PROBABILITY-BASED APPROACHES FOR INCORPORATING UNCERTAINTY INTO WATER RESOURCE MODELS

Joseph Nicholas LoBuglio

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Environmental Engineering in the Department of Environmental Sciences and Engineering, UNC Gillings School of Global Public Health.

Chapel Hill
2012

Approved by:
Dr. Gregory W. Characklis
Dr. Lawrence E. Band
Dr. David H. Moreau
Dr. Kenneth H. Reckhow
Dr. Marc L. Serre
Dr. Stephen C. Whalen
Abstract

JOSEPH NICHOLAS LOBUGLIO: Probability-based Approaches for Incorporating Uncertainty into Water Resource Models
(Under the direction of Dr. Gregory W. Characklis)

Uncertainty in information used to make decisions is unavoidable; however it can be reduced by integrating information from multiple sources, and model techniques incorporating uncertainty and variability can produce more useful probabilistic outcome estimates. This work demonstrates the use of methods for decreasing uncertainty and for using probabilistic outcome data effectively in understanding the water quality and quantity in the Catawba River system in western North Carolina.

Sparse monitoring data and error inherent in water quality models makes the identification of waters not meeting regulatory standards difficult. This work uses the Bayesian Maximum Entropy (BME) method of modern geostatistics to integrate water quality monitoring data together with model predictions to determine the likely status of a water (i.e. impaired or not impaired) and to estimate the level of monitoring needed to characterize the water for regulatory purposes. Although the model predictions used to augment the measured data has a high degree of uncertainty, their inclusion reduces the uncertainty in chlorophyll a estimates enough that the likely impairment status of all sections in all but one reservoir can be determined. For the remaining reservoir, probabilistic predictions of future chlorophyll levels are used to illustrate how monitoring costs can be reduced using a BME framework.

Rainfall-inflow models used for analyzing water availability often have complex forms that can inhibit a thorough analysis of uncertainty in model results because of long model run times and the large number of parameters that are not known with precision. This work demonstrates a rainfall-inflow model that uses reduced spatial and temporal resolution to
facilitate model construction and to allow for a robust assessment of model uncertainty. Uncertainty is captured in 2000 116-year inflow scenarios generated using Markov Chain Monte Carlo methods and scenario-specific estimates of model residual error. These scenarios were incorporated into a multi-reservoir management model. Although the median system behavior agrees with prior work that did not include uncertainty, including a distribution of possible outcomes results in a doubling of the estimate of the number of times reservoirs fall below target minimum levels and an increase in the likelihood of reaching critical levels.
Acknowledgement

I am indebted to my doctoral committee for their commitment and encouragement over the extended course of this study, and especially my advisor, Dr. Gregory Characklis, who set high standards and helped me understand the researcher's obligation to thoroughly and soberly examine contributions to a body of knowledge. I am grateful for Greg’s scholarly discernment, patience, and unwavering commitment, especially in light of the unconventional path I had taken during my tenure as a graduate student.

I am deeply grateful to the administrative staff in the Department of Environmental Sciences and Engineering. I received only encouragement, efficiency, and warmth from the staff, and they deserve special recognition for their efforts. I extend my gratitude to Dr. Jamie Bartram and the staff of the Water Institute at UNC, where I currently serve as research manager, for their encouragement and accommodating my absence while I was working on this research.

I am thankful for those who funded my education and research, both directly and indirectly. These include Dr. Thomas S. Royster and Mrs. Caroline H. Royster, whose generosity made the Royster Fellowship program possible. Many small grants and programs provided additional funding, including from the Graduate School, the Department of Environmental Sciences and Engineering, and numerous smaller funds awarded through my committee. I also would like to thank the people of the State of North Carolina for their commitment to the University and affordable education. Duke Energy provided data and technical support without which this work would not be possible.

Over the 9-years as a student I have been joined by a wonderful wife and four children, now five, five, eleven, and fourteen. I am thankful for their faith in my work and the sacrifices they have made without complaint. Moreover, their growth and learning are
inspirational and have reignited a passion in me for improving our world. My wife, Shannon Jordan, has my undying gratitude for her support and for taking on a greater load in raising our family while my efforts were directed elsewhere.

I would also like to acknowledge my parents, Angelina and Armand LoBuglio, who both passed away in 2001, just prior to the start of my degree. As first-generation Americans, born to poor Italian immigrants and living through the depression and World War II, they recognized the value of education and the importance of effective social institutions. As a school teacher and nurse, they instilled in each of their five children a love of learning and a strong social conscience. My parents’ legacy is ongoing and includes the example of their loving marriage of 43 years and a close-knit family that continues to support each other.
# Table of Contents

List of Tables.................................................................................................................................................. viii

List of Figures................................................................................................................................................ ix

1. Introduction................................................................................................................................................ 1

2. Cost-effective water quality assessment through the integration of monitoring data and modeling results ................................................................................................................. 5
   - Introduction ............................................................................................................................................. 5
   - Methods .................................................................................................................................................. 8
   - Application .......................................................................................................................................... 14
   - Results and Discussion ........................................................................................................................ 23
   - Conclusions ......................................................................................................................................... 31

3. Using Reduced-Parameter Empirical Models to Estimate Inflows into Cascading Reservoir System and Capture Model Uncertainty .................................................................................. 33
   - Introduction ........................................................................................................................................ 33
   - Methods .............................................................................................................................................. 35
   - Analysis and Results ........................................................................................................................... 48
   - Conclusions ........................................................................................................................................ 56

4. Using a Stochastic Cascading Reservoir Model to Estimate the Frequency and Extent of Water Resource Conflicts ........................................................................................................ 58
   - Introduction ........................................................................................................................................ 58
   - Methods .............................................................................................................................................. 61
   - Results ................................................................................................................................................ 68
   - Conclusions ........................................................................................................................................ 73

5. Summary .................................................................................................................................................... 75

Appendix A: Runoff Coefficients and Effects on Model Performance ......................................................... 78
List of Tables

Table 2.1: R-squared values of predicted versus observed chlorophyll a values ..........18
Table 2.2: Probability of at least one segment exceeding 40 µg/L in 2000...............27
Table 2.3: Summary of proposed monitoring plans ..................................................30
Table 3.1: Evaluation of model fitness using NS, RMSER, and PBIAS measures. ..........54
Table 4.1: Variable Descriptions for Linear Programming Model ...............................65
Table 4.2: Monthly Target Storage Index..................................................................66
Table 4.3: Summary of LIP trigger points .................................................................66
Table 4.4: Changes to Operating Targets According to LIP Stage..............................67
Table 4.5: Reservoir Parameters for the six reservoirs entirely within North Carolina...68
Table A.1: Values for runoff coefficients for land types identified in the model.........78
Table B.1: Description of meteorological stations. ....................................................80
Table C.1: Physical Reservoir Properties (Duke Energy, 2006).................................83
Table C.2: Outflows used in reservoir inflow calculations .........................................83
List of Figures

Figure 1.1: The Catawba River Basin (Duke Power, 2003)................................................... 4

Figure 2.1: Illustration of BME methodology ......................................................................10

Figure 2.2: Density of hard and soft data...........................................................................15

Figure 2.3: Covariance of Chlorophyll a over space and time ...........................................16

Figure 2.4: Chlorophyll a prediction model from BATHTUB ............................................18

Figure 2.5: Observed versus predicted chlorophyll a values using model 2. ......................19

Figure 2.6: Lake Wylie Monitoring Stations .....................................................................21

Figure 2.7: Distribution of future estimate of chlorophyll a along with random value from the distribution. .................................................................23

Figure 2.8: Chlorophyll a estimates for reservoir segment 17 for all years .......................24

Figure 2.9: Maps of mean Chlorophyll levels and uncertainty in 2000. ...............................25

Figure 2.10: Probability of reservoir segments exceeding 40 ug/L in 2000 if only hard data are used and when both hard and soft are used. .............................27

Figure 2.11: Possible monitoring plans for Lake Wylie and distribution of number of stations assessed.................................................................29

Figure 2.12: Number of segments assessed with 90% confidence versus cost factor ..........31

Figure 3.1: Inputs and activities for the creation of 116-year inflow record and associated uncertainty .........................................................................................36

Figure 3.2: Location of 19 meteorological stations used in the inflow model.....................37

Figure 3.3: Representation of data available for each meteorological station. .................38

Figure 3.4: Partitioning of vector of means, $\mu$, and covariance matrix, $\Sigma$, according to pattern of missing values in vector of monthly rainfalls, $x$ ..................39

Figure 3.5: Representation of historical dataset for each meteorological station..............48

Figure 3.6: Measures of fitness for imputed data using EM and linear regression. ...........49

Figure 3.7: Multidimensional scaling result based on covariance matrix of monthly station rainfall generated by EM algorithm. .................................................50

Figure 3.8: Scatter plot of Nash-Sutcliffe efficiency measure for three of 24 runoff coefficients ......................................................................................................................52

Figure 3.9: MCMC burn-in using two chains for two of the 90 variables showing convergence of the parameter values. ............................................................53
Figure 3.10: Posterior distributions Lake Norman $\beta_{10}$ showing typical result.........................54
Figure 3.11: Model results versus known flows, million cubic meters .................................55
Figure 3.12: Median predicted values versus actual values for each reservoir .........................56
Figure 4.1: Number of expiring FERC hydroelectric permits by year (FERC, 2011) ...............59
Figure 4.2: Schematic of Reservoir Management Model .....................................................61
Figure 4.3: Target and other operationally important reservoir levels for Lake James ..........64
Figure 4.4: Percentiles of reservoir levels for Lake James by month under 2008 demand scenario ..............................................................................................69
Figure 4.5: Frequency of Lake James reservoir levels being below the target minimum......70
Figure 4.6: Percent of Time LIP in Place for 2008 and 2058 Demand Scenarios ....................71
Figure 4.7: Hydroelectric power generation under 2008 and 2058 demand scenarios .........72
Figure A.1: Dotty plots of Nash-Sutcliffe efficiency measure for each runoff coefficient. ......79
Figure A.2: Cumulative distribution of Nash-Sutcliffe value for 5000 Monte Carlo runs ......79
Figure B.1: Normal probability plots of transformed rain data for each of the 19 stations .....81
Figure B.2: Convergence monthly rainfall means during EM algorithm ...............................82
Figure B.3: Values for linear transformation of MDS plot to geographic coordinates ............82
Figure C.1: Average Pan-Evaporation in Mountain Region of North Carolina .........83
Figure E.1: Distribution of Monthly Reservoir Levels for Six Reservoirs without Inflow Uncertainty .........................................................................................89
Figure E.2: Distribution of Reservoir Levels for Six Reservoirs with Inflow Uncertainty ......90
Figure E.3: Fraction of Times Reservoir Level Below Target Minimum Level, Baseline .......90
1. Introduction

Water supply and water quality challenges are increasing worldwide as humans demand an ever greater fraction of renewable fresh water resources (Postel, et al., 1996). Water quality and quantity are primary considerations when making management decisions and the broader value of water is recognized in a 1986 federal law covering the relicensing of hydropower facilities which requires that these new licenses give "equal consideration" to power production and non-power benefits, such as wildlife habitat, recreation, and water quality (ECPA, 1986).

One impediment to making water management decisions is the imperfect knowledge we have about the human and natural systems that effect water quality and supply and the resulting uncertainty in the estimates of current and future conditions. This is a result of the inherent complexity of these systems and the natural variability in meteorological and human behaviors. A report from the National Research Council (2001) recognizes that scientific uncertainty is a reality that cannot be entirely eliminated.

This work furthers the understanding of how uncertainty can be better dealt with in model development and decision making pertaining to surface water quantity, quality, and allocation. Furthermore it demonstrates ways of including data typically omitted from consideration because it is incomplete or has high uncertainty but which, if used, can contribute important insight to decision makers.

Presented here are three chapters focusing on different aspects of the use of uncertain information. The first chapter demonstrates the use of spatial/temporal random field theory to combine various forms of existing water quality monitoring and modeling information, some of which may have associated uncertainty, to generate probabilistic estimates for water quality parameters where quality monitoring data does not exist. These
estimates can be used to reduce the cost of assessing waters and for developing lower cost monitoring plans.

Work in the first section is motivated by the idea that the comprehensive assessment of water bodies (e.g. for TMDLs) is hindered by sparse historical monitoring data, the high uncertainty inherent in water quality models, and the cost burden of implementing extensive monitoring programs. Knowledge related to a water body is often available in various forms with differing accuracies, including precise and imprecise monitoring data, uncertain model results, and physical laws governing a system. These sources of information can be combined within a rigorous statistical framework (Bayesian Maximum Entropy Methods or BME) to calculate a probability distribution for the parameter values needed for assessment. These distributions can then be used to determine the probability of being out of attainment and hence if a water body should be classified as impaired, not impaired, or requiring further study. Although BME methods have been applied in several water-related contexts employing these methods as a means of integrating both monitoring and modeling results represents an unexplored area of research.

The second and third sections of this work focuses on creating a probabilistic empirical model relating basin rainfall to receiving water inflows and, in turn, the evaluation of a reservoir management strategy. These sections demonstrate the use of the Expectation Maximization algorithm, and Multidimensional Scaling to create and validate a 116-year rainfall dataset based on historical data. Markov Chain Monte Carlo techniques on aggregate-level models allow estimation of reservoir inflow values along with estimates of associated output uncertainty from model uncertainty and input variability. This uncertainty is propagated through a management model of a cascading reservoir system and demonstrates that including uncertainty can reveal a higher expected frequency of infrequent events.

Because the current level of understanding and observational evidence related to complex natural systems is often inadequate to make accurate detailed predictions; an
aggregate-level probabilistic model can provide a better approach when the outcome
variables do not require high resolution data. This model form sacrifices some temporal and
spatial resolution in favor of greater accuracy by focusing on aggregate-level results
(Reckhow, 1999). It relies on empirical relationships between a reduced number of system
variables as a means of using limited data and knowledge more efficiently. The model form
also recognizes that uncertainty can enter from many sources and propagates this
uncertainty through to model results. This type of model is useful for evaluating outcomes,
such as droughts, that have timescales of months or years and less useful for phenomenon
requiring greater temporal or spatial resolution, such as flooding.

Reservoir systems are highly managed and their performance is dependent on the
regulatory rules governing the various uses of the water resources. Understanding the effect
of management strategies on the frequency and extent of inflow shortfalls can be important
for maximizing the economic and societal benefits derived from the in-system uses and
withdrawn water. The Catawba reservoir system is governed under the terms of a Federal
Energy Regulatory Commission permit that includes a low-inflow protocol to modify
operating rules during times of low inflow. The modeling for this permit did not include a
rigorous analysis of uncertainty and therefore underestimated the likelihood of low-flow and
critical events.

The methods described are applied to the Catawba River watershed in western
North Carolina. The mainstem of the Catawba River is impounded by seven hydropower
reservoirs and runs from the foot of the Blue Ridge Mountains to the point where it enters
South Carolina. Dams predominate on the mainstem; only 13 miles of free-flowing river exist
in North Carolina, and these reservoirs serve not only to provide hydroelectric power, but
also water supply and recreational opportunities. Five of these seven reservoirs supply
about 85% of the drinking water in the Catawba Basin, which includes Charlotte, the state’s
most populous city. This is the fastest growing river basin in North Carolina, with a growth in
the Charlotte area alone exceeding 250,000 people by 2020, and has experienced both
hydroelectric and municipal water supply reductions as a result of a drought from 1998 to 2002. Water quality is also a concern in several of the reservoirs and this concern will grow with the increase in waterfront development. An 11 year water quality monitoring dataset is used as the basis for this analysis. Population and land use projections will allow to water management scenarios under future conditions to be evaluated.

![Figure 1.1: The Catawba River Basin (Duke Power, 2003)]](image-url)
2. Cost-effective water quality assessment through the integration of monitoring data and modeling results

Introduction

Under the Clean Water Act states are required to assign rivers, lakes, and estuaries a designated use (e.g. drinking water supply, contact recreation), and assess whether their water quality is sufficient to meet designated uses (National Research Council, 2001). For those not meeting these standards, the states must establish pollutant budgets, total maximum daily loads (TMDLs), that will bring these waters into attainment. As of 1998 only 23 percent of rivers and streams and 42 percent of lake area were assessed (USEPA, 2000), yet around 20,000 water bodies were reported as being in violation of at least one standard with over 41,000 violations in all (USEPA, 2002).

