

Finally, one of the benefits of using a probabilistic framework is the ability to have statistics of the error associated with each estimation. The predicted error is evaluated and compared to the actual error at known locations to ascertain model reliability. Additionally, a 95% confidence interval associated with each location along the river gives useful information as where additional sensors may be needed.

The major contributions of this work are: (1) Investigating the advantages and disadvantages of a variety of models used to predict flow temporally, spatially, and spatial-temporally, including hidden Markov models (HMMs), Gaussian processes (GP), and time series models; (2) Addressing the problem of having data that is sparse both spatially and temporally by utilizing observed data along with previously estimated data

within a Bayesian network, using minimal data to achieve optimal estimates of flow; and (3) Analyzing confidence of estimates based on the probability of estimated values versus the spread of probability associated with other flow values.

CHAPTER 1
INTRODUCTION

1.1 Motivation

For watershed monitoring and management, it is important to be able to predict with measurable accuracy the flow rates of major streams as a function of time. Flow estimation is critical for planning for and mitigating floods and droughts, as well as assessing the transport of sediments and pollutants through the watershed. To date, most of the work on stream flow estimation has fallen into one of two paradigms. One approach is to use deterministic models of hydrologic processes that attempt to parameterize the functional relationships between quantities, such as rainfall, surface topography, soil moisture, and stream discharge. This approach, however, is inherently limited because it is typically impossible to specify all the necessary boundary conditions at all times and locations with reasonable accuracy. Thus, such approaches require often unreasonable assumptions of uniformity on various hydrologic parameters over the watershed. The other common approach to the problem is to use classical methods of statistical time series analysis, such as regression of measured flow with rainfall and terrain features. Such methods are not adequate in general because the estimated regression parameters for one site or time period typically do not hold for other sites or time periods. This approach does not provide a principled way to predict how the regression parameters will vary in time or space, thus making the generalization of results problematic.

When handled probabilistically, complex hydrologic problems often involve the estimation of the likelihood of an event based on numerous features that influence the event process. This in turn requires fusing multiple disparate data sets from the

temporal and spatial domains. The fusion can be carried out using physical models, where multiple input parameters produce an output value, or by using data-driven approaches (statistical analysis, probabilistic learning). Some data-driven methods mitigate the increase in computational complexity as the number of features increases by using a reduced feature data set, thus tolerating greater error, while others accept the computational burden of a large feature set by sacrificing speed.

Less traditional approaches to the flow prediction problem include the use of graphical models. Bayesian networks (BNs) can be used to fuse both spatial and temporal features of the watershed. The causal nature of stream flow easily lends itself to being modeled by a tree structured BN with relationships between upstream and downstream locations. Given spatial-temporal rainfall and groundwater data, and temporal data from sparsely distributed *in situ* measurements of stream flow, stream flow can be predicted with measurable uncertainty due to the BN's probabilistic framework. In addition, the methodology is readily generalizable to other watersheds because the primary prediction parameters (likelihoods in this case) are learned from the data itself. However, there are open questions to the applicability of Bayesian networks to this problem. The first is that the spatial data probability distribution function (pdf) does not have an explicit representation for time dependencies, relying on the probability of temporal transitions over the training period. This makes it difficult to forecast extreme changes in flow, so it is important to use appropriate features to help indicate future variations in flow. The second is that Bayesian networks require large amounts of data for proper training which normally do not exist in practice. The third is

that if the statistics change over the river basin, then a proper data segmentation is necessary to reliably quantify the underlying flow statistics.

An alternate graphical model that is more suited to capturing temporal dependencies is the hidden Markov model. It is based on the Markov assumption that the current state, given the single previous state, is independent of all other previous states. Commonly used in speech and handwriting recognition applications, this model is potentially well suited for stream flow forecasting. Due to the Markov assumption, however, there is a strong stationarity requirement, which might make it less applicable.

Stream flow is highly correlated within stream, so data from surrounding locations can be helpful in predicting flow locally. Strictly forecast models can be implemented independently at each location within the stream that one wishes to forecast at, but this does not take advantage of the stream's natural correlation. Including inputs at other locations is possible, but increases the computational complexity of the model and does not necessarily incorporate the full statistics of the relationship. A Bayesian network connected both spatially along the stream and temporally to forecast can more fully exploit these relationships through the use of message passing to perform inference at the unknown locations in the network.

I study the estimation of stream flow in the Santa Fe Watershed, located in north-central Florida, USA. The test site exhibits diverse hydrologic conditions due to the existence of the Hawthorne Clay confining layer in the eastern two thirds of the watershed. In the unconfined portion, where the confining layer is not present, groundwater flow is a significant contributor to surface flow. In the confined portion, the influence of groundwater on stream flow is less significant due to the confining layer and

the major driving force of flow is rainfall and surface runoff. Due to these differences in the watershed, some features will contribute more to stream discharge in certain areas than in others. In this work I investigate four methods of estimation for either temporal forecasting, within stream spatial prediction, or a combination of spatial-temporal estimation within the watershed.

1.2 Previous Work

Traditional approaches to stream flow estimation include time series analysis and physical models. The most common time series methods include autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models, popularized by Box and Jenkins [1]. These methods have been accepted as standard in hydrologic applications, but they fail to model the often non-linear and non-stationary attributes of stream flow prediction [2]. Additionally, regression methods rely on second order statistics, so do not work well when the process being modeled cannot be well characterized in terms of its mean and variance.

Physical models provide an alternative estimation/forecast approach in hydrologic and geologic applications. These models are able to incorporate numerous feature data from different domains (spatial, temporal) into a unified framework and parameterize the functional relationship between the data [3]. Boundary conditions must be specified and are often chosen arbitrarily to ensure adequate performance of the system. Additionally, including the large amount of features that can be used in these models increases the runtime, which can be especially difficult for a large study site.

Current forecasting agencies, such as the National Weather Service (NWS) branch of the National Oceanic and Atmospheric Administration (NOAA) and the United States Department of Agriculture (USDA) National Water and Climate Center (NWCC),

use Ensemble Streamflow Prediction (ESP) to produce stream flow forecasts [4]. This involves generating multiple future stream flow scenarios and using them together to obtain forecasts and uncertainty. Running multiple models automatically calls for higher computational power and time than relying on a single model.

Probabilistic methods use the full probability distribution function of the random variables, extracting information beyond second order statistics, and therefore reducing the probability of error. Estimating a high dimensional pdf is difficult due to the large data requirements that it entails and the computational requirements involved (curse of dimensionality). Bayesian networks can alleviate this problem by modeling the interaction between sets of fewer variables, artificially limiting the dimension of the pdf [5]. Since the inference is dependent solely on the data available, large amounts of training data are still needed for acceptable performance.

Bayesian networks have grown in popularity in recent times and have been applied to a wide range of problems. NASA's Ames Research Center uses BNs to automatically interpolate raw data from interplanetary probes without imposing a preconceived structure on the data [6]. BNs are often used for pattern classification, outperforming other methods such as *k*-NN [7]. [8] has used Bayesian Markov random fields to capture non-Gaussian statistics in hyperspectral images and use them to classify terrain. Additionally, they have been successfully applied to molecular biology (represent ensembles of proteins), computer graphics (texture analysis, synthesis, and classification), and computer vision (model framework to infer information about 3D space from 2D images) [9].

1.3 Contributions and Outline

In any of the traditional forecasting methods discussed in Section 1.2, dense data is needed to achieve reasonable results. In this document, I implement a variety of methods in the application of estimating and forecasting stream flow. The study site is a watershed with complex geologic/hydrologic interactions with sparse data available. A physical model has previously been calibrated to this site, and its predictions will be incorporated into the models as training data where true data is not available.

The major contributions of this work are: (1) Investigating the advantages and disadvantages of a variety of models used to predict flow temporally, spatially, and spatial-temporally, including hidden Markov models (HMMs), Gaussian processes (GP), and time series models; (2) Addressing the problem of having data that is sparse both spatially and temporally by utilizing observed data along with previously estimated data within a Bayesian network, using minimal data to achieve optimal estimates of flow; and (3) Analyzing confidence of estimates based on the probability of estimated values versus the spread of probability associated with other flow values.

Chapter 1 contains a basic overview of the problem presented in this work, as well as a brief review of existing approaches. Chapter 2 explains some of the issues involved in modeling a process that varies in both space and time, working with non-stationary processes such as stream flow, and also describes the data and test site. Chapter 3 describes the methods used and gives details of a feature selection method useful in building a limited dimensionality model. Chapter 4 contains an analysis of the results, while Chapter 5 states the conclusions and describes potential future work.

CHAPTER 2
PROBLEM DESCRIPTION AND THEORETICAL MODELING

2.1 Test Site

Stream flow is highly correlated when moving from one location along the river to another location downstream. Variations in geologic and hydrologic conditions can have a significant impact on this correlation (i.e. rainfall, soil drainage, groundwater levels). The Santa Fe River Basin in north central Florida, USA is a prime example of how the geologic makeup in the surrounding area affects river flow. The eastern two thirds of the basin have a confining layer (the Hawthorne formation) below the surface soil. This confining layer is a base for the surficial aquifer system and hinders the penetration of rainwater into the groundwater system. Therefore, the Santa Fe River and New River in this region are mainly determined by upstream flow and rainfall [10]. The western third of the basin is missing this confining layer (unconfined area), causing the Santa Fe River in this region to be defined more significantly by groundwater levels and springs throughout the basin to become tributaries to the main river flow. The United States Geographic Survey (USGS) has several monitoring stations throughout this basin. Figure 2-1 shows the location of USGS stations in the Santa Fe Basin, along with the geologic makeup of the region.

2.2 Data Description

Daily stream flow data is available from USGS at several locations throughout the Santa Fe Basin. Data from five of the sites along the New River and Santa Fe River are being used in this study; two in the confined region (2321000, 2321500) and three in the unconfined region (2321975, 2322500, 2322800). Data is available for varying periods of time at these locations, but all have records on the order of decades. Data from

October 2000 through July 2008, measured in cubic feet per second, is used in the study presented in this document.

Training data is used for learning the parameters (weights, transitional probabilities) of the models implemented. A limited amount of observed data is available, so flow predictions from WAM, a physical model previously calibrated for prediction of flow in the Santa Fe Basin, are incorporated in some aspects of model training. Within stream flow in WAM is calculated based on kinematic wave equations [11], which is an accepted reliable practice in hydrologic routing [12]. The flow estimates of WAM are too persistent in time, however, so are not ideal for use in training the temporal models, but contain trusted information about the spatial relationship of stream flow [13]. The locations of the WAM estimates are where each reach, or portion, of the stream becomes an outlet of each subbasin along the main river. The region studied consists of one reach upstream from USGS station 2321000 on the New River through one reach downstream of USGS station 2322800 on the Santa Fe River, consisting of a total of 23 reaches. Of these reaches, 15 are in the confined area and 8 are in the unconfined area of the basin. These reach outlets are the 23 locations considered in the spatial models. For training temporal models, USGS flow data from October 2000 through December 2006 is used, while the models are tested on USGS flow data from January 2007 through July 2008. The spatial models are trained on WAM data from October 2000 through July 2008 and tested with USGS data over the same period. The spatial relationships in the spatial-temporal model are trained using WAM data and the temporal relationships are trained using USGS data where available, all from October 2000 through December 2006. Where USGS data is

not available, the temporal relationships are trained on WAM data. Testing is done from January 2007 through July 2008. This enables the use of separate data for the training and testing phases of the model while still having a reasonable amount of samples available for testing. WAM data from 10/1/2000 through 7/31/2008 calculated in cubic meters per second is converted to cubic feet per second for conformity to USGS flow.

Rainfall data is used to aid in forecasting flow in the models with a temporal component. Daily measurements of point source rain at High Springs, FL and Starke, FL are obtained from the National Climatic Data Center (NCDC) from October 2000 through July 2008 and are measured in millimeters. Raw rainfall data is very noisy as the amount of rain on any given day can vary greatly from rainfall on the surrounding days. A small amount of rainfall today will not dramatically affect the amount of flow several days in the future, but a large amount of rainfall will. In order to more accurately capture this behavior, rain was smoothed with a ten day moving average filter and then clustered into three classes via *k*-means. This enables the rainfall to indicate minor variability in future flow on low precipitation days and indicate a larger increase in flow after major storms. Many of the temporal methods in this study assume the data being modeled are Gaussian, so the natural logarithm of both the WAM flow and the USGS flow was used in an attempt to comply with this. The graphical models (HMM and Bayesian network) have discrete, rather than continuous, formulations. For these models, flow was clustered into sixteen classes at each spatial location separately via *k*-means. Sixteen classes give adequate variation of flow levels while maintaining a satisfactory number of samples in each class [14].

The temporal models presented perform a 3 day forecast of flow at the five USGS locations along the river. A forecast of 3 days was chosen based on the lag time between each point in the basin and the river. This lag was calculated based on the terrain slope and the Euclidean distance between each USGS station and points throughout the basin. For stations in the confined region, the average lag length was 3 days, while the average lag length in the unconfined region was 6 days. Flow in the confined region is more driven by rainfall than the unconfined region, so a 3 day forecast should be far out enough to capture the effect of rain on the river flow. The spatial models predict within stream flow at each of the 23 reaches on the current day. The spatial-temporal model predicts within stream flow on the current day and calculates a 3 day forecast at each of the 23 reaches. In the forecasts, rainfall at Starke is used as a contributing feature to the locations within the confined region and rainfall at High Springs is used for the locations in the unconfined region. Figure 2-1 shows the locations of the USGS observation stations and rainfall sources and Figure 2-2 shows the 23 reach outlet locations in the basin.

2.3 Modeling Spatial-Temporal Processes

From the description of the Santa Fe River Basin, one can realize that the information to model flow resides both in the geographical extent of the basin and over time. Assuming a stochastic model for the dependences, river flow prediction is a good example of a spatial-temporal process. When one seeks a stochastic description for the measurements, then the joint distribution over space and time contains all the necessary information to capture the statistical relationships that produce the flow. Due to the dimension of the problem (23 nodes over 2861 days), some assumptions will need to be made to practically predict flow from a finite number of measurements. Here

we will treat the general case first and then the special cases that are normally applied to predict flow.

2.3.1 General Case

Given a set of measurement stations $X_i(n)$ $\{i = 1,\ldots,M, n = 0,\ldots, N\}$ the joint distribution of the multidimensional stochastic process is $p(X_1(0), X_2(0),..., X_i(n),\ldots, X_M(N))$. This probability density function contains all the information to quantify the statistical dependencies between the M x N random variables $X_i(n)$, but in order to estimate this pdf a huge number of measurements are needed, which is impractical. Therefore one seeks approximations based on more or less reasonable assumptions supported by knowledge of the process.

2.3.2 Special case I - Temporal Models with Spatial Independence

Perhaps the most obvious assumption from a signal processing point of view is to assume Gaussianity and stationarity over time, i.e. the statistical properties of the random variables $\{X_i(n)\}$, $n = 0,\ldots, N$ do not change over time. This means that their mean value is the same across the time interval and their first order dependence over time, the autocorrelation function, is only dependent upon the lag between time instances. But this is still a multivariate random variable with dimension M. Again, conventionally this is not pursued in this form because of the need to estimate joint moments over space, so an independence assumption over space is imposed. This allows one to work basically with a combination of marginal distributions. These assumptions reduce the problem to the conventional MIMO (multiple input multiple output models found in time series analysis [15]).

2.3.3 Special case II – Spatially Coupled Models with IID Assumptions

Let us now look at the problem considering additional realism to quantify the spatial information. Indeed, for river basin flow modeling, instead of assuming spatial independence it is possible to create a realistic dependence graph by pruning the spatial graph so that only adjacent connections between the measurement stations along the river remain. This is the realm of graphical models which can be fully exploited here and will quantify the statistical properties of the joint pdf over space without any approximations. The difficulty however, is that the measurements at each station are not IID as in conventional graphical models since there are dependences over time provided by the physical structure of the flow. This dependence will concentrate the measurements in the bisector hyperplane of the joint space, distorting the estimation of the joint distribution and producing poor results.

An approximation is needed to decorrelate the random variables over time. One approach is to use dynamical system theory concepts and reconstruct the attractor of the time series at each measurement station with a delay embedding that minimizes the correlation amongst the samples. This means that the input to each node in the graph will be a vector of dimensionality K given by the correlation dimension of the attractor. The disadvantage of this is that it increases the dimensionality of the estimation problem.

Alternatively, one can avoid the dimensionality increase by decorrelating the input data, which is an extra approximation that is very sensitive to noise. One way of accomplishing this is to implement a predictor at the input of each spatial node and work with the residuals. We can expect that the residuals are much less correlated and the graphical model will be working with the spatial statistical structure of the residuals. The

problem with this approach is that any noise will be amplified by the preprocessing step, but there will be no increase in dimension.

In this dissertation these techniques will be implemented in their entirety or with approximations in hybrid form because of the finite data restrictions. The goal is to include as much spatial-temporal information as possible and find out the impact of the approximations in the quality of the estimations. We are now going to briefly review the available signal processing and machine learning techniques.

Often researchers are interested in knowing the behavior of processes that vary only in time (such as speech, stock prices, sales data) or processes that vary only in space (such as image segmentation, topography). Models which are well suited for handling these types of data can be modified to work with processes that are spatial and temporal, but often through making unsuitable assumptions and producing unsatisfactory results. In practice, spatial-temporal models can be broken up into their spatial components and their temporal components and modeled separately, but this ignores the effect that each component may have on the other [16]. For instance, stream flow is a highly correlated spatial and temporal process. Flow at an upstream location can impact knowledge of flow at a downstream location both today and in the future. To fully understand the stream flow process, it is in our best interest to keep the inherent spatial-temporal coupling intact. This is not always feasible, given constraints on computational complexity and available data, so models which are temporal, spatial, and spatial-temporal are analyzed in this work and their respective advantages and disadvantages weighed.

2.4 Advantages and Limitations

The four models used in this work are the Gamma model, Gaussian process, hidden Markov model, and Bayesian network. The Gamma model and the Gaussian process both take continuous input data and output continuous data. The HMM has continuous observed variables, but discrete hidden variables. In my application of the HMM, an estimate of the observed variables is desired, so the output is continuous data. The Bayesian network has been formulated as a discrete problem in which nodes are estimated to be one of *N* states. Each of these models has its own advantages and limitations corresponding to their structure and underlying assumptions.

A simple statistical method of estimating and predicting flow is the autoregressive moving average (ARMA) model. Methods such as the ARMA model have been accepted as standard in hydrologic applications, but they fail to model the often non-linear and non-stationary attributes of stream flow prediction [2]. The gamma model is an ARMA model with one free feedback parameter and trivial stability [17]. In this sense, it is easier to calibrate than a strictly ARMA model, but has many of the same characteristics. Since it is autoregressive, it is a linear model and assumes the process is stationary. It is, however, an adaptive filter, so the weights can be optimally set for given training data without extensive knowledge of the process being modeled. The gamma model can easily handle multiple inputs by training separate filters on each input process and combining the results at the output, although this only combines the marginal information available at each one of the probed locations.

In the traditional linear regression model, a predictive distribution of the output function is obtained by learning a set of weights on the input based on a prior distribution of the weights. In Gaussian process regression, a prior probability is defined

over the output function directly, rather than over the input weights. Therefore, the Gaussian process describes a distribution over functions directly, and gives a full predictive distribution, rather than just an average value estimate. This directly provides a measure of uncertainty of the estimate along with the prediction itself. Because Gaussian processes do produce an output distribution, it is assumed that the functions being modeled can be characterized jointly by a Gaussian distribution. The properties of the GP inference (stationarity, characteristic length scale) are dependent on the covariance function used, so both the type of covariance function as well as its hyperparameters must be chosen with care [18].

A hidden Markov model is a type of graphical model that is suited for dealing with processes with an inherent temporality, or processes in which the value at one time is influenced by the value at the previous time step [19]. These models have an added layer of complexity in modeling the process behavior than the previous methods. The HMM structure includes a dependence between the observation and a latent variable, or hidden state, from which the observation was produced, along with an explicit relationship between sequential states. The structure of the HMM can be seen as a single probabilistic dependency between the hidden state and the observation that is unfolded in time, due to the Markov assumption that the current state is independent of all others given the previous state. Naturally, this assumes the process being modeled is stationary in time. The probabilistic structure of precipitation has been modeled using HMMs by [20] for input into stream flow models, so despite the stationarity assumption, it is reasonable to extend this use of HMMs to modeling stream flow directly.

One of the benefits of Bayesian networks is also one of their limitations. Bayesian networks estimate the value of a random variable based on the conditional probability of the RV given its "parents," or other RVs that it is dependent on. BNs work with these conditional probabilities as transitional probabilities of moving from the state at the parent RV to the state at the child RV (explained further in the next section). These transitional probabilities can be learned from the training data, making them easy to calibrate and apply to different applications. If there is not enough training data available, however, the BN will perform poorly due to forming pdfs that are not representative of the true behavior of the random variables. The BN will not be able to handle situations with testing data that are not represented in the training data, specifically when dealing with discrete problems [6].

Another problem arises from the curse of dimensionality [21]. It is possible to manipulate the structure of a BN so that there is a maximum dimensionality at each node, but it may be necessary to enter into higher dimensions to appropriately model certain problems. With each added dimension, the number of training data samples required to form the joint pdf grows exponentially [19].

Finally, a full Bayesian formulation includes the prior probabilities of each random variable. Choosing the wrong priors can affect the ultimate inference of the network in that it may cause the posterior to make decisions that are not supported by the structure of the data. Conversely, using uninformative priors may put too much emphasis on the data when there is actually some structure to the data [19]. When a known structure is present in the data, however, including priors in the posterior calculation can improve inference to take into account all knowledge of the process being learned.

2.5 Message Passing

In the 1980s, various message passing algorithms were developed to perform inference in Bayesian networks, one of which was Pearl's message passing algorithm [22], [23]. This algorithm allowed known data to be input into the Bayesian network and used to infer values at unobserved nodes through passing Bayesian conditional probabilities throughout the network. Any amount of data input into the network helps improve estimates at all nodes, since information is passed from each node to all the other nodes. This is a very beneficial quality to have in an estimator when working with real data, which often times may be incomplete or sparse. It is better to have as much data as possible when performing inference, but message passing makes the maximum use of available data, even over short time records.

