ENVIROMENTAL MODELLING: LEARNING FROM UNCERTAINTY

JOHN M. JUSTON

November 2012
HAIKU

rain falls
uncertainty links data and models
still much can be said

Juston
Greece, October 2011
Abstract

Environmental models are important tools; however uncertainty is pervasive in the modeling process. Current research has shown that understanding and representing these uncertainties is critical when decisions are expected to be made from the modeling results. One critical question has become: how focused should uncertainty intervals be with consideration of characteristics of uncertain input data, model equation representations, and output observations? This thesis delves into this issue with applied research in four independent studies. These studies developed a diverse array of simply-structured process models (catchment hydrology, soil carbon dynamics, wetland P cycling, stream rating); employed field data observations with wide ranging characteristics (e.g., spatial variability, suspected systematic error); and explored several variations of probabilistic and non-probabilistic uncertainty schemes for model calibrations. A key focus has been on how the design of various schemes impacted the resulting uncertainty intervals, and more importantly the ability to justify conclusions. In general, some uncertainty in uncertainty ($u^2$) resulted in all studies, in various degrees. Subjectivity was intrinsic in the non-probabilistic results. One study illustrated that such subjectivity could be partly mitigated using a “limits of acceptability” scheme with posterior validation of errors. $u^2$ was also a factor from probabilistic calibration algorithms, as residual errors were not wholly stochastic. Overall however, $u^2$ was not a deterrent to drawing conclusions from each study. One insight on the value of data for modeling was that there can be substantial redundant information in some hydrological time series. Several process insights resulted: there can be substantial fractions of relatively inert soil carbon in agricultural systems; the lowest achievable outflow phosphorus concentration in an engineered wetland seemed partly controlled by rapid turnover and decomposition of the specific vegetation in that system. Additionally, consideration of uncertainties in a stage-discharge rating model enabled more confident detection of change in long-term river flow patterns.

Keywords: Models, data, error, uncertainty, hydrology, soil carbon, wetlands, phosphorus
ACKNOWLEDGMENTS

I would like to acknowledge my supervisor, Per-Erik Jansson, and current co-supervisors David Gustafsson and Jan Seibert. Per-Erik, I have long enjoyed my habit of encountering unique and interesting people, and you certainly rank amongst them. Thank you for the opportunity at KTH, thank you for our many useful dialogues, and thank you for the freedom you gave me. I don’t think either of us will soon forget our experience in the field in west Kenya in 2010. I am grateful to you, and look forward to a continued friendship. The continuous dialogue with David over the last years on uncertainties, methods, and understandings has been extremely beneficial (and fun) toward solidifying my ideas. And I have enjoyed every opportunity that has presented itself to work and collaborate with Jan. I hope each of these trends continues. I also thank Olle Andrén for many good discussions and an introduction to east Africa.

I have enjoyed discussions and good company with many colleagues at the LWR Department at KTH, GeoCentrum in Uppsala, Physical Geography at Stockholm University, and various groups at SLU. In particular, Ida, Magnus, Christoffer, and Steve (all from various institutions) have almost always brightened my days when we met. Ida and I, especially, have shared many academic adventures, good laughs, and valuable discussions. At the LWR Department, there have been many good (and fun) dialogues with colleagues, and very good support from the Staff and Faculty. Thank you to all. Special thanx to Andrew, Juan, and Sofie for thoughtful suggestions to the Kappa.

I also gratefully acknowledge financial support from the Swedish Development Cooperation Agency (SIDA), grant number SWE-2006-514, for partial funding of this work.

I have had the fortune to come across some very good men over the years, including Howard, Allan and Carlos. I am very grateful to Tom and Tony for support outside of academia, and for our future endeavors. Former academic mentors have influenced both who I am and how I think, especially Harry Hull Robershaw and Mark Brown. Lastly, I acknowledge my technical muses, for example Keith Beven, Bob Kadlec, and Matt Cohen. These distinguished thinkers have, at various times, provided the little voices in my head (unknownto them) to dig a little deeper, think a little harder. Keith in particular has provided hours of provocative discussion and support. Thank you.

I have also had the fortune of coming across some very good women as well, most especially my lovely wife Susie. Discussing life, culture, and politics with our son August (he’s 12, you know) during these weeks of preparation was as enlightening and stress reducing as always. Lastly, my parents have always supported my academic endeavors. Hopefully, I am done now.

This thesis is dedicated to Harry, my professor, mentor, and friend. RIP amigo.

Juston
Stockholm
2012
# Table of Contents

- Haiku  iii
- Abstract  v
- Acknowledgments  vii
- Table of Contents  ix
- Papers Included in the Thesis  xi
- Additional papers  xi
- 1. Introduction  1
  - 1.1. Objectives  2
- 2. Background  3
  - 2.1. Conceptual process models  3
  - 2.2. Paradigm for expressing uncertainties  3
  - 2.3. Characteristics of uncertainty in environmental modeling  5
  - 2.4. Likelihood functions  6
    - 2.4.1. Formal  6
    - 2.4.2. Informal  7
  - 2.5. Equifinality  8
  - 2.6. Bayes Theorem  8
  - 2.7. Sampling  8
  - 2.8. Calibration assessment with uncertainty estimation  9
    - 2.8.1. Bayesian Markov chain Monte Carlo algorithm  9
    - 2.8.2. Generalized Likelihood Uncertainty Estimation  9
- 3. Methods  10
  - 3.1. Overview  10
  - 3.2. Study I  10
    - 3.2.1. Research questions:  10
    - 3.2.2. Study site and data characteristics  10
    - 3.2.3. Model rationale and structure  12
    - 3.2.4. Calibration and uncertainty estimation  13
  - 3.3. Study II  13
    - 3.3.1. Research questions  13
    - 3.3.2. Study site and data characteristics  13
    - 3.3.3. Model rationale and structure  14
    - 3.3.4. Calibration and uncertainty estimation  15
  - 3.4. Study III  15
    - 3.4.1. Research question  15
    - 3.4.2. Study site and data characteristics  15
    - 3.4.3. Model rationale and structure  15
    - 3.4.4. Calibration and uncertainty estimation  16
  - 3.5. Study IV  17
    - 3.5.1. Research question  17
    - 3.5.2. Study sites and data characteristics  17
    - 3.5.3. Model structure  17
    - 3.5.4. Calibration and uncertainty estimation  17
    - 3.5.5. Hypothesis testing  18
- 4. Results  18
  - 4.1. Data  18
  - 4.2. Models  18
4.3. Residual Errors

4.4. Information and uncertainty
4.4.1. Study I
4.4.2. Study II
4.4.3. Study III
4.4.4. Study IV

5. Discussion
5.1. Learning about uncertainty schemes
5.1.1. Informal likelihoods
5.1.2. Formal likelihoods
5.1.3. Equifinality in intervals?
5.2. Learning about processes
5.3. Learning about data
5.4. Posterior analysis

6. Summary and Conclusions

References
PAPERS INCLUDED IN THE THESIS


III. Juston JM, DeBusk TA, Grace KA, and Jackson SD. *In Revision*. A model of phosphorus cycling to explore the role of biomass turnover in submerged aquatic vegetation wetlands for Everglades restoration. Accepted for publication in *Ecological Modelling*.

IV. Juston JM, Jansson P-E, and Gustafsson D. *In review*. Rating curve uncertainty and change detection in discharge time series: case study with 44-years historic data from the Nyangores River, Kenya. Submitted to *Hydrological Processes*.

ADDITIONAL PAPERS


1. INTRODUCTION

Sustainable natural resources and ecosystem services are the foundation to the continuation of human development (Odum, 1996). Accordingly, there are many critical scientific questions that must be addressed by the natural sciences to support conservation and management of natural resources. Some issues concern the global carbon cycle, climate change, decreasing soil fertility, altered nutrient cycles, ecosystem restoration, impacts of land use change, and waste management (Rockström et al., 2009; Fleishman et al., 2011). Progress on these issues will inevitably employ the scientific method, which describes a "procedure that has characterized natural science since the 17th century, consisting in systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypothesis" (OED, 2012). It has been referred to as a “motivated iteration”, wherein practice confronts theory and theory, practice (Box, 1976).

Modeling is an important tool for advancing scientific understanding as it poses hypothesis on the mathematical form of physical processes and/or relationships. Hypotheses in model structures can be tested against their ability to describe available data observations. Thus, the process of model development and calibration has basis in the scientific method (Fig. 1). Mathematical methods for quantifying uncertainties have developed in support of testing hypothesis and advancing science via the scientific method (e.g. Gauss 1809, Fisher, 1922; Zadeh, 1964; Box, 1980). For example, statistical methods in combination with probability theory can in some circumstances provide numeric expressions of confidence to reject (or not) hypotheses, as in testing a linear regression (Fisher, 1922). Borrowing words from the esteemed statistician George

![Figure 1. Environmental models are important tools; however uncertainty is pervasive in the modeling process. A key question has become: how focused should the final representation of uncertainty be with consideration of errors and uncertainties in input forcing data, approximate mathematical representations, and response observations used for calibration? The selection of the likelihood function and associated acceptance criteria plays a determining role during calibration with repetitive evaluations. The three feedback pathways in this diagram are consistent with the “motivated iteration” philosophy of the scientific method (Box 1976). Roman numerals refer to areas of research in the Papers.](image-url)
Box (1976), an explicit expression of uncertainty “increase(s) greatly the probability that the investigator will be led along a true rather than a false path.”