In 2003, the EPA withdrew its final TMDL rule in response to court challenges, congressional legislation against implementation, and a National Research Council report (2001) detailing technical issues with the rule (USEPA, 2003). Many of the concerns were based on the huge cost burden that the TMDL program would impose. The EPA (USEPA, 2001) estimates that cost of monitoring to support the TMDL program is $17 million per year while costs of developing and implementing TMDLs will cost between $900 million and $4.3 billion annually. Many of the technical issues are related to a lack of monitoring data that prevents evaluating water quality with sufficient confidence to determine whether or not a water is "impaired". To accommodate this uncertainty the NRC report (2001) recommends replacing a single list of impaired waters with two lists; a "preliminary list" of those that are likely impaired, but for which there is not enough confidence to make a definitive designation, and an "action list" of waters known to be impaired with a high level of confidence and for which a TMDL must be developed and implemented. This report also
suggests that agencies use statistical approaches to reduce costs in the areas of data analysis, assessing waters, and designing monitoring programs.

Classification of waters is always subject to uncertainty because the spatial and temporal coverage of monitoring programs is generally limited. When monitoring data is available, the parameter of interest (e.g. chlorophyll a, fecal coliforms) may not have been measured directly because monitoring programs often focus on specific questions (e.g. the effect of a wastewater effluent) and only a limited number of parameters are measured to reduce costs. Even when historic data are available, such as those used to assess dissolved oxygen and/or eutrophication during the 1960s and 1970s, they are often accompanied by greater uncertainty than those evaluated using current analytical and modeling techniques (Reckhow, 1999; Chapra, 2003).

Parameter predictions from models, both those which incorporate physical/chemical processes and those that are empirically based, can be combined with knowledge of spatial and temporal variation in parameter values to improve estimates, provided that model uncertainty can be quantified. Fuentes and Raftery (2005) use Bayesian methods to improve spatial estimates for atmospheric sulfur dioxide by combining monitoring and modeling data, but this technique does not appear to have been implemented for spatial and temporal estimates in water quality models. Further improvements in water quality estimates can be obtained by using statistical methods to include information from data sources that might otherwise be omitted because of their level of uncertainty. In this work, information from temporal/spatial statistics, monitoring data, and modeling results are combined within a rigorous statistical framework to generate parameter estimates along with estimates of their uncertainty. This combining of knowledge from multiple sources results in the reduction of uncertainties in parameter estimates (Christakos, et al., 2002).

Uncertainty in water quality models is unavoidable given the inherent complexity of natural systems and limits on data availability (Reckhow, 1999; Stow, et al., 2003; Sincock, et al., 2003). Models have become increasingly more sophisticated in order to capture more
of this complexity but this sophistication does not necessarily provide better results (Vreugdenhil, 2002). Understanding and calculating uncertainty in model results can be more difficult than running the model itself. Uncertainty assessment in process models is often limited by the time and resources required for a large number of model runs. For example, uncertainty in a mechanistic water quality model (CE-QUAL-W2) was investigated as part of an effort to model the Neuse River Estuary (Bowen & Hieronymus, 2003; Bowen & Hieronymus, 2000; Bowen, 2000). This analysis included variation in only 7 of the 115 model parameters; testing 115 parameters would require $10^{54}$ model runs, while testing 7 parameters reduced that to about 2000 model runs.

When first- and second-order derivatives can be estimated, uncertainty in model results based on given uncertainties in model parameters can be calculated using the mean-value first-order second-moment method and the advanced mean-value first-order second-moment method (Mailhot & Villeneuve, 2003). Other techniques, such as generalized likelihood uncertainty estimation (Freer, et al., 1996), Bayesian Monte Carlo methods, and Markov Chain Monte Carlo methods (Qian, et al., 2003) can be used generally although the large number of simulations required for these techniques often leads to computational constraints. Empirical/statistical approaches, such as SPARROW (Smith, et al., 1997; McMahon, et al., 2003) and other nonlinear regression techniques (Borsuk, et al., 2004), can facilitate an uncertainty analysis, but their predictive power is limited by the extent of their underlying data sets (Chapra, 2003).

For both mechanistic models and empirical/statistical models, measured values are usually used for model calibration after which uncertainty in model results are calculated based solely on the uncertainty inherent in the model; the measured values, along with their (usually) associated low uncertainty, are set aside. Bayesian Maximum Entropy methods allow systematic incorporation of general knowledge (such as a parameter's spatial and temporal covariance), monitoring data, and model predictions and preserve certainty where quality information exists while making estimations and uncertainty estimates at other
temporal and spatial locations (Christakos, 2000). Although statistical approaches such as BME provide estimates of uncertainty for a given set of information, they can also be used to determine the amount and type of data needed to achieve a specific level of uncertainty. Because of its ability to assess uncertainty under a variety of monitoring and modeling strategies, BME can be used to help choose more accurate and cost-effective monitoring programs for impairment assessment. For example, the number of highly accurate, but expensive, laboratory analyses might be reduced in favor of lower-cost, but less certain, data.

This work involves the application of BME methods to evaluate chlorophyll a levels on the Catawba River system. Results describe estimates of chlorophyll a concentrations and their uncertainty throughout the Catawba system, and a subsequent exercise is undertaken to illustrate the value of this approach as a means of reducing the extent of the monitoring systems required to assess water quality. Results are placed in a form that should be useful to decision makers charged with evaluating water quality and developing monitoring programs.

**Methods**

**Bayesian Maximum Entropy (BME)**

BME provides a formal framework to combine information and provide parameter estimates, along with uncertainty information, at any point throughout a system. Information on how a parameter varies over space and time is captured using spatiotemporal random field theory, which then is used to determine defined ranges of plausible estimates for the parameter of interest (Christakos & Li, 1998). The ranges are then restricted to make them consistent with site-specific hard (exact) and/or soft (uncertain) data. Figure 2.1 illustrates how a covariance relationship (general knowledge) is combined with hard information (carefully obtained monitoring data) and soft data (modeling results and monitoring data with uncertainty) to generate parameter estimates at a given monitoring location over time.
The BME approach incorporates the theory of spatiotemporal random fields (S/TRF) to model natural variability and uncertainties in a variable over space and time. A spatial random field (SRF) $X(s)$ is a random variable that is a function of location $s$ (i.e. $X(s) = x$ where $x$ is the random variable and $s$ is the vector of coordinates at a location). Another point, $s'$, would have a different corresponding random variable: $X(s') = x'$. Both $x$ and $x'$ are random variables that can take a range of plausible values, the distribution of which is represented by a cumulative density function (CDF), or its derivative (i.e. a probability density function (PDF)). The space/time covariance, $\text{cov}(x, x')$, between these two random variables is denoted as $c_X(s, s')$, in order to emphasize the relationship between covariance and the locations $s$ and $s'$. This covariance relationship characterizes the variability of the SRF across space, and it can be derived empirically or it can be analytically generated using physical laws governing the system. By incorporating a temporal coordinate, $t$, the framework can be extended to S/TRF theory.

The BME approach consists of three main stages illustrated in Figure 2.1. At the structural (or prior) stage the information theory concept of entropy maximization is used to process the general knowledge available and derive a prior PDF characterizing the mean trend and variability of the S/TRF. The general knowledge usually describes the distribution of field values in space and time by their mean function and covariance relationship. When the covariance depends only on the distance (spatial or temporal) between two points and not on the coordinates of the two points themselves, and when the mean function is unchanging, then the S/TRF is said to be spatially homogeneous and temporally stationary. When this is the case, as with the application in this paper, the BME analysis is simplified.
<table>
<thead>
<tr>
<th>Stage</th>
<th>Step</th>
<th>Illustration</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Structural Stage</strong></td>
<td>A</td>
<td>Prior PDF derived from general knowledge G, which includes the covariance relationship. Relationship determined analytically via physical laws or empirically via an existing dataset.</td>
</tr>
<tr>
<td><strong>Specificatory Stage</strong></td>
<td>B</td>
<td>Hard data (if any) identified. Illustration shows monitoring data for one location over time.</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>Soft data identified for space/time locations where hard data is absent. Illustration shows three measurements with uncertainty expressed as probability distribution.</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>Soft data can also be obtained from model predictions. Illustration shows mean estimate along with a confidence interval.</td>
</tr>
<tr>
<td><strong>Integration Stage</strong></td>
<td>E</td>
<td>General knowledge reconciled with site-specific data. With only empirically-based general knowledge and hard (monitoring) data, BME reduces to simple Kriging.</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>Integration of model results (soft data) with hard data. Uncertainty is reduced and greater resolution is obtained.</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>Integration of uncertain monitoring data with hard data and model predictions. Uncertainty is reduced near soft data space-time locations.</td>
</tr>
</tbody>
</table>

Figure 2.1: Illustration of BME methodology

The specificatory (or meta-prior) stage identifies both exact measurements (hard data, Figure 2.1, panel B), which might arise from rigorous monitoring, and inexact estimates (soft data, Figure 2.1, panels C and D), such as might come from historical measurements or model predictions. The vector of hard data, $\mathbf{X}_{\text{hard}} = (\mathbf{X}_1, \ldots, \mathbf{X}_m)$ contains exact measurements at the corresponding specific space/time points, $\mathbf{p}_i = (\mathbf{s}_i, t_i), (i=1, \ldots, m)$. The...
S/TRF at these points takes the values $\chi_{\text{hard}}$ with probability 1. For the soft data set, $\chi_{\text{soft}} = (\chi_{m+1}, \ldots, \chi_n)$, each value has an associated probability distribution. The value of the S/TRF at the single space/time location where an estimate is sought is denoted as $\chi_k$. Together the value of the S/TRF at all the $n+1$ mapping points of interest is expressed as $\chi_{\text{map}} = (\chi_{\text{hard}}, \chi_{\text{soft}}, \chi_k)$, where the space/time mapping points correspond to $p_i, (i=1, \ldots, n+1)$.

The integration (or posterior) stage assimilates the information from the first two stages to yield a posterior PDF ($f_k$) for the estimation point. The prior PDF $f_0(\chi_{\text{map}})$ obtained from general knowledge at the structural stage (Figure 2, panel A) is a joint PDF characterizing the plausible estimates for the S/TRF at all the space/time mapping points. The PDF for soft data $f_s(\chi_{\text{soft}})$ (identified at the specificatory stage; panels C and D), characterizes the uncertainty associated with the site specific knowledge. Assimilation (panels E through G) is then accomplished using the operational Bayesian conditionalization rule (Christakos, 2000; Christakos, et al., 2002), such that

$$f_k(\chi_k) = A^{-1} \int d\chi_{\text{soft}} f_s(\chi_{\text{soft}}) f_0(\chi_{\text{map}})$$

where $A = \int d\chi_{\text{soft}} f_s(\chi_{\text{soft}}) f_0(\chi_{\text{map}})$ is a normalization constant.

The posterior PDF, $f_k(\chi_k)$, contains probabilistic information of the parameter of interest at the estimation point, $p_k$. This information can be used to generate a mean estimate, a confidence interval, or to calculate the probability that the variable being estimated is above or below some threshold of interest.

In the simplest case, when only hard data are available, BME reduces to the Simple Kriging method of classical geostatistics. The confidence interval increases as the location of the point where the prediction is made becomes more distant from the measured points (Figure 2, panel E). BME methods improve on simple Kriging by allowing the use of uncertain data to improve the estimate near these soft data points (panels F and G) and
reduce the overall uncertainty of the estimate. BME methods combine information from the
covariance derived from the hard data with the uncertainty of, and covariance among, soft
data. In this work BME methods are used to combine monitoring data (hard data) with model
predictions from an empirical model (soft data) to make probabilistic parameter estimates at
regular intervals in the space/time region of interest.

Choosing a cost effective monitoring program

Probabilistic information can be used to calculate the probability of a water body not
being in attainment (i.e. in violation of a water quality standard). However, there is no
guarantee that this probability would be low enough to conclude that the water is attaining its
designated use or high enough to conclude that it is not; in which case further monitoring
may be required. BME can be used to help determine the spatial resolution and the quality
of a monitoring program required to assess a water body's attainment status with a desired
confidence level. By minimizing the amount of hard monitoring data needed by incorporating
less exact (and less expensive) soft data, assessment may be achieved at lower cost.

The degree to which measurement uncertainty affects the ability to assess a water
body depends on the true values of the parameter, which are unknown before they are
measured. If the true values are much higher or lower than the allowable limit, then even an
imprecise measurement can be useful in determining if a water body should be designated
as impaired or not. As the true value approaches the allowable limit, greater discrimination is
required. In the extreme, when the true value is equal to the limit, only data without
uncertainty can be used to determine if the water is in compliance.

A probability density function of the true value of the parameter at a future time can
be calculated with BME. Since there are no monitoring data available for future years, the
estimated values will have relatively high levels of uncertainty. The key to assessing a
potential monitoring plan that uses both hard and soft data is the combination of the
uncertainty in the future true value with uncertainty in the estimate of the parameter value at
the time of monitoring. The combined uncertainty can then be used to calculate the probability of having enough information to classify the water body (e.g., there is an 80% probability that the water can be classified with 90% confidence).

The procedure for selecting a cost effective monitoring program can be generalized into three steps: developing candidate monitoring programs, assessing the efficacy of the monitoring programs, and choosing the best monitoring program from among the candidates. The first step begins with identifying the locations where parameter estimates are required and the types of monitoring that are available along with estimates of their measurement uncertainty and their relative costs. These various types of monitoring methods are incorporated into a range of candidate monitoring schemes which include a variety of hard, soft, and unmonitored locations.

In the second step, BME is used with the existing dataset to generate posterior PDFs for the parameter of interest at selected monitoring locations for the future monitoring cycle. This information is used as the starting point for a Monte Carlo simulation of each monitoring plan. One iteration of the simulation begins by picking a value from the posterior PDFs for each monitoring location. For each iteration, the chosen value is considered the true parameter value. Monitoring is simulated by assigning these values to the location wherever hard data will be obtained. Where soft data will be collected, a value is selected from a distribution that has the expected value equal to the true value and which has the same PDF as the measurement method to be employed. Locations without any monitoring are assigned no values. BME is then used with general knowledge (i.e. spatial and temporal covariance relationships) and the synthetic hard and soft data to generate PDFs at all locations.

Using these PDFs, the probability of attaining the water quality standard of interest can be calculated for each location by integrating the distribution from its lowest bound up to the value of the standard. If this value for a particular location meets the confidence required to declare a water impaired or not (such as having less than 10%, or more than 90%,
confidence of exceeding a standard respectively), the location is considered "assessed". This method is similar to that employed by McMahon et al. (2003) who use nonlinear regression model results as a means of classifying stream reaches probabilistically according to hypothetical nitrogen standard and work by Garcia and Froidevaux (1997) in determining the classification of possibly contaminated soils. The number of locations that can be assessed, which is the measure by which a program is deemed successful or not, is then computed. This procedure, beginning with the selection of simulated true values, is repeated many times to develop a distribution of the number locations that can be assessed for any single monitoring program. The set of simulations is then repeated for each monitoring program.

Selection of a monitoring program proceeds by ordering them by cost and summarizing each program's probability distributions for the number of locations assessed. A decision maker can then choose the most appropriate program based on cost constraints and their tolerance for uncertainty.