In order to explain Pearl's message passing algorithm, it is first necessary to define some terms used in describing graphical models. Each variable is represented as a circle (called a node) in the graph. Arrows point away from the node towards its "child" nodes. The original node is referred to as a "parent" node. Each parent node may have multiple child nodes and vice versa. The arrows (or arcs) between nodes represent conditional probability distributions between the represented variables. The joint probability of each child node is defined as the product of the conditional probabilities given each of its parents. The topology of a Bayesian network refers to the way in which nodes are connected throughout the graph. The diagram in Figure 2-3 shows the relationship between three variables (X, Y, and Z) with the probabilities as defined by the arcs of the graph. X is a parent of Y, and Y is a parent of Z. Conversely, Z is a child of Y, and Y is a child of X.

A “leaf” node is a node with no children (*Z*, in Figure 2-3), and a “root” node is a node with no parents (*X*, in Figure 2-3). When values of the random variables are known, these nodes are called “evidence” or “observed” nodes and are instantiated with their known value with probability 1. Nodes for which the variable values are not known are called “unobserved” nodes and are assigned probabilities of each possible data value according to their conditional distribution.

In Pearl’s message passing, each parent node sends “pi messages” to each of its children, and the child nodes send “lambda messages” up to their parent nodes. Consider a graph consisting of a parent node, *X*, and a child node, *Y*. The lambda messages sent from each child node (*Y*) up the tree to a parent node (*X*) contain the sum over the possible child node values of the product of the probability of the child given the parent node and the child’s lambda value (Equation 2-1).

$$\lambda_Y(x) = \sum_y P(y \mid x)\lambda(y) \tag{2-1}$$

The lambda values ($\lambda(y)$) for each child *Y* of *X* are defined as:

- 1 if *Y* is an evidence node. *y* is the instantiated value of *Y* and 0 otherwise.
- 1 for all values of *y* if *Y* is a leaf node and not an evidence node.
- $\prod_{U \in CH_Y} \lambda_U(y)$, where CH_Y is the set of children of *Y*, if *Y* is a nonleaf node and not an evidence node.

The pi messages sent from each parent node (*X*) down the tree to its child nodes (*Y*) contain the product of the lambda messages received from all children of *X* excluding *Y* and the pi value of *X* (Equation 2-2).

$$\pi_Y(x) = \pi(x) \prod_{U \in CH_X - \{Y\}} \lambda_U(x) \tag{2-2}$$

The pi values ($\pi(x)$) for each parent X of Y are defined as:

- 1 if X is an evidence node. x is the instantiated value of X and 0 otherwise.
- $P(x)$ for all values of x if X is a root node and not an evidence node.
- $\sum_z P(x \mid z)\pi_X(z)$ where Z is the parent of X, if X is not a root node and not an evidence node.

Given the above lambda and pi messages, the probability of any variable, X, given the set of instantiated variables, **A**, is defined as

$$P(x \mid \mathbf{A}) = \alpha\lambda(x)\pi(x) \tag{2-3}$$

where α is a normalizing constant. The message passing algorithm is implemented using the Pearl inference engine included in the Bayes Net Toolbox for Matlab downloadable at http://code.google.com/p/bnt/ [24].

There are many inference schemes available for use in Bayesian networks. Traditional Bayesian inference is often performed simply by computing the maximum *a posteriori* (MAP) estimate. This method consists of computing the posterior probabilities of the classes over all the variables, and choosing the maximum posterior as the class estimate. This form of inference uses the prior probabilities and the likelihood of the classes learned from the training data. It does not take into account any current knowledge from adjacent variables in the network, but treats each node as a separate problem. The estimates are based on the current feature evidence and the training data at the current variable.

Message passing considers not only information available at the current node, but information from all nodes in the network. A commonly used message passing method is the junction tree algorithm. This algorithm gives an exact inference, rather than an approximation, on any graph topology, regardless of the presence of loops. Belief

propagation, or Pearl's message passing, only gives exact inference on directed acyclic graphs (DAGs), or graphs without loops. It can be applied to graph structures containing loops, but it is not guaranteed to converge and gives only approximate inference. The junction tree algorithm performs graphical operations to manipulate the joint probability distribution over the tree in order to interchange sums and products so that partial sums can be calculated [25]. This avoids working directly with the joint distribution, which can be unreasonably large due to the number of variables. As mentioned, the junction tree algorithm gives exact inference for all graphs, but it can be computationally expensive. The algorithm divides the graph into cliques, or subsets of the graph in which all nodes are connected to all other nodes in the subset, and works with the joint distribution of each clique. The more variables in the clique, the larger the computations become, and this computational cost grows exponentially with the maximum number of variables in a clique.

Several topologies of my Bayesian network were considered, some having loops and others being true trees. Ultimately, the topologies presented here are acyclic graphs. Rather than dealing with the increased computations involved in obtaining an exact inference given a graph with no loops, I chose to use Pearl's message passing algorithm. This algorithm has the ability to obtain exact inference in a tree structure, but can also be used to achieve approximate inference (via loopy belief propagation), in networks with loops. Loopy belief propagation simply applies Pearl's message passing either until convergence is reached or for a given number of iterations. The results of approximate inference, due to the possibility of non-convergence, may be poor but are often accurate [26].

2.6 Utilizing Alternate Model Estimates

A physical model, WAM, has been previously calibrated to the Santa Fe Basin study site. While this model is not perfect, it does contain valuable information based on strong interactions within the basin. When performing stream flow predictions, WAM tends to overestimate flow in temporal regions of low flow and underestimate flow at times of sudden high peaks. This is largely due to the method in which WAM performs rain runoff calculations. WAM divides the basin into grid cells and for each cell assumes that a certain percentage of runoff within the cell drains into groundwater and another percentage becomes above ground runoff and is transferred into adjacent cells. Due to the complex nature of the underlying geologic system in the Santa Fe Basin, this is not always very accurate. Within stream flow in WAM is calculated based on kinematic wave equations [11], which is an accepted reliable practice in hydrologic routing [12]. In order to utilize these within stream relationships, WAM flow predictions are used as training data when modeling the spatial relationship of flow (spatial HMM and spatial BN). This also allows all of the true flow measurement data to be used as for testing in the models.

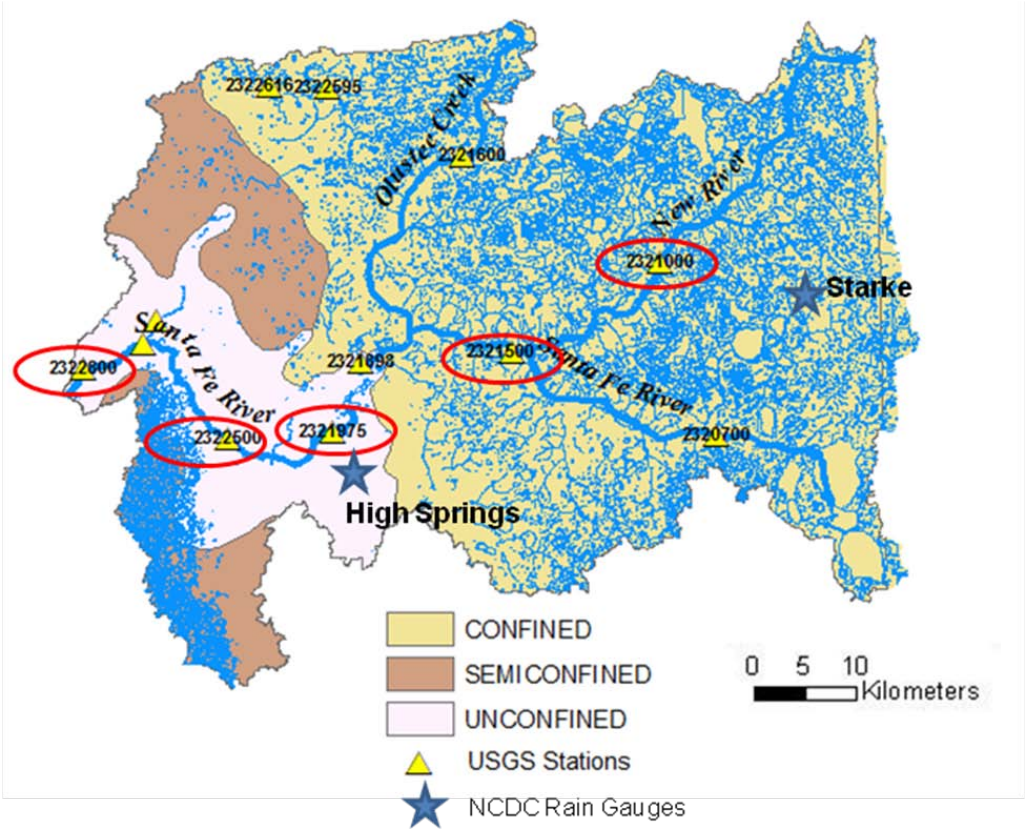

Figure 2-1. Santa Fe River Basin, Florida.

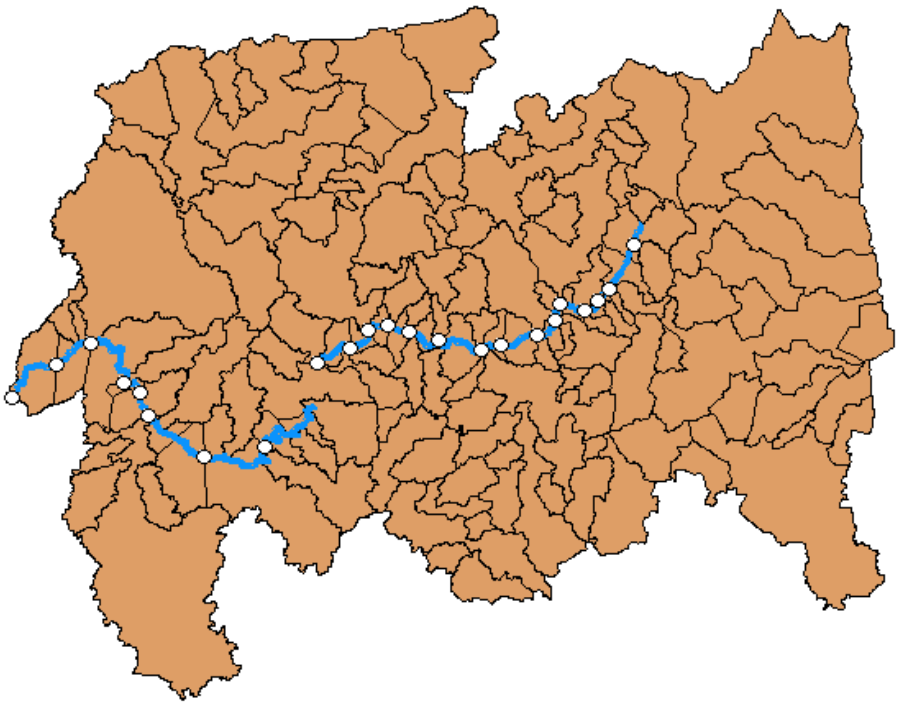

Figure 2-2. Reach outlets along the stream.

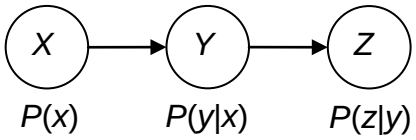

Figure 2-3. Example of a graphical model. X, Y, and Z are nodes representing random variables with conditional probabilities P(x), P(y|x), and P(z|y), as inferred from the arcs connecting the nodes.

CHAPTER 3
METHODS

3.1 Temporal Models

At any given location along the stream, future knowledge of flow is important in order to know which areas might be susceptible to floods. Three methods of forecasting flow are presented in this section: the Gamma model, Gaussian process regression, and the hidden Markov model. These models are applied independently at each spatial location along the river and their temporal relationships over a period of 3 days are modeled across the time record (3 day forecast). They approximate a solution to prediction in the full joint space by assuming spatial independence and modeling the temporal relationships with an assumption of some temporal stationarity.

3.1.1 Gamma Filter Model

The gamma filter is an adaptive feedforward infinite impulse response (IIR) filter which incorporates some of the advantages of a finite impulse response (FIR) filter for training. It is characterized by the following equations

$$y(n) = \sum_{k=0}^{K} w_k x_k(n) \tag{3-1}$$

$$x_k(n) = (1-\mu)x_k(n-1) + \mu x_{k-1}(n-1), \quad k = 1,...,K \tag{3-2}$$

where K is the number of taps, $y(n)$ is the filter output, $x_0(n)$ is the filter input, and μ is the adaptive memory parameter. Looking at Equation 3-2, it can be seen that the tap to tap transfer function can be written as

$$G(z) = \frac{\mu}{z-(1-\mu)} \tag{3-3}$$

which is a first order IIR filter. The filter structure is shown in Figure 3-1.

The filter only has one free recursive parameter, μ, which can be set to the value which gives the lowest mean square error during training. The order of the model, K, controls the number of FIR parameters and a higher number of taps will incorporate more of the historical data in the computation of $y(n)$. Finally, the weights are optimized for the training data as

$$W = R^{-1}P \tag{3-4}$$

where $R = E\left[X_n X_n^T\right]$, $P = E\left[d(n)X_n\right]$, X_n is the vector of tap variables at time n, and $d(n)$ is the desired signal at time n.

The filter in Figure 3-1 can be extended to incorporate multiple inputs. To include the contribution of rainfall to forecasting flow, a second gamma filter was trained and added to the output as in Figure 3-2. The filter was trained and tested on observed flow and rainfall data at each of the five USGS stations separately for the time periods specified in Section 2.2. After extensive testing, the number of taps was set to 10 and the optimal values of μ for the forecasting filter at each USGS location are given in Table 3-1. The values of μ at all five locations are less than or equal to 0.5, but are generally closer to the 0.1 – 0.2 range. The filter is stable for $0 < \mu < 2$, so the determined μ values are in the lower range of values that μ can take on. This means that in calculating the value of $x(n)$ at tap k in Equation 3-2, most of the contribution comes from the value of that tap at the previous time instance rather than from the previous tap. This results in a filter output that is more or less a weighted combination of the last K days input. Having a higher dependency on past values of flow indicates that the flow is more highly correlated in time. Higher values of μ indicate that additional contribution to $x_k(n)$ comes from the value at tap k-1 as well as the current tap's past

value, indicating that the temporal autocorrelation of the process being modeled is not as strong. Indeed, we see that at the most upstream station, 2321000, that the value of μ is the highest of the stations (0.5) where flow has a shorter autocorrelation length, and values of μ downstream are smaller where flow is more highly autocorrelated. Flow downstream contains information about external inputs to the system (local rain) as well as upstream flow. Additionally, the most downstream stations in the Santa Fe Basin are in the unconfined region where groundwater acts as a sort of low pass filter, making flow more slowly varying and therefore more correlated.

3.1.2 Gaussian Process Model

The Gaussian process can be used for both regression and classification. Here Gaussian process regression is used as a temporal model to forecast flow. In Gaussian process regression, a prior probability is defined over the output function directly, rather than over the input weights. A Gaussian process is defined [18] as "a collection of random variables, any finite number of which have a joint Gaussian distribution." Therefore, it can be defined by its mean and covariance function. The Gaussian process model of a function is

$$\begin{aligned} m(x) &= E\left[y(x)\right] \\ k(x_n, x_m) &= \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x_n - x_m)^2\right) \\ y(x) &\sim GP\left(m(x), k(x_n, x_m)\right) \end{aligned} \tag{3-5}$$

where the covariance functions has been chosen to be the squared exponential covariance. The hyperparameters σ_f^2 and l^2 are the signal variance and characteristic length scale, respectively, and are set by optimizing the marginal likelihood.

In the current application, current flow and rainfall at each of the five USGS locations over the testing period described in Section 2.2 are used to train the Gaussian process (learn the hyper parameters). Rain and flow values for each day of the test period are then used as test inputs and the predictive mean and variance are output. The forecasted flow value is taken to be the output mean.

To implement the Gaussian process model, the Gaussian Process for Machine Learning (GPML) toolbox developed by Rasmussen and Williams (available at http://www.gaussianprocess.org/gpml/code/matlab/doc/index.html) is used with a squared exponential covariance function, the exact inference engine, and a Gaussian likelihood function for regression with standard deviation of 0.1.

3.1.3 Hidden Markov Model

Hidden Markov models (HMMs) are graphical models that are suited for dealing with processes with an inherent temporality, or processes in which the value at one time is influenced by the value at the previous time step [19].

Learning in an HMM is done through the forward-backward algorithm. The probability of producing a certain sequence of visible observations is calculated based on the transitional probabilities between unobserved states. Figure 3-3 shows an example of a simple HMM. The transition probabilities between unobserved states are defined as

$$a_{ij} = P(\omega_j(t+1) \mid \omega_i(t)),\ \sum_j a_{ij} = 1 \text{ for all } i \tag{3-6}$$

where $\omega_i(t)$ is the hidden state at time t. The probability that a given unobserved state will produce a visible state (observation) is given by

$$b_{jk} = P(v_k(t) \mid \omega_j(t)),\ \sum_k b_{jk} = 1 \text{ for all } j \tag{3-7}$$

where $v_k(t)$ is the visible state associated with unobserved node $\omega_j(t)$ at time t.

Because there is a probability associated with the unobserved and visible state transition, the HMM model of a process is not a deterministic one in which the two visible states are directly linked. In an application of modeling a strongly correlated time series, such as stream flow, the added complexity of having a visible and hidden state for each observation may increase accuracy of prediction due to the decoupling of sequential observations.

HMMs are often used in speech processing to determine the unobserved state, rather than the observation [27]. In estimating stream flow, it is the actual observation that is desired. To achieve this, the method in [28] is implemented, which deals with the problem of having missing observations. A missing feature/observation at time $n+k$ is estimated as

$$\hat{x}_{n+k} = \sum_{i=1}^{K} \mu_i P(s_{n+k} = i \mid x_n, x_{n+N}) \tag{3-8}$$

where

$$P(s_{n+k} = i \mid x_n, x_{n+N}) = \kappa \frac{1}{\pi(i)} \left[\mathbf{A}^{(k)} \mathbf{e}_{q_n}\right]_{(i)} \left[\mathbf{B}^{(N-k)} \mathbf{e}_{q_{n+N}}\right]_{(i)} \tag{3-9}$$

given that observed data (x) is available at nodes n and $n+N$ and everything in between is missing, and $n < k < N$. s_m is the state at time m and has a pdf with mean μ_m. Transition matrices **A** and **B** are defined as

$$\mathbf{A}(i,j) = a_{ij}, \; a_{ij} = P(s_m = i \mid s_{m-1} = j) \tag{3-10}$$

$$\mathbf{B}(i,j) = b_{ij}, \; b_{ij} = P(s_m = i \mid s_{m+1} = j). \tag{3-11}$$

Additionally q_n is defined as $q_n = \arg\min_j \left\| x_n - \mu_j \right\|^2$ and $\mathbf{e}_i$ is an indicator vector of length 16 (corresponding to 16 states) consisting of all zeros except for a 1 at index *i*.

If there are no future observations available, the problem becomes one of strictly forecasting features. In this case, the probability in Equation 3-9 (and Equation 3-8) becomes

$$P(s_{n+k} = i \mid x_n) = \hat{\kappa}[\mathbf{A}^{(k)}\mathbf{e}_{q_n}]_{(i)}. \quad (3\text{-}12)$$

The values κ and $\hat{\kappa}$ need not be evaluated as they are normalizing factors.

The HMM for stream flow estimation is depicted in Figure 3-4. A separate HMM is trained at each of the five USGS locations along the stream. A unimodal Gaussian is used to model the probabilities between the hidden states and the observations. The hidden states are chosen to have 16 states to correspond to the number of classes used in the Bayesian networks. Each observation is taken to be two dimensional, corresponding to rainfall and flow on the current day. The transitional matrix **A** is learned from the temporal training data defined in Section 2.2. For each day *t* in the testing period, flow and rain are presented at time *t* and Equations 3-8 and 3-12 are used to forecast flow on day *t*+3 (for a 3 day forecast).

3.2 Spatial Models

Accurate forecasting is important, but if *in situ* sensors are distributed too sparsely, forecasts at the sensor locations have limited usefulness. Spatial estimation, or within stream prediction, can provide a different kind of information than forecasting methods. Two methods of spatially predicting flow are presented in this section: the Bayesian network and the hidden Markov model. Rather than modeling the spatial relationship of each node to all other nodes, just the conditional probabilities between neighboring

nodes are used. This is a reasonable simplification due to knowledge of the physics of the stream flow problem in that flow at one location contains information of flow at upstream locations, so can be considered independent of all upstream nodes given its immediate upstream neighbor. These methods link the stream flow data spatially, but lose the sequential relationship of flow over time. The transitions from flow at one node to a node downstream are based on the most probable transition over the time record of the training data, and so an assumption of temporal stationarity is imposed.

3.2.1 Bayesian Network

The causal nature of stream flow easily lends itself to being characterized by a directed graphical model. Locations of interest along the stream become nodes in the graph network, connected in the direction of flow from upstream locations to downstream locations. The spatial network of flow is shown in Figure 3-5 and consists of 23 nodes, one located at each of the reaches defined in WAM. The root node is the most upstream flow reach and the leaf node is the most downstream reach.

To form the conditional probabilities between the nodes, WAM flow predictions were used from October 2000 through July 2008, as described in Section 2.2. Because WAM adequately models the within stream relationships, this leaves all the available USGS data over the same time period to be used for testing. Gaussian process regression is used to compute the transitional probabilities between the nodes as a way of working with a true posterior (*likelihood* x *prior*), rather than just a likelihood. The Gaussian process is trained with the class labels of clustered WAM data at a parent-child pair of nodes and then each possible class label (1 through 16) of the parent node is presented as a test input. A Gaussian with the output mean and variance then determines the probability of the child class given the parent class.

To perform inference at unobserved nodes in the Bayesian network, there are a variety of algorithms available (see Section 2.3). I chose to implement the sum-product belief propagation algorithm, also known as Pearl's message passing algorithm [22], [23]. At each node in the network, probabilities in the form of "messages" are passed to its directly surrounding nodes. Once each node has sent and received a message to/from all of its surrounding nodes, the node infers its value (belief) based on the information passed to it from the surrounding nodes.

We do not have true USGS flow data at all of the spatial locations, but would like to use it for testing the Bayesian network. Therefore, when performing inference, true USGS flow (clustered into 16 classes) is presented as evidence in the network at 3 stations (2321000, 2321500, and 2322500) while 2 stations are reserved for validation (2321975 and 2322800).

3.2.2 Hidden Markov Model

The hidden Markov model was previously described in Section 3.1.3 for temporal forecasting. Any of the models in Section 3.1 can be applied as a spatial model following the same principles here. Rather than implementing the gamma filter and the Gaussian process spatially as well, the hidden Markov model is chosen as an example. Because it is another graphical model and is a discrete model, it is more apt for comparing to the spatial Bayesian network.