Environmental systems are intrinsically complex with open boundaries, uncontrolled conditions, and substantial feedbacks and interactions (Odum, 1983). These intrinsic complexities bring serious challenges to connecting model representations and predictions to field-collected data observations (Beven, 2009). For example, environmental models often require mathematical complexity beyond linear relationships, particularly process-based models that simulate time series data. Box (1979) also expressed the well-known opinion that “all models are wrong, but some are useful”. Following this logic, environmental model equations, no matter how many or detailed (e.g., Andren and Kätterer, 1997; Seibert, 1997; Gerten et al., 2004; Jansson, 2012), are at best simplified numerical approximations of complex natural processes. Field-collected environmental data add substantial additional complexity to the picture, as these data are often difficult (if not impossible) to measure with precision and adequate spatial and/or temporal representativity (e.g., rainfall, streamflow, water quality; McMillan et al., 2012). Thus, it is well-recognized that uncertainty permeates the environmental modeling process from the input data used to drive models, thru the inadequacies of equations themselves, and again in the data used to calibrate and validate model performance (Janssen and Heuberger, 1995; Walker et al., 2003; Loucks et al., 2005; Refsgaard et al., 2005; Beven, 2009). With these considerations, the entire undertaking can seem to encompass a rather “significant approximation” (Di Baldassarre and Montanari, 2009).

As in linear regression, the key to quantifying the uncertainty in environmental modelling (i.e., just how significant an approximation is this?) lies in interpreting information in the residual error between model simulations and data observations (Fig. 1). However, given the inherent complexities, uncertainties can no longer be estimated with classic textbook methods only. Consequently, new methods emerged during the 1990's to address these concerns. The field of hydrological sciences has been a particularly active arena for methodological advancement, including substantial developments both inside and outside of probability theory (e.g., Warwick and Cale, 1988; Beven and Binley, 1992; Kuczera and Parent, 1998). These have in fact been polarizing developments, as they have yielded considerable discussion and debate in the literature in recent years (e.g., Beven, 2006; Mantovan and Todini, 2006; Beven et al., 2008; Stedinger et al., 2008; Vrugt et al., 2008; Montanari et al., 2009; Beven et al., 2011; Clarke et al., 2011).

One critical question in these discussions has been: how focused should the final representation of uncertainty be with consideration of characteristics in uncertain input forcing data, model structure representation and output data observations? The key to this question lies in the selection of an appropriate evaluation measures to justify the conclusions drawn from comparing model output with independent observational data. For model calibration, these likelihood functions, as they will be called henceforth, represent the likelihood that a model (a hypothesis about the real world) is supported by data. A likelihood function consolidates the information in a set of residual errors in order to expedite repetitive and consistent evaluation of model performance (Fig. 1). Achieving an appropriate level of focus is important to properly address pressing environmental research questions (Fig. 1), when decisions are expected to be made from the modeling results. There is a continued need for further theoretical, experimental, and applied research on this issue.

1.1. Objectives

The research in this thesis addresses three questions about the value of information in net residual errors from model-data comparisons for quantifying uncertainties:

- How should the nature of uncertainties in data and models, and the information in residuals, guide
the selection of appropriate likelihood functions (i.e., evaluation measures)?

- How do different likelihood functions affect uncertainty intervals and hypothesis testing?
- How do different characteristics and/or types of data act to constrain model parameters, predictions, and uncertainties?

These objectives were addressed from an applied perspective by surveying experiences generated with four diverse case studies. These studies focused on developing appropriate models and uncertainty schemes with consideration of site-specific input and calibration data observations. In this way, this thesis considered relevant methodological issues through application to a suite of relevant environmental issues. Site-specific questions included:

- To what extent could a simple catchment hydrology model be expanded to more fully describe the near-surface hydrology at a nuclear fuel repository site in coastal Sweden? (Paper I)
- Could a simple soil carbon model developed for Swedish conditions be improved and used to clarify differences in decomposition processes in Sweden and arid Kenya? (Paper II)
- Is a process oriented model for the apparent limiting phosphorus concentration in a $1+$ billion constructed wetland system for Everglades restoration (Juston and DeBusk, 2011) supported by a diverse array of (uncertain) field data? (Paper III)
- Has historic deforestation in the Mara River basin in Kenya (home to the Serengeti ecosystem) manifested in any detectable change in the river’s response with consideration of uncertainty in the available stream flow data? (Paper IV)

2. BACKGROUND

This section introduces the ingredients in “the stew”, including background on conceptual process models, expressing uncertainty, the nature of error and uncertainty in environmental modelling, likelihood functions, and automated sampling strategies for model calibration and uncertainty analysis.

2.1. Conceptual process models

Ecological processes can be modeled as deterministic systems of discrete state variables and fluxes (Odum, 1983). This approach is followed in this thesis. Here, a storage is defined as a state variable that is constrained by conservation of mass. A flow is defined as a time dependent flux in or out of storages (Fig. 2). Conceptual process models are composed from these building blocks. Flow equations do not necessarily need to be complex, although sometimes this is necessary, as the function of the whole often “becomes more than the sum of its parts when there are interactions” (Odum, 1983). Here, the term model development refers to the process of adding storage and/or flux elements to an existing model in order to advance or improve a process description (Fig. 2). It is well-recognized that process understanding can benefit from interactive considerations of model structure and available data (e.g., Chapra and Canale, 1991; Krueger et al., 2010; Juston and DeBusk, 2011; McMillan et al., 2011).

2.2. Paradigm for expressing uncertainties

The paradigm for expressing uncertainties in environmental modeling follows a template from regression analysis (Schoups and Vrugt, 2010). Thus, the intent is to express marginal and conditional density functions for model parameters and credibility bounds for model predictions (Fig. 3), independent of the formulation of the likelihood function. It might be anticipated that different uncertainty schemes based on different assumptions might yield different uncertainty intervals.
Figure 2. a) A schematic of a conceptual process model with two storages and five flows. This model was existing and used as a starting place for further development, which included the schematic b) for a period before it was ultimately rejected for not adequately describing the available data, and c) which performed more satisfactorily and is elaborated in Paper III.

Figure 3. Linear regression provides a starting point for considering uncertainties in systems with more complex datasets and models. In this example, residual errors were independent, stochastic, and normally distributed (by design), thus uncertainties could be estimated from probability theory. The fit of the line itself was uncertain due to uncertainty in parameter values. Total prediction bounds account for the full data scatter around the uncertain line fit (based on an error model). Parameter uncertainty can be represented with marginal and conditional probability distributions (here, a density function for slope, and cumulative function for intercept), where the dots indicate conventionally reported 5-95% confidence intervals.
2.3. Characteristics of uncertainty in environmental modeling

Walker et al. (2003) provided a general definition of uncertainty in modelling as: “any deviation from the unachievable ideal of completely deterministic knowledge of the relevant system”. Such deviations can lead to an overall “lack of confidence” in the obtained results based on a judgment that they might be “incomplete, blurred, inaccurate, unreliable, inconclusive, or potentially false” (Refsgaard et al., 2007).

Various sources of error can contribute to uncertainty in modelling results. Data measurements inevitably contain error (Taylor, 1997). Inadequacies in process model equations are referred to as model structural errors (Beven, 2005). The net deviation (i.e., difference) between an observed and simulated environmental response is referred to as residual error (Fig. 1), even though there might be no fixed point of reference in this signal (i.e., net deviations can originate from either observational and/or simulation error in environmental modelling).

Walker et al. (2003) proposed a useful three-dimensional classification matrix for uncertainties according to level, location, and nature. Level refers to the magnitude of uncertainty as a progression from low to high, “know” to “no-know”. Location refers to the originating source of uncertainty including input forcing data, the model concept and implementation, model parameters, and/or model output. Nature refers to if the uncertainty is due to inherent variability or is epistemic (Walker et al., 2003).

Epistemic uncertainty is defined as originated from a false, limited, or imperfect knowledge, independent of inherent variability (Walker et al., 2003; Refsgaard et al., 2007; Beven, 2009; Spiegelhalter and Riesch, 2011).

Variability uncertainty in environmental modelling is considered natural and non-reducible. This can include classic random error in field or lab measurements (e.g., Larson and Peck, 1974; Sauer and Mayer, 1992; Ståhl et al., 2004) and intrinsic spatial variability of some environmental variables (e.g., Conant and Paustian, 2002; Jager and King, 2004). Stochastic variables and residual errors can often be treated with probabilistic models and methods (e.g., Fig. 3).

Epistemic uncertainty in environmental modelling can originate from several sources, including: systematic errors in field measurements (e.g., Larson and Peck, 1974; Sauer and Mayer, 1992; Taylor, 1997; Neyroud and Fisher, 2003; Ståhl et al., 2004); representation errors of field-collected data relative to model variables (e.g., Freer et al., 2004) or unknowable patterns (e.g., the “true” spatial distribution of rainfall; Lebel et al., 1987; McMillan et al., 2012); artifacts in data from non-stationary physical phenomena (e.g., Westerberg et al., 2011); erroneous data (Beven et al., 2011); and scientific uncertainty and/or inadequacy in model concepts and equations (Spiegelhalter and Riesch, 2011; Beven et al., 2011). All of

![Figure 4. A typical simulated response of an environmental variable from a conceptual process model. Here, data is shown as a deterministic signal, although uncertainty existed. Residual errors are considerably more patterned than the example in Figure 3, with evidence of serial dependence and non-stationary variance. Residual errors also had bimodal distribution.](image-url)
these sources are not random and have a base quality of imperfect or inadequate knowledge.