**Application**

To illustrate the usefulness of the BME framework, annual average chlorophyll a levels are estimated for the seven reservoirs in the Catawba River Basin. Eutrophication is a major reason for impairment of lakes and reservoirs, accounting for 25% of the impaired area of lakes and reservoirs in North Carolina (NCDENR, 2006) and annual average chlorophyll a level is often used in classifying the eutrophic state of a lake or reservoir (Reckhow & Chapra, 1983; Heiskary & Walker Jr., 1995). It is also a parameter that has been estimated by several established water quality models (Vollenweider & Kerekes, 1980; Walker, 1985; Rechhow, 1988). The importance of incorporating uncertain data is demonstrated by using the technique with general knowledge and monitoring data ("hard") alone and comparing the results to those generated with the addition of uncertain ("soft") data from model predictions.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>Lake James 01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Lake James 02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Lake James 03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Lake James 04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Lake James 05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Lake James 06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Lake James 07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>Lake Rhodiss 01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>Lake Rhodiss 02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>Lake Rhodiss 03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>Lake Rhodiss 04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>Lake Rhodiss 05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>Lake Rhodiss 06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>Lake Rhodiss 07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Lake Hickory 01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Lake Hickory 02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Lake Hickory 03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Lake Hickory 04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Lake Hickory 05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Lake Hickory 06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Lake Hickory 07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Lake Hickory 08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Lake Hickory 09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Lake Hickory 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Lake Hickory 11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Lake Hickory 12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Lake Hickory 13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Lake Hickory 14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Lookout Shoals Lake</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>Lake Norman 01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>Lake Norman 02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>Lake Norman 03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>Lake Norman 04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Mt Island Lake 01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>Mt Island Lake 02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>Mt Island Lake 03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>Mt Island Lake 04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>Mt Island Lake 05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>Mt Island Lake 06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>Mt Island Lake 07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>Mt Island Lake 08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>Mt Island Lake 09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Mt Island Lake 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Mt Island Lake 11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Mt Island Lake 12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Mt Island Lake 13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>Mt Island Lake 14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>Lake Wylie 01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>Lake Wylie 02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>Lake Wylie 03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>84</td>
<td>Lake Wylie 04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>Lake Wylie 05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>86</td>
<td>Lake Wylie 06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>87</td>
<td>Lake Wylie 07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>88</td>
<td>Lake Wylie 08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>89</td>
<td>Lake Wylie 09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>Lake Wylie 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>Lake Wylie 11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>Lake Wylie 12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>Lake Wylie 13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>Lake Wylie 14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>Lake Wylie 15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>Lake Wylie 16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>Lake Wylie 17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>Lake Wylie 18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>Lake Wylie 19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.2: Density of hard and soft data.

Black cells indicate the presence of hard data, gray cells soft data, and white cells no data. Rows are roughly ordered by station location.

Each of the seven Catawba reservoirs is divided into as many as nineteen segments, based on its morphology. With 11 years of data and 66 reservoir segments there are 726 space/time locations where average annual chlorophyll values could exist. Of these, 219 have hard information based on monitoring data from Duke Power and an additional 54
have secondary data (i.e. model inputs) that allow estimate predictions (soft data) from an empirical model. BME can be used to provide estimates for the remaining 453 space/time locations. Figure 2.2 shows the density of the hard and soft data in both space and time. The rows are roughly ordered by the location of their corresponding station along the river. Each cell indicates the existence of either hard data (black), soft data (gray) or no data (white).

**General Knowledge**

A space/time covariance model was constructed using the existing monitoring data. Figure 2.3 shows the covariance empirically derived from the dataset (circles) as a function of spatial distance (upper plot) and time lag (lower plot) along with the model of the covariance (solid line). The spatial covariance is calculated as the covariance of pairs of monitoring values that are measured in the same year from stations having a spatial separation close to the value of the spatial lag for which the covariance is sought. Likewise, the data pairs used for calculating the temporal covariance at each value of temporal lag share the same location, but are separated by a number of years equal to the temporal lag.

![Figure 2.3: Covariance of Chlorophyll a over space and time](image)

Circles indicate values calculated from monitoring data; solid lines show covariance model.

An exponential model (1) is used for the spatial component (left-most term in equation 1) while a combination of two exponential models (terms in square brackets in equation 1) was used for the temporal component. The model has an overall variance (var_x)
= 0.8510 and a spatial range (a_s)=25.6 km. The temporal range of the exponential models are (a_t1)=1.75 years and (a_t2)=30 years with a weighting of (\omega)=0.5.

\[
c_x(r, \tau) = \text{var}_x \exp\left(-\frac{3r}{a_r}\right) \left[ \alpha \exp\left(-\frac{3\tau}{a_{t1}}\right) + (\alpha - 1) \exp\left(-\frac{3\tau}{a_{t2}}\right) \right]
\]

(BATHTUB model predictions of Chlorophyll a)

Mean chlorophyll a concentration can be predicted via a range of models, five of which are presented by Walker (1985) and used in the BATHTUB empirical eutrophication model. The models range from a simple linear relationship between chlorophyll a and phosphorus concentration to one that includes nitrogen and phosphorus concentrations, light parameters, and reservoir flushing rates (i.e. inverse of residence time).

Reservoir level and release data are used to calculate the annual flushing rate for each reservoir. Surface data (depth \leq 3 m) on chlorophyll a, phosphorus, and nitrogen are aggregated by month and then by year for April through September in each reservoir segment. Mean depth is calculated for each segment based on stage/area information. The depth of the mixing zone is estimated using an expression from the BATHTUB model that is calibrated on a set of Army Corps of Engineers reservoirs from around the United States (R^2 = 0.93, SE = 0.0026) (Walker 1985).

\[
\log (Z_{mix}) = -0.06 + 1.36 \log (Z) - 0.47 [\log (Z)]^2
\]

where \(Z\) is the total depth in meters and \(Z_{mix}\) is the mean depth (meters) of the mixing layer. Data on non algal turbidity is not available and the BATHTUB model default of 0.08 m\(^{-1}\) is used. Seasonal chlorophyll a concentrations are calculated using four of the five BATHTUB models and regressions are performed to evaluate the accuracy of the models at 162 space/time locations where direct measurements of chlorophyll a are collocated with model input data. Correlation coefficients (Table 2.1) reveal that these models have modest...
capabilities to predict chlorophyll a and model 2 was chosen because it has the highest $R^2$.

The steps involved in computing chlorophyll a via model 2 are described in Figure 2.4.

<table>
<thead>
<tr>
<th>Walker model number</th>
<th>Variables involved</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Phosphorus, Nitrogen, Light Attenuation, Flushing Rate</td>
<td>0.09</td>
</tr>
<tr>
<td>2</td>
<td>Phosphorus, Light Attenuation, Flushing Rate</td>
<td>0.24</td>
</tr>
<tr>
<td>4</td>
<td>Phosphorus only (linear model)</td>
<td>0.20</td>
</tr>
<tr>
<td>5</td>
<td>Phosphorus only (exponential model)</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Table 2.1: R-squared values of predicted versus observed chlorophyll a values

---

<table>
<thead>
<tr>
<th>Step</th>
<th>Variable Definition and Source of Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculate the Phosphorus-Potential Chlorophyll a Concentration</td>
<td>$B_p = \text{Phosphorus-Potential Chlorophyll a Concentration (mg/m}^3\text{)}$ [calculated]</td>
</tr>
<tr>
<td>$B_p = P^{1.37/4.88}$</td>
<td>$P = \text{Total Phosphorus Concentration (mg/m}^3\text{)}$ [obtained from data set]</td>
</tr>
<tr>
<td>Calculate the Kinetic Factor Used in Chlorophyll a Model</td>
<td>$G = Z_{mix} (0.19 + 0.0042 F_s)$</td>
</tr>
<tr>
<td>$G = \text{Kinetic Factor Used in Chlorophyll a Model [calculated]  }$</td>
<td>$Z_{mix} = \text{Mean Depth of Mixed Layer (m) [estimated from total depth, equation 2]}$</td>
</tr>
<tr>
<td>$F_s = \text{Summer Flushing Rate (year}^{-1}\text{) [calculated from reservoir release information]}$</td>
<td></td>
</tr>
<tr>
<td>Calculate the Chlorophyll a Concentration</td>
<td>$B = K B_p / [(1 + b B_p G) (1 + G a)]$</td>
</tr>
<tr>
<td>$B = \text{Chlorophyll a Concentration (mg/m}^3\text{)}$ [calculated]</td>
<td>$K = \text{Calibration factor [value of 1 used]}$</td>
</tr>
<tr>
<td>$a = \text{Non-Algal turbidity (m}^{-1}\text{) [value of 0.08 1/m used]}$</td>
<td></td>
</tr>
<tr>
<td>$b = \text{Algal Light Extinction Coefficient (m}^{-1}\text{) [default value of 0.025 used]}$</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.4: Chlorophyll a prediction model from BATHTUB

The log transformed data for model 2 is shown in Figure 2.5, along with a linear model fit. This figure shows the observed data as a function of the model prediction value, which is the relationship required to integrate the modeling information with the monitoring data. Although the $R^2$ is low, both the slope and intercept are significantly different from zero ($p < 0.001$).
Uncertainty in the model prediction value was derived from the empirical fit of the observed data versus the chlorophyll $a$ model prediction. If we designate the model prediction as $X$ and the observed value as $Y$, we come up with a relationship for the expected value of $Y$ given a particular model prediction value $X_0$ that can be expressed as

$$E\{Y \mid X_0\} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$  

(3)

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the regression coefficients for the intercept and slope, respectively. The standard error (SE) of the prediction can be calculated from

$$SE[E\{Y \mid X_0\}] = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S_x^2}}$$  

(4)

where $\hat{\sigma}$ is an estimate of the population standard deviation in the measured chlorophyll $a$ values, $n$ is the number of data points and $S_x^2$ is the variance in the predicted chlorophyll $a$ values. The expected value and standard error of the prediction can be used to create a PDF for each predicted value from the chlorophyll $a$ model using a t-distribution. Because the t-distribution approaches a normal distribution for large values of $n$, a normal distribution was used in this work ($n=162$).
Integration of hard and soft data

Pre and post processing of the hard and soft data is accomplished with numerical tools developed to extend the BMElib implementation of BME methods within the MATLAB programming platform (Christakos, et al., 2002). Annual average chlorophyll values (hard data) are obtained by averaging the monitoring results by month and then by year (from April through September) for each of the 219 stations where chlorophyll a is measured. BATHTUB model 2 (Figure 2.4) is used to obtain chlorophyll a prediction values at 54 space/time locations where phosphorus and flushing rate data are available, but where chlorophyll is not measured. The prediction variance (equation 4) is used to generate the uncertainty (expressed as a Gaussian PDF) associated with these prediction values, thereby providing the soft data. This information, along with the general knowledge (covariance relationships) are used to generate chlorophyll estimates for all 726 space/time locations twice, both with and without soft data, using BME methods. Results are presented for both approaches to demonstrate the improvement in estimates that can be achieved by including soft data.

Evaluation of Monitoring Plans

Lake Wylie was selected as the site to evaluate monitoring programs due to its higher levels of chlorophyll a and spatial extent. There are 19 historical monitoring stations on the lake (Figure 2.6) and it is assumed that they are located to adequately represent their corresponding lake segments and, in aggregate, provide enough information to allow assessment of the entire lake. This work, however, does not depend on using locations that were monitored historically; BME can generate estimates at any location and thus allows evaluation of plans with any number of different monitoring sites. The dataset used for this work ends in 2003 so monitoring plans for 2004 are evaluated. Simeonov et al. (2003) discuss how cluster analysis can be used to locate sampling sites when a well-populated historic dataset exists and monitoring will proceed with the same methods.
An exhaustive program would monitor all 19 locations using conventional measurement techniques. For instance, samples might be acquired, transported to a laboratory, and then undergo bench-scale analysis. This would allow assessment of each reservoir segment with an assumption of no uncertainty; however it would also be an expensive program. Cost could be reduced by monitoring fewer stations and using less certain (soft) techniques to estimate concentrations at some locations, or a combination of both. A soft monitoring technique could refer to different measurement methods (e.g. in-situ probe data versus laboratory analyses), the use of model predictions, or less frequent measurements in computing an annual average (e.g. once per month instead of once per week). The tradeoff with using a lower cost monitoring program is that a greater degree of uncertainty is introduced. Using a combination of BME and Monte Carlo methods this uncertainty can be quantified and presented as a probability distribution of the number reservoir segments that will be assessed within a specified confidence level. For example, given a particular monitoring program for Lake Wylie, there might be a 90% probability that 17 or more of the 19 reservoir segments will be assessed (i.e. designated as either in, or out of, compliance) with at least 90% confidence.

Figure 2.6: Lake Wylie Monitoring Stations
The development of lower cost monitoring begins by choosing a set of candidate monitoring programs that include subsets of the complete set of monitoring stations and possible monitored methods for each station. This reduces the number of programs that must be evaluated to a manageable number. This is important since, even with only 19 stations monitored with one of three monitoring options (hard, soft, or not at all), there are over one billion possible scenarios. Future work may allow sampling schemes to be evaluated more holistically, as is done by Reed et al. (2001) who use genetic algorithms for optimizing the tradeoff between sampling costs and estimation errors in an existing groundwater monitoring network. For this work it is assumed that chlorophyll $a$ at candidate locations will be monitored with one of two hypothetical methods or not monitored at all.

Prior to evaluating the candidate monitoring programs, BME methods are used to generate estimates of the 2004 chlorophyll $a$ values prior to 2004 monitoring by adding the spatial/temporal locations for 2004 to the BME model used previously in this work and computing the posterior PDFs. These PDFs are shown with dashed lines on Figure 2.7 with a dot (●) at the mean estimate. For each Monte Carlo iteration evaluating the candidate program proceeds by selecting a value from each of the 2004 chlorophyll $a$ posterior distributions (one realization is shown with open circles on Figure 2.7). Monitoring is simulated by assigning this value to the monitoring results for the hard data stations (squares on Figure 2.7) and choosing a distribution near the chosen value for the soft data station based on the uncertainty of soft measurement technique (PDFs shown with solid lines in Figure 2.7).

Using the simulated monitoring data for the Monte Carlo iteration, BME is used to calculate the posterior PDFs for chlorophyll $a$ at all stations. These are used to calculate the probability of exceeding the chlorophyll $a$ limit in each reservoir segment. If the probability of exceeding the limit is high enough (e.g. 90%) to classify the water as being in violation of the standard or low enough (e.g. 10%) to reasonably conclude it is below the standard, the
A large number of Monte Carlo iterations are performed in this manner to create a distribution for the total number of assessed segments for each plan.

Figure 2.7: Distribution of future estimate of chlorophyll a along with random value from the distribution. Also shown are two hard and three soft data points representing a particular monitoring program.

**Results and Discussion**

Seasonal chlorophyll a values are estimated for all reservoir segments over the historical period covered by the data. The results illustrate the probabilistic nature of BME estimates and the value of using soft data to reduce uncertainty. The estimates are compared with the North Carolina chlorophyll a standard of 40 µg/L to demonstrate using probabilistic information to guide assessment decisions. Lastly, monitoring plans for a single reservoir are evaluated to investigate how monitoring programs can be made less costly by including general knowledge and soft data.

Figure 2.8 shows chlorophyll a estimates for all years in one reservoir segment in Lake Hickory. When using only the hard data we obtain estimates characterized by the BME posterior PDFs shown with solid lines. Alternatively, when using both hard data and soft data derived from the BATHTUB predictions, we obtain estimates characterized by the BME posterior PDFs shown as dashed lines. In the five years where no hard data exist, BME provides estimates based on values nearby in space and time. In years where hard data
exists (e.g. 1993, 1995) there is no uncertainty as these data are considered exact. In some instances (e.g. 1992) the posterior PDFs with and without soft data are similar since there is an absence of soft data at stations and times nearby. Compared to the case where only hard data is used, there is an improvement in the uncertainty where soft data exists (e.g. 1998, 1999) as can be seen by the narrower distributions. These improved estimates do not simply take on the mean value of the soft data (shown as open diamonds in 1998, 1999 and 2000), but are a result of combining general knowledge, hard monitoring data, and soft modeling data within the BME framework.

Another useful way to display the data is to show a map of chlorophyll estimates for all stations during a given year. Since the uncertainties of each estimate are critical to interpreting the estimates, two maps are needed for each year to describe both the mean estimate and the size of the prediction interval. Figure 2.9 contains four figures; those on the left describe the chlorophyll estimate for 2000 while those on the right describe the magnitude of the corresponding 68% prediction interval (for the natural log of the estimate). Maps A and B were generated using only hard data while maps C and D were generated using both hard and soft data.
The mean estimates derived from both hard and soft data (C) show noticeable changes in chlorophyll in Lakes Hickory and Wylie relative to estimates based on hard data alone (A). More importantly, the inclusion of soft data leads to a reduction of uncertainty in Lakes Rhodhiss and Wylie (8B vs. 8D). Figure 2.9 gives a broad overview of the entire reservoir system and is the sort of approach that might be used to identify reservoir segments where chlorophyll levels may be of concern. Because mean estimates of chlorophyll \( a \) levels are less than half the standard, one can be confident that segments with low uncertainty are highly likely to be in compliance with the standard. However, it is not clear from figure 2.9 that segments with high chlorophyll estimates and high uncertainty
(Lakes Rhodhiss, Hickory, and Wylie) can be confidently determined to be in compliance. This can be done more rigorously by using the posterior PDF to compare the estimates against probabilistic water quality standards.