For this method an additional preprocessing step has been applied. Stream flow in the Santa Fe basin is not only non-stationary in time, but also spatially due to the Cody Scarp, which is the boundary between the confined and unconfined regions. Flow in the unconfined region tends towards much higher levels than flow in the confined region, both because of the different hydrogeology, and because it is the portion of the

river that is more downstream. The HMM has a single transitional probability matrix between states due to the Markov assumption. In order to make flow along the river appear more spatially stationary, flow at each location is “normalized” by its contributing area (square miles). The contributing area of each location consists of the physical area through which water drains into the stream reaches upstream of the current location. This normalizing is done prior to the preprocessing mentioned in Section 2.2.

The spatial HMM is trained with WAM data as described in Section 2.2 and estimation of the missing observations is approached as in Section 3.1.3. Now, however, if USGS flow at stations 2321000, 2321500, and 2322500 are presented as known observations, Equation 3-9 is used in Equation 3-8 for estimating flow at locations with known upstream and downstream observations and Equation 3-12 is used in Equation 3-8 at locations with either an upstream or a downstream known observation only. The model is validated at stations 2321975 and 2322800. Figure 3-6 shows an illustration of the spatial HMM setup.

3.3 Spatial-Temporal Models

A model which is able to incorporate both within stream dependencies and forecast temporally will be able to capture the more complete behavior of the stream flow process than just spatial predictors and temporal forecasters alone. To this end, a Bayesian network distributed spatially for the current day and a forecast day is presented in this section. This model attempts to model the spatial relationship between nodes along the stream, yet also incorporate the temporal relationships at each node in order to forecast flow. The model attempts to more fully capture the joint space information, but in order to do so would need explicitly represent the temporal dependencies at each location. To keep the dimensionality of the problem low, a single

relationship of flow on day *t* and flow on day *t*+3 is modeled at each location. This assumes flow is stationary over time, which is not the case. Nevertheless, this assumption is made in order to see the potential benefit the Bayesian network can have on the problem of estimating flow spatially and temporally.

3.3.1 Bayesian Network

The spatial Bayesian network of Section 3.2.1 is modified to include an additional child node for each spatial node. These new nodes represent the three day forecast at each spatial location, while the original nodes contain the current day's flow information. The topology is shown in Figure 3-7. To aid in the forecasting of stream flow, a rain node with flow on the current day is added as a parent to each of the forecast day nodes. The network is trained as described in Section 3.2.1 with WAM data for the transitional probabilities on the nowcast day and USGS data (where available) for the transitional probabilities for the forecast day. Where USGS flow is not available, WAM is used to determine the forecast day's pdfs. Because USGS data is being used in the training, the training and testing periods cannot overlap as they did in the purely spatial Bayesian network. Data from October 2000 through December 2006 is used for training and data from January 2007 through July 2008 is used for testing. Additionally, evidence is presented at 2321000, 2321500, and 2322500 on the nowcast day, as well as at the rain nodes. The model is validated at 2321975 and 2322800 on the nowcast day, and at all five USGS locations on the forecast day.

3.3.2 Bayesian Network Modifications

The spatial-temporal Bayesian network of the previous section models the temporal relationship of flow at each location over a three day period, but this relationship in the model is static. The pdfs for the forecast day nodes contain the joint

probability of rainfall, current flow, and forecast flow, but these are generalized over the training period so do not capture the more intricate day to day interactions of flow and rainfall. In order to more explicitly represent the time dependency of future flow on past flow and rainfall, modifications can be made to the Bayesian network. Two ideas are discussed here.

One idea is to preprocess the flow data that is input to the Bayesian network to be flow conditioned on previous flow rather than the raw flow data itself. To accomplish this, the autocorrelation length of flow at each of the USGS stations was calculated and found to range from 9 days at the most upstream stations to around 1000 days at the downstream stations. As a first step, a correlation length of 9 days was chosen for conditioning flow at all stations. Conditioning flow on past flow that is too distant can lose forecasting information, and the day to day relationship of flow at upstream locations is likely to be present downstream, albeit in a muted or altered form. Rather than clustering daily flow data into 16 classes, flow on the current day and flow 9 days prior was jointly clustered into 16 classes. This quantized conditional flow is then used as input to the Bayesian network as discussed in Section 3.3.1. The inferred classified values output at the validation nodes appears to be reasonably well matched to the desired classified data, but the problem lies in relating the class values back to true flow values (ft^3/s). Previously, the output flow class was assigned the flow center value for the forecast day of its class cluster. Now, however, since the class value represents conditional flow, it is difficult to know if the class value should represent the flow class center value on the current day of flow, or if it should be dependent on the previously

estimated day's flow value as well. The output flow class must be related back to an actual flow value in order to be of much benefit to researchers.

Alternatively, the form of the inputs of the model can remain the same while the model topology is changed. In order to have a better idea of if flow has been increasing or decreasing, additional layers of historical flow data can be included. The network in Figure 3-8 shows one topology that includes day (t-1), day t, and day (t+3), or a past day, current day, and forecast day of flow. Rainfall was omitted from this network both for simplicity and to see how well flow could be forecasted merely from flow itself should rainfall data not be available, but it could be included if desired.

These are both ways of approximating what a dynamical system does. A dynamical system maps the time series of flow from one dimension to a higher dimensional space whose coordinates are defined by the previous K samples. Flow can be mapped this way as a preprocessing step and then used in the current Bayesian network without adding explicit time dependencies, because the temporal information is contained in the data points themselves. This mapped flow can then be clustered and used in the Bayesian network. Since the amount of data is reduced in performing the temporal embedding, however, it is possible that the resulting data will be too sparse to cluster appropriately. In this sense, the previous two methods discussed are a good approximation to including the full temporal information in the input data themselves.

3.4 Information Contribution of Features

It is known that many different processes are driving forces in the behavior of stream flow, such as precipitation, groundwater level, topography, evapo-transpiration, soil temperature, etc., but it is desirable to incorporate only the most useful features in a model to keep the complexity feasible. This raises the question of if there is a

methodical way to determine the most informative features contributing to the estimation system. Consider the Bayesian network for both spatially predicting and temporally estimating flow. All available features could be added to the Bayesian network, but this increases the computational complexity and the runtime of the model, mainly due to forming the joint pdfs of increasing dimensions (curse of dimensionality). Therefore, the number of features used is often reduced to the most informative one or two features. A framework is developed here that quantitatively determines the information gained by incorporating a new feature so that well-informed tradeoffs can be made regarding computation complexity and accuracy.

In [29] and [25], a spatio-temporal Bayesian network (STBN) was developed to estimate flow at specific points within the Santa Fe Watershed utilizing a combination of spatial, temporal, and spatial-temporal data. For the remainder of this discussion, the STBN will be analyzed and is shown in Figure 3-9. In this formulation of the STBN, stream flow (discharge) is modeled as a multinomial distribution with each flow regime corresponding to a class C_i, and feature likelihoods $p(X \mid C_i)$ for each flow class are estimated at instrumented points along the rivers via the histogramming method of Parzen windowing. When estimating the unobserved flow at a location with no *in situ* data, it is important to know which features will contribute the most information, to limit the dimensionality and complexity of the problem. One way to pick the most informative features is a system of feature ranking based on measures such as mutual information (MI) or conditional entropy (CH) amongst the flow classes. The feature combinations that provide the most separation among all the classes are considered best. The MI gives a quantitative way of measuring this separation. Given a feature set X and class

C, the best subset of features X is the set that gives the largest MI between the feature pdfs and class membership based on the training data.

MI based on Shannon's entropy, the Kullback-Leibler divergence, can be computationally intensive and difficult to estimate, since the integral of a logarithm must be computed. Quadratic MI (QMI) has been proposed in [30], based on Renyi's α-order entropy ($\alpha = 2$), defined as

$$H_R(X) = \frac{1}{1-\alpha}\log\left(\int_{-\infty}^{\infty} f_X(z)^{\alpha} dz\right) = -\log\left(\int_{-\infty}^{\infty} f_X(z)^2 dz\right). \quad (3\text{-}13)$$

This moves the integral inside the logarithm and allows MI calculations to be simplified by utilizing the information potential [31]. The QMI then becomes a measure of the separation between the distributions of the classes based on a given feature(s). It is a useful measure for feature ranking because it not only measures the separability among all the classes, but also the compactness of individual classes.

The form of QMI based on the Cauchy-Schwartz divergence (QMI-CS) for 2 distributions and 2 classes is

$$I_{CS} = \log\frac{\int p(x_i)^2 dx_i \int q(x_i)^2 dx_i}{\left(\int p(x_i)q(x_i)dx_i\right)^2}, i = 1, 2. \quad (3\text{-}14)$$

QMI-CS is used to rank features in order of most informative to least informative so as to determine which subset is most predictive for developing the flow prediction network. This involves maximizing the QMI-CS between features and stream flow classes. From the above equation, $I_{CS}(C;F_i)$ is obtained as,

$$\log \frac{\left\{\sum_{k=1}^{N}\left[P(C_k)^2 \int f_{F_i}^2(F_i \mid C_k)dF_i\right]\right\}\left\{\sum_{k=1}^{N}\left[P(C_k)^2 \int f_{F_i}^2(F_i)dF_i\right]\right\}}{\left\{\sum_{k=1}^{N}\left[P(C_k)^2 \int f_{F_i}(F_i \mid C_k) f_{F_i}(F_i)dF_i\right]\right\}^2} \tag{3-15}$$

where N is the number of classes, C is the set of classes, and F_i is the feature set under evaluation.

Another way to approach this problem is to minimize CH of flow classes with respect to the feature set as

$$H(C \mid X) = \sum_{x,c} p(x,c) \log \frac{p(x)}{p(x,c)}. \tag{3-16}$$

The Shannon MI, expressed as $I(X;C) = H(C) - H(C \mid X)$, reveals a dependence on conditional entropy. When relatively few training samples are available for one or more classes the estimate of $H(C)$ can become biased. Using the CH of flow class C given feature set X avoids this dependence on the sole entropy of C in MI methods.

In the approach presented here, feature nodes are ranked using both QMI-CS and CH and the performances of the networks using the suggested features from both ranking schemes are compared. The hydrologic makeup of the study site is non-uniform, making it difficult to find a single Bayesian network to accurately predict stream flow at all locations in the watershed. The three test sites used are USGS stations 2321000, 2321500, and 2322500. The feature ranking algorithm is applied separately to each of the three stations, thus obtaining the optimal structure at each location. The dimensionality at each node is limited to a maximum of three. The information content of the nodes in the previously defined STBN [29] is measured and the potential of additional features is analyzed.

The features considered are:

- Rainfall, Runoff (260m spatial, daily temporal resolution)
- Watershed Assessment Model (WAM) flow estimate along the river (daily temporal resolution)
- Ground Water Levels (GWL) at fixed well locations (daily temporal resolution)
- Measured flow at the nearest neighboring (TNN) station (daily temporal resolution)
- Curve Number (260m spatial resolution), CN.

First, the information metrics (i.e. CH and QMI-CS) are used to determine the single most informative feature at each of the test stations (rankings shown in "First Feature" column of Tables 3-2, 3-3 with the best ranked feature in row "Rank 1," etc.). Next, the remaining features each paired with the best feature from the previous step are ranked ("Second Feature" in Tables 3-2, 3-3). This process is continued until the output node (node F) has three input feature nodes. Each time a feature is added to the network, the information metrics are computed for the entire network. The decrease in values along the rows in Table 3-2 (and increase in Table 3-3) indicates features with additional information are being added to the network, reducing CH and increasing QMI-CS.

Runoff is a spatial-temporal process that depends on rainfall and CN (determined from landcover type). But rainfall and CN data over the contributing area for each test site cannot be directly added to a BN with purely temporal features, so, as described in [29], the STBN is constructed and spatial runoff is mapped into temporal runoff using temporal lags. The same feature ranking process is then applied to the spatial level in the STBN. When evaluating the temporal version of runoff, the information contained in its parent nodes must be accounted for, namely spatial rainfall and curve number. This

is accounted for by multiplying the information metric of the spatial network (cross-hatched nodes in Figure 3-9) by the metric for runoff evaluated in the temporal network. This accounts for the total information gain of using runoff.

The features were ranked and added to the Bayesian network according to CH and QMI-CS separately, as shown in Tables 3-2 and 3-3. For both measures, the set of features ranked among the top three follow similar trends. QMI-CS gives a slightly different ranking order at 2321500 than CH. Stream flow is divided into 15 classes, making the highest CH value of class separability 2.7 ($\log_e 15$). The features in the first row in Tables 3-2 and 3-3 are those chosen as the best three features at each training station. To gauge the performance of the networks using these rankings, the Nash-Sutcliffe Efficiency (NSE) measure from the hydrology literature [32] is used, which has an upper bound of unity (NSE = 1 for perfect agreement). It is suggested in [32] that NSE > 0.6 is satisfactory for monthly flow samples, implying slightly lower values are still satisfactory for daily samples. From Table 3-4, CH and QMI-CS perform comparably at 2321000, while CH is better at 2321500 and QMI-CS is best at 2322500. It appears that CH can be used confidently in the confined region. In the downstream unconfined area where flow is less transient, compactness of individual class pdfs should be accounted for and hence QMI-CS is suggested. Although the same three features were chosen at several locations, the information metrics are different due to the use of different feature values at each location (i.e. different TNN values at each station, rain over the contributing area for each individual station, etc.). Additionally, the order of the features selected is important. The kernel size in Parzen windowing was chosen to be the smallest such value while there is a non-zero probability for each flow

class at each of the three stations. This ensures that there will be a non-zero probability of flow given test inputs that are not represented in the training data. The kernel size for each feature dimension is set given the previously chosen feature, so if features are selected in a different order, they will have a different joint pdf and produce different flow estimates.

At stations 2321000 and 2321500, most of the information is contained in feature TNN. This is because these stations are located in the upstream confined area of the watershed where rainfall has an immediate and short-lived effect on discharge. In such cases where groundwater does not contribute significantly to stream flow, upstream flow is a very good indicator of flow at a downstream station. Adding groundwater and WAM contribute only a 1.4% reduction in CH, while TNN eliminates 41.6% of the uncertainty at 2321000. The same set of features is suggested by both metrics, and the NSE values confirm the validity of using either metric. At 2321500, TNN and GWL eliminate 33.8% uncertainty and adding WAM eliminates only 0.5% more according to CH. QMI-CS chooses a slightly different set of features, but the NSE values in Table 3-4 suggest that features selected by CH are more reliable. Low CH values and smaller incremental information gains at these two stations indicate that a complex STBN is probably not necessary.

Station 2322500 lies in the unconfined region, where groundwater is the dominant contributor to stream discharge. Here flow at station 2321500 is used as the TNN feature, but this station lies in the confined region. Due to this relationship, TNN is not chosen as the best feature for this station. Stream discharge in the unconfined region is affected by a number of variables making it hard for one feature to adequately

characterize it (net reduction of 30.2% in CH values with 3 features). QMI-CS is a more appropriate metric in this region (as indicated by NSE values in Table 3-4) because measuring compactness of individual discharge classes is important in the unconfined area since there is less temporal flow variation.

Spatial analysis of runoff impacts on stream discharge is less warranted at stations fed by small contributing areas (e.g., 2321000). In the case of large contributing areas (e.g., 2322500), nonstationarities observed in the spatial features are significant and performance can benefit from including runoff in the network (as indicated by the QMI-CS value for 2322500 in Table 3-3 that exceeds unity). NSE values were also computed from flow estimates using just the best two features in the temporal portion of the STBN. NSE, averaged over all three stations, decreased dramatically from almost 0.50 to less than zero for both CH and QMI-CS.

Table 3-1. Adaptive memory parameter for Gamma filter at each forecasting location.

Station	μ
2321000	0.5
2321500	0.2
2321975	0.1
2322500	0.2
2322800	0.2

Table 3-2. Feature ranking based on conditional entropy (CH). The set of available features reduces by one with each column, represented by "-" entries.

		2321000			2321500			2322500		
	Rank	First feature	Second feature	Third feature	First feature	Second feature	Third feature	First feature	Second feature	Third feature
Temporal	1	TNN: 1.577	GWL: 1.548	WAM: 1.540	TNN: 1.898	GWL: 1.787	WAM: 1.774	WAM: 2.010	GWL: 1.945	TNN: 1.885
	2	WAM: 1.831	WAM: 1.559	rain: 1.546	WAM: 1.903	WAM: 1.844	rain: 1.782	GWL: 2.072	TNN: 1.947	rain: 1.905
	3	GWL: 1.876	rain: 1.574	runoff: 1.635	GWL: 1.924	rain: 1.877	runoff: 1.884	TNN: 2.171	rain: 1.963	runoff: 2.018
	4	rain: 1.943	runoff: 1.660	-	runoff: 2.014	runoff: 1.896	-	rain: 2.372	runoff: 2.064	-
	5	runoff: 2.057	-	-	rain: 2.047	-	-	runoff: 2.516	-	-
Spatial	1	CN: 0.232	rain: 1.059	-	CN: 0.259	rain: 1.059	-	CN: 0.143	rain: 1.059	-
	2	rain: 0.232	-	-	rain: 0.260	-	-	rain: 0.153	-	-

Table 3-3. Feature ranking based on Mutual Information (QMI-CS). The set of available features reduces by one with each column, represented by "-" entries.

		2321000			2321500			2322500	
Rank	First feature	Second feature	Third feature	First feature	Second feature	Third feature	First feature	Second feature	Third feature
Temporal 1	TNN: 0.194	WAM: 0.218	GWL: 0.229	GWL: 0.286	WAM: 0.308	rain: 0.329	WAM: 0.329	TNN: 0.367	GWL: 0.395
Temporal 2	WAM: 0.194	GWL: 0.206	rain: 0.223	runoff: 0.273	rain: 0.292	TNN: 0.304	GWL: 0.322	runoff: 0.346	runoff: 0.378
Temporal 3	GWL: 0.169	rain: 0.199	runoff: 0.050	WAM: 0.244	TNN: 0.291	runoff: 0.301	TNN: 0.250	GWL: 0.341	rain: 0.362
Temporal 4	rain: 0.146	runoff: 0.047	-	rain: 0.191	runoff: 0.235	-	rain: 0.173	rain: 0.327	-
Temporal 5	runoff: 0.035	-	-	TNN: 0.152	-	-	runoff: 0.156	-	-
Spatial 1	rain: 0.034	CN: 0.294	-	rain: 0.212	CN: 0.821	-	rain: 0.377	CN: 1.073	-
Spatial 2	CN: 0.003	-	-	CN: 0.004	-	-	CN: 0.198	-	-

Table 3-4. Actual flow error residuals (NSE values) at each station for the STBN using the features selected from the CH rankings and those selected from the MI rankings.

	2321000	2321500	2322500
CH	0.476	0.635	0.300
QMI-CS	0.502	0.274	0.661

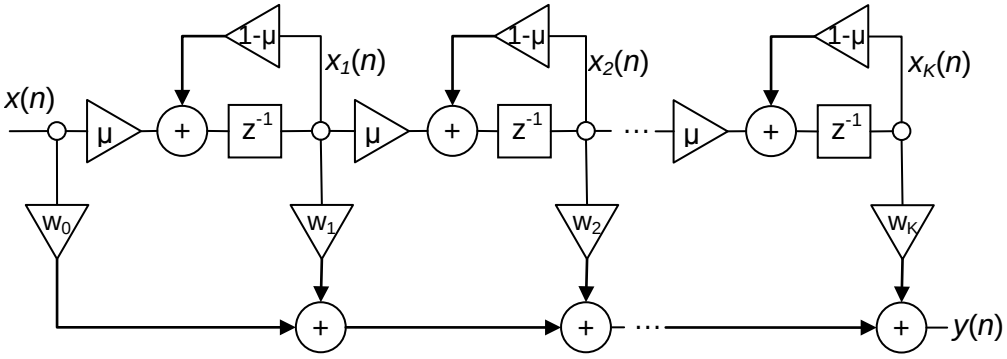

Figure 3-1. Gamma filter structure with *K* taps.

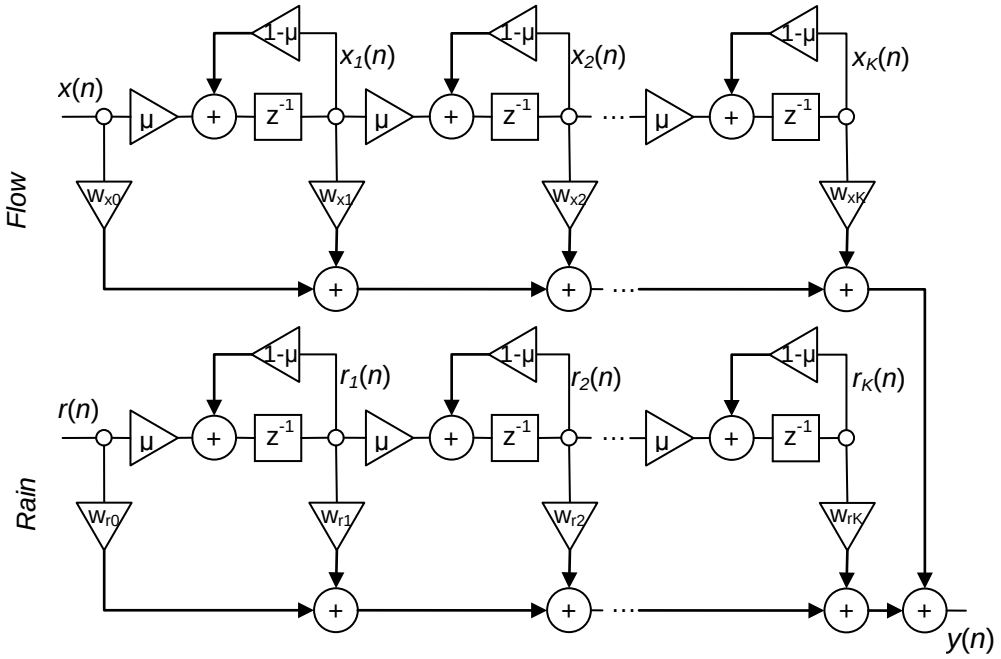

Figure 3-2. Gamma filter for flow forecasting with flow and rain inputs.