The net effect of epistemic error in modelling can produce patterned and/or structured residual errors (Beven et al., 2011; Fig. 4), but the influence can be less obvious as well (e.g., Taylor, 1997; Westerberg et al., 2011). Epistemic uncertainty is often considered reducible by more study, but such studies may not be practical or possible (e.g., for some historic data sets). Epistemic error can also include elusive unknown-unknowns (Spiegelhalter and Riesch, 2011; Beven et al., 2011). There is “continued argument” on the influence and treatment of epistemic factors in model uncertainty estimation (Spiegelhalter and Riesch, 2011; Beven et al., 2011).

A categorization framework for uncertainties in field-collected environmental data and their relationship to a specific model framework is suggested in Figure 5; this will be recalled later to summarize data and characteristics used in this thesis. This framework focuses on the data “location” in the level-location-nature uncertainty matrix (Walker et al., 2003). However, the data location is sub-categorized for uncertainties originating from field (point) measurement and uncertainty originating from the representivity of this data to specific model variables. This second character can only be assessed in the context of specific model usage of the data, as different models might use the same data in different ways. For example, groundwater levels may be rather precisely measured at several points within a catchment, but there might be some (epistemic) uncertainty that accompanies how this data is interpreted for constraining a single groundwater storage variable in a hydrological model.

2.4. Likelihood functions

A series of residual errors (Figs. 3 and 4) is calculated:

\[ \varepsilon_i = O_i - M_i(\Theta, I) \]  

(Eq. 1)

where \( \varepsilon \) is the set of residuals for \( i \) observations, \( O_i \) are the data observations and \( M_i \) represents model output with parameters, \( \Theta \), and input forcing data, \( I \). The combined errors, uncertainties, and/or inadequacies of input data, model equations, and calibration data are embodied in this single error signal. In some cases, a model application may have multiple calibration objectives and thus multiple error series to consider. However, two key points are that a residual error series, \( \varepsilon \), is an aggregated measure of the net errors and uncertainties in the modeling process, and that it can be difficult, if not impossible, to disentangle relative contributions of various sources in this signal (Beven, 2009).

Likelihood measures further consolidate the information in a residual error series so that it can be used to guide model parameter estimation (Fig. 1). Likelihood functions for environmental modelling can be characterized as either formal or informal (Smith et al., 2008; Schoups and Vrugt, 2010). A formal likelihood function is one that is defined within the context of statistics and probability theory. An informal likelihood is one formulated outside of probability theory.

2.4.1. Formal

A formal likelihood is based on an assumed statistical model for residual errors, known as an error model. For clarity, the error model would be in addition to the conceptual process model, so there are in effect two
models being considered. For example in linear regression, the linear model can be accompanied with a statistical error model of residuals, which might assume (and thus require) that residuals be independent, normally distributed, with zero mean, and constant variance, $\sigma^2$ (e.g., Fig. 3). The parameters in the error model can be used to formulate a formal statistical likelihood function (Aldrich, 1997):

$$L(O|M(\Theta, I)) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} \sigma} \exp \left[ -\frac{1}{2 \sigma^2} e_i^2 \right]$$

(Eq. 2)

where $L(O|M(\Theta, I))$ is defined as the likelihood that a model $M$ with parameters $\Theta$ and input $I$ describes the $n$ observations $O$. Model parameters that maximize this likelihood function (or any other for that matter) are considered optimal. It should be noted that maximizing Eqn. 2 will yield the same solution as minimizing the sum of square error (SSQE), provided errors are indeed random and Gaussian. The advantage of a formal likelihood formulation is that the optimal solution is also the most probable from a statistical perspective, again provided the assumptions in the error model are valid and confirmed (e.g., Fig. 3). Validation of the assumed error model is in fact the key step that imbues the formal approach with a sense of objectivity (Schoups and Vrugt, 2010; Stedinger et al., 2008). However, the opposite can also be true: an incorrect error model might lead astray the objectivity (Beven et al., 2008). There are limited examples of extended error model formulations that attempt to address the complicated nature of errors that can occur in time series simulations (e.g., Fig. 4), such as serial dependence, non-normality and non-stationary variance in residuals (e.g., Yang et al., 2007; Schoups and Vrugt, 2010).

### 2.4.2. Informal

Informal likelihood measures do not employ a statistical error model, although they may in some cases specify non-parametric distributions for acceptable residuals. SSQE is itself an informal likelihood if it used without statistical assumptions on the structure and distribution of residuals (Smith et al., 2008), as is the related root mean square error (RMSE). Another common informal measure is the so-called model efficiency parameter (Nash and Sutcliffe, 1970):

$$R_{df} = 1 - \frac{\sigma^2_{residuals}}{\sigma^2_{observations}}$$

(Eq. 3)

This measure normalizes the variance in errors to the variance in the observations, such that the maximum possible score for a perfect simulation is 1.0. Scores from sequential evaluation of measures such as these can be compared to one another, but cannot be assessed with any objectivity since there is no commonly agreed framework as to what values from these measures mean. Thus, individual judgment can come into play with informal measures if one wishes to compare the merit of one score relative to another. This subjective aspect of informal measures has been criticized in some hydrological modelling applications (Kuczera and Parent, 1998; Mantovan and Todini, 2006; Stedinger et al., 2008).

Other informal likelihood functions attempt to mitigate this subjectivity by establishing a means to validate if residuals errors are within acceptable limits. In the limits of acceptability (LOA) approach (Beven, 2006; Liu et al., 2009), acceptable ranges for residuals are specified for a simulation based on analysis of uncertainties in the output observational data (i.e., the observations used to compare to model output)(Fig. 6). This is not a comprehensive treatment of all possible errors, since the effect of input and model equation errors are difficult to encapsulate in this measure. However, it does allow means to explicitly check if model outputs are at least within the limits of observational accuracy, and investigate possible patterns and causes when they are not. LOA can be set with consideration of uncertainties in a single series of field observations (e.g., Westerberg et al., 2011) and/or with consideration of intrinsic spatial variability in multiple series of observations (e.g., Freer et al., 2004).
2.5. Equifinality

The concept of *equifinality* derives from general systems theory and suggests an inability to uniquely differentiate the pathway by which a final state in an open system is achieved (von Bertalanffy, 1968). Beven was the first to use the word in the context of model calibration and uncertainty estimation (Beven, 1993). Beven (1993, 2006) argued that equifinality is intrinsic in many environmental modelling applications; specifically, that there can be several different model constructions and/or many different parameter sets that produce similar empirical output, and that there is no unproblematic way to know which realization is closer to nature. The principle can also be stated from a rejectionist perspective: there is often inadequate evidence to reject multiple hypotheses of acceptable models and/or parameter sets given inherent uncertainties and errors in environmental observations and model equations (Beven, 2006). Equifinality and optimality can be viewed as competing philosophies in model calibration; thus, it is not a universally held belief (e.g., Mantovan and Todini, 2006; Stedinger et al., 2008; Clarke et al., 2011). However, the concept has found wide adoption and utility.

2.6. Bayes Theorem

The likelihood measures of the previous section can be integrated into a Bayesian framework to provide a theoretical framework for model parameter calibration and uncertainty estimation. Bayes Theorem (Bayes, 1763) provides a means to refine *prior* information (i.e., beginning distributions) for model parameters to narrow(er) *posterior* distributions through repeated evaluation of an appropriate likelihood function:

\[
P(\Theta | O) = \frac{L(O | M(\Theta)) \cdot P(\Theta)}{C}
\]  

(Eq. 4)

where \(P(\Theta)\) denotes the prior joint probability density for model parameters, \(P(\Theta | O)\) denotes a posterior probability density after conditioning to observations via the likelihood function, \(L\), and \(C\) is a scaling constant so that cumulative probabilities sum to one. Bayesian inference has been adapted as a general calibration (i.e., learning) framework with either formal or informal likelihood functions (e.g., Freer et al., 1996).

2.7. Sampling

Monte Carlo methods provide a means for repetitive model evaluation of randomly drawn parameter sets, followed by evaluation

![Figure 6. A hypothetical example of limits of acceptability applied to a modeled series. This example illustrates a trapezoidal function with a core acceptance range (inner error bars) and fuzzy boundaries (full extent) defined for each observation. The model is compared to the acceptance limits at each step and assigned a 0-1 score. Scores can be averaged over a simulation to yield a single-valued likelihood function, and/or evaluated individually or by subset. Fuzzy boundaries are optional.](image-url)
of likelihood measures in comparison to observations. So-called prior distributions of model parameters establish a feasible parameter space. Monte Carlo methods sample this parameter space. Sampling can be truly random, or partly structured. Markov chain Monte Carlo (MCMC) analyses structure the search with clever algorithms (e.g., Metropolis et al., 1953) for generating random samples in focused regions of the parameter space based on feedback and learning from likelihood evaluations. The principal feature of MCMC algorithms is iterative sampling coupled to probabilistic acceptance criteria. In this way, the resulting sample density can be used to approximate complex probability distributions that cannot be analytically solved. This can be contrasted to a simple random sampling algorithm with no “learning”. Both have utility.

2.8. Calibration assessment with uncertainty estimation

Model calibration methods can draw upon a number of ingredients in regards to sampling and evaluation. Two methods have become prevalent in recent years that integrate calibration with uncertainty estimation, but draw upon different ingredients to do so (Table 1).