A probabilistic water quality standard combines information contained in the mean and variance maps as is done in Figure 2.10. The shading in the map on the left of Figure 2.10 corresponds to the probability that a reservoir segment would have exceeded a 40 µg/L standard in 2000; i.e. \( P[\chi_k > \ln(40)] = \int_{\log(40)}^{\infty} d\chi_k f_k(\chi_k) \). This was done by integrating the PDF for each estimated value between the natural log of 40 (since it is log transformed) and infinity. Most of segments have a less than 5% chance of having violated the standard in 2000 (lightest shade) but parts of Lake Hickory shows a 5-10% chance of a violation (medium gray) and parts of Lake Wylie show a 10%-25% chance (dark gray) of being in violation. Map A includes just information from hard data while map B shows the effect of including soft data. The results in map B indicate that violations are less likely in both lakes Hickory and Wylie, with the probability of violating the standard in Lake Hickory falling below the 5% level.

It is also possible to calculate probabilities pertinent to the reservoirs as a whole. Table 2.2 shows, for each reservoir, the probability (Pr) that at least one of its segments has an annual chlorophyll a level in excess of the 40 µg/L limit for 2000. This is calculated as

\[
P_r = 1 - \prod_{i=1}^{n} P[\chi_{k_i} \leq \ln(40)]
\]

where \( n \) is the number of segments in a reservoir. In this application the calculation is simplified for computational efficiency by assuming that each posterior PDF for each segment is independent; otherwise multi-dimension integration of the joint BME posterior PDF could be implemented.
Figure 2.10: Probability of reservoir segments exceeding 40 ug/L in 2000 if only hard data are used (left) and when both hard and soft are used (right).

These results could be useful for classifying reservoirs. A decision maker might decide that those reservoirs with a less than 10% chance of at least one segment exceeding the 40 µg/L limit need not be listed for nonattainment (e.g. Lookout Shoals and Lake Norman); those with a probability between 10% and 50% (e.g. Lake Rhodhiss and Lake Hickory) be placed on a preliminary list reflecting some concern over compliance, and those with a 50% chance or greater be placed on an action list requiring immediate attention (e.g. Lake Wylie).

<table>
<thead>
<tr>
<th>Reservoir</th>
<th>Hard Data Only</th>
<th>Hard and Soft Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lake James</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>Lake Rhodhiss</td>
<td>13%</td>
<td>10%</td>
</tr>
<tr>
<td>Lake Hickory</td>
<td>44%</td>
<td>27%</td>
</tr>
<tr>
<td>Lookout Shoals</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Lake Norman</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>MI Lake</td>
<td>6%</td>
<td>5%</td>
</tr>
<tr>
<td>Lake Wylie</td>
<td>80%</td>
<td>51%</td>
</tr>
</tbody>
</table>

Table 2.2: Probability of at least one segment exceeding 40 µg/L in 2000.

Figure 2.10 shows that some regions in Lake Wylie have up to a 25% probability of being in violation of the chlorophyll standard or, conversely, a 75% chance that the water body does not violate the standard. A decision maker may wish to be more certain of the
attainment status and request further monitoring. In this work it is assumed that 90% confidence of either attainment or nonattainment is sufficient to make a designation, although any confidence level could be chosen. Nine monitoring plans, each involving different levels of hard and soft data, are evaluated with respect to their ability to assign an "assessed" designation (either in or out of compliance) to each of the segments in Lake Wylie with at least 90% confidence.

Figure 2.11 shows the hard and soft stations for each monitoring plan as well as a boxplot summarizing the distribution of the number of segments (out of a total of 19) that would be classified as assessed under each plan. For example, the distribution of the number of assessed segments for plan 5, which has 3 hard and 4 soft monitoring sites, has a median of 17 and an interquartile range that extends from 15 to 19. From left to right in the figure the number of hard data monitoring stations increases; the number of soft data stations increases from top to bottom of the figure.

Plans in the upper left of figure 2.11 (1, 2, and 4) have a wide spread in the number of assessed segments. For plan 1, for example, one can expect to have anywhere between 3 and 19 segments assessed. As the monitoring efforts get more extensive the range of the expected number of assessed regions decreases while the median increases. In the most extensive plan evaluated (plan 9), one would expect to be able to assess 18 or 19 segments with 90% confidence.

The choice of a monitoring plan will surely be a function of cost as well as the regulators' tolerance for not having some segments assessed at the desired level of confidence. For example, choosing plan 5 would require that regulators accept the risk that 10% of the time 7 or more segments could not be assessed with 90% confidence. Plan 9 would involve acceptance of the risk that 10% of the time 2 or more segments could not be assessed.
Table 2.3 summarizes each plan, including the number of stations, distribution information for the number of segments assessed, and a cost factor. The cost factor is the ratio of the variable cost of a candidate plan to those of a plan having hard data obtained for all 19 segments. The relative costs of hard versus soft monitoring must be taken into account when calculating the cost factor; in this example it is assumed that hard data is twice as expensive to acquire as soft data. The higher cost of obtaining a hard monitoring data point could be because a particular test is inherently more expensive (e.g. a laboratory analysis versus an in-situ measurement) or because the measurement is an aggregate measure based on fewer individual measurements when only soft data is required (e.g.
estimating a seasonal average with monthly samples instead of weekly or biweekly samples.)

<table>
<thead>
<tr>
<th>Plan</th>
<th># of Hard Monitoring Stations</th>
<th># of Soft Monitoring Stations</th>
<th># Segments Assessed with 90% Confidence</th>
<th>Cost Factor ($ hard: $ soft = 2:1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10th %ile</td>
<td>Median</td>
<td>90th %ile</td>
<td></td>
</tr>
<tr>
<td>Plan 1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>Plan 2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>Plan 3</td>
<td>1</td>
<td>5</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>Plan 4</td>
<td>3</td>
<td>0</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>Plan 5</td>
<td>3</td>
<td>4</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>Plan 6</td>
<td>3</td>
<td>6</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>Plan 7</td>
<td>8</td>
<td>0</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>Plan 8</td>
<td>8</td>
<td>5</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>Plan 9</td>
<td>8</td>
<td>10</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>Full</td>
<td>19</td>
<td>0</td>
<td>19</td>
<td>19</td>
</tr>
</tbody>
</table>

Table 2.3: Summary of proposed monitoring plans including number of each type of monitoring station, distribution information for the number of assessed stations (with 90th confidence), and cost factor

Figure 2.12 shows how the assessment uncertainty varies with cost factor using the data from Table 2.3. As might be expected, decreasing uncertainty comes with increasing cost. Plan 3, with its 1 hard and 5 soft monitoring stations, is likely to allow the assessment of more segments than Plan 4, which has 3 hard monitoring stations, even though both plans are about the same cost. Plan 5 is also a dramatic improvement over Plan 4 because of the additional 4 soft data stations. Including soft data stations within the BME framework allows greater flexibility in determining the appropriate tradeoff between cost and the number of segments that will be assessed. The most comprehensive plan (plan 9) would save roughly one third of the monitoring costs compared to the baseline scenario of obtaining hard monitoring data at all stations. Plan 8 would cost about half of the baseline scenario and might also have an acceptable level of uncertainty.
Figure 2.12: Number of segments assessed with 90% confidence versus cost factor assuming a 2:1 ratio between the variable costs of acquiring hard data versus soft data.

Conclusions

BME methods allow information of various types and qualities to be combined to provide parameter estimates at desired locations and times. These estimates are expressed probabilistically with the degree of uncertainty reflecting the uncertainty inherent in the underlying information. By including spatial and temporal covariance information, BME allows the estimation of chlorophyll $a$ values at locations and times where they were not measured. Including uncertain data from a simple chlorophyll $a$ model, even one with highly uncertain results, improves the chlorophyll estimates. The resulting PDFs allow for the quantification of the probability of a water not meeting a water quality standard and, subsequently, a means of identifying which ones should be listed as impaired, which ones require more monitoring, and which ones should be designated as unimpaired.

As a consequence of its ability to propagate information from one space/time location to another, BME can be used to lower the costs of monitoring programs by permitting the use of less expensive measurement methods and a reduced number of monitoring locations. More expensive monitoring methods can be used preferentially where future estimates indicate values close to a regulatory threshold and the results of these monitoring methods, augmented with less certain data, allow the computation of parameter estimates.
everywhere they are needed with an acceptable level of uncertainty. Although these methods were demonstrated with chlorophyll a concentrations in a reservoir system, they can be applied to the estimation any surface water quality parameter that satisfies the requirements of a S/TRF for lake, reservoirs, rivers, and streams. Given that assessment and monitoring costs are a major impediment to a more complete implementation of surface water quality initiatives (e.g. TMDL program), this work may allow clean-water regulations to move forward more quickly by lowering the costs of assessing surface waters and allowing the design of more cost effective monitoring plans.
3. Using Reduced-Parameter Empirical Models to Estimate Inflows into Cascading Reservoir System and Capture Model Uncertainty

Introduction

Because the number of suitable dam sites has declined in the US, environmental concerns have grown, and costs have increased, there is a growing need to use existing reservoir systems more effectively. This need is even more acute in multipurpose reservoirs, such as those used for hydropower production, where strategies must be developed to properly manage low-inflow periods. There are over 330 hydroelectric dams with Federal Energy Regulatory Commission licenses expiring between 2012 and 2026, representing one of two peak relicensing periods (FERC, 2011). The 1986 federal law covering the relicensing of hydropower facilities (ECPA, 1986) requires that these new licenses give “equal consideration” to power production and non-power benefits, such as wildlife habitat, recreation, and water quality.

Hydrologic and reservoir management models allow reservoir performance to be calculated under a variety of management strategies. Although such models are important tools, they can be expensive to develop because of model complexity, the need to develop supporting datasets, and the high level of skill needed to properly construct, calibrate, and validate a model. Furthermore, it is critical that model uncertainty be effectively managed and reported (Beven, 2006) even though uncertainty analysis can add substantially to the modeling effort.

This work proposes an aggregate-level empirical model for a cascading reservoir system based on a century-long, but incomplete, rainfall dataset. This model sacrifices temporal and spatial resolution in favor of greater accuracy by focusing on aggregate-level results (Reckhow, 1999). This relaxes the challenges to uncertainty estimation due to
scaling and over parameterization while allowing iterative techniques to be used to robustly estimate model uncertainty. By reducing the model complexity and incorporating uncertainty the model can be applied where data scarcity or modeling resources would otherwise have prohibited model development.

The model estimates inflow in a system of six cascading reservoirs based on rainfall from 19 discrete meteorological stations. Markov Chain Monte Carlo techniques are used to characterize model parameter uncertainty using a 141-month record of monthly rainfall, reservoir withdrawals, and reservoir heights for model calibration and validation. Separately, the Expectation Maximizing algorithm is used to complete a rainfall history going back to 1893 to capture long-term model input variability. Uncertainty and variability information is retained throughout the model development, resulting in an inflow dataset that is represented probabilistically.

Reservoir management models can use inflow datasets directly, bypassing the need for hydrologic models for estimating inflows from rainfall. Historic datasets of inflow, however, are often of limited duration and include effects from land use conditions and water abstraction patterns that may no longer be relevant. Some corrections can be made using mass-balance calculations, but such corrections rely on knowing accurately withdrawals and other historic conditions over a long time period.

Using rainfall as the primary input to a model for estimating reservoir inflows is a viable alternative strategy because there is often a long historical rainfall dataset available and using rainfall reduces the need to understand historical withdrawals and reservoir operations. However this introduces uncertainty because the model relating rainfall to inflow is imperfect. Rainfall measurements are only taken at discrete points and error is introduced as they are applied to a watershed as a whole, thereby excluding known spatial variability in rainfall patterns. Recent models are able to use more spatially distributed precipitation data, such as the National Weather Service Next Generation Weather Radar NEXRAD and data from satellites but these data have only been readily available over the past decade (Tobin
& Bennett, 2008). Although the application of discrete rain gage data to a watershed is imperfect (Morrissey, et al., 1995), it is typically the only rainfall data available for long-duration historical datasets.

Despite a high level of complexity in distributed hydrologic models (perhaps requiring thousands of parameter values), these models may not closely represent natural processes (Beven, 2001; Beven, 1989; Vreugdenhil, 2002) because of issues related to nonlinearity, scale, and uniqueness. Nonlinearity is present in the mechanistic description of many models but cannot be fully implemented as the model form is necessarily imperfect. Errors from nonlinearity also occur from the application of mean parameter values developed for local phenomenon over a wide area to a model with finer resolution. Similarly, scaling requires using mechanisms and parameters developed on the element scale at much larger scales; such uses are not appropriate when assumptions are not transferable or there are spatial arrangements and flow paths that simply cannot be captured in a reasonable model (Bloschl, 2001). Uniqueness is most apparent when models developed and calibrated in one location are applied to other locations. If such models, in fact, modeled real processes, the model and parameter values could be applied to different watersheds; however it is found that the “uniqueness of place” (Beven, 2001) plays a strong role in model performance.

The model in this work addresses some of the above limitations by either reducing the number of required datasets, lowering uncertainty through aggregation, explicitly recognizing model limitations by removing the assumption that the model is mimicking real processes, and comprehensively including uncertainty in the model results.

**Methods**

A means by which the long-term inflow dataset is created is demonstrated in Figure 3.1. The inflow model structure captures fast- and slow-flow processes with parameters that vary by calendar month and also incorporates land use and land cover information. A limited dataset of inflows and corresponding precipitation is divided into even and odd years for
model calibration and validation. A long-term rainfall dataset is created and applied to the resulting validated model to obtain estimates of long-term inflow with uncertainty.

**Development of an Long-Term Rainfall Dataset**

Missing data for 19 meteorological stations (Figure 3.2) from 116-year incomplete rainfall dataset (Figure 3.3) are imputed using the Expectation Maximization (EM) algorithm. This algorithm estimates missing parameters by generating an underlying multivariate
probability distribution that maximizes the likelihood of the observed data. The probability distribution is then used to generate the expected values of the missing data and associated confidence intervals. This method is compared to a regression model that predicts missing data based on values at other stations. The EM algorithm, by generating a multivariate probability model consistent with the dataset as a whole (rather than a subset of data available for a regression model), produces better estimates for missing data.

Figure 3.2: Location of 19 meteorological stations used in the inflow model

Generally EM requires that the missing data be missing at random (Dempster, et al., 1977) and an appropriate underlying distribution be chosen. The former requirement is not met in this dataset as blocks of data are missing due to changes in monitoring programs over time; this is less critical as long as there is no correlation between data being missing and rainfall amount. Such a correlation would result in greater errors in the estimates of the mean and variance. The choice of a multivariate normal, if appropriate, facilitates the process as the underlying parameters, the vector of means and covariance matrix, are easily calculated and the expected values can be readily generated from these and the measured data.
Using the notation of Schneider (2001), we consider our dataset as an $n$ by $p$ matrix $X$, where $n$ is the number of months and $p$ is the number of meteorological stations. The vector of station overall rainfall means, $\mu$, is of length $p$ and the rainfall covariance matrix, $\Sigma$, is a $p$ by $p$ square symmetric matrix with ones along the diagonal. A particular record in $X$ is $x_i$, a 1 by $p$ vector containing the rainfall for a particular month, which can be decomposed into two vectors, one of actual rainfall values ($x_a$) and one where rainfall values are missing ($x_m$). These vectors vary in size from record to record according to the number of stations for which rainfall data is missing. For each month, the vector of means is divided corresponding into $\mu_a$ and $\mu_m$. Similarly, the covariance matrix can be decomposed into four submatrices by assembling elements according to the presence or absence of rainfall data in the record under consideration. The covariance between variables that have available data in the record is denoted $\Sigma_{aa}$, that between variables with missing data is $\Sigma_{mm}$, and two matrices describe the covariance between missing and available data: $\Sigma_{am}$ and $\Sigma_{ma}$. While the first two
submatrices are square, the later matrices often are not. An example for p=7 is shown in Figure 3.4.

\[
x_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \langle x_{i,3} \rangle \\ x_{i,4} \\ \langle x_{i,5} \rangle \\ x_{i,6} \\ x_{i,7} \end{bmatrix} \rightarrow x_a = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \langle x_{i,3} \rangle \\ x_{i,4} \\ \langle x_{i,5} \rangle \\ x_{i,6} \end{bmatrix}, \quad x_m = \begin{bmatrix} \langle x_{i,3} \rangle \\ \langle x_{i,4} \rangle \\ \langle x_{i,6} \rangle \end{bmatrix}
\]

where \( \langle \rangle \) denotes a missing rainfall value. The mean and covariance matrix are partitioned according the the pattern of missing rainfall data \( x_m \):

\[
\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \end{bmatrix} \rightarrow \mu_a = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{bmatrix}, \quad \mu_m = \begin{bmatrix} \mu_3 \\ \mu_4 \\ \mu_6 \end{bmatrix}
\]

\[
\Sigma = \begin{bmatrix} \Sigma_{1,1} & \ldots & \Sigma_{1,7} \\ \vdots & \ddots & \vdots \\ \Sigma_{7,1} & \ldots & \Sigma_{7,7} \end{bmatrix} \rightarrow \Sigma_{aa} = \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} & \Sigma_{1,5} & \Sigma_{1,7} \\ \Sigma_{2,1} & \Sigma_{2,2} & \Sigma_{2,5} & \Sigma_{2,7} \\ \Sigma_{5,1} & \Sigma_{5,2} & \Sigma_{5,5} & \Sigma_{5,7} \\ \Sigma_{7,1} & \Sigma_{7,2} & \Sigma_{7,5} & \Sigma_{7,7} \end{bmatrix}, \quad \Sigma_{mm} = \begin{bmatrix} \Sigma_{3,3} & \Sigma_{3,4} & \Sigma_{3,6} \\ \Sigma_{4,3} & \Sigma_{4,4} & \Sigma_{4,6} \\ \Sigma_{6,3} & \Sigma_{6,4} & \Sigma_{6,6} \end{bmatrix}, \quad \Sigma_{am} = \Sigma_{ma}^T
\]

Figure 3.4: Partitioning of vector of means, \( \mu \), and covariance matrix, \( \Sigma \), according to pattern of missing values in vector of monthly rainfalls, \( x \).