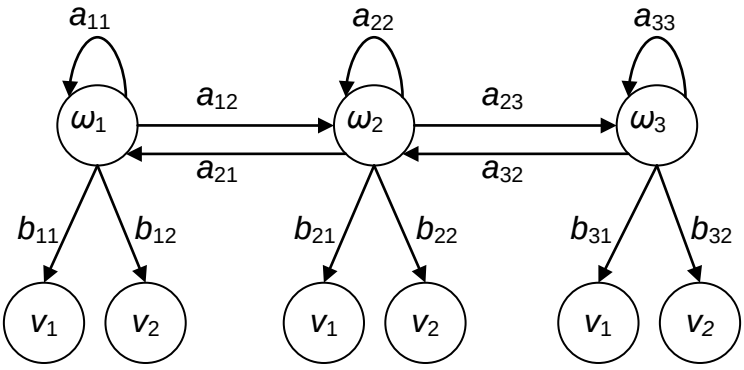

Figure 3-3. Example hidden Markov model (HMM).

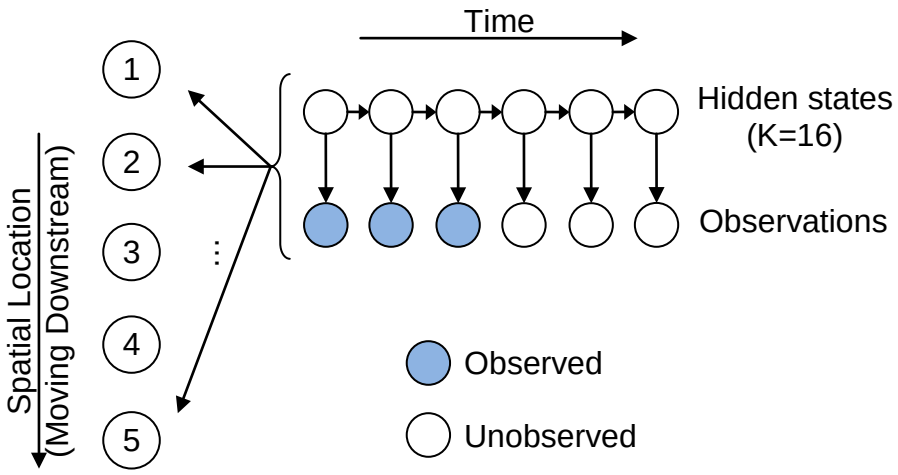

Figure 3-4. Hidden Markov model for temporal stream flow estimation.

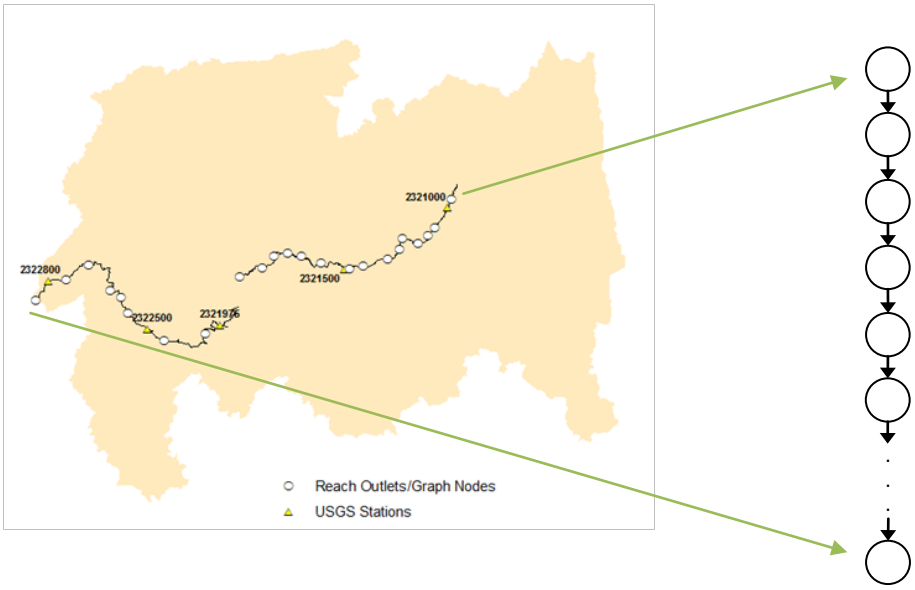

Figure 3-5. Along stream graph topology as it corresponds to 23 reach outlets along the stream (spatial BN).

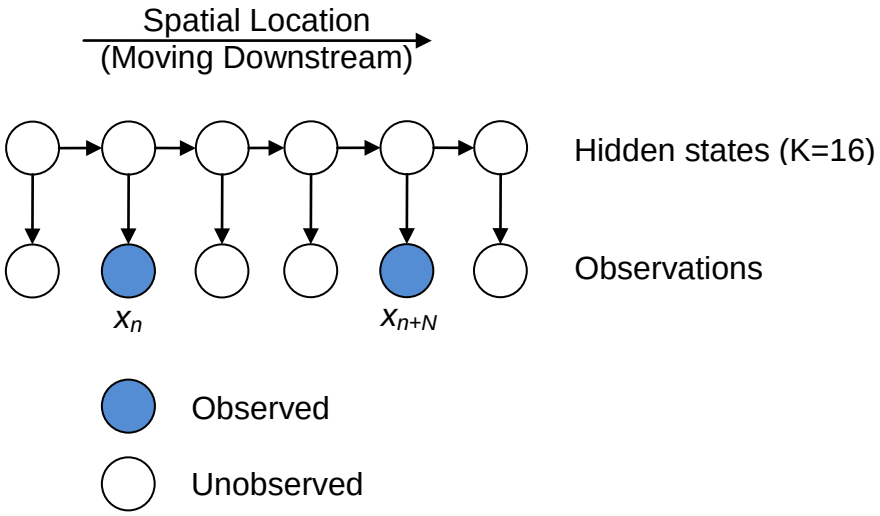

Figure 3-6. Spatial HMM illustration.

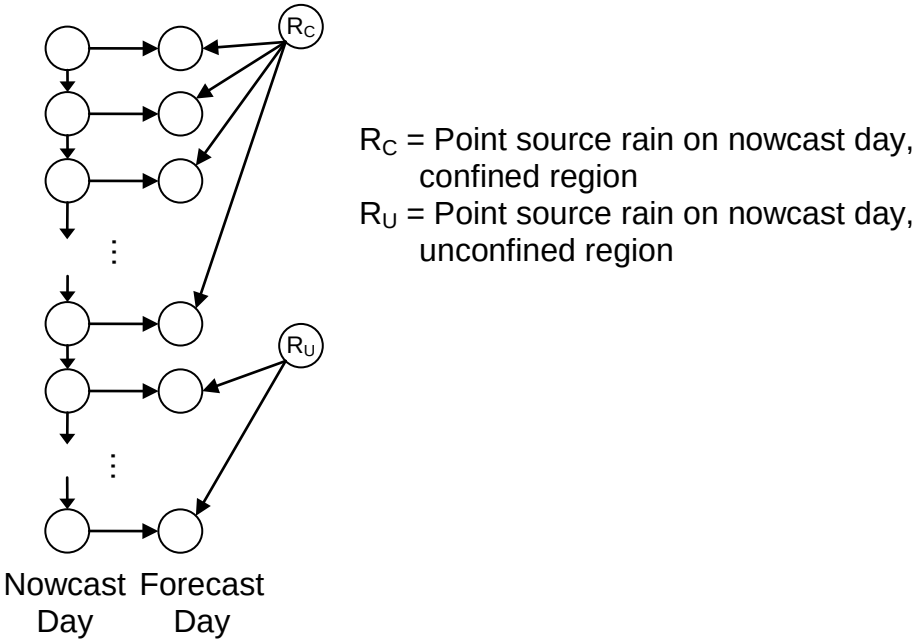

Figure 3-7. Bayesian network topology for nowcasting flow and forecasting flow with point source rain data.

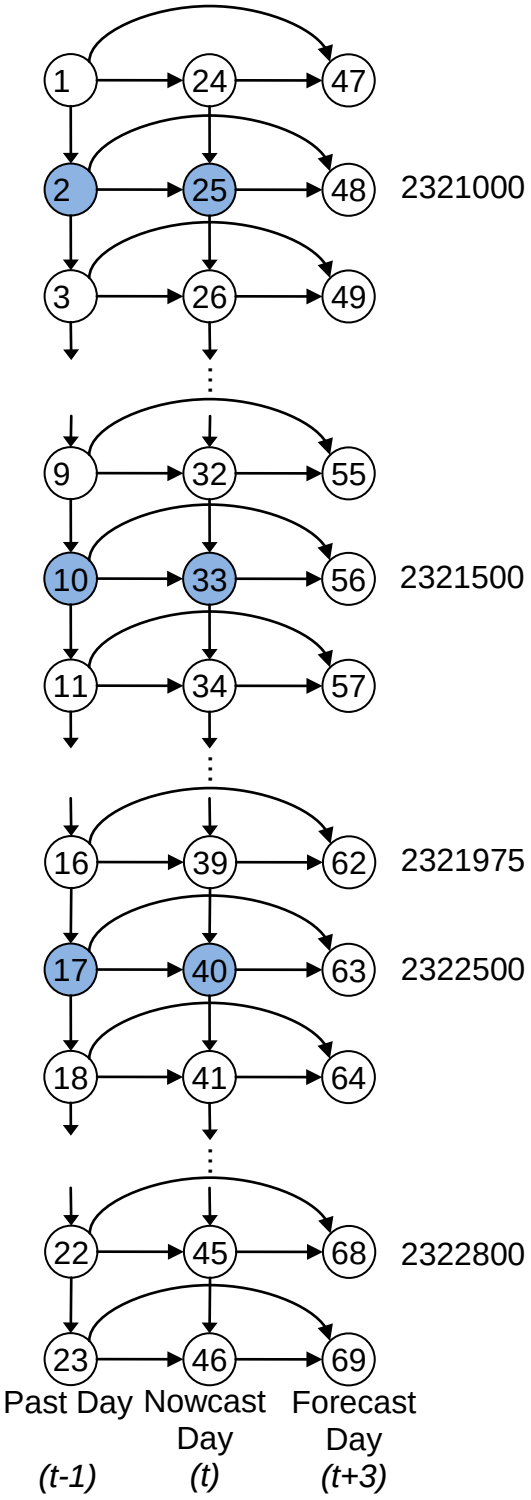

Figure 3-8. 23 spatial node Bayesian network with additional past day of flow. Blue nodes are evidence nodes.

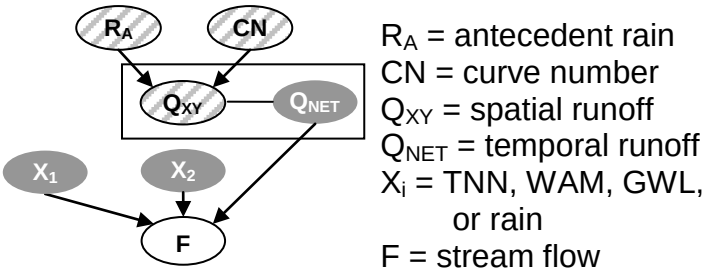

Figure 3-9. Spatio-temporal Bayesian network (STBN).

CHAPTER 4
RESULTS AND ANALYSIS

4.1 Performance Metrics

In order for any of the models presented to be worthwhile, they must do better than a trivial predictor. A temporal trivial predictor forecasts future flow as the current value of flow. Spatially, a trivial predictor estimates flow at a given location as current flow at the nearest upstream observed location. Before any model is deemed worthwhile, it should beat the trivial predictor.

The Nash Sutcliffe Efficiency Coefficient (NSE), the root mean square error (RMSE)-observations standard deviation ratio (RSR), and the percent bias (PBIAS) were computed as error metrics. These were chosen as they are indicative of how well the model performs on average and are commonly used metrics in the hydrologic field [32]. The error metrics are defined as

$$NSE = 1 - \frac{\sum_{i=1}^{N}\left(Y_i^{obs} - Y_i^{est}\right)^2}{\sum_{i=1}^{N}\left(Y_i^{obs} - \overline{Y^{obs}}\right)^2} \tag{4-1}$$

$$RSR = \frac{\sqrt{\sum_{i=1}^{N}\left(Y_i^{obs} - Y_i^{est}\right)^2}}{\sqrt{\sum_{i=1}^{N}\left(Y_i^{obs} - \overline{Y^{obs}}\right)^2}} \tag{4-2}$$

$$PBIAS = \frac{\sum_{i=1}^{N}\left(Y_i^{obs} - Y_i^{est}\right)}{\sum_{i=1}^{N} Y_i^{obs}} * 100 \tag{4-3}$$

where Y_i^{obs} is the observed value at time *i*, Y_i^{est} is the estimated value at time *i*, $\overline{Y^{obs}}$ is the mean of the observed values, and *N* is the total number of data points. NSE has

values on the range of (-∞, 1], with 1 indicating perfect prediction. RSR can have values on the range of [0, ∞), with a value of zero indicating perfect prediction (RMSE = 0). RSR normalizes the RMSE by the standard deviation of the observation data. This makes it unnecessary to qualify what a low RMSE value is based on the data. Percent bias is an indicator of if the model is generally over-predicting or under-predicting, and by what percentage. A negative PBIAS implies the model tends to over-predict, while a positive PBIAS implies the model tends to under-predict. A PBIAS of zero would indicate no bias within the model. According to [32], satisfactory values for stream flow estimation on a monthly time scale are: NSE > 0.50, RSR ≤ 0.70, and PBIAS < ±25%. Daily estimation is considered in this work, so in reality these cutoffs can be lowered a little, but I leave them as they are. Table 4-1 summarizes the ranking of the performance metric ranges as “Very Good,” “Good,” “Satisfactory,” and “Unsatisfactory”.

Values of these performance metrics are given for the temporal trivial predictor at all five USGS stations in Table 4-2 and for the spatial trivial predictor at 2321975 and 2322800 (using true flow as flow at 2321500 and 2322500, respectively) in Table 4-3. Plots of original flow measurements at each of the five USGS stations from October 2000 through July 2008 are shown in Figure 4-1. There are several areas of data dropout in the observations at 2322800, and these regions are not considered in any performance statistics. Specifically, data from 10/1/2000 through 1/31/2001 and 10/4/2005 through 9/30/2007 are missing at 2322800. Data to the left of the dashed line in Figure 4-1 is used for training in temporal models and data to the right is used for testing.

The temporal trivial predictor performs Unsatisfactory to Satisfactory in the confined region, and Very Good in the unconfined region, indicating that on average flow today is a good indicator of flow three days from now in the unconfined region, but that there is a higher variation of flow on a day to day basis in the confined region. Spatially, upstream flow is a Satisfactory to Good predictor of downstream flow, being better when both stations are in the same geologic regime (2322500 and 2322800). There is room for improvement in these estimates, so modeling flow spatially and temporally could contribute valuable information.

4.2 Temporal Model Forecasting

Models can be designed to forecast flow for any number of days into the future, but the performance is a function of the number of estimation steps ahead and the quantity of available training data. The further out one attempts to forecast, the less reliable results will be. All forecast work presented in this document is for a 3 day forecast so as to be enough of a forecast to be more useful than today's flow and not so far out that uncertainty becomes unreasonable. In this section, the performance of the three temporal models presented in Section 3.1 are analyzed.

4.2.1 Gamma Filter Model

A gamma filter was trained to forecast flow at each of the five USGS stations. Hydrographs of the observed and estimated forecast flow values on the testing period are shown in Figure 4-2 through Figure 4-6 for each of the five stations, along with a plot of the daily error. The performance statistics for each of the filters is summarized in Table 4-4. There is an improvement in forecasting over the trivial model in the confined region, but similar performance in the unconfined. Performance rankings are Satisfactory or better at all locations and consistently Very Good with minimal bias in the

unconfined region. Although the flow forecast at 2322800 is very poor if the hydrograph is analyzed, the NSE, RSR, and PBIAS values indicate the model is working very well, yet tends to underestimate flow (positive bias). This comes from the statistics being reduced to a single value over the entire time period rather than a daily statistic. The error plot of observed minus estimated flow is a better indicator of daily performance.

At all locations the most significant error occurs in the regions of peak flow. The positive initial error at a peak indicates that the model is having difficulty predicting an increase in flow (underestimating). There is less error on the falling edges of peaks, which indicates that although the model has trouble predicting higher flow, it does know that once flow goes high, it does not tend to stay high for very long. There is more constant error throughout the time period in the unconfined region than in the confined area. This seems to be because the model expects the flow to have more variation in it (has noisier estimates) than it actually does. This is due to including rainfall in the forecast procedure when flow in the unconfined region is not as affected by rainfall than in the confined region. The physics of flow in the Santa Fe Basin are such that rainfall has a more immediate effect on flow in the confined region than in the unconfined. Base flow level is higher in the unconfined region (~500 ft^3/s) and flow reacts more slowly to external inputs such as rainfall. The gamma filter model, in its current design, uses the same value of μ for the flow contribution as the rainfall contribution. The low values of μ indicate that for both flow and rainfall, the filter outputs are largely a weighted combination of the previous K days of flow and rain. This does not account for the delayed impact of rain compared to the impact of previous flow in the unconfined

region, so it contributes negatively to the flow forecast evidenced by the more constant noise throughout the estimates at stations 2321975, 2322500, and 2322800.

4.2.2 Gaussian Process Model

A Gaussian process for forecasting flow at each of the five USGS stations is analyzed next. While the gamma filter produces a weighted sum (average) of past and current inputs, the Gaussian process behavior is defined more by the covariance of the inputs. The squared exponential covariance that I use contributes to a smooth output since it imposes that if the inputs are close together in time, the corresponding outputs should not be dramatically different from each other. Hydrographs of the observed and estimated forecast flow values on the testing period are shown in Figure 4-7 through Figure 4-11 for each of the five stations, along with a plot of the daily error. The performance statistics for each of the filters are summarized in Table 4-5.

Immediately it can be seen from the PBAIS values that there is virtually no bias in the estimates. Looking at the hydrographs, the Gaussian process does estimate the actual peak flow values well where the gamma filter tends to underestimate. The Gaussian process models outperform the trivial predictor everywhere except at 2322800, and outperform the gamma filter everywhere except at 2321000. Indeed, the performance metrics lie in the Very Good range for all locations except 2321000. This is most likely due to the overestimation of the first main peak along with being too persistent in maintaining the high flow. It is still difficult for the model to anticipate peaks, but gives smoother estimates than the gamma filter due to the squared exponential covariance function used. Similar trends in the daily error exist as with the gamma filter in that there is more constant error throughout the unconfined region due

to the non-representative rainfall contribution, and the error is more highly concentrated around the peaks in the confined region.

4.2.3 Hidden Markov Model

Finally a hidden Markov model for forecasting flow at each of the five USGS stations designed and implemented. The explicit dependence of the current state on the previous state makes HMMs well suited for modeling temporal processes, but this includes a stationary assumption of the data. Hydrographs of the observed and estimated forecast flow values on the testing period are shown in Figure 4-12 through Figure 4-16 for each of the five stations, along with a plot of the daily error. The performance statistics for each of the filters is summarized in Table 4-6.

The HMM seems to be reasonable at anticipating an increase in flow. It tends to overestimate the rate of increase of flow, as indicated by the initial negative error at the location of some of the major peaks, particularly at 2321500 and 2322500. The performance metrics for the HMMs in the unconfined region are worse than the trivial predictor, and perform Unsatisfactory at all locations, but with reasonably low bias. The hydrographs show that although the HMM estimate is noisy, the flow trend is more or less captured in the forecast and it models peak flow well with a tendency to overestimate. It remains difficult for the HMM to capture the true nonstationary nature of stream flow.

If the pattern of the noise in the unconfined region forecasts is compared to the clustered rain data in the same region, it becomes clear that rain is causing the noise. Stream flow processes in the unconfined region are driven more by groundwater rather than rainfall, but ground water is a fairly constant level over time. The variations and more immediate impact of rainfall was chosen over groundwater for use in the

forecasting models, but as previously stated, flow in the unconfined region does not respond immediately to rainfall. The HMM in this region is affected negatively by the rain contribution, however, so it is suggested that a different indicative feature of future flow be found. This is also true of the estimates in the unconfined region produced by the gamma model and the Gaussian process model, but the effect of rainfall is dampened by the input weighting (in the gamma model) and the covariance function (in the Gaussian process model).

4.2.3 Seasonal Forecasts

Predictable times of the year for a particular geographic location are prone to more precipitation, and therefore higher flow values, than others. A plot of the normalized mean monthly flow values at the five USGS locations along the Santa Fe River from October 2000 through July 2008 is shown in Figure 4-17. This plot indicates that there are typically higher flows from January through March and July through October, with the remaining months being drier. In an attempt to make flow appear more stationary, the training and testing periods were divided into their “wet seasons” and “dry seasons” and the HMM trained and tested on these times alone. Figure 4-18 is an example of the divisions of flow into its wet and dry seasons at 2321500. As can be seen, most of the higher peaks are captured in the wet season, while lower flow dominates the dry season.

The model performances improved in the wet seasons at all locations except 2321500, and improved in the dry seasons at 2321975 and 2322800. Performance is still mostly Unsatisfactory, even though there was improvement. Performance improved most at 2322500 and 2322800, especially during the wet seasons. It could potentially be worthwhile to run the model on wet and dry data separately in the unconfined region,

but it would be best to use one of the other temporal models entirely. Performance metrics at the five stations for the wet and dry seasons are shown in Table 4-7.

4.3 Spatial Model Prediction

For the spatial models analyzed, true flow was presented at 2321000, 2321500, and 2322500, and prediction was done at the 20 remaining locations along the stream. Out of these 20, true flow is only available at 2321975 and 2322800, so results for these two locations are evaluated in this section.

4.3.1 Bayesian Network

Figure 4-19 and Figure 4-20 show the hydrographs of observed and predicted flow at 2321975 and 2322800, respectively. As with the forecast results, a plot of daily error is also shown. Additionally, the right of each figure has the estimated probability of flow in each flow class for each day. The predicted flow class is the maximum *a posteriori* (MAP) flow class, so the spread of probability values on each day around this MAP value give an indication of how sure or uncertain the model's prediction is.

As was mentioned in Section 2.1, the stream flow process has different characteristics in the confined and unconfined regions. Table 4-8 lists the mean flow value and the autocorrelation length (number of lags for autocorrelation to fall below exp(-1)) for each of the five USGS stations. The correlation length, as is to be expected, continually increases as you move downstream, since there is a component of upstream flow in each location downstream, but increases by approximately 1100 days when moving from 2321975 to 2322500. Additionally, the mean flow value increases from on the order of 10^2 to 10^3 when moving from 2321975 to 2322500. This demonstrates the nonstationarity of flow within the stream.