2.8.1. Bayesian Markov chain Monte Carlo algorithm

A probabilistic ideology frames the BMCMC algorithm. Most often, BMCMC analysis is aimed at producing a sample to approximate the posterior probability density using formal likelihood measures (e.g., Eqn 2) and Bayesian inference (Eqn. 4) (Smith and Roberts, 1993; Gelman et al., 1995; Kuczera and Parent, 1998). Accordingly, this approach requires a statistical error model for residuals, which might include data transformations and attributes to address serial dependence and non-stationarity in residuals, if necessary. As stated above, the key step that imbues objectivity and probabilistic meaning to results is the posterior affirmation of error model characteristics (Gelman et al., 1995; Schoups and Vrugt, 2010). In practice, the algorithm works flawlessly for many posterior probability densities (e.g., Gelman et al., 1995; Figure 3). However despite strong-minded advocates, there is considerable ongoing debate, particularly for modelling hydrological time series (e.g., Stedinger et al., 2008; Beven et al, 2008; Clarke et al., 2011), as to if the BMCMC algorithm yields reliable probabilities for model applications substantially affected by non-random uncertainty (e.g., Beven et al., 2008; Beven et al., 2011).

2.8.2. Generalized Likelihood Uncertainty Estimation

On the other hand, an equifinality ideology frames the GLUE method. GLUE applications tend to explicitly avoid the use of statistical models to describe the structure of model residuals. Thus, GLUE utilizes informal likelihood measures in combination

| Table 1. Differences in philosophy, theory, and practice for Bayesian Markov Chain Monte Carlo (BMCMC) and Generalized Likelihood Uncertainty Estimation (GLUE) methods |
|---------------------------------|-----------------|-----------------|
| BMCMC                          | GLUE            |                 |
| **Philosophical basis**        | Optimality, errors can be modeled with parametric distributions | Equifinality, immunity from parametric constraints on residuals |
| **Error model**                | Parametric assumptions for structure of residuals errors          | No fixed assumptions; nonparametric distributions are an option |
| **Likelihood measures**        | Formal          | Informal        |
| **Monte Carlo sampling**       | Markov chain (e.g., Metropolis) | Random |
| **Acceptance criteria**        | Probabilistic   | Thresholds and/or LOA |
| **Representation of uncertainty** | Parameters: posterior densities Predictions: 1) parameter uncertainty bounds, 2) total prediction bounds | Parameters: posterior densities Predictions: GLUE prediction bounds |
with threshold criteria to accept sets of model simulations that represent equally plausible representations. Selection of threshold criteria can introduce subjectivity to a GLUE assessment (Beven and Binley, 1992). Early efforts with GLUE tended to use informal measures (e.g. $R_\alpha$) with user-selected threshold criteria (e.g., $R_\alpha >0.8$)(e.g., Freer et al., 1996; Beven and Freer, 2001). These can be referred to as soft informal likelihoods, as it can be difficult to defend the basis for selecting either the likelihood or the acceptance criteria. More recent efforts have focused on LOA methods to define what could be referred to as informed informal likelihoods, as these are based on analysis of the scatter in data observations and can provide a more rigorous means to validate residuals and investigate deviations.

The GLUE methodology is most often implemented by generating a large number of purely random samples of a parameter space. Since GLUE does not employ a formal model of residual errors, this approach does not separate predictive uncertainty bounds into contributions from model parameters and data scatter, as in linear regression (Fig. 3) and BMCNC output. Thus all predictive uncertainty is embodied in one interval which will be referred to herein as GLUE prediction bounds (Table 1).

### 3. Methods

#### 3.1. Overview

The four Papers encompassed a variety of study sites (Fig. 7), data, model applications (Table 2), and uncertainty schemes (Table 3). This thesis presents only snapshots from each Paper; henceforth these snapshots are called Studies, where each Study relates to a corresponding Paper. By necessity, methodological details are lacking in some descriptions; the interested reader is referred to the Papers. The brief methodological overviews for each Study begin with a statement of research questions that relate to the overall Thesis Objectives (Section 1.1).

#### 3.2. Study I

**3.2.1. Research questions:**

- Could a simple catchment hydrology model be expanded to describe near-surface groundwater levels in addition to discharge?
- Are all data in daily discharge and groundwater time series equally informative towards calibrating the model?

**3.2.2. Study site and data characteristics**

The 5.6-km² catchment at the Forsmark site bordered the Baltic Sea in central Sweden (Fig. 7). This was a low-lying region, with

---

### Table 2. Overview of data and models. Symbols and units for data are as follows: PPT=precipitation (mm/d), T=temperature (C), PET=potential evapotranspiration (mm/d), Q=discharge(mm/d), GW=groundwater depth (mm), $C_i$=carbon input to soil (kg/m²/yr), $C_s$=carbon storage in soil (kg/m²), $Q_0$= water inflow rate (mm/d), $P_i$=inflow phosphorus concentration ($\mu g/l$), $V$=distribution of vegetation species, $P_w$=P concentration in water ($\mu g/l$), $P_s$=P storage in sediment (g/m²), GH=gauge height (m). The “storages” and “parameters” columns indicate the total count in the model structures and equations.

<table>
<thead>
<tr>
<th>Study</th>
<th>Study sites</th>
<th>Model type</th>
<th>Input data</th>
<th>Calibration data</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>storages (#)</td>
<td>parameters (#)</td>
</tr>
<tr>
<td>I</td>
<td>Forsmark, SE</td>
<td>Catchment hydrology</td>
<td>PPT, T, PET</td>
<td>Q, GW</td>
<td>3</td>
</tr>
<tr>
<td>II</td>
<td>SLU, SE</td>
<td>Soil carbon</td>
<td>$C_i$</td>
<td>$C_s$</td>
<td>3</td>
</tr>
<tr>
<td>III</td>
<td>Everglades USA</td>
<td>Wetland P cycle</td>
<td>$Q_0$, $P_i$, V</td>
<td>$P_w$, $P_s$</td>
<td>4</td>
</tr>
<tr>
<td>IV</td>
<td>Mara River, KE</td>
<td>Rating Curve</td>
<td>GH</td>
<td>Q</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 3. Overview of calibration and uncertainty schemes. For Study II, the second method was not documented in Paper II.

<table>
<thead>
<tr>
<th>Study</th>
<th>Calibration Method</th>
<th>Likelihood Type</th>
<th>Likelihood functions</th>
<th>Acceptance criteria</th>
<th>MC samples (#)</th>
<th>Error analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>GLUE</td>
<td>Informal</td>
<td>Q: $R_{eff}$ GW: LOA</td>
<td>Top 200 scores; combined index</td>
<td>4E6</td>
<td>visual</td>
</tr>
<tr>
<td>II</td>
<td>GLUE</td>
<td>Informal</td>
<td>RMSE</td>
<td>RMSE &lt; 0.19 g/m$^2$</td>
<td>50E6</td>
<td>visual</td>
</tr>
<tr>
<td></td>
<td>BMCMC</td>
<td>Formal</td>
<td>Eqn. 2</td>
<td>Probabilistic</td>
<td>30E3</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>GLUE</td>
<td>Informal</td>
<td>$P_s$: LOA $P_w$: LOA</td>
<td>+/- IQR of scatter</td>
<td>100E6</td>
<td>Posterior validation</td>
</tr>
<tr>
<td>IV</td>
<td>BMCMC</td>
<td>Formal</td>
<td>Eqn. 2</td>
<td>Probabilistic</td>
<td>30E3</td>
<td>Posterior validation</td>
</tr>
</tbody>
</table>

Figure 7. Study sites for the four modelling studies: a) the 5.6 km$^2$ watershed boundary for the catchment hydrology study in Paper I, b) two of the 4-m$^2$ plots from the SLU Frame Trial for the soil carbon dynamics study in Paper II, c) the 920-ha engineered wetland for the phosphorus cycling study in Paper III (dots indicate internal sampling locations), and d) the 650-km$^2$ catchment boundary for the discharge rating model study in Paper IV.
small-scale topography, dominated by Quaternary deposits and an average depth of ~5 m to underlying granite rock. A rather fine-resolution network of meteorological, hydrological, and hydro-geological monitoring was initiated on site and in the surrounding region in 2003 (Juston et al., 2007).

Daily time series data for precipitation (PPT), calculated potential evapotranspiration (PET), stream discharge at the catchment outflow, and a distributed network of near-surface groundwater wells were available since 2003 and assumed to be of high quality (i.e., low level uncertainty). Temporal correlation in daily precipitation from two stations with 2-3 km of catchment boundaries was very high ($R^2=0.85$). The stream was instrumented with a pair of long-throated flumes with different measurement ranges. Each flume was equipped with pressure transducers coupled to data loggers for automatic recording of water levels. A theoretical rating curve was applied that had good agreement to periodic flow measurements (Johansson and Juston, 2007).

There were 10 on-land groundwater wells in the surficial deposits within the catchment boundary. Both stream height and groundwater elevations were recorded at high frequency and reduced to daily average values for modeling purposes.