The relationship between the existing and missing values can be described by Equation 3.1 where \( B \) is a matrix of regression coefficients and \( e \) is an error, or residual, term that is distributed as a multivariate normal distribution with means of zero and a covariance matrix \( C \). The maximum likelihood estimates of \( B \) and \( C \) can be calculated from the covariance submatrices conditional on the assumption that the overall covariance matrix is accurate.
\[ x_m = \mu_m + (x_a - \mu_a) B + e \]  
Equation 3.1

\[ B = \Sigma_{aa}^{-1} \Sigma_{am} \]  
Equation 3.2

\[ C = \Sigma_{mm} - \Sigma_{ma} \Sigma_{aa}^{-1} \Sigma_{am} \]  
Equation 3.3

Since the error has a mean of zero and is distributed normally, the expected values for the missing data is calculated from Equation 3.1 without the error term

\[ x_m = \mu_m + (x_a - \mu_a) B \]  
Equation 3.4

The expectation maximization algorithm begins with an initial estimate of the means and covariance matrix for the \( p \) stations generated using the known rainfall values. Record by record, the mean and variance are parsed according to the pattern of missing data and the regression coefficients, \( B \), and residual variance, \( C \), are calculated. The missing data are then estimated using Equation 3.4 and the process is repeated for each month until the missing data for all months are estimated. All months with identical patterns of missing data can be processed in a single step since the \( B \) and \( C \) matrices depend only on the pattern of missing data.

The complete rainfall dataset, with both the original and the imputed data, is then used to calculate a new vector of means and a covariance matrix. Because the imputed data are simply the expected values of a distribution, the variance in the missing data is underestimated; to get a better estimate of overall covariance, the covariance of the residuals, \( C \), is included. The imputed data is subsequently deleted and the missing data is again estimated with the new mean and covariance information. This is repeated until the estimates in the mean no longer change significantly from iteration to iteration.

Since the monthly rainfall data is not distributed normally, as is required for the implementation described here, the data are transformed by adding a constant and then taking the logarithm such that:

\[ R^* = \log(R + \alpha) \]  
Equation 3.5

Where \( R^* \) represents the transformed data, \( R \) the original data, and \( \alpha \) a constant chosen to maximize normality. This offset also allows the log-transformation of zero values while the
log-transform increases the importance of the accuracy in the estimates of low rainfall values. Once the imputed data are calculated, they are back-transformed to obtain monthly rainfall values.

**Simplified Rainfall-Inflow Model**

The empirical reservoir inflow model assumes two major components of reservoir inflow: that which has a lag (sometimes called “slow flow”, analogous to baseflow) and that which occurs quickly (“fast flow”, analogous to runoff). Runoff occurs during a precipitation event and may continue for a short period thereafter, typically up to a day or two. Baseflow is the contribution from groundwater or snowmelt and occurs outside of precipitation events and may have a dependence on rainfall with a lag of weeks or months as it travels through the subsurface.

Runoff can be related to rainfall through a runoff coefficient (Novotny, 2003) which depends on several factors, including land cover/land use, soil type, and slope. For this model, the application of the runoff coefficient to rainfall produces what is termed “potential runoff”, an estimate of runoff that is an independent variable in the overall empirical inflow model.

The model divides the Catawba Basin into six large catchments, which are divided into 261 subcatchments. For each subcatchment the fraction of land in each of several types of landcover, average slope, and total area is known. Each subcatchment is also associated with up to three meteorological stations within 30 kilometers of the subcatchment centroid and which are used to estimate subcatchment rainfall using an inverse-distance weighted average. The potential runoff from a subbasin is calculated as

\[
\text{Potential Runoff} = \text{Subbasin Area} \cdot \text{Precipitation} \sum_{n=1}^{n\text{Types}} (f_n \cdot C_n) \quad \text{Equation 3.6}
\]

where the summation is over the types of land use present in the subbasin and, for each land use, the fraction of the basin area (f) and corresponding runoff coefficient (C, Figure
A.1) are known. The use of runoff coefficients in this model essentially scale the rainfall contribution based on land use and land type, providing higher runoff values for subcatchments where impervious landcover dominates or slopes are high. The potential runoff from the subcatchments is aggregated over the six reservoir catchments to provide overall values for each of the Catawba reservoirs.

The reservoir inflow model combines regression models for both runoff and base flow. Runoff is assumed to be directly related to potential runoff, with the constant of proportionality allowed to vary by month. This captures variation due to average temperature (including precipitation falling as snow versus rain), evaporation, and unaccounted-for uses that vary seasonally. Eleven dummy variables (isMonth_i, i=2 to 12) are used to capture the monthly variation. The value of isMonth_i is 1 if i corresponds to the simulation month (Feb=2, March=3, etc.) and zero otherwise.

\[ \text{runoff} = \text{Potential Runoff} \left( a_1 + \sum_{i=2}^{12} (a_i \text{isMonth}_i) \right) \]  
Equation 3.7

After exploring several models for baseflow, it is determined that the prior month’s rainfall and rainfall two months prior are the dominant predictors of baseflow. The baseflow model is shown in Equation 3.8.

\[ \text{baseflow} = \gamma_0 + \gamma_1 \text{Rain}_{m-1} + \gamma_2 \text{Rain}_{m-2} \]  
Equation 3.8

These two models are combined to create an overall inflow model, at which point the regression coefficients can be combined into a common nomenclature. The resulting model (Equation 3.9) has 15 parameters and includes an error term (e) that is normally distributed with a zero mean.

\[ \text{inflow} = \beta_1 + \beta_2 \text{Rain}_{m-1} + \beta_3 \text{Rain}_{m-2} + \text{Potential Runoff} \left( \beta_4 + \sum_{i=2}^{12} (\beta_{i+3} \text{isMonth}_i) \right) + e \]  
Equation 3.9
Creating a stochastic model and updated parameter uncertainty

A Markov Chain Monte Carlo method is chosen to specify the model parameter distributions and capture parameter uncertainty in the inflow model. The concept of equifinality (Beven & Binley, 1992) asserts that because of the imperfections in modeling, there is no one ideal or optimal set of model parameter values; rather there are many sets of parameters that describe the modeled system. This concept underlies the creation of the generalized likelihood uncertainty estimation (GLUE) methodology which combines Monte Carlo techniques with Bayesian-like updating (Beven & Binley, 1992) to estimate parameter and model output uncertainty. Using GLUE, models are run many times, each time with a different set of parameters drawn from a specified prior distribution. Each model result is assessed based on an informal likelihood measure, typically either the ratio of explained to observed variance (similar to Nash-Sutcliffe (Nash & Sutcliffe, 1970)) or a related exponential form (Lamb, et al., 1998). Some parameter sets are removed if the results are considered non-behavioral; Beven and Freer (2001), for example, use a Nash-Sutcliffe value below 0.6 as the removal criterion. Posterior distributions for the parameters and outcome variables can then be generated with the remaining parameter sets by creating cumulative distribution functions by summing the goodness of fit values in order of increasing parameter value and normalizing the sum to a value of one.

When a more formal likelihood function, based on statistical principles and Bayesian concepts (Kuczera & Parent, 1998; Vrugt, et al., 2009), is used this procedure leads to Bayesian Monte Carlo (BMC) and Markov chain Monte Carlo (MCMC) methods, but there remains some controversy as to if formal likelihood functions are always an improvement (Mantovan & Todini, 2006; Beven, et al., 2007).

Criticisms of BMC include the large number of samples that are required to adequately sample a joint probability distribution of many dimensions. These distributions can have tight local maxima as a result of correlated parameters and these maxima can be overlooked with inadequate sampling (Qian, et al., 2003). When models become large
maintaining the sample density quickly becomes very difficult and more so with sophisticated models having relatively long runtimes.

Markov Chain Monte Carlo (MCMC) methods diverge from other Bayesian methods in that they do not begin with picking parameter sets from a prior distribution. Instead MCMC generates parameter sets in such a way that the distribution of the parameter sets eventually converges to their posterior distribution. The assembly of the selected parameter sets selected during the MCMC process is then used to describe the posterior distribution of the parameters and outputs and calculate any desired summary statistics. Marshall et al. describe the application of MCMC methods to inflow models (2004) in detail for an 11-year dataset for a single catchment in Australia.

A Markov chain process is one where the probability of moving to the next step in the chain relies solely on the current state variables (in this case, the parameter set.) The MCMC process begins with an arbitrary state, Y, of initial parameter values. A candidate state, Y*, is chosen from a somewhat arbitrary distributions, Q(Y*|Y), that looks like the posterior distribution, if possible, but from which it is possible to draw samples. The key to the method is that the probability of accepting the candidate state depends on two Bayesian-like updating steps. The first calculates the kernel of Bayes rule for the likelihood of obtaining each of the current and candidate parameter sets given the data (D). These likelihood values are expressed as \( f(Y|D) = P(Y)P(D|Y) \) and \( f(Y^*|D) = P(Y^*)P(D|Y^*) \). Second, these likelihoods are used to update the distribution used for selecting \( Y^* \), \( Q(Y^*|Y) \), giving

\[
R(Y^*|Y) = f(Y^*|D) Q(Y|Y^*)
\]

and

\[
R(Y|Y^*) = f(Y|D) Q(Y^*|Y)
\]

The candidate parameter is then chosen based on the ratio of the updated probability of moving to \( Y^* \) from \( Y \) to that for moving to \( Y \) from \( Y^* \). That is to say, the move to \( Y^* \) is done with a probability \( g(Y^*) \) shown in Equation 3.10.
\[ g(Y^*) = \min \left( 1, \frac{R(Y^*|Y)}{R(Y|Y^*)} \right) = \min \left( 1, \frac{f(Y^*|D)Q(Y|Y^*)}{f(Y|D)Q(Y^*|Y)} \right) \quad \text{Equation 3.10} \]

In this way the "walk" in the Markov Chain is more likely to proceed in directions of parameter sets that better fit the data, and thus regions of good model fit are sampled more often than regions of poor fit.

MCMC is implemented using WinBUGS (Spiegelhalter, et al., 2003) and the reduced parameter set permitted all reservoirs to be modeled simultaneously; with 15 parameters for each reservoir model and six reservoirs being modeled, there are 90 parameters in total. The reservoirs were fit simultaneously because the reservoir flows are interdependent and it is therefore important to generate cohesive parameter sets.

**Estimating Reservoir Inflows from Reservoir Management Data**

Daily reservoir levels, stage-reservoir area relationships, and managed reservoir releases for hydropower, spillway flows, and municipal withdrawals are used to calculate the input to each reservoir using mass balance approach (Equation 3.11).

\[ \Delta \text{Storage} = \text{inflows} - \text{outflows} \quad \text{or} \quad \text{inflows} = \Delta \text{Storage} + \text{outflows} \quad \text{Equation 3.11} \]

The change in storage (\( \Delta \text{Storage} \)) was calculated directly from daily reservoir level and height-volume relationships. Outflows include municipal and industrial withdrawals, releases for hydropower production, minimum required flows, and flows over the spillway. Evaporation losses were also estimated from monthly average pan evaporation data and included as an outflow.

**Validation of Imputed Rainfall Data**

In this work, two methods are used to validate the rainfall imputation method: cross-validation and multidimensional scaling (MDS). For the cross-validation analysis, 1% of the observed data are removed at random from the dataset prior to the imputation of missing
rainfall data. Rainfall estimates corresponding to the removed data are compared with the original values using the three fitness criteria used for the model results: NS, RMSER, and PBIAS (Equation 3.13 through Equation 3.15). This is repeated 100 times to provide a distribution of fitness measures for the EM algorithm and, for the sake of comparison, again for imputation using linear regression.

The linear regression method uses existing data to build regression models predicting the missing rainfall value based on other data available in the given month. Regressions models using all possible combination of stations are evaluated and the regression model with the highest explanatory power (as measured by $R^2$) is used to predict the missing value as long as the $R^2$ was above 0.4.

MDS uses difference or similarity information to generate plots where more similar objects are closer together than less similar ones; it is related to cluster and principal component analysis. In this case a correlation matrix, calculated from the covariance matrix generated from the EM algorithm, was used as the similarity matrix with the goal of comparing the resulting MDS plots to the spatial distribution of meteorological stations. The result from the EM algorithm is compared with that based on a correlation matrix generated from only the existing data.

The scale and the orientation of the raw MDS results are arbitrary so for this analysis the coordinates are modified with a linear transform to make them correspond to latitude and longitude values. The transformation, shown in Equation 3.12, allows an offset of the origin, a rotation of the orthogonal axes, and scaling along each axis. In this equation, $x$ is the original coordinate, $M$, a 2x2 matrix that scales and rotates, and $b$ is a length-2 vector containing the offset values.

$$x_i = b + M \times$$

Equation 3.12

The values of $M$ and $b$ are chosen to minimize the difference between the MDS results and the actual station coordinates.
Quantification of Inflow Model Fitness

Overall model results are quantified using metrics described by Moriasi et al. (2007): Nash-Sutcliffe Efficiency (NS) (Nash & Sutcliffe, 1970), root mean square error observation standard deviation ratio (RMSER), and percent bias (PBIAS). PBIAS is also referred to as mass-balance error (Tobin & Bennett, 2008). These values are defined as:

\[
NS = 1 - \frac{\sum_{i=1}^{n}(Q_{obs}^i - Q_{sim}^i)^2}{\sum_{i=1}^{n}(Q_{obs}^i - Q_{obs,ave}^i)^2}
\]  
Equation 3.13

\[
RMSER = \sqrt{\frac{\sum_{i=1}^{n}(Q_{obs}^i - Q_{sim}^i)^2}{\sum_{i=1}^{n}(Q_{obs}^i - Q_{obs,ave}^i)^2}}
\]  
Equation 3.14

\[
PBIAS = \frac{\sum_{i=1}^{n}(Q_{obs}^i - Q_{sim}^i)}{\sum_{i=1}^{n}Q_{obs}^i}
\]  
Equation 3.15

Where \(Q_{obs}^i\), \(Q_{obs,ave}^i\), and \(Q_{sim}^i\) are the individual observed, average observed, and individual simulated inflow values respectively.

NS describes the amount of variation explained by the model and is analogous to an \(R^2\) value for a regression model. It can take on values between negative infinity and one. Acceptable values of NS generally are above 0.5 (Moriasi, et al., 2007; Tobin & Bennett, 2008), although values as low as 0.36 are sometimes characterized as marginally acceptable (Motovilov, et al., 1999). RMSER is a normalized measure of root mean square error and acceptable model values should be below 0.7 (Tobin & Bennett, 2008). PBIAS measures the bias in the estimate, is ideally zero, and generally should not exceed 25% (Moriasi, et al., 2007).
Analysis and Results

Development of an Extended Rainfall Dataset

Missing rainfall data are imputed using the EM algorithm with 1386 records (the number of months) and 19 variables (the number of meteorological stations). The value for $\alpha$ of 1.8 (Equation 3.5) is determined by maximizing the normality of the transformed data; QQ plots are shown in Figure B.1 for all 19 stations. The EM algorithm is initialized using the mean of the existing data and a covariance matrix estimated using whatever paired meteorological station rainfall data existed for each corresponding matrix element.