For both the locations shown, the model has minimal uncertainty, yet performance is subpar at 2321975. The model has very high probability values for the estimates here, yet flow is grossly overestimated. The next downstream station from 2321975 in the network is an evidence node at 2322500. Flow at 2322500 is much higher than at 2321975, and since it is an evidence node it passes very high probability messages to 2321975 for the upper range of its flow classes. The true behavior of flow at 2321975 is somewhat in between that of its upstream station 2321500 in the confined region and its downstream station 2322500 in the unconfined. Were 2322500 more representative of flow at 2321975, performance would be very good. It would seem reasonable that since the baseflow at 2322500 is more or less continually higher than that at 2321975 that this increase in flow would be captured in the transitional pdfs. The WAM data that the pdfs are trained on do not exhibit this true relationship, however, and instead has the baseflow level gradually increase from node to node when moving downstream.

The performance statistics for these locations are shown in Table 4-9. In spite of the overestimated baseflow, 2321975 performs in the Satisfactory range, and not much worse than the spatial trivial predictor. 2322800 performs in the Very Good range and is more indicative of network performance when working with well formed pdfs. The model does not consistently overestimate or underestimate flow, but falls victim to both errors while remaining close to the true trend of flow.

4.3.2 Hidden Markov Model

Figure 4-21 and Figure 4-22 show the hydrographs of observed and predicted flow at 2321975 and 2322800, respectively, as produced by the spatial HMM. In order to make the data appear more stationary when moving from the confined region to the unconfined region, flow at each location was divided by its contributing area as a

preprocessing first step. There are occasional spurts of significant error through the time record at 2321975, but the model tends to overestimate flow in areas of low flow and underestimate the peak flows. Both 2321975 and 2322800 have noisy estimates, but the model performs better at 2322800, where it is in the Very Good performance category, as shown in Table 4-10.

It is interesting to note that although the NSE and RSR values at 2321975 are almost identical to those given by the Bayesian network in Table 4-9, the heavy bias in the Bayesian network result is not present in the HMM estimate. This could be due to the states being uniform across the spatial locations in the HMM due to the Markov assumption, whereas in the Bayesian network a separate transitional pdf is learned between each of the nodes so it is more sensitive to individual pdfs misrepresenting the true data relationships.

4.4 Spatial-Temporal Model Prediction and Forecasting

The Bayesian network for spatially predicting within stream flow and forecasting flow with the contribution of rainfall uses evidence at 2321000, 2321500, and 2322500 on the current day as well as rainfall at two locations. The network is validated at 2321975 and 2322500 on the current day (for spatial prediction) and at all five USGS locations on the forecast day. Plots of spatially predicted current flow at 2321975 and 2322800 are similar to those produced by the purely spatial Bayesian network (Figures 4-19 and 4-20) only over the reduced time period of January 2007 through July 2008, so will not be repeated. Figures 4-23 through 4-27 show hydrographs of the observed and forecasted flow at the five USGS stations, as well as plots of the daily error and the probability beliefs of each day. Table 4-11 contains the performance metrics for the spatial predictions and forecasts.

The performance metrics are in the Satisfactory range at best. This is partly because they are for a much shorter time period (compared to the spatial models) and so each error is more significant in the total performance metric. 2321975 in particular is very poor, and this is due to the ill-formed pdf discussed in the spatial Bayesian network analysis in Section 4.3.1.

Forecasts at the remaining four stations are quite reasonable, performing as high as the Very Good range at 2321500. 2322500 and 2322800 are able to estimate the high flow peak in March 2008, but also give noisy estimates during the rest of the period. Again, comparing the noise pattern in the unconfined region forecasts to clustered rain shows that rainfall is causing the oscillatory forecast error. The BN would benefit from using an alternative feature to rainfall in the unconfined region to help predict future flow, as with the temporal HMM.

The confidence of estimates on the forecast day is dramatically decreased when compared to the spatial estimations in Figures 4-19 and 4-20. No evidence is provided on the forecast day and there is more uncertainty in the probability beliefs that are passed through the network as they move further away from evidence nodes. The transitional probabilities for forecasting flow are also less deterministic than those for spatially predicting flow. In spatial prediction, all estimation is done on the same day and there are limited possibilities of flow transitions moving along the stream due to the high correlation of within stream flow. When forming a joint pdf of current and future flow (and rainfall), however, the probability distribution is more spread out amongst the classes since flow can just as easily increase, decrease, or stay the same. These pdfs are static in time, so the forecasting is not as good as those models which rely on a

moving window of past flow to produce forecasts, such as the gamma filter or even the trivial temporal predictor. The spatial-temporal Bayesian network was also run over the "wet" and "dry" time periods as was done with the HMM in Section 4.2.3, but results were not significantly improved over the fully sampled Bayesian network. Specifically, the performance metrics for forecasts at 2321000 and 2322800 were improved slightly (NSE values of 0.583 and 0.407, respectively for the combined wet/dry periods) but were degraded for all other locations.

As was mentioned in Section 3.3.2, modifications could be made to this Bayesian network to more explicitly represent the nonstationary nature of the flow time series at each location. Given the poor results and low confidence of the network's forecasts, it would be reasonable to expect some improvement if more temporal information were added. Even including one additional day would provide more certainty in what the future flow class should be, since there would be fewer classes with high probability resulting from a 2 day sequence of observations than from a single day's observation. The more days in the sequence included, however, the more specific the transitional probabilities become and more data is needed to accurately represent the true temporal behavior of flow.

4.5 Model Uncertainty

Knowing the uncertainty of a model's estimates is just as important as knowing the estimates themselves. It is desirable that a model will always perform well, but the user needs to know when it can be trusted and when it cannot. Information can be gained from a poorly performing model if it knows when its estimates are good and when they are bad. This can be quantified by the notion of predicted error, which is discussed in Section 4.4.1.

Another way a model's uncertainty of its estimates can be useful is to indicate where new observations need to be made, if able. The 95% confidence interval associated with the spatial Bayesian network predictions are utilized in Section 4.4.2 to suggest locations for new sensors.

4.5.1 Predicted Error

One of the main benefits of using a probabilistic framework such as the Bayesian network is that the model may provide statistics of how confident it is of its predictions. Using Parzen windowing with a Gaussian kernel, the pdf of the actual error of the spatial-temporal Bayesian network estimates were computed, along with their predicted error pdf. Estimated flow is taken to be the maximum *a posteriori* (MAP) class, and so error is simply $\sum_{i \neq MAP} f_i P_i$, where f_i is the flow value of class i and P_i is the probability of class i. In other words, it is the probability of flow not being the MAP flow class, converted to actual flow values. Figures 4-28 and 4-29 show plots of the actual and predicted error pdfs at all five USGS locations on the current day and the forecast day, respectively. The Cauchy-Schwartz divergence between the pdfs was calculated and is given in Table 4-12. The pdfs coincide exactly at 2321000, 2321500, and 2322500 on the current day since they were included as evidence in the network. There is a slight divergence between the pdfs at 2321975, with actual error being higher than expected. This is due to the high probability values of the estimates passed to 2321975 from evidence node 2322500, whose flow is much higher than that at 2321975.

There is also a non-zero divergence at 2322800, this time with the predicted error being larger than the actual error. Figure 4-29 shows that this is also the case at 2322500 and 2322800 on the forecast day. The model is performing better than it

expects to. This is due to the higher class values in the unconfined region than in the confined region. There is a larger baseflow in the unconfined region, so even low flow classes have high flow values. Due to the way predicted error is calculated, if there is any non-zero probability in the non-MAP flow class at these locations, the high flow class values will artificially increase the predicted error. The predicted and actual flow pdfs coincide nicely at 2321000, 2321500, and 2321975 on the forecast day. Even though error is higher at 2321975, the model mostly expects this due to its low confidence in the estimates. The peak around 800 ft^3/s in the actual flow error pdf at 2321975 is due to a combination of the noise in the estimates and the flow class levels. As can be seen in Figure 4-25, the flow estimate oscillates between ~500 and ~800 ft^3/s. These are actually adjacent class values, so when the model is not very confident and has low MAP values in these two classes, the range of actual error increases, but the predicted error does not.

In general, the spatial-temporal Bayesian network can give a good indication of where it is uncertain in the confined region, but a different approach may be needed to achieve accurate indications in the unconfined region.

4.5.2 Spatial Uncertainty

In situ sensors are often costly and funds are often limited. If a new sensor needs to be placed along the stream, it is important to place it where it will be most beneficial to the user and provide the most information. The Bayesian methods in this work are able to provide a measure of uncertainty of their estimates along with the actual flow estimates. This allows the user to see where along the stream the model is most uncertain of its predictions and suggest a reasonable location for a new sensor. To demonstrate, the spatial Bayesian network was implemented first with three evidence

nodes (2321000, 2321500, and 2322500) and then again with four evidence nodes (2321000, 2321500, 2322500, and 2322800). In the first scenario, the model consistently has high uncertainty in its forecasts at 2322800. Evidence was added at 2322800 to see the effect a new sensor could have on its surrounding locations along the stream. Adding evidence, or a sensor, at this node both reduces uncertainty at 2322800 as well as uncertainty at the nodes between sensors (locations between 2322500 and 2322800). Figure 4-30 shows plots of the estimates at spatial locations along the stream for a single day (12/30/2003) along with the 95% confidence interval. On the left are estimates using three evidence nodes and on the right are estimates using four evidence nodes. Specifically, it can be seen that the upper and lower bounds of the confidence interval at nodes between 2322500 and 232280 are tightened (narrower confidence interval) when a sensor is added at 2322800.

Table 4-1. General performance ratings for recommended statistics for a monthly time step

	NSE	RSR	PBIAS (%)
Very Good	0.75 < NSE ≤ 1.00	0.00 ≤ RSR ≤ 0.50	PBIAS < ±10
Good	0.65 < NSE ≤ 0.75	0.50 < RSR ≤ 0.60	±10 ≤ PBIAS < ±15
Satisfactory	0.50 < NSE ≤ 0.65	0.60 < RSR ≤ 0.70	±15 ≤ PBIAS < ±25
Unsatisfactory	NSE ≤ 0.50	RSR > 0.70	PBIAS ≥ ±25

Table 4-2. Performance metrics for temporal trivial model

	2321000	2321500	2321975	2322500	2322800
NSE	0.380	0.587	0.846	0.915	0.975
RSR	0.787	0.643	0.392	0.291	0.157
PBIAS	1.60	0.42	-0.52	-0.12	0.88

Table 4-3. Performance metrics for spatial trivial model

	2321975	2322800
NSE	0.598	0.678
RSR	0.634	0.567
PBIAS	38.72	30.12

Table 4-4. Performance metrics for temporal gamma filter models

	2321000	2321500	2321975	2322500	2322800
NSE	0.542	0.805	0.795	0.903	0.823
RSR	0.677	0.442	0.453	0.312	0.421
PBIAS	21.40	23.53	13.77	0.52	0.86

Table 4-5. Performance metrics for temporal Gaussian process models

	2321000	2321500	2321975	2322500	2322800
NSE	0.452	0.814	0.855	0.933	0.856
RSR	0.740	0.431	0.380	0.258	0.380
PBIAS	-2.78	3.38	-0.43	-0.02	3.69

Table 4-6. Performance metrics for temporal hidden Markov models

	2321000	2321500	2321975	2322500	2322800
NSE	-0.182	-0.030	-0.728	0.012	-1.863
RSR	1.087	1.015	1.315	0.994	1.692
PBIAS	-9.64	-19.91	-13.42	-7.65	-34.04

Table 4-7. Performance metrics for temporal hidden Markov models designed for wet and dry seasons separately

	2321000	2321500	2321975	2322500	2322800
NSE – wet	0.242	-0.281	0.015	0.683	0.692
NSE – dry	-0.325	-2.505	0.283	-1.725	0.527
RSR – wet	0.871	1.132	0.992	0.564	0.555
RSR – dry	1.151	1.872	0.847	1.651	0.688
PBIAS – wet	0.43	0.77	-7.87	-3.36	-7.01
PBIAS – dry	-6.30	-41.26	-10.16	-7.77	-1.15

Table 4-8. Mean and autocorrelation length of flow at USGS stations

	2321000	2321500	2321975	2322500	2322800
Mean Flow (ft^3/s)	114	260	424	1062	1125
Autocorrelation length (days)	9	27	64	1170	793

Table 4-9. Performance metrics for spatial Bayesian network

	2321975	2322800
NSE	0.531	0.821
RSR	0.685	0.423
PBIAS	-101.12	2.54

Table 4-10. Performance metrics for spatial hidden Markov model

	2321975	2322800
NSE	0.552	0.713
RSR	0.669	0.536
PBIAS	0.90	7.82

Table 4-11. Performance metrics for spatial-temporal Bayesian network

	Current 2321975	Current 2322800	Forecast 2321000	Forecast 2321500	Forecast 2321975	Forecast 2322500	Forecast 2322800
NSE	-27.201	0.640	0.355	0.758	-45.26	0.581	0.385
RSR	5.310	0.600	0.803	0.492	6.801	0.648	0.784
PBIAS	-328.90	8.20	14.25	1.80	-496.25	-1.87	3.26

Table 4-12. Cauchy-Schwartz divergence of actual and predicted error pdfs for spatial-temporal Bayesian network

	2321000	2321500	2321975	2322500	2322800
Current Day	0	0	5.07	0	36.98
Forecast Day	0.03	0.02	0.20	12.33	6.74

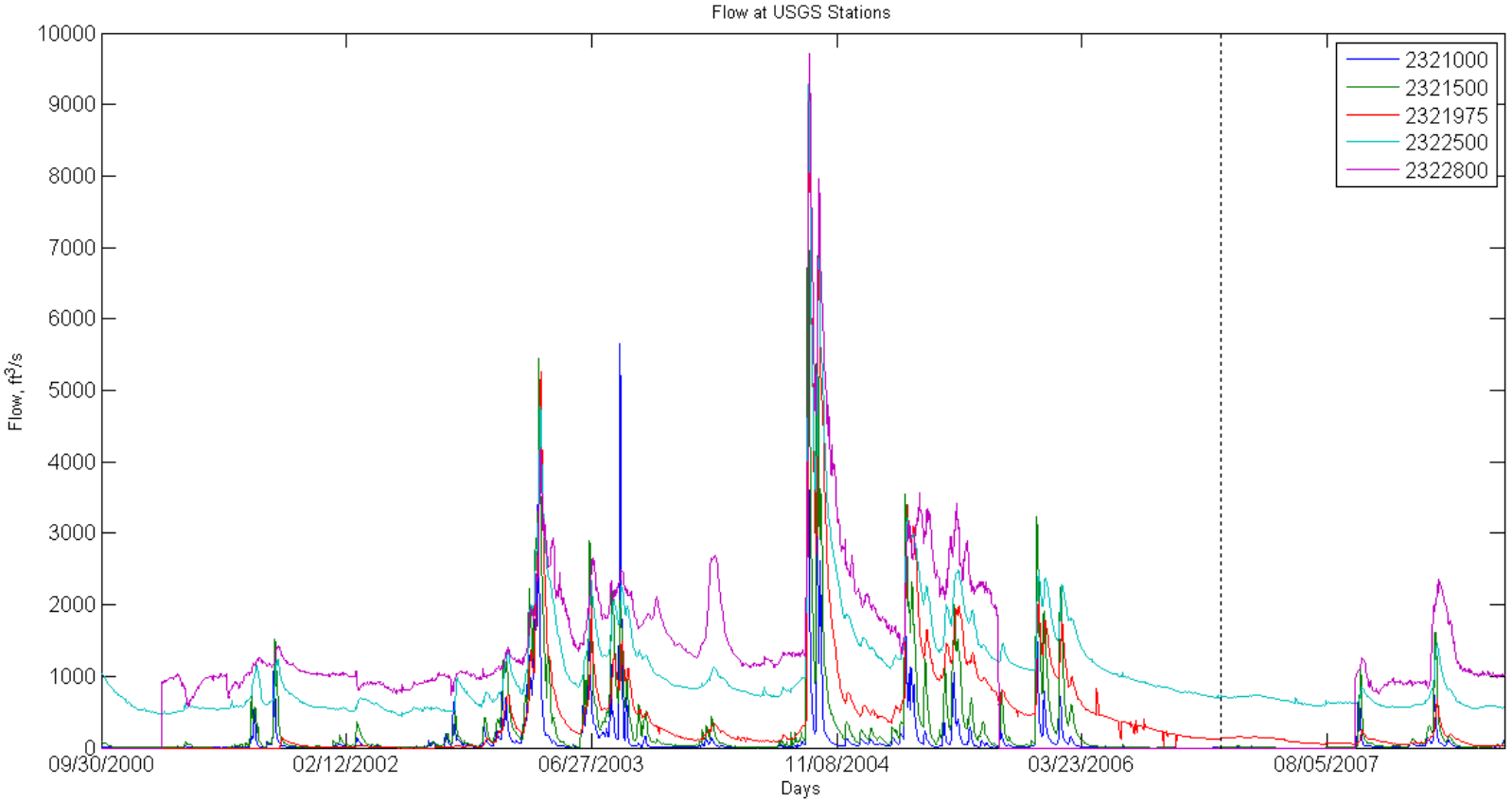

Figure 4-1. Flow at USGS stations from October 2000 through July 2008.

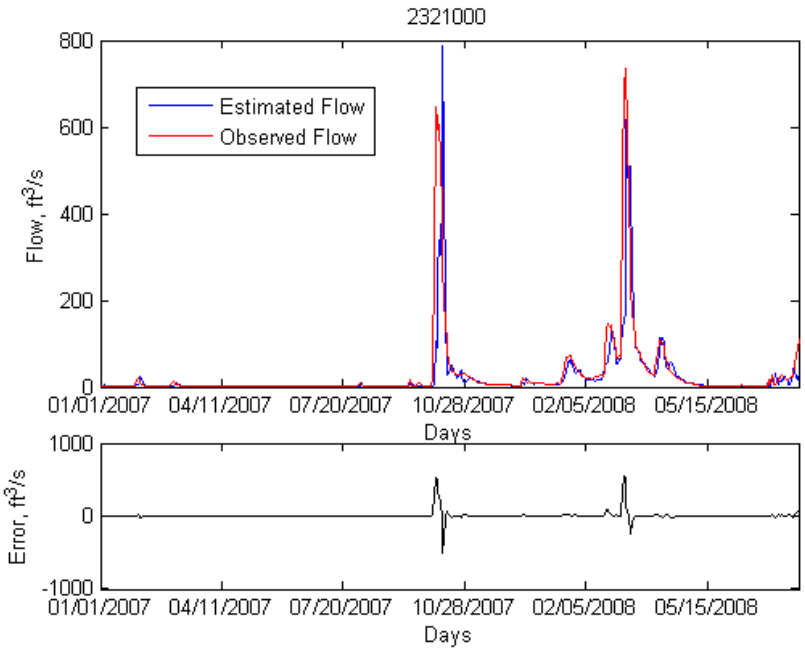

Figure 4-2. Gamma Filter forecasted and observed flow, along with daily error at 2321000.

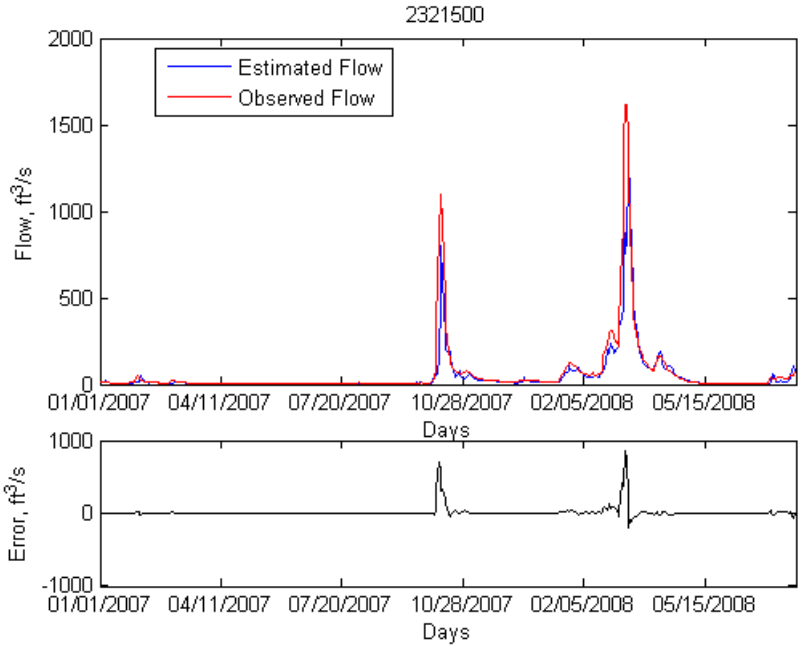

Figure 4-3. Gamma Filter forecasted and observed flow, along with daily error at 2321500.

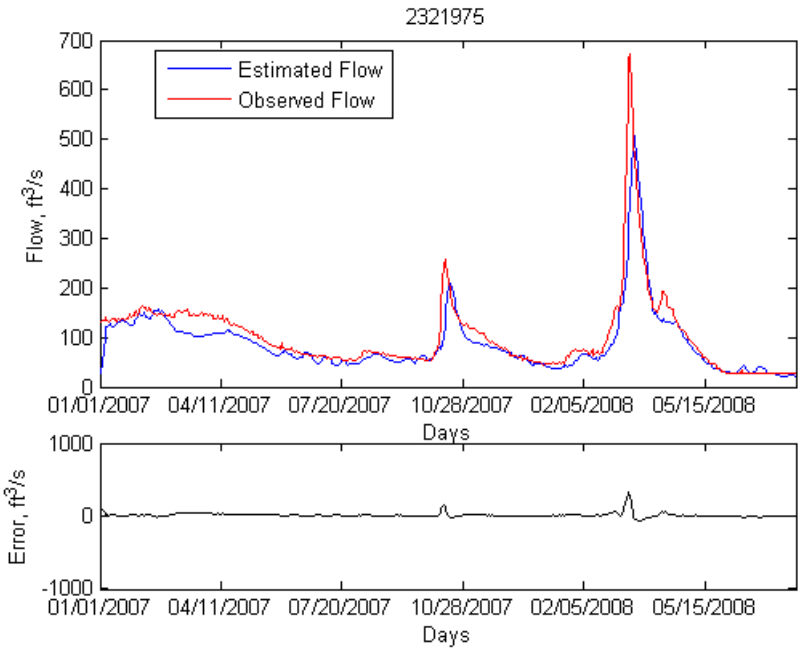

Figure 4-4. Gamma Filter forecasted and observed flow, along with daily error at 2321975.