**3.2.3. Model rationale and structure**

The model used in this study was developed from the HBV model, a well-known lumped conceptual catchment hydrology model (Bergström 1976; Siebert, 1997). One standard feature of the HBV structure is fixed field capacity storage for representing the unsaturated zone of the soil profile. However, groundwater measurements in the study site indicated near-surface levels generally within the upper 1-m of the soil profile that challenged this conceptualization (Juston et al., 2007). Thus, HBV was adapted to site conditions by elaborating more detailed interactions between saturated and unsaturated zone (Fig. 8). Specifically, the storage capacity of the unsaturated zone became time-dependent and several new physical variables were specified to

![Figure 8. The HBV model was adapted to site-specific conditions at Forsmark, namely close coupling between the unsaturated zone and near-surface groundwater (green circles indicate enhancements). There were 11 model parameters that were calibrated (indicated in bold red).](image-url)
approximate the water holding capacity of the till. The resulting model retained the capacity to simulate catchment discharge, but had a new capacity to link the groundwater storage variable to measured depth to groundwater from the monitoring network (Fig. 8).

3.2.4. Calibration and uncertainty estimation
The model was calibrated with the objective of matching both discharge and groundwater dynamics in the small catchment. The informal likelihood, $R_{\text{eff}}$, was used to evaluate residual errors in discharge simulation. The 10 near-surface wells in the study catchment indicated different mean elevations and seasonal amplitudes but generally similar temporal co-variance (Juston et al., 2007). An LOA approach was used to constrain groundwater simulations in the model, but with recognition that the true catchment response was not knowable from this limited sample of information. A time-varying core acceptance range for simulations was defined from the 95% confidence interval of groundwater observations in the region; and a time-varying fuzzy boundary from the 99.9% interval (see Fig. 5 for definitions). The likelihood function for groundwater simulations was the average LOA score from the series of scores at each time step. A single pooled likelihood was defined to guide calibration by averaging $R_{\text{eff}}$ and LOA scores (both having maximum possible values of 1.0). Uncertainty intervals for model parameters and simulations were calculated from the top 200 scores from $4E6$ random samples of the model parameter space.

After a baseline calibration was established, several calibration experiments were performed using sampled subsets of the available data (Fig. 9). The resulting parameter sets from each subset calibration were evaluated in their ability to simulate the full daily data series. The purpose of these experiments was to utilize the revised model and site data as a platform to explore more general questions about the information value of data for model calibration (e.g., Gupta et al., 1998; Seibert and Beven, 2009).

3.3. Study II
3.3.1. Research questions
- Can a modeling approach provide new insights on soil carbon decomposition processes in a long-term plot-scale agricultural experiment?
- How dissimilar are GLUE and BMCMC uncertainty bounds in comparative calibration?

3.3.2. Study site and data characteristics
The “Frame Trial” experiment (Fig. 7) at the Swedish University of Agriculture in Uppsala has maintained 15 various plot-scale agricultural treatments since 1956 (Persson and Kirchmann, 1994). It is one of the longest running agricultural trials in the world (e.g., Richter et al., 2007). Six treatments from the Frame Trial were selected for study; each of these had received consistently different organic carbon “inputs” to the upper soil horizon over the duration of the study. Carbon inputs occurred via direct organic amendments (e.g.,

![Figure 9. In Paper I, several calibration experiments were conducted using weekly, monthly, quarterly, and “informed” data subsets. Subsets were sampled from the daily series, not interval-averaged. For reference, the complete daily time series had 1065 observations; the informed observer subset had 53.](image-url)
straw and manure) in three treatments and/or continuous belowground root contributions (all treatments). Aboveground biomass was harvested and thus not a major source of carbon to the underlying soil. The six treatments included a null treatment (i.e., fallow), a basic crop cycle (no fertilizer or amendment), fertilization (increased productivity), straw application, straw plus fertilization, and manure application. Each treatment had four replications at the study site.

Data on carbon inputs were either measured or estimated. Crop residue and manure applications were normalized with lab procedures to ~0.38 kg-C/m²2 every two years; this was considered a low level uncertainty in the modeling (although certainly not without uncertainty). Below-ground inputs were not sampled but were calculated with allometric equations based on estimates of above-ground production. High uncertainty and error accompanied these calculations (Bolinder et al., 2007; Kätterer et al., 2011); yet these calculations accounted for 20-100% of net carbon forcing for the different treatments in the model (Kätterer et al., 2011). The concentration of soil organic carbon (SOC) in the treatments was sampled biannually. The average coefficient of variation amongst treatment replicates was ~10%, suggesting a moderate level of stochastic variability in observations. Some systematic bias was identified in lab processing techniques over time that contributed visible artifacts in the output series (Kätterer et al., 2011). Conversion of SOC concentrations to estimates of SOC storage in the upper 20 cm (kg/m²) were based on interpolated time series of bulk density from four measurements in each treatment over the 51-year duration. Thus, this introduced further uncertainty in the soil storage time series due to unknowns in gap-filling soil physical properties.

3.3.3. Model rationale and structure

The Introductory Carbon Balance Model (ICBM) was originally developed and parameterized specifically to describe these same six Frame Trial treatments (Andrén and Kätterer, 1997). The original ICBM simulated soil organic carbon dynamics with two storages, each with a conceptual “age” (Fig. 10). Soil carbon inputs (roots, crop residues, manure) are added to the young pool (with different sub-storages for different quality inputs). Some young carbon is oxidized to the atmosphere, and some is decomposed to more resistant states. Both processes occur in proportion to the storage turnover rate and a climate-dependent biological activity factor, r. This activity factor was normalized to approximately 1.0 at the Frame Trial site (Andrén et al., 2007), so it was not considered further in this study. Soil C pools were initiated to field measured values in 1956 (~4.3 kg/m²).

During a preliminary study of field trial from a site in arid Kenya, it was determined that the original ICBM structure was not adequate to describe trends in those data (Andrén et al., in review). One hypothesis was that the model did not adequately account for carbon with very long turnover times relative to the

Figure 10. The ICBM in Study II. Carbon inputs were crop materials (including roots) or manure. The young pool, Y, contained sub-storages for inputs of different quality, each with humification factor, h. A fraction of the initial soil carbon content was assumed effectively inert. Calibration parameters are indicated in bold. The climatic activity factor, r, was normalized to 1.0 in this study.
decade interval in the study. It was proposed to add an additional carbon pool in the model that was effectively inert over the period of study. This pool did not alter the conceptualized decomposition dynamics in the model but did alter the initial partitioning of C. The fraction of initial carbon assigned to the inert pool was treated as a calibration variable in this new formulation (Fig. 10).

3.3.4. Calibration and uncertainty estimation
This insight from the Kenyan experiments motivated a thorough re-investigation of the Frame trial dataset with the new ICBM framework. Two calibration intervals were considered, 1956-91 and 1956-07, using both informal GLUE and BMCMC algorithms (Table 3). The BMCMC results have not been previously reported (and were, in fact, produced specifically for the Kappa). The calibration objectives were to jointly simulate time series trends of the six Fame Trial treatments with single parameter sets. The GLUE experiments utilized a soft informal RMSE likelihood with acceptance threshold of 0.19 kg/m². This value was greater than the average standard deviation in replicate samples (0.14 kg/m²), but very close (as will be shown in Results) to the best achievable RMSE scores. The BMCMC calibrations were conducted rather ad hoc, with no detailed consideration or validation of assumptions in the error model, using a BMCMC algorithm similar to Juston and DeBusk (2011).

3.4. Study III
3.4.1. Research question
- Can a process modeling approach support, and provide additional insights on, the lowest achievable outflow phosphorus (P) concentration observed in an engineered wetland?

3.4.2. Study site and data characteristics
The study site was a 920 ha constructed wetland situated between agricultural lands and protected Everglades areas in south Florida USA (Fig. 7). The treatment wetland has been operated continuously for P removal since 2001, and is part of a larger system of treatment wetlands designed for stripping P from inflowing water to protected Everglades areas. The study site has been managed for a dense cover of submerged aquatic vegetation, which has been identified as having unique P removal capacity in the Everglades environment (in comparison to more conventional emergent vegetated wetlands). However, a recent study suggested that there may be a fundamental limitation to the lowest achievable outflow P concentration from these systems between 13-17 µg/l (Juston and DeBusk, 2011).

A wide variety of data were available from the study site to support a modelling study. Inflow and outflow rates have been measured continuously and P concentrations sampled ~weekly since startup. These data were considered to be of relatively high quality. Phosphorus concentrations have also been periodically sampled (n=28) at 45 regularly-spaced internal sites in the cell, providing snapshots of longitudinal P removal gradients; there was considerable temporal variability in these snapshots (Juston and DeBusk, 2011). Additionally, P in accrued sediment was sampled (n=54) in 2010 to help establish a spatial mass balance in the cell (g/m²). Some measurement error inevitably accompanies field-collected soil cores such as these; however this was likely dwarfed by the high intrinsic spatial variability in these samples, as well as longitudinal patterns, similar to what have been observed in other similar wetland systems (Grunwald et al., 2004). Additionally, composition of submerged vegetation species has been monitored annually since startup; here again, there were expectations of considerable spatial variability in these data (e.g., Jager and King, 2004).

3.4.3. Model rationale and structure
A P cycle model was developed as an extension of the simple and well-known steady-state NKC* model (Kadlec and Knight, 1996; Fig. 2). The NKC model simulates constituent removal from the water column with a 1st-order rate constant, k, limited by an intrinsic background
concentration, $C^*$, in a hydraulic system of $N$ tanks-in-series (TIS). The TIS formulation provides a means to simulate longitudinal gradients; accordingly NKC* has become a standard tool for modeling treatment gradients in a simple conceptual framework (Kadlec and Wallace, 2008). However, the NKC* model was not adequate to describe the sediment $P$ profile from this site.