The EM algorithm converged in 60 iterations (Figure B.2) and after the last iteration, in accordance with Equation 3.1, random errors ($e$) are added from a multivariate distribution with a mean vector of zeros and covariance $C$ (the residual variance). The data are back-transformed to obtain monthly rainfall data from the imputed values, which are represented in Figure 3.5.

Figure 3.5: Representation of historical dataset for each meteorological station including imputed data. Color indicates the monthly rainfall amount (cm).
As shown in Figure 3.6, the cross-validation results for the EM method exceed the criteria for acceptable model performance and are substantially better than those for the linear regression method. There is good explanatory power and little bias in the estimates. Since the uncertainty in the input dataset is propagated through the model and contributes to the uncertainty of the inflow estimate, such a level of performance is necessary to obtain usable results.

Figure 3.6 show the distribution of the fitness measures for the EM and linear regression methods. For each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. Outliers are those points falling more than 1.5 times the difference between the 25th and 75th percentiles from either edge of the box. This corresponds to approximately 99.3 percentile coverage if the data are normally distributed.

Figure 3.6: Measures of fitness for imputed data using EM and linear regression. X-axis labels describe the fitness measure (NS, RMSER, or PBIAS) and the imputation method (EM = Expectation Maximization, LR = Linear Regression).
Figure 3.7 shows the station locations (numbered) along with the MDS results (smaller circles) with corresponding locations connected by a line for the initial covariance matrix (left) and the covariance matrix after applying the EM algorithm (right). The transformation from MDS to geographic coordinates is shown in Figure B.3. It can be seen that information from the final correlation matrix was sufficient to capture most of the geographical information of the station locations and there is substantial improvement over the initial correlation matrix estimate. The root means square error distance is approximately 40 kilometers prior to the EM algorithm and 19 kilometers afterwards.

![Figure 3.7: Multidimensional scaling result based on covariance matrix of monthly station rainfall generated by EM algorithm. Plot on the left uses covariance matrix prior to EM, that on the right after EM](image)

**Estimating Reservoir Inflows**

A physical description of each reservoir was obtained from Duke Energy records (made available as part of the FERC renewal process) and from Duke Power's WARMF model. The parameters needed to model reservoir inflows include the stage/area relationships, and normal minimum and maximum lake surface levels (Table C.1). The reservoir working capacity was calculated by integrating the stage/area relationship between the normal inflows are estimated directly from 141 months of monitoring data supplied by Duke Power using Equation 3.11. Change in storage is calculated from the change in
reservoir height and known outflows include hydroelectric flows, non-hydropower dam releases, withdrawals for the towns and withdrawal from power stations shown in Table C.2. Average monthly pan evaporation rates from a long-term pan evaporation site in Coweeta North Carolina (Figure C.1) are used to estimate evaporation losses.

The inflows to the reservoirs are cumulative because of the cascading nature of the system. Attempts to model net reservoir inflows produce unreliable results because the reservoirs are of very different sizes and the calculation of the net inflows to the smaller reservoirs, which is done by subtracting upstream inflow from overall inflow, is dominated by errors in the calculation of the inflows to the larger upstream reservoirs. For example, the net basin area for Mountain Island Lake is only 4% of the total upstream drainage basin area (Table C.1) so errors as small as 4% in the upstream flow calculation would create an error of 100% for net Mountain Island Lake inflows. Because of this interdependence, the cumulative values were used in a model estimating all inflows simultaneously.

Potential Runoff

The model results are generally insensitive to the value of the runoff coefficients used to calculate potential runoff because developed land is only a small fraction of the total land use and the values for the runoff coefficients for less developed land types are similar. Accordingly, uncertainty in runoff coefficient specification is not included in the model in order to permit greater modeling efficiency to capture uncertainty from other sources. Instead a single value for each land type is used for all simulations.

Model insensitivity is shown with dotty plots (Figure 3.8 and Figure A.1Figure A.1) where model performance (Nash-Sutcliffe) variation is plotted against parameter values for 5000 Monte Carlo model runs. Although individual parameters do not drive model performance independently, dotty plots show regions where high or low performance is unlikely. The model is only weakly sensitive to runoff coefficients for three land types (reproduced in Figure 3.8): pasture with a slope greater than 0.03, cultivated land with slope
greater than 0.03 and low-intensity development, as seen by the increased number of low-performing models at either end of the input parameter distribution. However, the low-performing simulations would likely be removed as “non-behavioral” (Beven & Freer, 2001) and high-performance is possible at any runoff coefficient value, as seen by the consistently high values at the top of each plot.

The apparently hard limit seen at the top of each dotty plot approaches the maximum achievable value for the overall model. During calibration (specification of model parameters in Equation 3.9), the model parameters are adjusted to attain the best fit, compensating for variation in runoff coefficients. Table A.1 shows the cumulative distribution of NS values; if the hard limit was, in fact, uniform, the distribution would be vertical at the highest NS values instead of having a slope.

![Figure 3.8: Scatter plot of Nash-Sutcliffe efficiency measure for three of 24 runoff coefficients: pasture with a slope greater than 0.03, cultivated land with slope greater than 0.03 and low-intensity development](image)

**Empirical Inflow Model**

Six of the seven Catawba reservoirs in North Carolina are modeled using Equation 3.9. The seventh, Lake Wylie, is excluded because it boarders North and South Carolina and the sources and withdrawals are less well defined than for the other reservoirs.

The model was calibrated using MCMC on one-half of the dataset, composed of odd-numbered years, and validated using data from even-numbered years. The choice of odd/even years, as opposed to first-half/last-half, was made because the lowest inflow periods were all in the last half of the dataset so that using every-other year provided a wider range of rainfall and inflow values with which to calibrate and validate the model.
For calibration, two chains were run in order to check for convergence; one chain was initialized with all parameters set to zero and the other chain initialized with parameter values determined from a simple regression model. Convergence typically occurred in fewer than 50 iterations; however some parameters took up to 500 iterations for convergence; Figure 3.9 shows the chains for two of the 90 model parameters.

The first 1000 iterations are discarded and the model is run for 2000 iterations to generate the final parameter sets. Posterior distributions for one parameter are shown in Figure 3.10 and appear approximately normal (Appendix C shows all parameter distributions); the distribution on the left was generated with a normal prior and one on the right with a uniform prior. They show nearly identical results, indicating robustness in the posterior distributions. The parameter sets of the 2000 iterates are retained and used as draws from the joint distribution of parameter values for subsequent model runs.

![Figure 3.9: MCMC burn-in using two chains for two of the 90 variables (Equation 3.9) showing convergence of the parameter values.](image)
Figure 3.10: Posterior distributions Lake Norman $\beta_{10}$ showing typical result. The plot on the left column had a normal prior, that on the right a uniform prior.

The model fits the existing data well, with all three measures of fitness falling within the range of acceptability for the calibration and validation estimates (Table 3.1). The bias estimate remains at a low value for the validation dataset, but there is a reduction in the Nash-Sutcliffe value and a corresponding increase in the measurement of model error. Values for the individual reservoirs (not shown) all fall within the acceptable range and have similar values and trends as the overall measures shown.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Ideal Value</th>
<th>Acceptable Criteria Value</th>
<th>Calibration Value</th>
<th>Validation Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS</td>
<td>1</td>
<td>&gt;0.7</td>
<td>0.85</td>
<td>0.76</td>
</tr>
<tr>
<td>RMSER</td>
<td>0</td>
<td>&lt;0.7</td>
<td>0.39</td>
<td>0.49</td>
</tr>
<tr>
<td>PBIAS</td>
<td>0</td>
<td>&lt;0.25</td>
<td>0.01</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 3.1: Evaluation of model fitness using NS, RMSER, and PBIAS measures.

Figure 3.11 shows the measured values (dots) along with the model results represented as lines and showing the 95% credible interval. Figure 3.12 compares the model estimates with the actual values for the validation dataset.
Figure 3.11: Model results versus known flows, million cubic meters
Dots show measured data, the line model results, and the red dashed lines the 95% credible interval.
Figure 3.12: Median predicted values versus actual values for each reservoir using the validation dataset. A perfect model would have values falling along a line with slope one and beginning at the origin.

The extended inflow dataset needed for evaluating reservoir performance is created by applying the long-term rainfall dataset to the 2000 parameter sets and adding to each result error drawn at random according to the model error estimate (‘e’ in Equation 3.9). The result is a distribution of 2000 inflow values for each of 1386 months represented in the rainfall dataset capturing uncertainty in model results.

**Conclusions**

The role of rainfall-inflow models in assessing reservoir performance is limited by the required cost and expertise of creating accurate detailed models. Capturing uncertainty in distributed mechanistic rainfall-inflow models can be difficult because of long run times that limit the use of iterative techniques and the large number of parameters that need to be evaluated. Reservoir inflow estimates also require extended input datasets to capture historical variability.
In this work inflows into a system of six cascading reservoirs are estimated using an aggregate-level model that reduces the spatial and temporal resolution as well as the number of model parameters. This type of model allows for high model accuracy and the assessment of model uncertainty using iterative methods despite the complexity of the underlying system.

The monthly temporal resolution allows the analysis of outcomes, such as drought, that take place over months and years. Low-inflow conditions are a primary concern in many systems as they determine the frequency and extent of conflicts between the multiple uses of the water, which is often seen in systems providing water for municipal supply, hydropower production, and other industrial uses, such as the Catawba.

Model performance evaluated on a validation dataset with traditional measures of modeling efficiency, mean squared error, and bias, is high despite having only 15 model parameters per reservoir. Parameter uncertainty is captured in the 2000 parameter sets created using MCMC methods and additional model uncertainty is quantified in the estimate of model residual variance. The parameter sets and variance estimates can be applied to a rainfall dataset to produce a monthly distribution reservoir inflow estimates representing overall model uncertainty.

A coherent historical 116-year (1386 month) rainfall input dataset is created that is shown to have sufficient accuracy to provide meaningful inflow estimates. This is achieved through statistical techniques that impute missing rainfall values based on a multivariate-normal model of transformed rainfall data. This method is demonstrated to have a high explanatory power and little bias and to perform better than an alternative method using linear regression.

Because the extended inflow dataset generated from the aggregate-level model and extended rainfall dataset captures model uncertainty, it provides a more comprehensive basis for evaluating management practices and the performance of the Catawba reservoir system during low-inflow conditions, as is described elsewhere (LoBuglio, 2011).
4. **Using a Stochastic Cascading Reservoir Model to Estimate the Frequency and Extent of Water Resource Conflicts**

*Introduction*

Although there is a rich literature concerning modeling uncertainty in hydrologic models (Beven, et al., 2007; Freer, et al., 1996; Kuczera & Parent, 1998), there are fewer examples of studies on the effect of uncertainty in reservoir inflows on reservoir performance. Much of the existing literature focuses on creating stochastic inflow datasets that match statistical parameters of observed data (Efron, 1979; Srinivas & Srinivasan, 2005) or look at how to use uncertain weather predictions to make immediate operational decisions (Georgakakos, et al., 1998; Georgakakos & Graham, 2007). It has been decades since Wood (1978) discussed the importance of including uncertainty as applied to the problem of reservoir-storage-yield analysis, however the recent example of a study of the Catawba-Wateree Federal Energy Resource Regulatory Commission (FERC) relicensing study (HDR Engineering, 2006) shows that uncertainty can be omitted for even high-profile studies.

Including uncertainty in the estimate of reservoir inflows adds both higher- and lower-flow events to the scenarios being evaluated without significantly affecting the mean inflow. However, because of the asymmetry in storage behavior, including uncertainty in the inflow estimates results in an increase in frequency of low reservoir levels. The Catawba-Wateree reservoir system, like many systems designed for hydropower production, are run with operational guide curves that set the target reservoir heights close to the full pond level whenever flood control is not a concern. This maximizes the head available for hydropower production and allows some cushion for providing minimum release flows and municipal supply during low-flow times. When there is abundant flow, excess water is used for
hydropower production or otherwise released, typically leaving the reservoir height where it started regardless of the amount of excess flow. However when there is an unexpected low-flow event, the reservoir height can end lower than the target height and this deficit is carried over to the next time period. In this way low-flow events persist in the storage system while higher-flow events do not.

There are over 330 hydroelectric dams with licenses expiring between 2012 and 2026, representing one of two future high-activity relicensing periods (Figure 4.1). The 1986 federal law covering the relicensing of hydropower facilities (ECPA, 1986) requires that these new licenses give "equal consideration" to power production and non-power benefits, such as wildlife habitat, recreation, and water quality. It is therefore critical to better characterize reservoir performance by accounting for inflow uncertainty in performance models.

Figure 4.1: Number of expiring FERC hydroelectric permits by year (FERC, 2011)

There have been water resource systems models that incorporate a probabilistic approach. Jacobs et al (1998) demonstrates a model using a summary of long term data in the form of linearized streamflow duration curves to maximize water withdrawals from an unregulated basin. Pereira and Pinto (1985) use stochastic methods to solve an optimization problem for a multi-reservoir hydroelectric system, and Cai et al. (2001) has used genetic algorithms to find optimal solutions for complex systems. These models incorporate
uncertainty in historical streamflow records, but do not account for many uncertainties that would be important to decision makers, such as those associated with future water demand.

Evaluating uncertainty of complex systems can be done with Monte Carlo tools; however the number of iterations required for robust results is high and predicates the need for models that can be evaluated quickly, computers that are very fast, or limits to the length of the dataset being evaluated. Empirically-based aggregate-level modes are good candidates for this work because they reduce the number of system variables and the spatial and temporal resolution while still providing accurate aggregate-level results. In this work, applying a management strategy using a linear programming methodology on a 116-year simulation for 2000 predetermined sets of reservoir inflows at a monthly time step took about 8 hours on a modern desktop computer. Models with much greater spatial or temporal resolution would quickly become impractical. The probability-based output provides a broader understanding of possible outcomes, including extreme events, and the frequency with which different management strategies come into play.

The Catawba River is an interesting system to model because of its importance, complexity, and the existence of competing uses for its water. The Catawba basin has the largest population density of North Carolina's 17 basins, in large part because it includes more than 70% of the Charlotte metropolitan area. Nearly 830,000 people are supplied with surface water by public water supply systems that make withdrawals from 5 of the Catawba reservoirs (about 146 million gallons per day (MGD) in 1997). The total net water withdrawals for the six North Carolina reservoirs is expected to grow from the 2008 rate of 217 MDG to 461 MGD in 2058. The net public water supply component of these withdrawals is projected to grow disproportionally, accounting for 32% of net withdrawals in 2008 and projected to rise to 52% by 2058 (HDR Engineering, 2006).
**Methods**

**Reservoir Management Model**

Figure 4.2 shows the components of the reservoir management model; the details of each box are developed later in this section. The model begins each month with the current reservoir level and an estimate of reservoir inflows and catchment-area rainfall from a stochastic reservoir inflow model developed previously (LoBuglio, 2011). Reservoir level provides information needed to estimate evaporation losses and to decide if the reservoir is being operated under normal rules or under rules governing low inflow conditions (see Low Inflow protocol (LIP) later in this section.) Rainfall is an input to a stochastic model determining water demand. Evaporation losses are subtracted from inflow to generate the net monthly inflow while the management rules (normal or low-inflow conditions) determines if and how the demand and reservoir operating parameters (target levels and minimum release requirements) are modified. The net inflows, adjusted demand, and operating rules are used to decide the amount of water withdrawn and the distribution of water among the reservoirs.

![Figure 4.2: Schematic of Reservoir Management Model](image-url)
Stochastic Inflows

Stochastic reservoir inflows are generated from an empirical model taking rainfall inputs from 19 geographically distributed stations over 116 years, as described in section 3. The original rainfall dataset was far from complete; missing values are imputed using the Expectation Maximization algorithm, with imputed data drawn from a distribution of possible values. The model relating rainfall to reservoir inflow uses Markov Chain Monte Carlo (MCMC) to generate distributions in the underlying parameter values based on 140 months of detailed observed rainfall and inflow data. The rainfall imputation method and inflow model show high levels of modeling efficiency, with Nash-Sutcliffe values of 0.80 and 0.75 respectively. The inflows used in this work are a set of 2000 116-year monthly inflow values generated from the extended rainfall dataset and rainfall-inflow model.