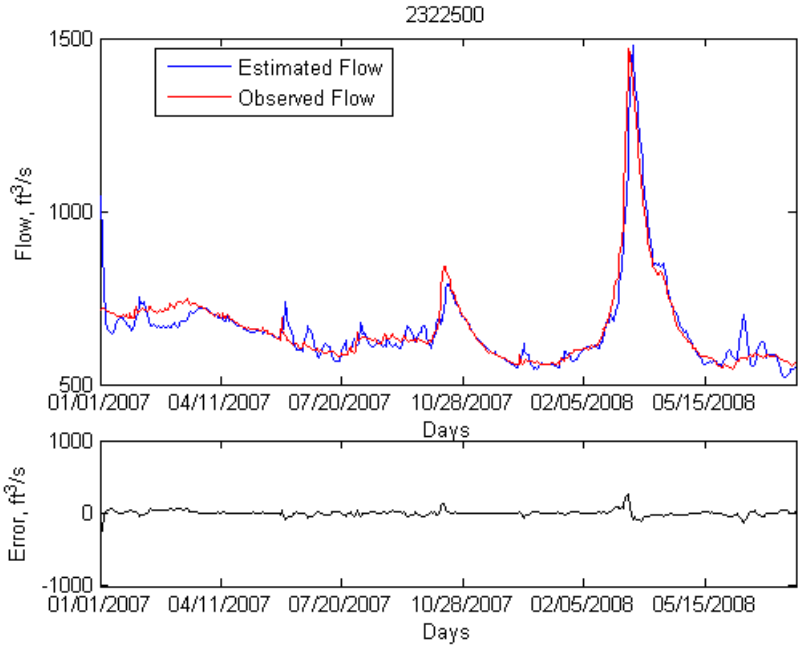

Figure 4-5. Gamma Filter forecasted and observed flow, along with daily error at 2322500.

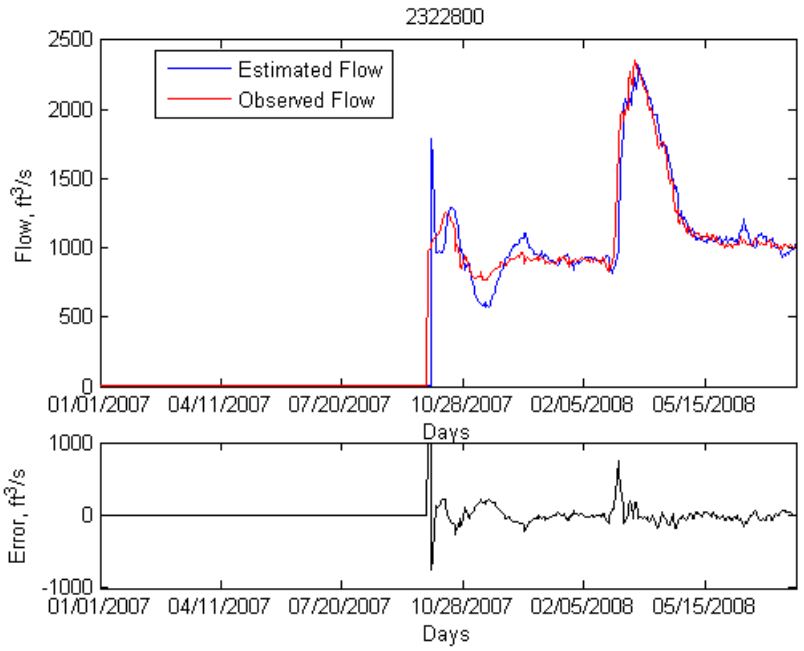

Figure 4-6. Gamma Filter forecasted and observed flow, along with daily error at 2322800.

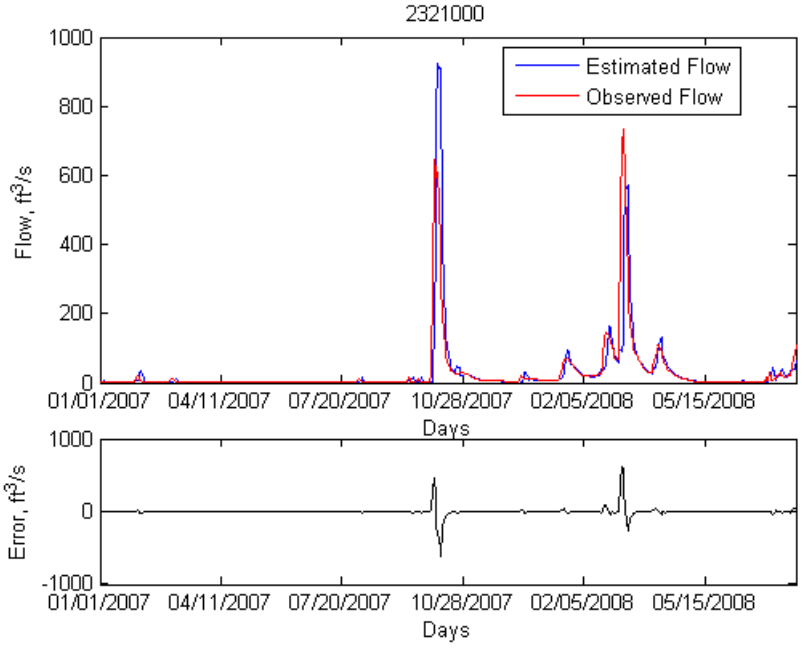

Figure 4-7. Gaussian process forecasted and observed flow, along with daily error at 2321000.

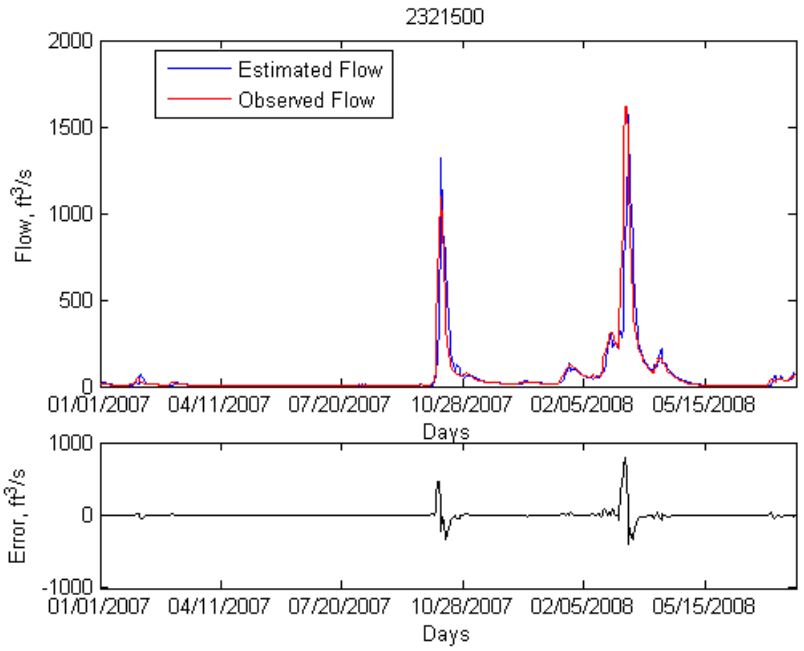

Figure 4-8. Gaussian process forecasted and observed flow, along with daily error at 2321500.

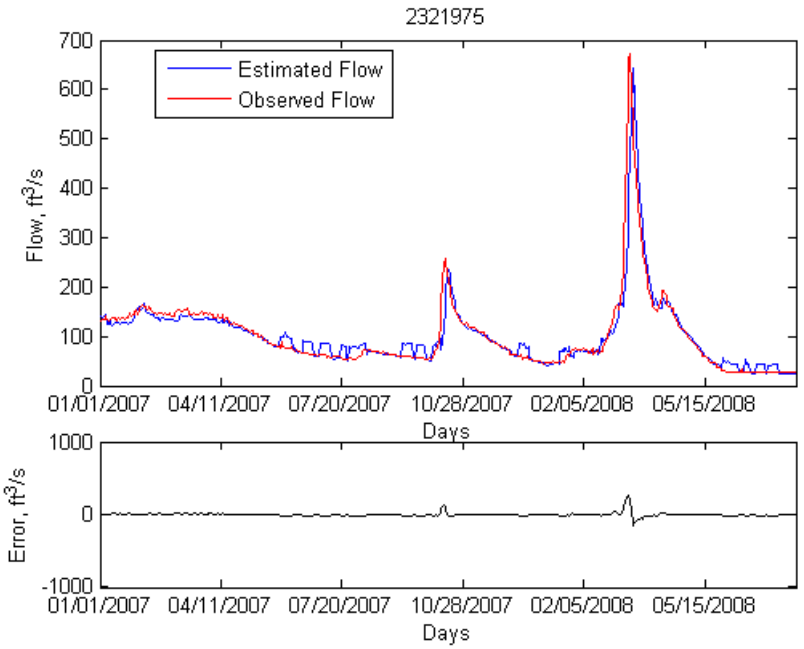

Figure 4-9. Gaussian process forecasted and observed flow, along with daily error at 2321975.

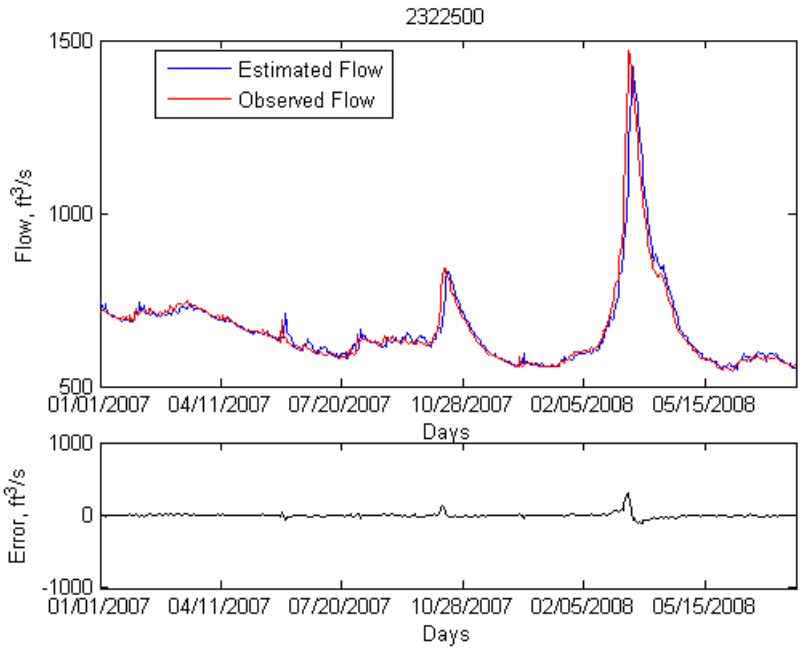

Figure 4-10. Gaussian process forecasted and observed flow, along with daily error at 2322500.

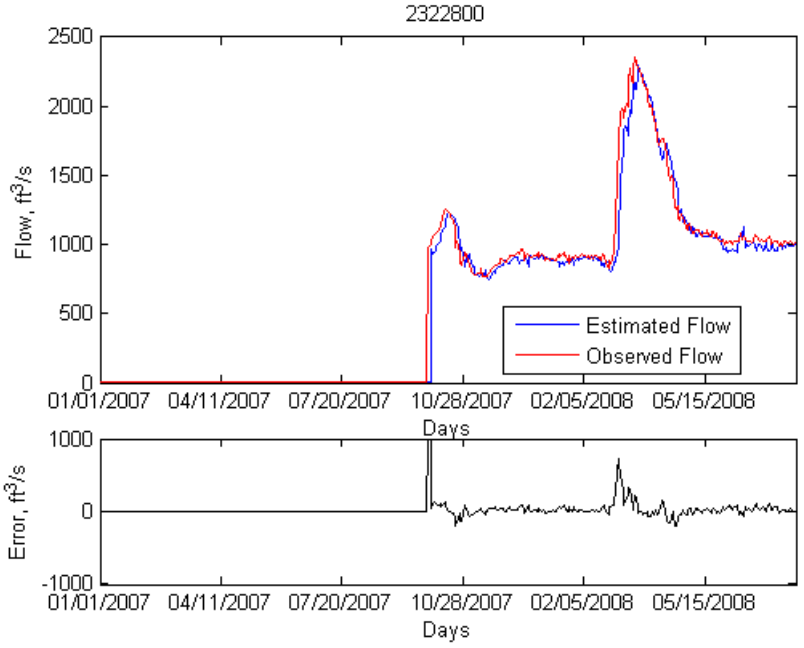

Figure 4-11. Gaussian process forecasted and observed flow, along with daily error at 2322800.

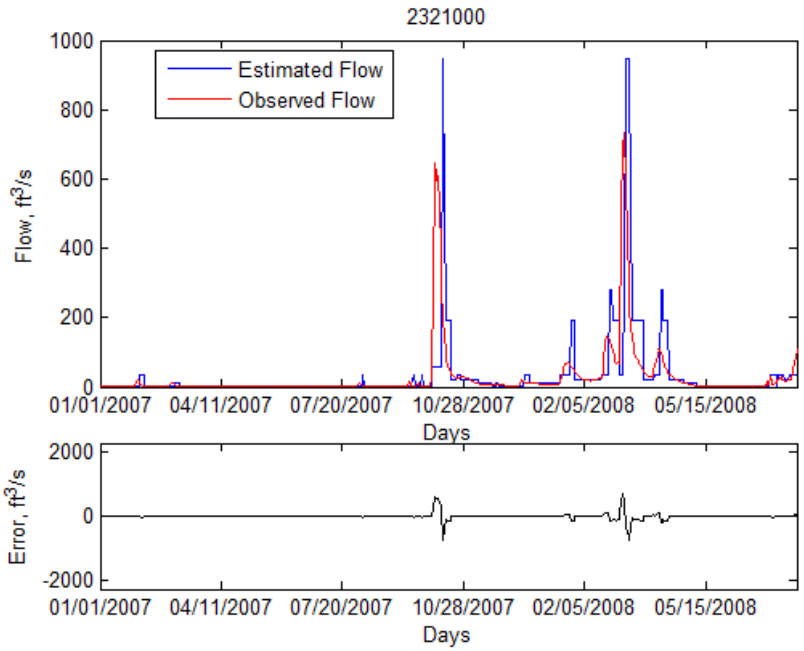

Figure 4-12. HMM forecasted and observed flow, along with daily error at 2321000.

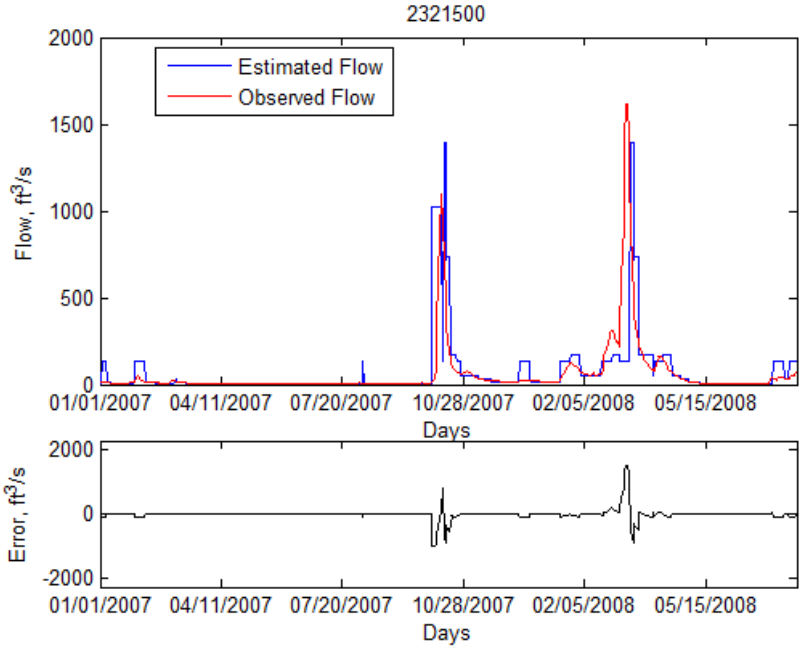

Figure 4-13. HMM forecasted and observed flow, along with daily error at 2321500.

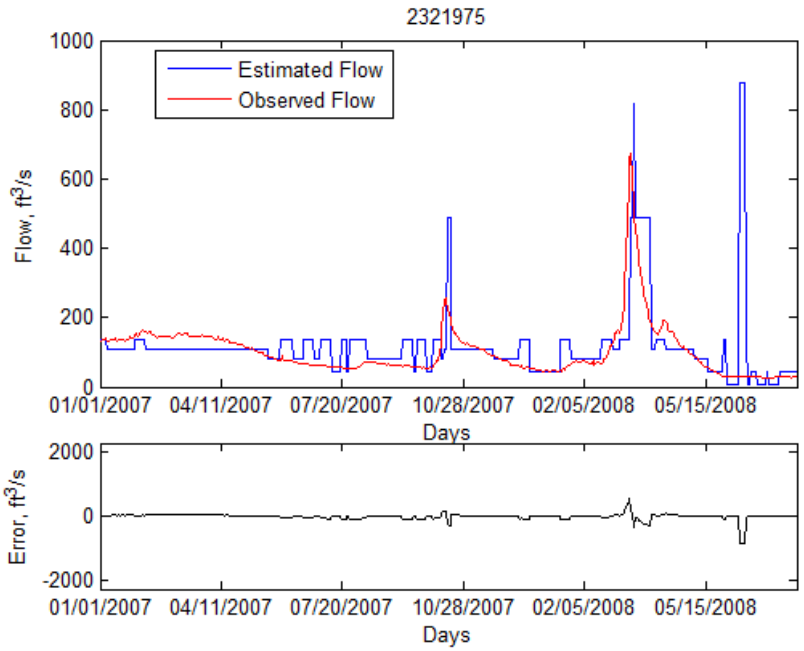

Figure 4-14. HMM forecasted and observed flow, along with daily error at 2321975.

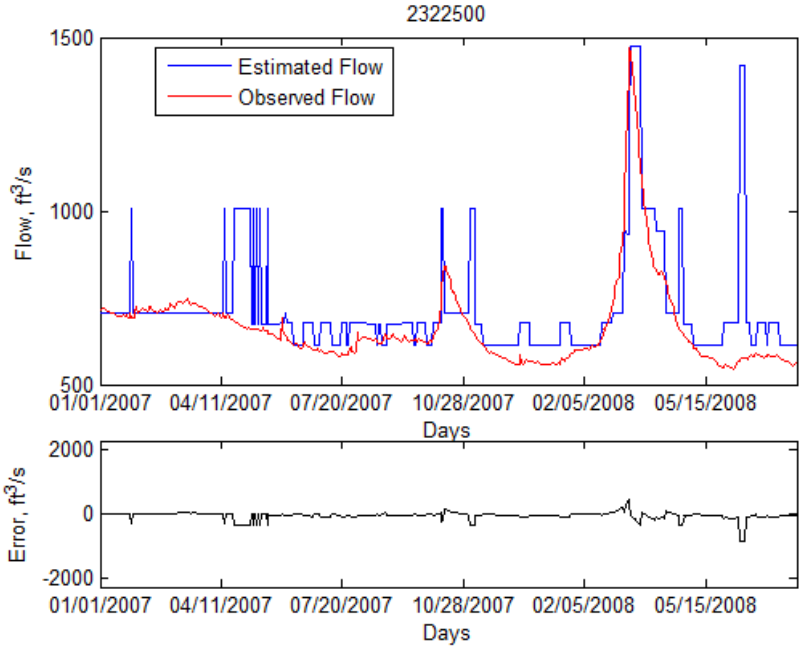

Figure 4-15. HMM forecasted and observed flow, along with daily error at 2322500.

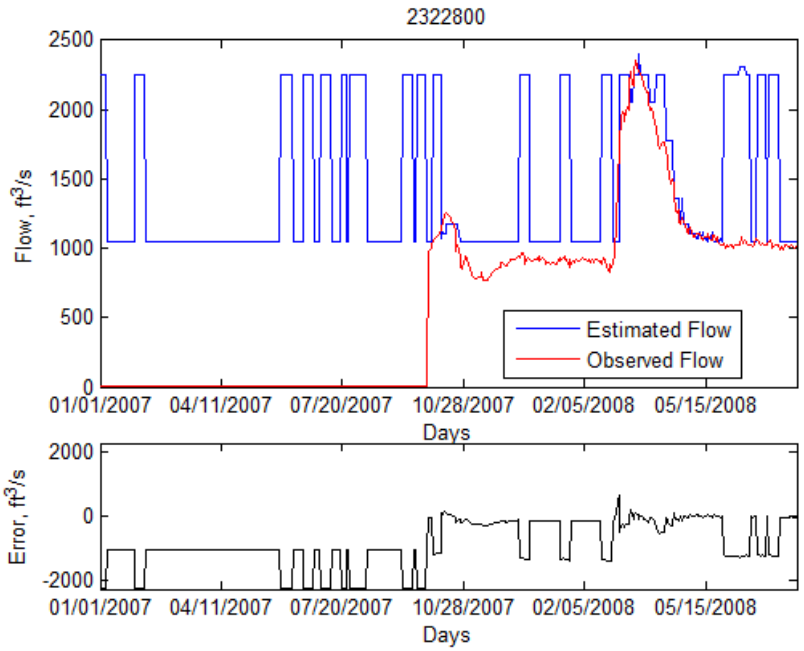

Figure 4-16. HMM forecasted and observed flow, along with daily error at 2322800.

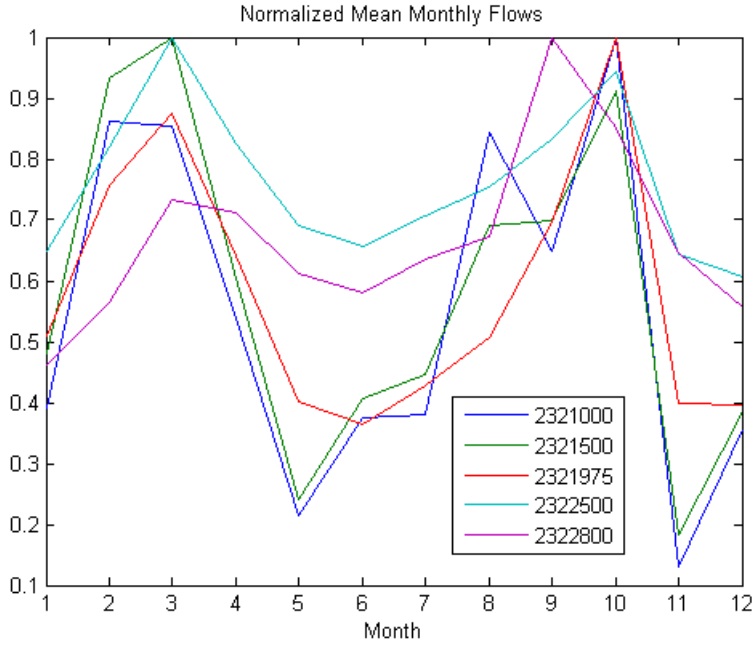

Figure 4-17. Normalized mean monthly flow at five USGS stations.