The process basis for the extension of the NKC* model was hypothesized from an extensive survey of literature regarding the $P$ cycle in SAV-dominated lake systems. The new model structure was based on the following notions: decomposition of SAV tissue associated with biomass turnover is rapid and acts as a $P$ source to the water column; rooted SAV can effectively “pump” $P$ from sediment to the water column via this mechanism; the presence of SAV mediates efficient $P$ removal processes separate from the $P$ cycling thru SAV plant tissue itself. In this way, it was hypothesized that $P$ removal in SAV systems is self-limited due to an internal load created by the biomass $P$ cycle, and $C^*$ was from as-yet unexplained (residual) processes.

3.4.4. Calibration and uncertainty estimation

The calibration scheme utilized limits of acceptability (LOA) to constrain model-predicted water column and sediment $P$ gradients with recognition of temporal and spatial uncertainties in the field-measured observations. Data treatment assumed a physical correspondence between longitudinal position in the direction of flow and mathematic representation of state variables in TIS; thus, for example, the field samples in the first one-sixth of the study site (Fig. 7) were used to constrain predictions in the first of the six TIS in the model (Fig. 11). The threshold criterion for acceptable

Figure 11. The wetland $P$ cycle model in Study III. Green circles indicate new features added to the NKC* framework. Calibration parameters are in bold red; three additional were required to specify biomass $P$ turnover as a function of position in the wetland treatment gradient.
simulations was defined from the interquartile range (IQR) of observations corresponding to each model. There were a total 20 criteria (Table 4); however one criterion was eventually relaxed due to physical considerations, as described in Results. All parameter sets that achieved a score of 19 were treated as equally likely to describe observations.

3.5. Study IV

3.5.1. Research question

- Was there evidence of a change in flow patterns over time in a 45-year discharge series, with consideration of flow duration curves and rating model uncertainty?
- Are issues of epistemic error ever really that important for hydrological hypothesis testing?

3.5.2. Study sites and data characteristics

The 650 km$^2$ study catchment was located west Kenya (Fig. 7). Loss of forest cover and other land use transitions in the basin have become an issue of concern in regards to sustaining historic flow patterns to the downstream ecosystem. River levels (i.e., stage) have been measured daily at the gauge station that defines the catchment outflow with good consistency since 1964. These data were considered high quality in comparison to uncertainties in stream rating (i.e., discharge) measurements. There were 114 discharge measurements available for 1964-07. A low concrete weir was constructed across the river channel in the early 1990’s; 88 of the ratings were from before and 33 from after construction. Discharge (i.e., stream flow) was measured using the area velocity method, from which both random and systematic errors can be expected (Sauer and Mayer, 1992). There was a moderate degree of scatter (i.e., uncertainty) in stage-discharge data pairs from this site. There were several site-specific issues that raised particular concern over non-repeating, non-random error in these data, including: a high number of outliers, suggesting at least periodic systematic difficulties in the field or reporting; multiple observers (from site records); multiple stream cross-sections for observations (from site records); and high potential for seasonal and longer-term variability in physical characteristics of the stream bed. Thus, significant random and non-random uncertainty was suspected in the data scatter, although a reduction of contributions was not possible. Therefore, this study proceeded assuming both required consideration and neither was negligible.

3.5.3. Model structure

The rating curve was based on an empirical power law model, which transformed to a linear relationship when both stage and discharge observations were log-transformed (Fig. 12). Visual assessment of the data indicated that a two stage rating model was appropriate (i.e., a piecewise linear model). Furthermore, the slope of the lower segment in the rating model changed after the construction of the concrete weir.

3.5.4. Calibration and uncertainty estimation

The rating model was calibrated using a BMCMC algorithm. The error model treated the scatter of data around each segment of the rating model (Fig. 12) as-if it was random. The likelihood formulation assumed the data dispersion around each segment of the rating model...
follow-up investigations (beyond the scope of this study), such as: was this change due to deforestation and/or climate variability?

4. Results

Presentation of results is aimed at cross-study comparisons. The Papers elaborate substantially more detail.

4.1. Data

All data used in this thesis was field collected. The categorization framework introduced earlier (Fig. 5) was useful to synthesize field data characteristics and their relationship to model usage. The character of error and uncertainty in input and calibration data varied amongst modelling studies (Fig. 13). Study I had low-uncertainty input data, one (assumed) low-uncertainty calibration target ($Q$), and a second high quality calibration dataset that had high spatial representation uncertainty in reference to model usage ($GW$). Study II utilized high-uncertainty input and calibration data. Study III was driven by low- and high-uncertainty inputs, and calibrated to high variability-uncertainty outputs. Study IV had low-uncertainty input data, but utilized calibration data ($Q$) that was suspected of substantial random and non-random errors.

4.2. Models

Visual assessment of the calibration results suggested that model equations and parameter calibrations were generally successful at simulating the targeted observations (Fig. 14 and 15). The calibrated hydrological model of Study I captured seasonal trends in groundwater response with similar variance and covariance as the cloud of groundwater well data; seasonal dynamics in discharge, and peak events as well, were well-captured. The calibrated soil carbon model of Study II jointly captured central tendencies in carbon storage trajectories from the six field trial treatments. The calibrated wetland $P$ cycle model of Study III reproduced steady-state trends in water column $P$ and sediment $P$ storage gradients well-centered within the cloud of field data for each variable. The calibrated

Figure 12. The rating model with five parameters, indicated in bold red. The two lower limbs correspond to before and after a low concrete section control was constructed.

model was normally distributed and with steady variance, $\sigma_i^2$, such that:

$$L(O|\Theta, M) \propto \prod_{j=1}^{n} \prod_{i=1}^{k_j} \text{exp} \left[ -\frac{1}{2\sigma_j^2} \right]$$

(Eq. 6)

Procedural details for the BMCMC algorithm were essentially the same as in Juston and DeBusk (2011). Values for the three error variance terms, $\sigma_i^2$, were calibrated within the BMCMC inference.

3.5.5. Hypothesis testing

The calibrated rating model and uncertainty intervals were used to infer a discharge time series with uncertainty intervals from two eight-year periods in the daily gauge height series: 1964-71 and 2000-07. Flow duration curves (FDCs) were calculated for these intervals with accompanying uncertainty intervals, and compared to one another to test this null hypothesis: there is no evidence of altered flow conditions in the basin between these two eight-year data intervals. This hypothesis would be accepted if FDC uncertainty intervals were fully overlapping for the two periods. Separation of FDC uncertainty intervals (either partial or complete) would provide basis to reject the null hypothesis. A rejection might lead to
rating model in Study IV captured the central tendency in the stage-discharge data.

The estimated uncertainty bands in each calibration study reflected different likelihood measures and acceptance criteria (Table 3), and thus had different meaning. For the Study I results, the GLUE uncertainty bounds represented the simulated extremes from the top 200 likelihood scores from 4E6 random parameter realizations. This was an arbitrary criterion, aimed at elucidating equifinality at “the upper echelons” of the performance space; differently chosen threshold criteria (e.g., top 100 or 400) would yield narrower or wider boundaries. The top 200 scores contained discharge simulations with $R^2$ scores in the range of 0.79-0.87, groundwater scores in the range of 0.89-0.97, and net averaged scores of 0.88-0.90. These were indeed good scores (considering max possible = 1.0 in all cases), and provided basis to assess a “successful” calibration. However, the subjective basis of the threshold interfered, to some extent, with more detailed model evaluations. For example, one might assess that the model often lagged in predicting hydrograph recession (e.g., April 2004 and 2006, Fig. 14); however selection of a wider uncertainty band could erase that conclusion. This illustrates one problem with soft informal likelihood criteria.

The GLUE uncertainty bands in Study II were similarly selected. In this case, the acceptance threshold was set for simulations with RMSE <0.19 g/m$^2$ and the boundaries were derived from 250 simulations that met this criterion from 50E6 realizations. The lowest identified RMSE score in the Monte Carlo search was 0.184 g/m$^2$; so here again, this was a depiction of equifinality at the upper echelons of the parameter space. However, the uncertainty bands did not cover the data scatter in spite of the fact that the threshold value was in excess of the standard error in field replicates. Thus, although the central tendencies in the six treatments were well-simulated, there still appeared considerable unexplained phenomenon with shorter-term dynamics (Fig. 14).

The GLUE uncertainty bounds in Study III were determined by 550 simulations that satisfied 19 of the 20 predefined LOA...
(Table 4) from 100E6 random parameter set realizations. No simulations satisfied all 20 criteria, and all 550 failed on the same criterion: sediment P storage in the first tank of the model was consistently over-simulating in comparison to the IQR of data scatter in front region of the study site (Fig. 15). Was it the model or data that contributed to the consistency of errors? Further consideration suggested the data was most suspect; there was higher variability in sediment accrual in this region due to
physical factors (specifically, jetting and translocation of sediment from inflow culverts), thus suggesting a possibility that the region was under-sampled. In turn, this suggested a possibility that this particular acceptance criterion was perhaps based on inadequate information. Additional field data would be useful and important to confirm this. However in general, the new model structure was considered quite adequate to describe the available data, and this provided rationale to consider insights

Figure 15. Data observations, calibration results and uncertainty intervals for Studies III and IV.
on calibrated $C^*$ values in a broader context (discussed below).