Demand

Average annual demand associated with each reservoir has been calculated (Duke Power, 2003) for 10 year intervals from 2008 to 2058 as part of the Catawba-Wateree FERC relicensing effort. These data were assembled from public water supply plans, population growth estimates, and data from the U. S. Geological Survey. To account for demand variation from the annual average due to rainfall and season, a model was constructed and calibrated based on twelve years of daily demand data from the Charlotte-Mecklenburg water utility. The year 2058 is chosen as an endpoint because it is the year in which the FERC license expires.

The per-capita monthly demand is modeled using the current month’s rainfall, prior month’s rainfall, and dummy variables to capture monthly baseline variation due to seasonal changes (Equation 4.1). Variables are normalized by the long-term average values for demand and monthly rainfall and the regression is performed on half of the data (odd-numbered years) and validated on the other half of the dataset (even numbered years); validation showed that the model explains 85% of the demand variance.
Water Allocation

For each reservoir a guide curve specifies an average target water level and target minimum and maximum water levels throughout the year. There are also critical levels (Table 4.5) below which intakes to power plants or water treatment plants begin to be affected. Figure 4.3 shows the monthly variation in guide curves and critical reservoir levels for Lake James.

In this work, given the monthly reservoir inflows, demand, and release requirements, linear programming is used to determine how water is withdrawn from and distributed among the reservoirs (Labadie, 2004). The linear programming model uses mass balance constraints (Equation 4.2) and bounds on the variable values to ensure physically meaningful (feasible) solutions. The linear program includes six variables (Table 4.1) for each reservoir, for a total of 36 variables. Inflows, demand, and evaporation for each reservoir are generated from separate stochastic models.

\[
\frac{\text{Monthly Demand}}{\text{Average Monthly Demand}} = \alpha_1 + \sum_{i=2}^{12} \left( \alpha_i \text{ isMonth}_i \right) + \alpha_{13} \frac{\text{Monthly Rainfall}}{\text{Average Monthly Rainfall}} + \alpha_{14} \frac{\text{Prior Rainfall}}{\text{Average Monthly Rainfall}}
\]

Equation 4.1

Where:
- Monthly Demand = per-capita demand for a particular month.
- Average Monthly Demand = average per-capita demand for entire dataset
- isMonth = a dummy variable that is 1 for the month corresponding to month being estimated and 0 otherwise
- Monthly Rainfall = rainfall (mm) for the month being estimated
- Prior Rain Fall = rainfall (mm) for the month prior to the one being estimated
- Average Monthly Rainfall = average monthly rainfall (mm) for the entire dataset.

Water Allocation

For each reservoir a guide curve specifies an average target water level and target minimum and maximum water levels throughout the year. There are also critical levels (Table 4.5) below which intakes to power plants or water treatment plants begin to be affected. Figure 4.3 shows the monthly variation in guide curves and critical reservoir levels for Lake James.

In this work, given the monthly reservoir inflows, demand, and release requirements, linear programming is used to determine how water is withdrawn from and distributed among the reservoirs (Labadie, 2004). The linear programming model uses mass balance constraints (Equation 4.2) and bounds on the variable values to ensure physically meaningful (feasible) solutions. The linear program includes six variables (Table 4.1) for each reservoir, for a total of 36 variables. Inflows, demand, and evaporation for each reservoir are generated from separate stochastic models.
Figure 4.3: Target and other operationally important reservoir levels for Lake James. Target Storage Index is based on storage in all reservoirs; the estimated contribution from Lake James is shown.

Storage Volume \(_{\text{final}} - \) Storage Volume \(_{\text{initial}} = \) Inflow – withdrawals – releases – evaporation

Operating rules are implemented through having the linear program optimize a weighted sum of the uses for water (storage or release) (Equation 4.3); the highest weight is given to ensure the volume below the critical level is filled, followed by filling demand and minimum release requirement, then storage up until the minimum target level, then storage up to the target level. The specific values of the weights are arbitrary and were chosen to consistently provide the expected outcome over a large number of test cases.

\[
\sum_{res=1}^{6} (6V_{d,res} + 5D_{r,res} + 4MR_{r,res} + 3V_{c,res} + 2V_{w,res} + 0.5R_{o,res})
\]

The actual operation of the Catawba Reservoirs is informed by proprietary software (CHEOPS\textsuperscript{TM}) and so it is difficult to compare the method in this work with that implemented by the hydroelectric plant operator. However, for the FERC study the CHEOPS\textsuperscript{TM} model was
modified to include operation of the system under the new operating rules and to include the implementation and impacts of the Low Inflow Protocol (HDR Engineering, 2006). These rules and protocol are also captured in model developed for this work, so some comparison is justified.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Symbol</th>
<th>Description</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reservoir Dead Volume</td>
<td>$V_d$</td>
<td>Volume of water stored below the &quot;critical&quot; water level.</td>
<td>Dead volume of reservoir</td>
</tr>
<tr>
<td>DemandFilled</td>
<td>$D_f$</td>
<td>The amount of demand that is met by reservoir withdrawals.</td>
<td>Predicted demand for the month</td>
</tr>
<tr>
<td>Minimum Release Filled</td>
<td>MR$e$</td>
<td>Amount of water going towards meeting the minimum release requirement.</td>
<td>Minimum release requirement for the month</td>
</tr>
<tr>
<td>Reservoir Critical Volume</td>
<td>$V_c$</td>
<td>Volume of water between the dead volume and the minimum target reservoir level</td>
<td>Volume between minimum target and critical level</td>
</tr>
<tr>
<td>Reservoir Working Volume</td>
<td>$V_w$</td>
<td>Volume of water between the target reservoir height and the minimum working height</td>
<td>Volume between target and minimum reservoir levels.</td>
</tr>
<tr>
<td>Other release</td>
<td>$R_o$</td>
<td>Any release to a downstream reservoir above the minimum required release.</td>
<td>No upper bound</td>
</tr>
</tbody>
</table>

Table 4.1: Variable Descriptions for Linear Programming Model

**Low Inflow Protocol**

A Low Inflow Protocol (LIP) (FERC, 2006) was developed as part of the comprehensive relicensing agreement to provide a means of allocating water when storage levels are substantially below the target minimum operating levels. The frequency of being in the five increasingly severe stages of the LIP is used in this study as a measure of system performance.

Under the LIP, the status of the system is evaluated at the beginning of each month and, on the basis of the reservoir levels, drought status, and streamflow data, the stage of the low inflow condition is determined (Table 4.2 and Table 4.3). Entering each stage triggers operational changes (Table 4.4) and public notification. In this work only reservoir levels are used to determine the LIP stage because modeling the state of stream gages and drought index involve data that are not available. However, as seen in Table 4.3, storage is a primary determinant.
Being in any stage of the LIP is far more severe than simply being below the target minimum levels because the least severe LIP stage, stage 0, does not occur until the system-wide Storage Index (SI, the ratio of current usable storage to the total usable storage) falls below the monthly Target Storage Index (TSI). The TSI varies between 54% and 77% of the total usable storage compared to about 91% for the minimum target. For example, in January, stage 0 occurs when the total usable storage falls below 63% and stage 1 is triggered below 56.7% (90% of 63%).

<table>
<thead>
<tr>
<th>Month</th>
<th>Target Storage Index (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>63</td>
</tr>
<tr>
<td>Feb</td>
<td>54</td>
</tr>
<tr>
<td>Mar</td>
<td>63</td>
</tr>
<tr>
<td>Apr</td>
<td>68</td>
</tr>
<tr>
<td>May</td>
<td>77</td>
</tr>
<tr>
<td>Jun</td>
<td>77</td>
</tr>
</tbody>
</table>

Table 4.2: Monthly Target Storage Index

<table>
<thead>
<tr>
<th>Month</th>
<th>Target Storage Index (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jul</td>
<td>77</td>
</tr>
<tr>
<td>Aug</td>
<td>77</td>
</tr>
<tr>
<td>Sep</td>
<td>77</td>
</tr>
<tr>
<td>Oct</td>
<td>77</td>
</tr>
<tr>
<td>Nov</td>
<td>71</td>
</tr>
<tr>
<td>Dec</td>
<td>64</td>
</tr>
</tbody>
</table>

Entering the stages trigger a number of activities, including communications to officials governing and using the reservoir, public outreach, and, for stages above 0, changing operational targets. The changes affecting the model are shown in Table 4.4. For the purpose of the model, the target reductions in demand were assumed to be the high end of the target range shown in Table 4.4 and it is assumed that these targets are achieved.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Storage Index</th>
<th>Drought Monitor (3-month average)</th>
<th>Monitored USGS Streamflow Gages</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>90% &lt; SI &lt; TSI</td>
<td>0 ≤ DM</td>
<td>AVG ≤ 85%</td>
</tr>
<tr>
<td>1</td>
<td>75% &lt; SI ≤ 90%TSI</td>
<td>1 ≤ DM</td>
<td>AVG ≤ 78%</td>
</tr>
<tr>
<td>2</td>
<td>57% &lt; SI ≤ 75%TSI</td>
<td>2 ≤ DM</td>
<td>AVG ≤ 65%</td>
</tr>
<tr>
<td>3</td>
<td>42% &lt; SI ≤ 57%TSI</td>
<td>3 ≤ DM</td>
<td>AVG ≤ 55%</td>
</tr>
<tr>
<td>4</td>
<td>SI ≤ 42%TSI</td>
<td>DM = 4</td>
<td>AVG ≤ 40%</td>
</tr>
</tbody>
</table>

Table 4.3: Summary of LIP trigger points

The LIP leaves the specifics of demand reduction strategies to the individual water suppliers and consumers so the actual reductions will depend on the reaction of water consumers to water reduction programs.
The changes to the minimum reservoir level target are not uniform across all reservoirs. The greatest reductions occur in Lake James and Lake Norman. Lake James is one of the largest reservoirs and the most upstream and thereby serves an important role in regulating downstream reservoir levels. Lake Norman is the largest reservoir and so can accommodate large withdrawals.

Hydropower Production

Release and withdrawal information are used to calculate power production and understand the frequency and extent of resource conflicts. Hydroelectric parameters for each reservoir (Table 4.5) were obtained from Duke Energy records (made available as part of the current FERC renewal process) and from Duke Power’s WARMF model (Systech Engineering, 2005). Power generation is calculated using standard formula and depends primarily on hydraulic head and generator efficiency.

Operation of the hydropower turbines depends on a complex set of operating rules which were not fully modeled in this simulation. These rules depend upon, among other things, electricity market prices, predicted market prices, contractual obligations, and the availability of other peaking technologies. Because this model focuses on monthly results, the specifics of when turbines are used during the day are less important and a more complex model for hydropower operations is not justified.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Target Reduction in Demand from average monthly levels</th>
<th>Change to Minimum Reservoir Level Target</th>
<th>Reduction to Non-Critical Component of Minimum Release Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>1</td>
<td>0-5%</td>
<td>1 to 2 feet</td>
<td>60%</td>
</tr>
<tr>
<td>2</td>
<td>5-10%</td>
<td>2 to 4 feet</td>
<td>95%</td>
</tr>
<tr>
<td>3</td>
<td>10-20%</td>
<td>3 to 10 feet</td>
<td>100%</td>
</tr>
<tr>
<td>4</td>
<td>20-30%</td>
<td>critical reservoir level</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 4.4: Changes to Operating Targets According to LIP Stage
Results

The results of this work fall into two main themes. First is the demonstration of the importance of including uncertainty and variability in the analysis of reservoir performance; the likelihood of LIP implementation increases when inflow uncertainty and demand variability are included in the operations model when compared to using single inflow values each time period, as was done for the FERC relicensing work. The second theme is analyzing the performance of the reservoir system under two scenarios, 2008 baseline demand conditions and 2058 increased demand conditions.

Using Model Uncertainty and Input Variability Information

The effects of including uncertainty are demonstrated by looking at differences in the distribution of monthly reservoir levels for a model with and without uncertainty. Figure 4.4 shows the distribution of monthly reservoir levels for Lake James, the most upstream reservoir, under the 2008 demand scenario. The plot on the left is generated from 116 values for each calendar month (each calendar month appears once each year of the 116-year simulation) while the plot on the right represents the distribution of 232,000 values (2000 MCMC runs, each 116 years long).
The variation in the 50\textsuperscript{th} percentile (the top line) is a result of changes in target elevations over the course of the year, essentially showing the normal level guide curves for the reservoir; this curve is coincident with all percentile values above the 50\textsuperscript{th}. The increased frequency of low levels can be seen by comparing contours between the left and the right plots; when uncertainty is not included, 95 percent of the time the level in Lake James is above 261 to 262 meters throughout the year. When inflow uncertainty is included, 95\textsuperscript{th} percentile reservoir height includes values below 258 meters. The discrepancy in the percentile curves begins to be pronounced below the 25\textsuperscript{th} percentile, implying that about 25 percent of the time we would expect to predict lower levels with the model that includes uncertainty and variability versus the model that does not.

Of all the reservoirs, Lake James shows the most variation, primarily because inflow to Lake James comes from the smallest basin and there are fewer management strategies when there is no upstream reservoir with which to balance levels. Results for the remaining reservoirs are shown in Appendix E.

Another way to quantify the differences between the model with and without uncertainty is to directly count the fraction of times levels fall below the minimum target.
Figure 4.5 shows the results for Lake James, again showing that the analysis including uncertainty shows substantially higher frequency of low reservoir levels in most months.

![Graph showing reservoir levels comparison]

**Figure 4.5**: Frequency of Lake James reservoir levels being below the target minimum with 2008 demand, comparing the model with single monthly inflow values and a model capturing uncertainty with a distribution of monthly inflow values.

**Reservoir Performance Under 2058 Demand Scenario**

In this section, the frequency of having conditions that trigger each stage the LIP is used as a measure of reservoir performance for the two demand scenarios, 2008 and 2058. Figure 4.6 shows the percentage of time, or likelihood, that any stage of the LIP will be active for the model without (left) and with (right) uncertainty and input variability. When uncertainty and variability are not included, only stage 1 and stage 0 occur during the 116-years modeled; including uncertainty and variability reveals a significant likelihood of being in the more severe stages.

Focusing on the analysis with uncertainty (right plot of Figure 4.6), it can be seen that time in stages one through four increases with increasing demand (2058 scenario). The most severe stage, stage 4, occurs about 0.5% of the time with 2008 conditions (about one month in 16 years) and increases to about 1% under 2058 conditions (about 1 month in 8
years). Overall, the likelihood of being in any stage above 0 increases from about 2% to about 4% when demand increases from 2008 to 2058 levels.

The Catawba-Wateree Water Supply Study (HDR Engineering, 2006) does not characterize the frequency of operating under the LIP. Instead it analyzed safe yield based on critical intake elevation constraints (Table 4.5). This is a more severe condition than any of the LIP stages; below the critical intake elevation the total usable storage has been depleted while stage 4 of the LIP begins when between 20% and 35% of the usable storage remains. Although HDR’s results are not directly comparable to this work, there is value in reviewing them in the context of what would happen if HDR had included an extensive analysis of uncertainty and included more input variability.

Although the HDR model showed instances of end-of-day elevations below the critical elevation for the 2008 and 2058 demand scenarios, these were not counted as failures because there was assumed to be extra storage in upstream reservoirs. The results of the model in this work also show no critical failures in 2008 or 2058 when single monthly inflow values are used. However when a distribution of monthly inflow values is used in the model, critical failures are seen in at least one reservoir in about 1% of the simulations. It
may be that by omitting uncertainty in inflows and variability in demand, the HDR Water Supply Study failed to consider a non-trivial chance of critical reservoir failure.

*Power Generation*

This simulation estimates the maximum amount of power that could be generated given the amount of streamflow available (Figure 4.7). In almost every month hydropower was limited by the amount of available water so that an increase in demand withdrawals or decrease in inflow almost always results in a reduction of hydropower output. Since downstream reservoirs can utilize the releases from upstream reservoirs, hydropower production generally increases moving downstream in the system. The deviation from that trend for Mountain Island Lake is due to the large withdrawals there. The reduction from increased demand varies from 5% to 10% for the four most upstream reservoirs to about 20% for Lake Norman and Mountain Island Lake, where the increase in industrial and municipal demand is projected to be proportionally higher.