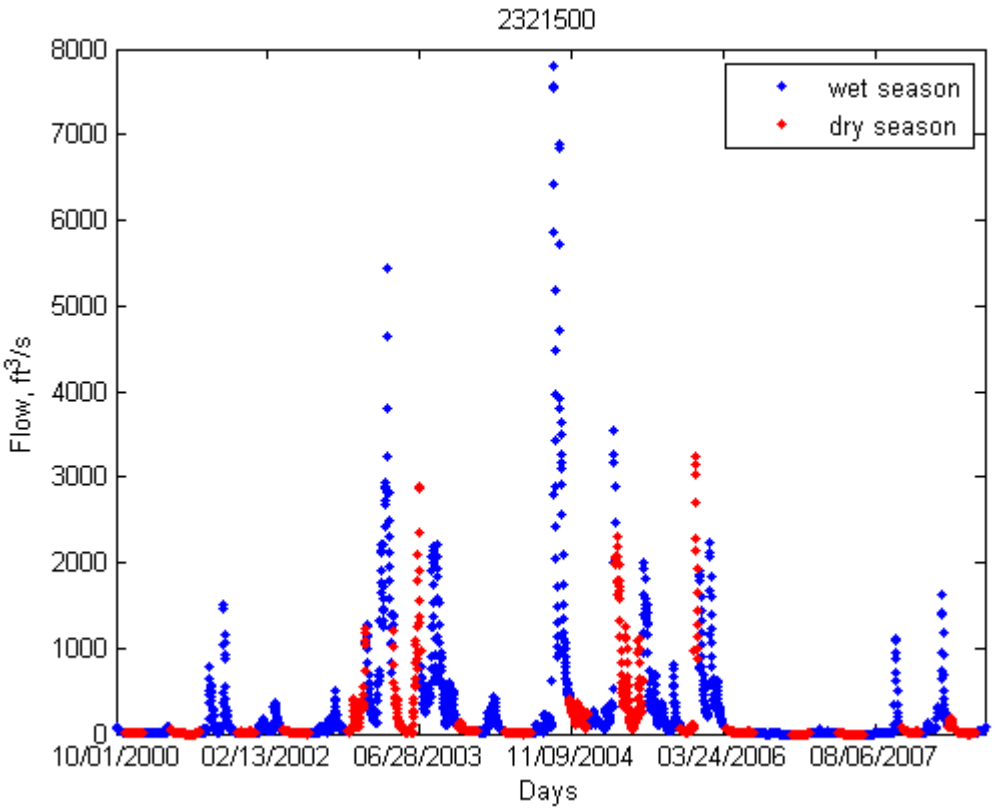

Figure 4-18. Flow at 2321500 divided into wet and dry seasons.

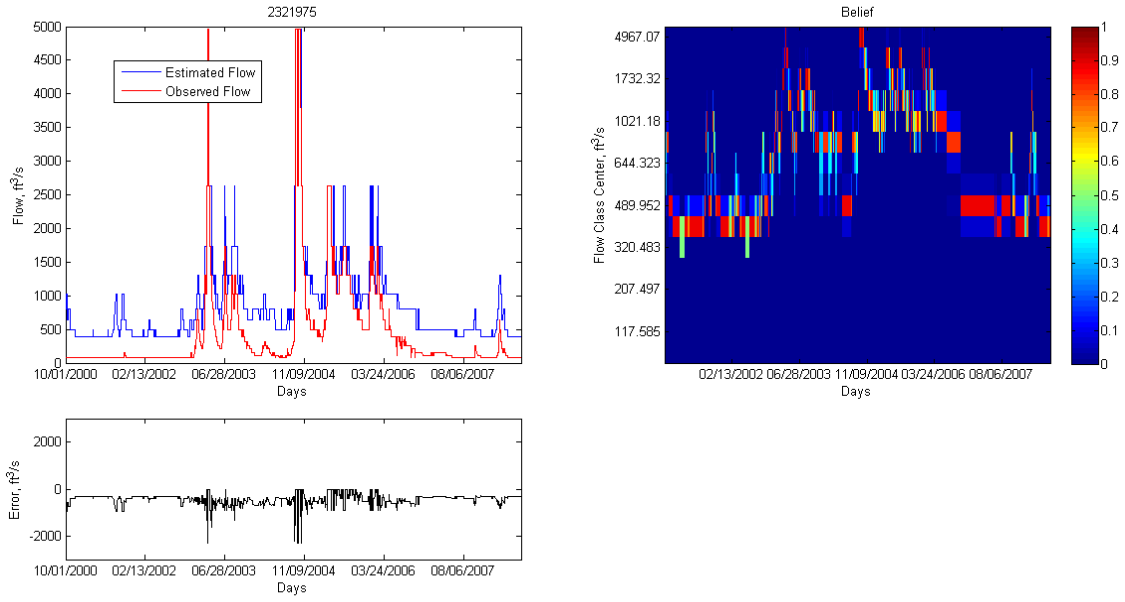

Figure 4-19. Bayesian network spatial predicted and observed flow, along with daily error and associated probability belief at 2321975.

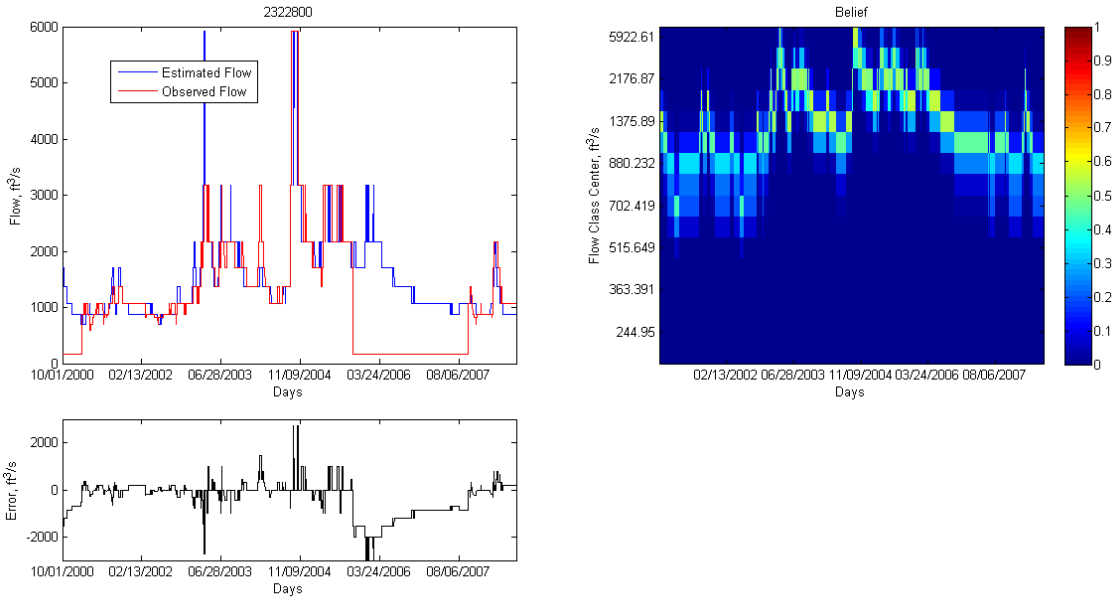

Figure 4-20. Bayesian network spatial predicted and observed flow, along with daily error and associated probability belief at 2322800.

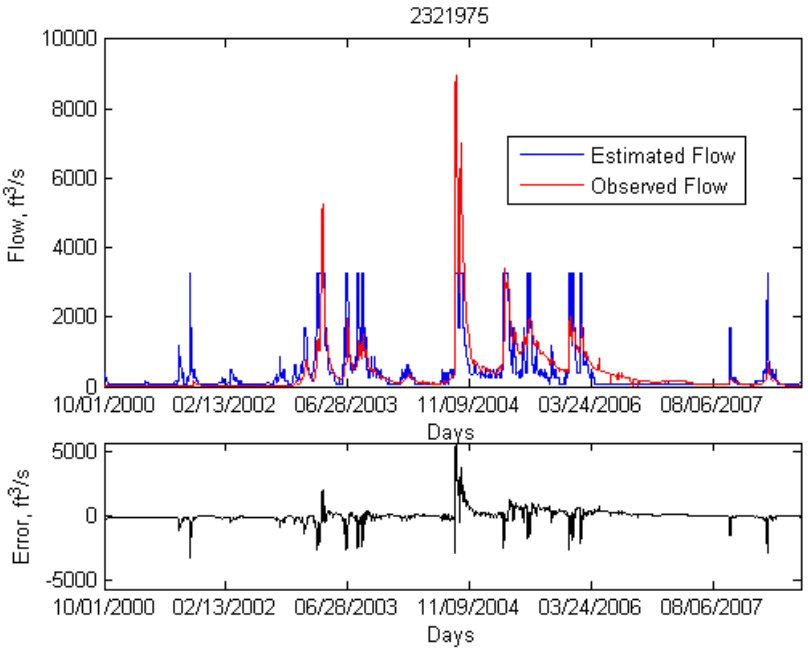

Figure 4-21. Hidden Markov model spatial predicted and observed flow, along with daily error at 2321975.

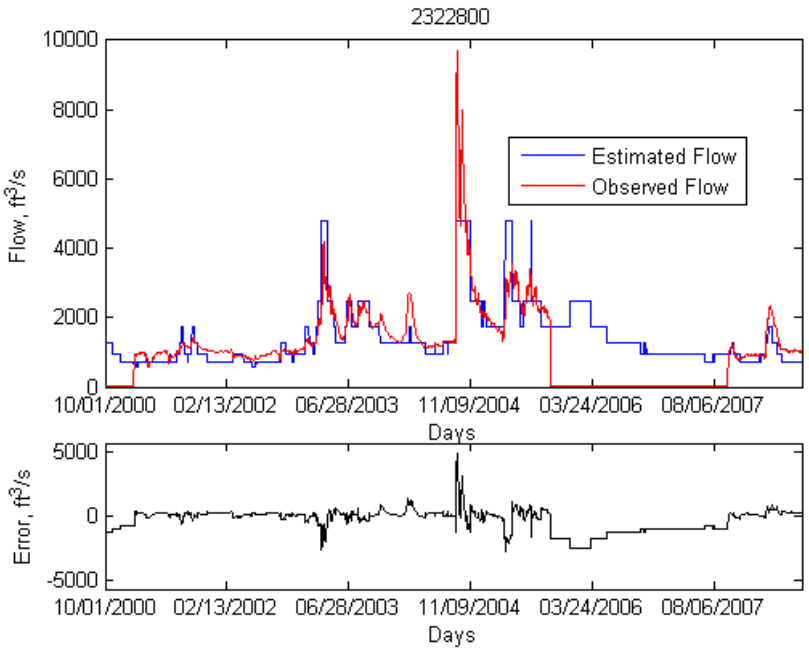

Figure 4-22. Hidden Markov model spatial predicted and observed flow, along with daily error at 2322800.

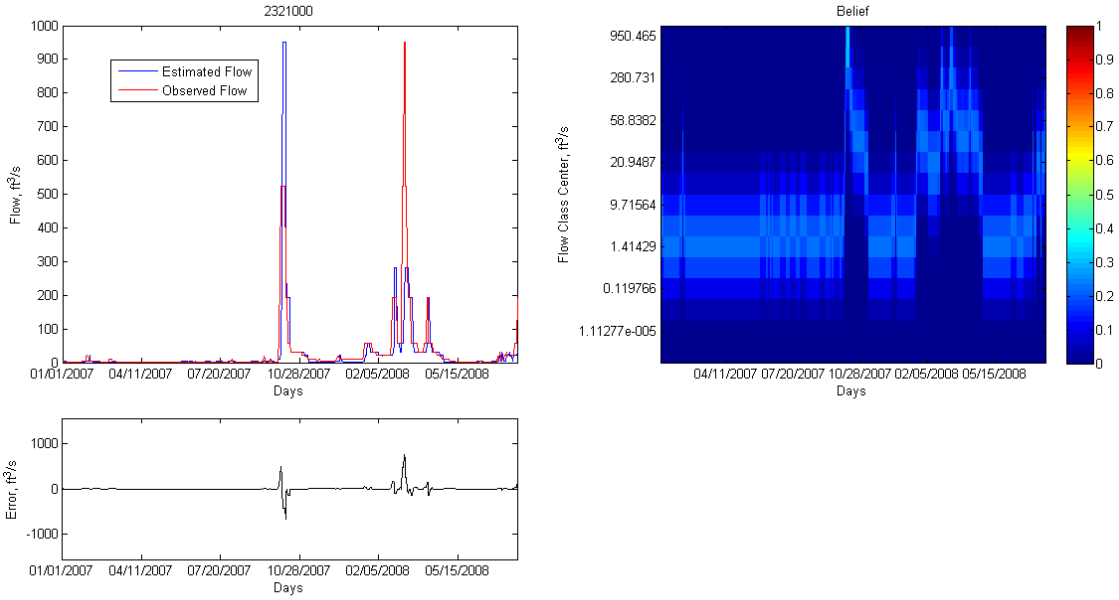

Figure 4-23. Bayesian network forecasted and observed flow, along with daily error and associated probability belief at 2321000.

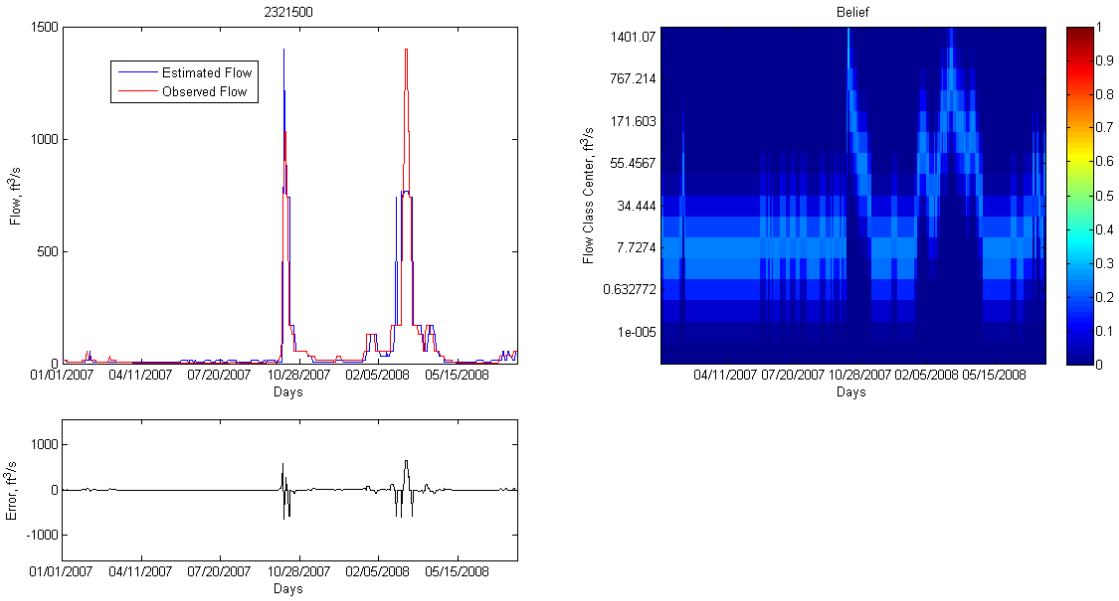

Figure 4-24. Bayesian network forecasted and observed flow, along with daily error and associated probability belief at 2321500.

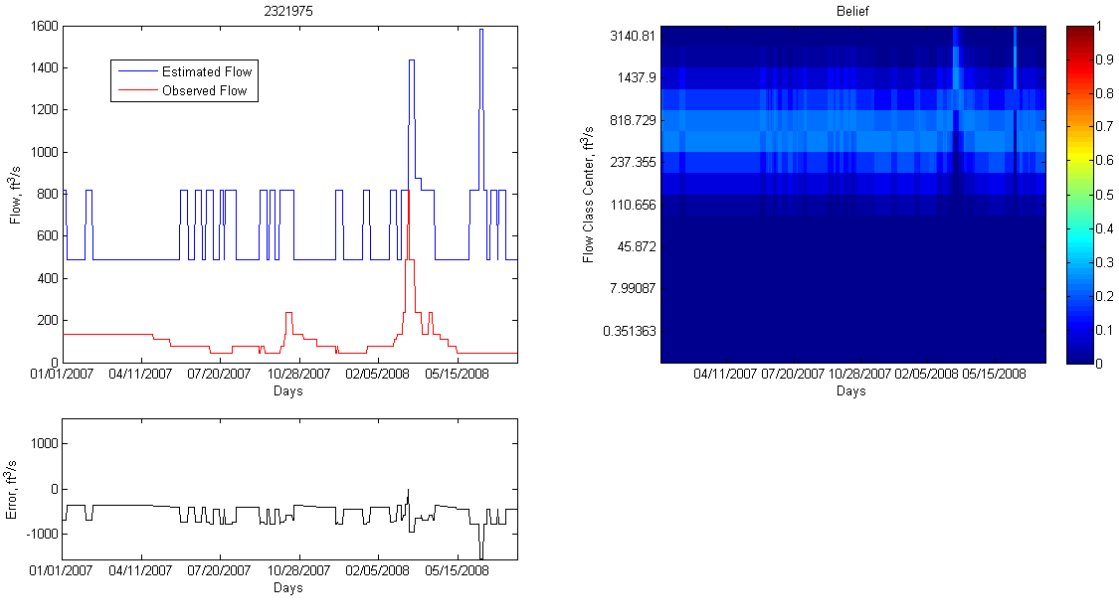

Figure 4-25. Bayesian network forecasted and observed flow, along with daily error and associated probability belief at 2321975.

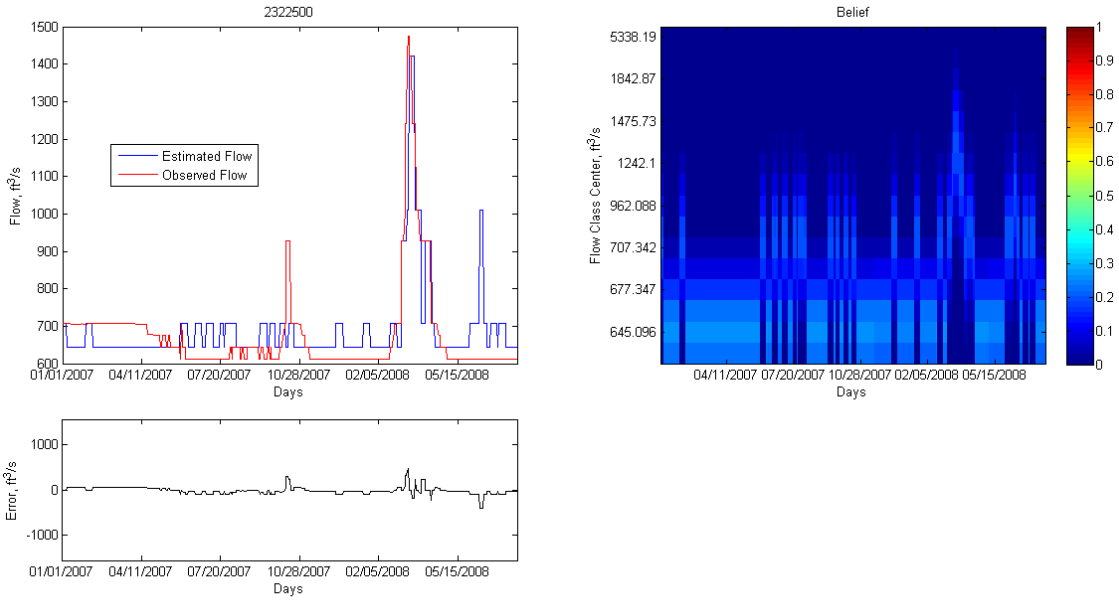

Figure 4-26. Bayesian network forecasted and observed flow, along with daily error and associated probability belief at 2322500.

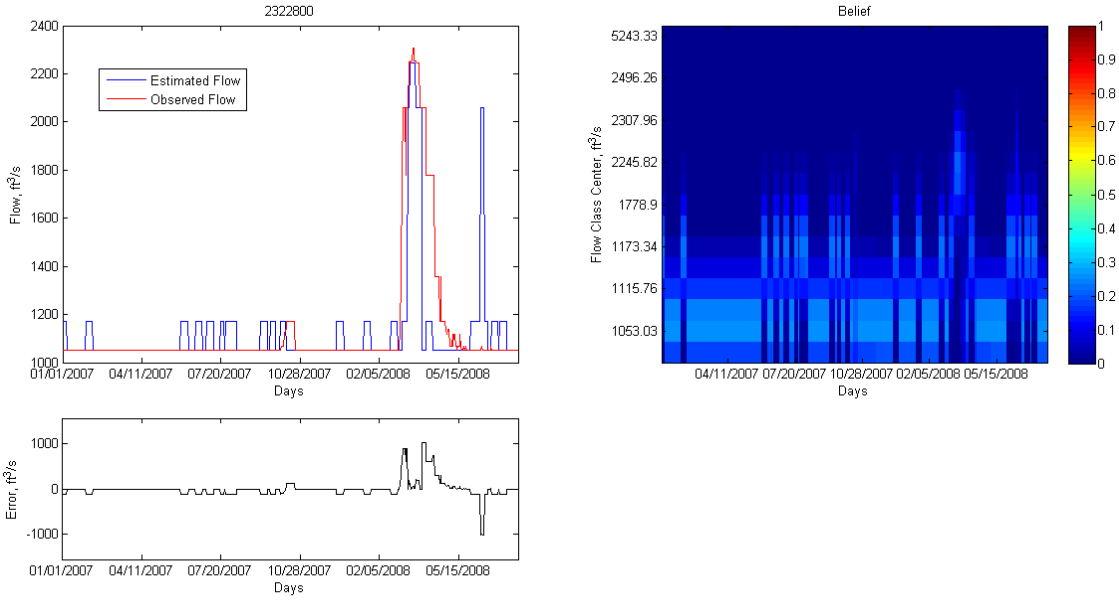

Figure 4-27. Bayesian network forecasted and observed flow, along with daily error and associated probability belief at 2322800.