The uncertainty bounds in the rating model calibration (Study IV) resulted from probabilistic acceptance criteria built in to the BMCMC algorithm. Model parameter and prediction uncertainty bounds were estimated (Fig. 15), which have direct analogy to parameter and prediction uncertainty bounds in linear regression under the assumption of random, Gaussian errors (e.g., Fig. 3). When this assumption is valid, the difference between model uncertainty bounds would be due to measurement noise with magnitude $\sigma$ (Eqn 6). However, it has been established that it was unlikely scenario in Study IV, since substantial non-random errors were suspected (Fig. 13). Under this condition, it has been recognized in other studies that the BMCMC algorithm inevitably represents “non-repeating systematic error… as part of the regression noise” (Petersen-Overleir et al., 2009). This calibration procedure could not explicitly account for these systematic error contributions to parameter uncertainty intervals. Thus, the resultant parameter uncertainty bounds from the Study IV BMCMC analysis must be interpreted (and used) with some care, as they were likely over-conditioned (i.e., too narrow) due to treating systematic error “as-if” it were random.

4.3. Residual Errors

Information in residual errors represented the cumulative filtering of input uncertainties thru model equations and comparisons to uncertain calibration data. The different models, data, and calibration methods yielded a wide variety of residual error series (Fig. 16).

There was no native stochasticity evident in the discharge simulation error from Study I. Input and discharge series in this study were considered amongst the highest quality data used in the studies (Fig. 13), yet the residuals for the discharge simulation were the most patterned of all (Fig. 16). This suggested that patterns were introduced largely by the model equations. It is however interesting to consider the July-August 2005 sub-interval in more detail (see circled regions for Study I in Fig. 14 and 16). For the most part, the model failed to simulate the moderate discharge event during this two-month interval. Was it the model or data that contributed to the consistency of errors? Three other regional gauge stations reported a similar discharge event during this interval; thus reasonably ruling out error in the discharge data. However, interestingly, the July-August interval coincided with a period of missing precipitation data (the only one in the on-site records), during which the series was gap-filled from stations 20-30 kilometers distant. Thus, it seemed most plausible that the gap-filled data might have yielded the simulation error. This suggested another example of how “knowledge errors” can introduce patterns in model residuals.

In general, residuals from Study II were more scattered (Fig. 16), but still indicated patterns both between and within treatment time series. For example, there is a curious commonality in response in the three treatments that received straw and manure amendments during 1986-1992. Interestingly, this corresponded to a suspected temporal interval of systematic error in lab procedures (Kätterer et al., 2011). Furthermore, there was net bias in some of the residual series for individual treatments. This suggested that this calibration had value for understanding the treatment responses in a unified framework, but less value for understanding the response of each individual treatment.

For the LOA likelihoods (Study I and III, Fig. 16), errors occurred when simulations exceeded the limits, while all values simulated within the core acceptance range were not in error. For example, residuals in groundwater simulations in Study I demonstrated patterns (relative to the mean)(Fig. 16), but much of these patterns were within the predefined representation uncertainty of the available groundwater data, and thus not in error.

The standardized residuals in Study IV appeared stochastic and were approximately normally distributed with relatively steady variance (Fig. 16). The stochastic appearance is interesting, considering the suspected non-
random error in these data. This highlights a need to consider not just the appearance of residuals, but their actual nature as well (e.g., Taylor, 1997; Beven et al., 2011; Fig. 5 and 13).

4.4. Information and uncertainty

4.4.1. Study I

Calibration experiments with the catchment model in Study I and sampled data subsets were illuminating in regards to understanding information contained in time series data for

Figure 16. Residual errors calculated from the difference of model simulations and data observations in each Study.
There existed parameter sets that were derived from calibration to as few as 53 samples of the complete 1065-day series that provided virtually indistinguishable performance from calibration to the full 1065 days (e.g., “informed” observer samples; Fig. 9 and 17). An additional calibration suggested that even a minimal amount of groundwater data (e.g., quarterly samples) could provide significant help constrain model parameters in comparisons to having no groundwater data at all (Fig. 17). These results suggested there can be much redundant information in time series data for model calibration. A comparison of performance measured from the full data (D-D), the discharge only (D-N), and groundwater only (N-D) calibrations corroborate insights of Fenicia et al. (2008) on the value of “orthogonal” data to challenge a model to simultaneously represent several aspects of catchment behavior.

4.4.2 Study II

Calibration trials with the soil carbon model of Study II and different likelihood functions (Table 3) illuminated several perspectives. The originally-published GLUE calibration (Paper II) presented marginal distributions for ICBM parameters from calibration to two data intervals (1956-91 and 1956-07) using an informal RMSE likelihood and an acceptance threshold of RMSE < 0.19 kg/m². Those distributions are reproduced in Figure 18. Concurrent with Kappa preparations, a second calibration was performed using a BMCMC algorithm (Table 3). Interestingly, the posterior distributions from the two methods were, for the most part, indistinguishable for
calibrations to both data intervals (Fig. 18). In one way, this was not an unexpected result since minimizing an informal RMSE likelihood is essentially the same as maximizing a formal likelihood function (assuming errors were random and Gaussian, as was assumed here). However, it is also well established that uncertainty bounds and parameter distributions from a GLUE analysis are dependent on the threshold criteria when using soft likelihood functions, such as RMSE (e.g., Fig. 19). Thus, this result was from another perspective completely surprising given that the subjective threshold criteria chosen three years prior provided near exact corroboration (and vice versa) to the uncertainty intervals estimated with the BMCMC algorithm (Fig. 18).

It is interesting to further consider the identification of the inert fraction model parameter in this study (Fig. 18). Posterior distributions for the inert fraction from the 1956-91 calibration suggested some possibility that this parameter could still equal zero and adequately explain the data. However, this changed with the addition of 16-years of subsequent data (Figure 14); thus the extended time series contained valuable information to support the new proposed process hypothesis in the model structure (Fig. 10). This was evident in both GLUE and BMCMC results.

It is also interesting to consider the covariant relationship between old pool turnover rate and inert fraction parameters in Study II (Fig. 18d). There has also been some debate as to if laboratory fractionation analyses can be useful to precisely initialize storage pools in soil carbon models such as ICBM (Paul et al., 2006). However, the results from this modelling study turn this argument around somewhat. Any additional constraint on possible values for the inert pool, beyond what was identified from the 51-year SOC time series, would automatically constrain the range of covariant turnover rates (Fig. 18d). Thus, even imprecise information from lab chemical analyses could be of value to reduce model parameter and hence long-term predictive uncertainties.

4.4.3. Study III

After the simulation capacity of the P cycle model in Study III model was established (Fig. 15 and 16), it was of high interest to evaluate the model predictions and insights for C*. Often, it is not prudent to compare parameters between model studies because they can lose meaning outside of their specific conceptual model. However, the background P concentration parameter, C*, in Study III had direct relevance to the same parameter estimated in a prior study (Juston and DeBusk, 2011). In fact, Study III was motivated specifically toward new insights on this parameter. The model and data basis in the previous study were very different than in Study III; the previous study employed a database of historic annual scale in-out P performance from several sites in the same project (including the Study III site), an empirical model, a formal likelihood function, and a BMCMC algorithm for calibration. The posterior distribution for C* from this study is reproduced in Figure 20 (5-95% C.I. = 13-17 μg/L).

![Figure 19. Width of parameter distributions and prediction bounds in Study II as a function of GLUE acceptance criteria.](image)
The posterior density for $C^*$ produced from the process model calibration in Study III had a very similar range, especially in consideration that the allowable range for $C^*$ in this calibration was 0-40 μg/L. The cumulative constraint of the in-out and internal gradient mass balances that were specified in this study (Table 4) provided for a relatively narrow estimation of plausible $C^*$ values (Fig. 15) with about the same range and distribution as identified in the previous study. This is interesting given the very different data and methods involved. The process model implementation in Study III also provided some “extra” insight on the nature of $C^*$ in these systems, suggesting that about a third of the observed $C^*$ might be explained by internal loading caused by biomass turnover and decomposition of the SAV (Fig. 20). This is interesting, as it suggests that the lowest achievable outflow $P$ from these systems might be self-limiting, even though these systems are highly efficient at treating $P$ to this background level.

4.4.4. Study IV
A discharge time series was inferred from the calibrated rating model (Fig. 15), along with estimates of discharge uncertainty intervals (Fig. 21). Two precedents have been established in the hydrological literature in regards to hypothesis testing with discharge uncertainty intervals inferred from BMCMC calibration of rating models: one suggested model parameter uncertainty intervals were appropriate for testing (Reitan and Petersen-Øverleir, 2008); the other that total prediction bounds were appropriate (Moyeed and Clarke, 2005). It has already been established that parameter uncertainty intervals were not suitable for hypothesis testing here since they were likely over-conditioned due to the treatment of residual errors as-if they were random (a suspected fallacy) in the BMCMC calibration. In fact in this case, evaluation of FDC uncertainty intervals based on parameter uncertainty might have led to some false conclusions.