![Figure 4.7: Hydroelectric power generation under 2008 and 2058 demand scenarios](image-url)
Conclusions

Evaluating uncertainty of complex systems can be done effectively with Monte Carlo tools as long as model run times are reasonably fast. Empirically-based aggregate-level modes are good candidates for this because they reduce the number of system variables and model resolution while still providing accurate aggregate-level results. The model developed in this work reveals greater concern about the performance of the reservoir system in low-inflow periods than was seen in the Catawba-Wateree relicensing study because the relicensing study assumed one inflow scenario and constant average demand. This work suggests that a probabilistic model that propagates the uncertainty and variability of input variables and model parameters can provide a more informed view of the likelihood of rare events, including the number instances where demand and release reductions are needed as well as the number of reservoir failures.

With respect to the Catawba-Wateree water supply system, a recent analysis did not incorporate uncertainty in the estimates of inflows that would have arisen from measurement error and unaccounted for flows. Although the results of that study and this work are not directly comparable, there is general agreement between the results of this previous study and the analysis presented here when uncertainty and demand variability were omitted. The increase in likelihood of events of concern when uncertainty was included in the model indicates that the Catawba-Watery Water Supply Study could have had different conclusions if uncertainty and additional variability had been captured.

The reservoir inflow model is run for both 2008 and 2058 demand scenarios and reveals an increase in the likelihood of being in a low-inflow stage by about 50%. Because low inflow conditions remain relatively rare, only modest reductions in hydropower generation were observed for most reservoirs.
Using a stochastic model to evaluate the performance of a complex reservoir system provides additional information about the frequency of low reservoir levels. Because low-inflow performance is an important criterion in evaluating management strategies, especially when there are competing uses for water, it is important to use such models to better capture both input variability and model uncertainty. Because of their lower computation requirements and need for less detailed supporting input datasets, aggregate-level models provide a means for implementing stochastic models when models with greater spatial or temporal resolution make doing so impractical.
5. Summary

Uncertainty in information used to make watershed and reservoir management decisions is unavoidable. This work shows that uncertainty can be reduced by rigorously integrating information from several sources and that there can be great benefit from using even highly uncertain information when uncertainty can be propagated and reported in the outcome estimates. It can be computationally intensive and cost prohibitive to create large mechanistic models to propagate uncertainty and variability, which is often done using techniques based Monte Carlo methods. Empirically-based aggregate level models provide a means of creating models that can be run more quickly while preserving accuracy at the lower spatial and temporal resolutions. This cannot be done for all outcomes of interest, for example flood conditions can develop over the course of hours or days, but it can be useful for evaluating persistent conditions, such as drought.

The work presented in this dissertation has five main summary points:

1) **Bayesian Maximum Entropy methods provide a rigorous framework for combining information of various level of uncertainty to allow a better understanding of water quality.**

   BME methods allow information of various types and qualities to be combined to provide parameter estimates at desired locations and times. These estimates are expressed probabilistically with the degree of uncertainty reflecting the uncertainty inherent in the underlying information. Including uncertain model data, even from one with highly uncertain results, can improve estimates and allow for the quantification of the probability of a water not meeting water quality standards.

2) **Bayesian Maximum Entropy methods can be used to design lower cost monitoring programs based on known water quality conditions and the underlying spatial and temporal covariance.**
As a consequence of its ability to propagate information from one space/time location to another, BME can be used to lower the costs of monitoring programs by permitting the use of less expensive measurement methods and a reduced number of monitoring locations. More expensive monitoring methods can be used preferentially where the expectation is for values close to a regulatory threshold and the results of these monitoring methods, augmented with less certain data, allow the computation of parameter estimates everywhere they are needed with an acceptable level of uncertainty.

3) An aggregate-level model that reduces the spatial and temporal resolution as well as the number of model parameters can permit high model accuracy and the assessment of model uncertainty using iterative methods.

The role of rainfall-inflow models in assessing reservoir performance is limited by the required cost and expertise of creating accurate detailed models. Capturing uncertainty in distributed mechanistic rainfall-inflow models can be difficult because of long run times that limit the use of iterative techniques and the large number of parameters that need to be evaluated. An aggregate-level model that reduces the spatial and temporal resolution as well as the number of model parameters allows for high model accuracy for some outcomes and the assessment of model uncertainty using iterative methods. The monthly temporal resolution allows the analysis of outcomes, such as drought, that take place over months and years. Low-inflow conditions are a primary concern in many systems as they determine the frequency and extent of conflicts between the multiple uses of the water, which is often seen in systems, such as the Catawba, providing water for municipal supply, hydropower production, and other industrial uses.

4) The Expectation Maximization algorithm performed well in imputing missing data in 116-year rainfall dataset with monthly rainfall data from 19 meteorological stations.

A coherent historical 116-year (1386 month) rainfall input dataset is created that is shown to have sufficient accuracy to provide meaningful inflow estimates. This is achieved through statistical techniques that impute missing rainfall values based on a multivariate-
normal model of transformed rainfall data. This method is demonstrated to have a high explanatory power and little bias and to perform better than an alternative method using linear regression.

5) Including uncertainty and model input variability in reservoir management models provides greater insight into the expected frequency of extreme conditions.

This work suggests that a probabilistic model that propagates the uncertainty and variability of input variables and model parameters can provide a more informed view of the likelihood of rare events, including the number instances where demand and release reductions are needed as well as the number of reservoir failures. Although the median system behavior agrees with prior work that did not include uncertainty, including a distribution of possible outcomes results in a doubling of the estimate of the number of times reservoirs fall below target minimum levels and an increase in the likelihood of reaching critical levels.
Appendix A: Runoff Coefficients and Effects on Model Performance

Table A.1 contains the range of values for runoff coefficients obtained from the literature. Figure A.1 shows overall model performance (Nash-Sutcliffe efficiency measure) over the range of parameter values selected from independent uniform distributions for 5000 Monte Carlo simulations of a 141 month dataset. Runoff coefficients are used to create “potential runoff” values which are then used to predict reservoir inflow in the overall model. Model performance is based on the calibrated model.

<table>
<thead>
<tr>
<th>Land Type</th>
<th>Slope</th>
<th>Low Estimate</th>
<th>Median Estimate</th>
<th>High Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1%</td>
<td>0.1</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Deciduous Forest</td>
<td>2%</td>
<td>0.2</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>3%</td>
<td>0.3</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>Evergreen Forest</td>
<td>1%</td>
<td>0.1</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>2%</td>
<td>0.2</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>3%</td>
<td>0.3</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>Mixed Forest</td>
<td>1%</td>
<td>0.1</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>2%</td>
<td>0.2</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>3%</td>
<td>0.3</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>Pasture</td>
<td>1%</td>
<td>0.2</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>2%</td>
<td>0.3</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>3%</td>
<td>0.4</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>Cultivated</td>
<td>1%</td>
<td>0.2</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>2%</td>
<td>0.3</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>3%</td>
<td>0.4</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>Recreational Grasses</td>
<td>Flat</td>
<td>0.05</td>
<td>0.11</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>0.1</td>
<td>0.16</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>Steep</td>
<td>0.15</td>
<td>0.24</td>
<td>0.33</td>
</tr>
<tr>
<td>Water</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Barren</td>
<td>0.5</td>
<td>0.6</td>
<td></td>
<td>0.7</td>
</tr>
<tr>
<td>Low Intensity</td>
<td>0.25</td>
<td>0.32</td>
<td></td>
<td>0.4</td>
</tr>
<tr>
<td>Development</td>
<td>High Intensity</td>
<td>0.4</td>
<td>0.55</td>
<td>0.75</td>
</tr>
<tr>
<td>Development</td>
<td>Commercial</td>
<td>0.5</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>Wetlands</td>
<td>0.75</td>
<td>0.85</td>
<td></td>
<td>0.95</td>
</tr>
</tbody>
</table>

Source: (Novotny, 2003) and (Reckhow, et al., 1992)

Table A.1: Values for runoff coefficients for land types identified in the model. Range shown accounts for values for all soil types.
Figure A.1: Dotty plots of Nash-Sutcliffe efficiency measure for each runoff coefficient. Coefficients drawn from independent uniform distributions for 5000 Monte Carlo simulations of inflow for a 141 month period. Parameter names include land type and slope information (see Table A.1).

Figure A.2: Cumulative distribution of Nash-Sutcliffe value for 5000 Monte Carlo runs where only runoff coefficients were varied.
### Appendix B: Meteorological Stations and Rainfall Transformation Results

<table>
<thead>
<tr>
<th>Station Label</th>
<th>ID used in CRONOS database</th>
<th>Description</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>None</td>
<td>McGuire Nuclear Station</td>
<td>35.44</td>
<td>-80.95</td>
</tr>
<tr>
<td>2</td>
<td>311081</td>
<td>Bridgewater Hydroelectric Station</td>
<td>35.74</td>
<td>-81.84</td>
</tr>
<tr>
<td>3</td>
<td>311579</td>
<td>Catawba 3 NNW</td>
<td>35.74</td>
<td>-81.08</td>
</tr>
<tr>
<td>4</td>
<td>311582</td>
<td>Catawba 3 Lookout Shoals</td>
<td>35.73</td>
<td>-81.07</td>
</tr>
<tr>
<td>5</td>
<td>311690</td>
<td>Charlotte Douglas Airport</td>
<td>35.22</td>
<td>-80.96</td>
</tr>
<tr>
<td>6</td>
<td>311695</td>
<td>Charlotte</td>
<td>35.23</td>
<td>-80.85</td>
</tr>
<tr>
<td>7</td>
<td>311990</td>
<td>Conover Oxford Shoals</td>
<td>35.82</td>
<td>-81.19</td>
</tr>
<tr>
<td>8</td>
<td>313565</td>
<td>Grandfather Mountain</td>
<td>36.11</td>
<td>-81.83</td>
</tr>
<tr>
<td>9</td>
<td>314020</td>
<td>Hickory Regional Airport</td>
<td>35.74</td>
<td>-81.38</td>
</tr>
<tr>
<td>10</td>
<td>314938</td>
<td>Lenoir</td>
<td>35.91</td>
<td>-81.53</td>
</tr>
<tr>
<td>11</td>
<td>314996</td>
<td>Lincolnton 4 W</td>
<td>35.46</td>
<td>-81.33</td>
</tr>
<tr>
<td>12</td>
<td>315340</td>
<td>Marion 2 NW</td>
<td>35.66</td>
<td>-82.03</td>
</tr>
<tr>
<td>13</td>
<td>315838</td>
<td>Morganton</td>
<td>35.73</td>
<td>-81.67</td>
</tr>
<tr>
<td>14</td>
<td>315913</td>
<td>Mt Holly 4 NE</td>
<td>35.33</td>
<td>-80.99</td>
</tr>
<tr>
<td>15</td>
<td>315923</td>
<td>Mt Mitchell</td>
<td>35.76</td>
<td>-82.27</td>
</tr>
<tr>
<td>16</td>
<td>316602</td>
<td>Patterson</td>
<td>36.00</td>
<td>-81.57</td>
</tr>
<tr>
<td>17</td>
<td>317229</td>
<td>Rhodhiss Hydroelectric Plant</td>
<td>35.77</td>
<td>-81.44</td>
</tr>
<tr>
<td>18</td>
<td>318448</td>
<td>Swannanoa 2 SSE</td>
<td>35.57</td>
<td>-82.39</td>
</tr>
<tr>
<td>19</td>
<td>318519</td>
<td>Taylorsville</td>
<td>35.92</td>
<td>-81.17</td>
</tr>
</tbody>
</table>

Table B.1: Description of meteorological stations.  
(Station label refers to numbers in Figure 3.2. CRONOS is the acronym of the database used by state climatology office.)
Figure B.1: Normal probability plots of transformed rain data for each of the 19 stations (R² of fit of transformed data to an ideal distribution is shown for each plot)
Figure B.2: Convergence monthly rainfall means during EM algorithm. Omitted are two downward spikes at iteration 1 with values of approximately -19 and -12.

<table>
<thead>
<tr>
<th>Individual Values Used in Coordinate Transformation</th>
<th>Matrix Form of Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td>Value</td>
</tr>
<tr>
<td>X offset</td>
<td>-81.5</td>
</tr>
<tr>
<td>Y offset</td>
<td>35.7</td>
</tr>
<tr>
<td>X scale</td>
<td>3.72</td>
</tr>
<tr>
<td>Y scale</td>
<td>2.53</td>
</tr>
<tr>
<td>Rotation</td>
<td>29.6</td>
</tr>
</tbody>
</table>

Figure B.3: Values for linear transformation of MDS plot to geographic coordinates
Appendix C: Withdrawal and Loss Information for Reservoir Mass Balance

<table>
<thead>
<tr>
<th>Reservoir</th>
<th>Drainage Basin Area</th>
<th>Overall Capacity</th>
<th>Working Capacity</th>
<th>Surface Area</th>
<th>Mean Depth</th>
<th>Mean Retention time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lake James</td>
<td>984.2</td>
<td>89710</td>
<td>8886</td>
<td>27.6</td>
<td>13.5</td>
<td>208</td>
</tr>
<tr>
<td>Lake Rhodhiss</td>
<td>1838.9</td>
<td>15150</td>
<td>3990</td>
<td>12.4</td>
<td>6.3</td>
<td>21</td>
</tr>
<tr>
<td>Lake Hickory</td>
<td>569.8</td>
<td>33660</td>
<td>3443</td>
<td>17.1</td>
<td>9.5</td>
<td>33</td>
</tr>
<tr>
<td>Lookout Shoals</td>
<td>362.6</td>
<td>8146</td>
<td>1064</td>
<td>5.3</td>
<td>7.3</td>
<td>7</td>
</tr>
<tr>
<td>Lake Norman</td>
<td>880.6</td>
<td>356400</td>
<td>63526</td>
<td>131.4</td>
<td>10.2</td>
<td>239</td>
</tr>
<tr>
<td>Mountain Island Lake</td>
<td>181.3</td>
<td>18670</td>
<td>2675</td>
<td>13.3</td>
<td>5.4</td>
<td>12</td>
</tr>
<tr>
<td>Lake Wylie</td>
<td>3004.4</td>
<td>74690</td>
<td>13150</td>
<td>54.4</td>
<td>7</td>
<td>39</td>
</tr>
</tbody>
</table>

Table C.1: Physical Reservoir Properties (Duke Energy, 2006)

<table>
<thead>
<tr>
<th>Reservoir</th>
<th>Outflows in addition to dam operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lake James</td>
<td>None</td>
</tr>
<tr>
<td>Lake Rhodhiss</td>
<td>Lenoir, Valdese, Granite</td>
</tr>
<tr>
<td>Lake Hickory</td>
<td>Hickory, Longview</td>
</tr>
<tr>
<td>Lookout Shoals</td>
<td>Statesville</td>
</tr>
<tr>
<td>Lake Norman</td>
<td>Mooresville, Lincoln, CFCC, Davidson, N. Mecklenburg, DPC-MNS, DPC-MSS</td>
</tr>
<tr>
<td>Mountain Island Lake</td>
<td>Charlotte-Mecklenburg, Mount Holly, Gastonia, DPC-RSS</td>
</tr>
<tr>
<td>Lake Wylie</td>
<td>Belmont, Rock Hill, DPC-CNS, DPC-ALSS</td>
</tr>
</tbody>
</table>

Table C.2: Outflows used in reservoir inflow calculations

Figure C.1: Average Pan-Evaporation in Mountain Region of North Carolina
Appendix D: Posterior Distributions for Rainfall-Inflow Model Parameters

Posterior parameter distributions are labeled using the following method: “alpha” indicates a model parameter. The letter following “alpha” denotes the reservoir (h=”Hickory”, j=”James, etc.). The number is the subscript of the parameter, corresponding to those shown in Equation 3.9. “Sigma” indicates the distribution of the variance used to measure model error.
Appendix E: Distribution of Reservoir Levels for Six Reservoirs

Using the 2008 demand model, Figure E.1 is generated by calculating the percentiles of 116 values of reservoir level for each calendar month (each calendar month appears once in each year of the 116-year simulation) while Figure E.2 represents the distribution of 232,000 values (2000 MCMC runs, each 116 years long). Note that the vertical axis for Lake James has increased to span 8 meters in Figure E.2 while the others remain at 4 meters. The expected greater frequency of low levels is observed. The differences are most clear in Lake James, Lake Norman, and Mountain Island Lake.

Figure E.1: Distribution of Monthly Reservoir Levels for Six Reservoirs without Inflow Uncertainty
Figure E.2: Distribution of Reservoir Levels for Six Reservoirs with Inflow Uncertainty
Note that the vertical scale for Lake James spans three times the range as the other plots in this figure.

Figure E.3: Fraction of Times Reservoir Level Below Target Minimum Level, Baseline
References