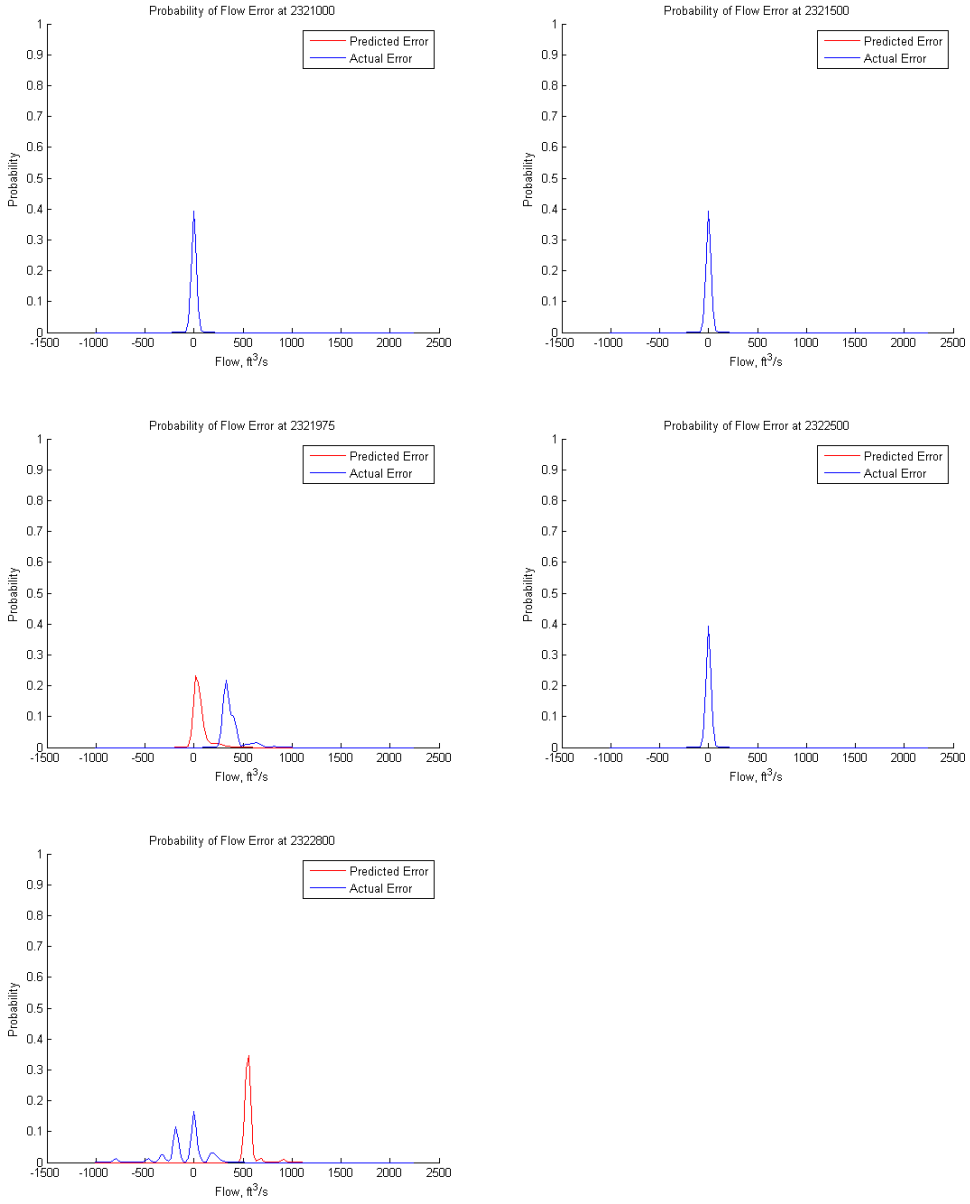

Figure 4-28. Predicted error and actual error pdfs for spatial-temporal Bayesian network on current day.

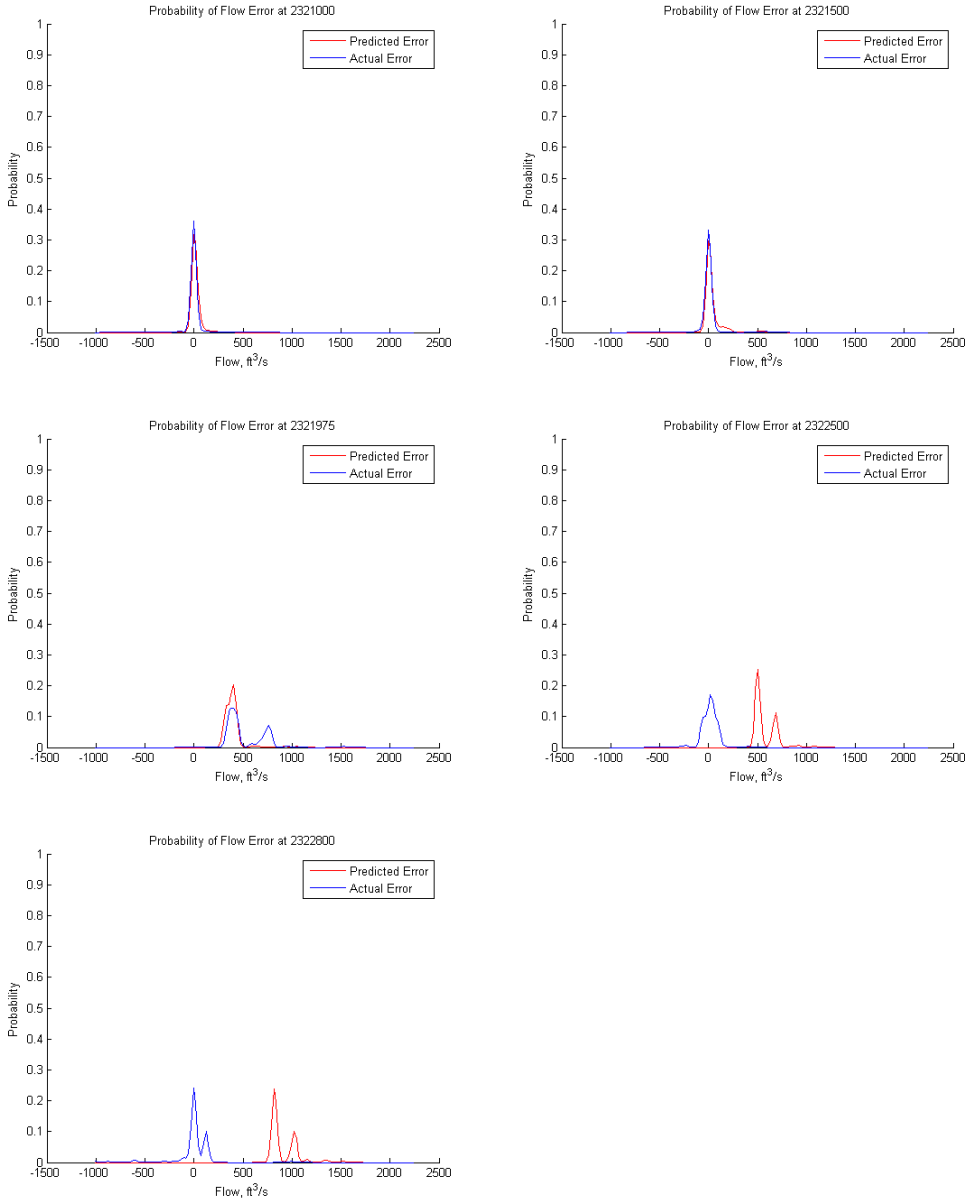

Figure 4-29. Predicted error and actual error pdfs for spatial-temporal Bayesian network on forecast day.

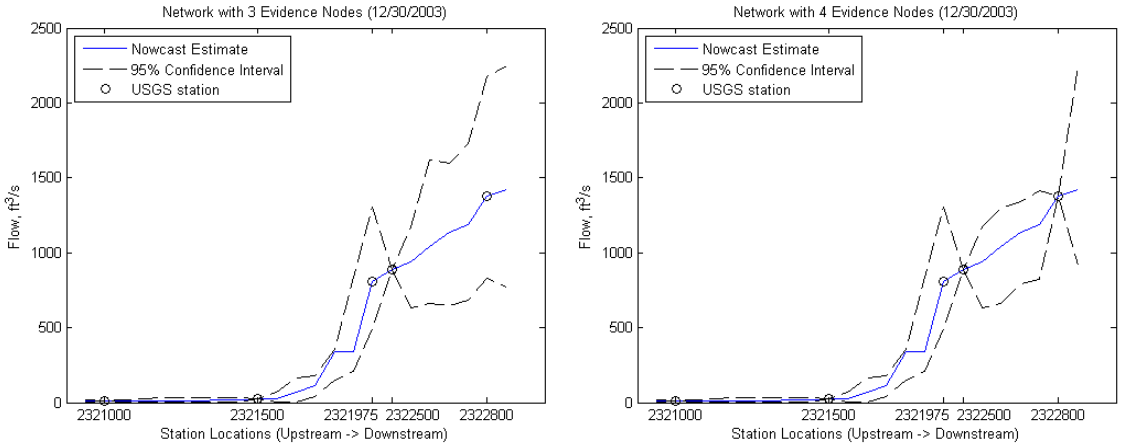

Figure 4-30. Spatial plots of forecast estimates on 12/30/2003) using three evidence nodes (left) and four evidence nodes (right). Dashed lines indicate 95% confidence interval.

CHAPTER 5
CONCLUSIONS AND FUTURE WORK

The primary goal of this work was to investigate a variety of models to approximate the joint space of modeling spatial and temporal flow within the Santa Fe basin. The models analyzed addressed portions of the larger problem, making assumptions that were not necessarily true of the flow process. The models were designed for estimating flow temporally (imposing spatial independence), predicting flow spatially (assuming temporal stationarity), and combined spatial-temporal estimation (assuming stationarity of flow on forecast day conditioned on flow today). A Gamma filter model, Gaussian process model, and hidden Markov model addressed the temporal forecasting problem. A Bayesian network and hidden Markov modeled were developed for predicting within stream flow, and another Bayesian network was designed to both predict within stream flow and forecast flow.

Graphical models including the hidden Markov model and Bayesian network model are used to estimate and predict stream flow in a watershed system. In order to take full advantage of sparsely available data, a framework has been implemented utilizing Pearl's message passing algorithm to improve inference at unobserved nodes in the Bayesian network, and flow predictions from a previously calibrated physical model have been incorporated for training data. Confidence of estimates based on the probability of estimated values versus the spread of probability associated with other flow values were computed and indicate how reliable the model estimates are. Additionally, a method for determining the most informative feature set has been analyzed and can aid in building the Bayesian network if a limited number of features is to be used. Spatial-temporal graphical models can be useful if the relationships

between the nodes can be accurately captured in the transitional probability distribution functions, but if not, an alternate method should be relied on.

The value of a probabilistic model such as the Bayesian network lies in the uncertainty measures included with the predictions. The model generates temporal predicted error pdfs at each node indicating how confident it is in its estimates (what the probability of the MAP value is versus the probability of all the other classes). Confidence intervals around the model estimates can be calculated and used to suggest potential site for new sensors along the stream.

Several options are available for potentially improving the Bayesian inference results in this work. If enough evidence nodes are available, larger Bayesian networks may be constructed for estimation over larger watersheds. If evidence is too sparse, the network can suggest sites for additional sensor deployment based on areas of largest uncertainty in the predictions. Rather than basing the transitional probabilities in the Bayesian network on estimates from another model, it would be more logical to base them on true measurement data along the stream. This can be done and would potentially improve the forecast estimates, but the number of spatial nodes in the network would be limited to only the observation locations, since there would be no information on which to model the relationship between intermediate nodes. This eliminates half of the usefulness of the BN presented in this work in that it would not be able to do spatial prediction.

The inputs to the Bayesian network could be modified to more explicitly represent the temporal relationship of flow. If the temporal dependencies are fully captured, the Bayesian network could closely approximate the joint space of flow in the basin. Using

clustering of flow today conditioned on a past day's flow was discussed as an option, but a mapping of the classes back to actual flow values needs to be determined for the estimates to be useful. This problem can be alleviated by implementing dynamical system methods of encoding the temporal dependencies in the transformed data and using proven methods to map the outputs to the original problem space. The dimensionality of the data (number of samples needed) is too high, however, there may not be enough data to accurately model the conditional probabilities of the BN. Alternatively, the topology of the network can be changed so as to include an additional layer of past flow values or other features that contribute to the forecasting of flow. It has been seen that rainfall is not the optimal feature for predicting flow in the unconfined region, so an alternate feature that will contribute less noise to the forecasts should be used in the graphical models. These modified Bayesian networks are expected to perform better than the Bayesian network analyzed in this work because they either eliminate or loosen the temporal stationary assumption that is currently imposed and will more fully model the joint space with minimal assumptions.

Rainfall is a noisy process in which most of its information is contained in the peak rain days. Low levels of sporadic rainfall do not have as much of an impact on the amount of flow in a river as much as upstream flow does. It is mainly when there are periods of extremely high rainfall over the course of several days that flow will increase dramatically. This is especially seen in the confined region where surface water is has a more immediate reaction to precipitation. Flow in the unconfined region is more slowly varying and does not show the affect of rainfall for a longer period of time after a precipitation event. The temporal models examined in this work can not accurately

represent the behavior of flow with respect to rainfall in the unconfined region due to the nonlinearity of the rainfall's contribution. These linear models attempt to capture the high rise in flow after a strong rainfall event has occurred, but in doing so exaggerate the effect rainfall has in times of lower precipitation. A nonlinear model is needed to more accurately represent the effect of rainfall on flow, especially in the unconfined region.

The output from a physical model has been used to form the transitional probabilities for the Bayesian network. Once the estimates from the BN have been calculated, the errors between them and the WAM predictions can be used to re-define the parameters governing stream flow estimation in WAM. WAM prediction can be reproduced and used to form new transitional probabilities in the BN. This process may be iterated until there is minimal change in the WAM parameters, potentially indicating the optimal values have been achieved. WAM flow is generated using one of four field scale models on each grid cell in the watershed. Then, flow is routed to the nearest body of water for each grid cell by estimating the travel time to determine when the water will get there. It is possible to update the delay coefficient in Equation 5-1 based on the difference between WAM flow and the BN flow estimate.

$$T_S(day) = k_{surfacetimedelay}(day) + \frac{D_S(m)}{F_S(m/day)} \tag{5-1}$$

where T_s is the travel time of surface flow from each cell to the nearest stream in days, $k_{surfacetimedelay}$ is the delay coefficient for surface flow from each cell to the nearest stream, D_s is the length in meters of the surface flow path based on topography, and F_s is the average surface flow velocity in meters per day between each cell and the nearest stream. If this update to current physical models is pursued, it will both help

improve current hydrological predictions and enable graphical models to gain popularity as a method of solving hydrological problems by first being introduced as a submodel within the physical models scientist are accustomed to using.

LIST OF REFERENCES

[1] G. E. P. Box and G. M. Jenkins, *Time series analysis: Forecasting and control*. San Francisco: Holden-Day, 1976.

[2] P. C. Nayak, K. P. Sudheer, D. M. Rangan, and K. S. Ramasastri, "A neuro-fuzzy computing technique for modeling hydrological time series," *Journal of Hydrology*, vol. 291, issues 1-2, pp. 52-66, May 2004.

[3] M. Yadav, T. Wagener, and H. Gupta, "Regionalization of Constraints on Expected Watershed Response Behavior for Improved Predictions in Ungauged Basins," *Advances in Water Resources,* vol. 30, pp. 1756 – 1774, 2007.

[4] Natural Resources Conservation Service, (n.d.), "Water supply forecasting -- a short primer." Retrieved October 24, 2011, from Natural Resources Conservation Service National Water and Climate Center, http://www.wcc.nrcs.usda.gov/factpub/wsf_primer.html.

[5] M. Datcu, F. Melgani, A. Piardi, and S. B. Serpico, "Multisource Data Classification with Dependence Trees," *IEEE Trans. Geosci. and Remote Sens.,* vol. 40, no. 3, pp. 609 – 617, Mar. 2002.

[6] D. Niedermayer, "An Introduction to Bayesian Networks and Their Contemporary Applications," *Innovations in Bayesian Networks: Theory and Applications*, Studies in Computational Intelligence, vol. 156, Springer, 2008, pp. 117 – 130.

[7] F. Pernkopf, "Bayesian network classifiers versus selective k-NN classifier," *Pattern Recognition*, vol. 38, issue 1, pp. 1 – 10, Jan. 2005.

[8] R. Neher, and A. Srivastava, "A Bayesian MRF Framework for Labeling Terrain Using Hyperspectral Imaging," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43 , issue 6, 2005.

[9] D. Rother, "Bayesian Network Applications in Molecular Biology, Computer Graphics, and Computer Vision," University of Minnesota PhD Dissertation, 2008.

[10] Suwannee River Water Management District (SRWMD), *Water supply assessment*, 2010. Retrieved March 24, 2012 from SRWMD, http://www.srwmd.org/DocumentView.aspx?DID=1759.

[11] B. M. Jacobson, N. B. Pickering, A. B. Bottcher, J. G. Hiscock, "Unique Routing Algorithm for Watershed Assessment Modeling," Presented at the 1998 ASAE Annual International Meeting, Orlando, Florida, USA, 12-16 Jul. 1998, pp. 10, Paper No. 982237. ASAE, 2950 Niles Road, St. Joseph, MI 49085-9659 USA.

[12] V. P. Singh, "Kinematic wave modeling in hydrology," Invited paper submitted to ASCE-EWRI Task Committee on Evolution of Computer Methods in Hydrology. Reston, VA., ASCE, 2002.

[13] S. E. Serrano, “Development and Verification of an Analytical Solution for Forecasting Nonlinear Kinematic Flood Waves,” *J. Hydrologic Engrg*., vol. 11, issue 4, pp. 347 – 353, 2006.

[14] K. Nagarajan, C. Krekeler, K. C. Slatton, and W. D. Graham, “A Scalable Approach to Fusing Spatiotemporal Data to Estimate Streamflow via a Bayesian Network,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48 , issue 10, 2010.

[15] K. W. Hipel, A. I. McLeod, “Time Series Modelling of Water Resources and Environmental Systems,” *Developments in Water Sciences*, vol. 45, Apr. 1994.

[16] P. W. Fieguth, W. C. Karl, A. S. Willsky, and C. Wunsch, “Multiresolution Optimal Interpolation and Statistical Analysis of TOPEX/POSEIDON Satellite Altimetry,” *IEEE Trans. Geosci. Remote Sens*., vol. 33, no. 2, pp. 280 – 292, Mar. 1995.

[17] J. C. Principe, B. deVries, P. G. de Oliveira, “The gamma filter: a new class of adaptive IIR filters with restricted feedback,” *Transactions on Signal Processing*, vol. 41, no. 2, pp. 649 – 656, Feb. 1993.

[18] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, Cambridge, MA, 2006.

[19] R. O. Duda, P. E. Hart, and D.G. Stork, *Pattern Classification*, 2nd edition, John Wiley & Sons, New York, 2001.

[20] J. P. Hughes, P. Guttorp, S. P. Charles, “A non-homogeneous hidden Markov model for precipitation occurrence,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 48, issue 1, pp. 15 – 30, 1999.

[21] M. Koeppen, “The Curse of Dimensionality,” 5th Online World Conference on Soft Computing in Industrial Applications (WSC5), held on the internet, Sep. 4-18, 2000.

[22] J. Pearl, “Fusion, Propagation, and Structuring in Belief Networks,” *Artificial Intelligence*, vol. 29, 1986.

[23] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo, CA, 1988.

[24] K. Murphy, “The Bayes Net Toolbox for Matlab,” *Computing Science and Statistics: Proceedings of the Interface*, vol. 33, 2001.

[25] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[26] K. Murphy, Y. Weiss, and M. Jordan, “Loopy belief propagation for approximate inference: an empirical study,” *UAI*, 1999.

[27] L. R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, Feb. 1989.

[28] B. J. Borgstrom, P. H. Borgstrom, and A. Alwan, "Efficient HMM-Based Estimation of Missing Features, with Applications to Packet Loss Concealment," *Proceedings. InterSpeech 2010*, pp. 2394-2397, Sep. 2010.

[29] K. Nagarajan, C. Krekeler, K. C. Slatton, “Probabilistic Fusion of Spatio-Temporal Data to Estimate Stream Flow via Bayesian Networks,” *Geoscience and Remote Sensing Symposium, 2007. Proceedings, IGARSS 2007. IEEE International*, pp. 23-27, Jul. 2007, invited.

[30] D. Xu and J. Principe, “Learning from examples with quadratic mutual information,” *Proc. IEEE Workshop on Neural Networks for Signal Processing '98*, pp. 155-164, 1998.

[31] D. Xu and J. C. Principe, “Feature evaluation using quadratic mutual information,” *Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on*, vol.1, pp.459 – 463, 2001.

[32] D. N. Moriasi, J. G., Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, “Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations,” *Transactions of the American Society of Agricultural and Biological Engineers*, vol. 50, no. 3, pp. 885 – 900, Mar. 2007.

[33] E. T. Jaynes, “Information Theory and Statistical Mechanics,” *The Physical Review,* vol. 106**,** issue 620, 1957.

[34] C. Krekeler, K. Nagarajan, K. C. Slatton, “A Method for Measuring the Incremental Information Contributed from Non-Stationary Spatio-Temporal Data to be Fused,” *Geoscience and Remote Sensing Symposium, 2008. Proceedings. IGARSS 2008. IEEE International Conference on*, vol. 2, pp. II-261 – II-264, Jul. 2008.

[35] C. Krekeler, K. Slatton, M. Cohen, “Multi-variate Bayesian Classification of Soil Drainage using Feature-level Fusion of Topographic and Hydrologic Data,” *Geoscience and Remote Sensing Symposium, 2006. Proceedings, IGARSS 2006. IEEE International*, pp. 2522 – 2525, Jul. 2006.

[36] R. E. Neapolitan, *Learning Bayesian Networks*, Pearson Prentice Hall, Upper Saddle River, NJ, 2004.

BIOGRAPHICAL SKETCH

Carolyn Krekeler received a B.S. degree in electrical engineering from Loyola Marymount University, Los Angeles, California in 2004, an M.S. degree in electrical and computer engineering from the University of Florida, Gainesville, Florida in 2005, and a Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, Florida in 2012. She was a Graduate Research Assistant in the Adaptive Signal Processing Laboratory (ASPL) from January 2005 to April 2010 and was a Graduate Research Assistant in the Computational NeuroEngineering Lab (CNEL) from April 2010 to May 2012. Her research interests include pattern recognition, information theory, and graphical models.