![Figure 20. Distributions for the background P concentration ($C^*$) in SAV wetlands from Study III and a past study (Juston and DeBusk, 2011). $C^*$ residual is defined in Eq. 5. The principal issue of comparison is that no $C^*$ estimate suggested possibilities for achieving outflow $P < \sim 13$ μg/L. This corroboration provides additional decision support for Everglades restoration.](image)

![Figure 21. Uncertainty in discharge time series inferred from rating curve uncertainty in Study IV (Fig. 15). The horizontal lines partition the flow according to rating model segments; “knee” defines the flow regime above the breakpoint in the two-segment rating model (Fig. 15).](image)
(e.g., Type 1 error) regarding the hydrologic response in the basin (i.e., overestimating differences between response intervals; Fig. 22b). At this site, the total uncertainty intervals in FDCs provided a more conservative basis for hypothesis testing due to the mixed nature of errors in rating data observations (Fig. 13). Comparison of these intervals suggested that the difference in flow patterns in recent years exceeded the uncertainty in the data only for flows with >98% exceedance probability (Fig 22c). This was a small range of the overall flow domain, but was nonetheless an indication that the basin had at least short intervals with lower flows during 2000-07 than were experienced in the historic record. More detailed analysis of the 2000-07 time series (not shown) provided supplemental support: there were three separate dry season intervals during 2000-07, each a few weeks long, which contributed to this altered flow pattern. Thus, hypothesis testing with consideration of uncertainty intervals provided a more confident basis to motivate further investigation of this response.

5. DISCUSSION

5.1. Learning about uncertainty schemes

Beven (2008) coined the phrase “uncertainty about uncertainty” in the context of unknown sources of uncertainty in modelling and as a characterization of the ongoing dialogue between advocates of GLUE and BMCMC methods. Here, the notion of uncertainty in uncertainty intervals (\(u^2\)) is introduced. \(u^2\) was endemic in the four studies, although it did not, for the most part, interfere with producing meaningful results.

5.1.1. Informal likelihoods

GLUE schemes that investigated equally-plausible sets of “best possible” simulations (Study I and II) had intrinsic \(u^2\), as the reported intervals (e.g., Fig. 14 and 18) were dependent on subjective threshold criteria; changing the criteria would have changed the
5.1.2. Formal likelihoods

Parameter and prediction uncertainty intervals from BMCMC algorithms using formal likelihoods are often associated with probabilistic interpretation (Gelman et al., 1995). However, this interpretation was somewhat corrupted in Studies II and IV because residual errors from calibrations in either study were not wholly random; thus these results needed to be interpreted with some care.

In Study II, the input data (i.e., deterministic calculations of below ground carbon inputs) were determined to contribute high uncertainty to the modelling process. These calculated data were understood to contain systematic and non-stationary errors in year-to-year carbon input estimates that could not be further refined at the time of study (e.g., Bolinder et al., 2008). Therefore, the posterior distributions from the BMCMC algorithm (Fig. 18) did not have an objective quality of determination because of the known and non-random errors they were based on; without this objective quality, the results required subjective interpretation (i.e., were they still valuable or not?). The relevant question became: was probabilistic interpretation really necessary to derive meaning from the BMCMC results in this study? This will be further addressed.

In Study IV, BMCMC analysis produced parameter and total uncertainty bounds for the rating model (Fig. 15). But a key question was: which of these intervals was “correct” for hypothesis testing of predicted responses (Fig. 21 and 22)? If residual errors in this study were (hypothetically) wholly random, then it is agreed that the parameter uncertainty bounds would have been appropriate for hypothesis testing (e.g., Reitan and Petersen-Øverleir, 2008; Fig. 3). In this case, scatter in rating curve residuals would have reflected noise that was not important to the “true” knowledge of stream flow (Reitan and Petersen-Øverleir, 2008). On the other hand, if residual errors in this study were (hypothetically) wholly non-random, then it is suggested that the total uncertainty bounds would have been most appropriate for inference and hypothesis testing (Fig. 15, 21, and 22). In this case, scatter in rating curve residuals would have contained information that was indeed important to the “true” knowledge of stream flow. However, given the suspected nature of residual errors in this specific study (i.e., a non-reducible blend of random and non-random contributions; Fig. 13), this again suggested evidence of some uncertainty in
uncertainty intervals, where the two hypothetical cases (wholly random and wholly non-random) provided “bounds”, in a conceptual sense, on the \( u^2 \). Fortunately though, this \( u^2 \) was not a limiting consideration in this study (either), since small changes in the river flow pattern were detected using the conservative total uncertainty intervals (Fig. 22c).

These studies suggested that the posterior distributions derived from BMCMC algorithms should not automatically be associated with probabilistic interpretations and objectivity. Similar views have been expressed in some other studies with BMCMC algorithms. For example, after exhaustive statistical error model development and a rather strong-minded probabilistic orientation, Yang et al. (2007) ultimately concluded: “This makes it difficult to interpret the posterior marginals as realistic uncertainty estimates of the parameters (despite the realistic uncertainty estimates of the predictions dominated by the error model).” This comment has direct analogy to the discussion and conclusion just given for change detection in Study IV, since predictions dominated by the error model were indeed those represented by the total uncertainty bounds (Fig. 15).

Despite the recommendation to interpret parameter distributions from Studies II and IV cautiously and not as probability densities, results from the BMCMC algorithm were still valuable to support outcomes in both studies. An important dialogue and debate continues in the scientific literature (e.g., Clark et al., 2011; Beven et al., 2012; Clark et al., 2012) specific to the appropriate use of BMCMC and/or GLUE methods for environmental model calibration and uncertainty estimation, particularly in the context of process modelling of high resolution time series data (e.g., daily hydrological series). Much of this exchange is based on issues surrounding data, models, and highly patterned residual errors similar to those found in Study I here (Fig. 8, 14, and 16); however BMCMC methods were not tested to Study I data in this thesis. Nevertheless, the views in this thesis are relevant to this discussion as they provide some alternative perspectives.

5.1.3. Equifinality in intervals?

Traditionally, the word equifinality has been used in the context of model structure and parameter identification (Beven and Freer, 2001). Vrugt et al. (2008) expanded usage to uncertainty estimation algorithms in a comparison of total prediction bounds from BMCMC and GLUE prediction bounds. Here, there is evidence to expand the principle of equifinality one step further. Consider the parameter distributions in Study II (Fig. 18). The intervals from GLUE and BMCMC algorithms were virtually identical; was there adequate evidence to reject either of these hypotheses for uncertainty estimation? Once the BMCMC results were “relieved” of probabilistic interpretation (by consideration of the nature of residual errors), than neither algorithm produced “better” results than the other in this case, or perhaps alternately stated, both approaches produced results that were at least equally wrong and/or subjective. This type of equifinality might be worth consideration in other studies as well.

5.2. Learning about processes

Despite \( u^2 \), to what extent were uncertainty estimations useful in these calibration studies toward learning about processes? In other words, was a “process insight” feedback (Fig. 1) activated by these methods? Two examples follow.

In Study II, the evaluation of the uncertainty intervals for model parameters (Fig. 18) was the key step to infer that the inert storage was a useful addition to the model to describe the data (conditional on the rather large uncertainties in the input data). However, the accompanying process insight was the deeper contribution here. Some soil carbon appears extremely resistant to decomposition in the SLU Frame Trials. This insight has only become apparent over time (Fig. 18). There exist only a handful of agricultural field trials in the world that have operated as long as the SLU Field Trial (Richter et al., 2007). The uncertainty estimation in Study II provided an insight
that (now) seems robust enough to consider possible implications for interpreting data from other systems as well.

In Study III, a key issue was not “is there one value of the C*-model parameter that fits the data”, but was instead “is there any evidence at all in the available data to support an hypothesis that this wetland systems can achieve lower outflow P”. Uncertainty estimation was not just useful to addressing this important question, but essential (Fig. 20). The corroboration of Study III distributions based on largely independent data with past results strengthens the evidence for C* in these systems. However, the process insight from this study might be the more useful result toward decision support for ecosystem restoration planning. If the lowest achievable P concentration is indeed self-limited by internal loading from the specific vegetation in these systems, then it might be necessary to transition to a different community type downstream in order to achieve lower P effluent from the engineered wetland treatment systems. This suggestion represents a new hypothesis that should be further investigated with new data and process insights.

In summary, it is argued that uncertainty estimation in the model calibration studies in this thesis have provided robust insights. In other words, results of this thesis have supported the opinion of Box (1976), that explicit expressions of uncertainty in modeling “increase(s) greatly the probability that the investigator will be led along a true rather than a false path.”

5.3. Learning about data

To what extent were methodologies in the Studies useful toward learning about data? In other words, was a “future data needs” feedback (Fig. 1) activated by these studies? The results of Study I demonstrated that subsets of time series data, some relatively sparse, can provide very nearly similar constraint on model calibration and parameter identification as complete daily series. Turning this around, calibration using a reduced amount of data might provide similar results as calibration on the full data set, provided the subsets are selected with some hydrological reasoning (e.g., the “informed” strategy, Fig. 9 and 17). This finding has implications for sampling design. For example, if budget is limited, Study I suggests a possibility that infrequent samples of stream flow, complimented with some extra samples during dominant events, could provide valuable information for model calibrations. Although not shown in this thesis, comparison of posterior parameter distributions provided additional support to this assessment (Paper I). Results in Studies I and II (Fig. 17 and 18d, respectively) reinforced other recent developments on the value of complimentary data to support multi-objective calibrations (e.g., Feni et al., 2008). As suggested above, the “new” hypothesis that evolved from Study III suggests a need for new and different types of field data. Lastly, Studies II and IV demonstrated that interpretation of results from existing probabilistic calibration algorithms can be limited by the nature of uncertainties in environmental data. This suggests a need for more cross-sectional studies about data, and the level and nature of uncertainties in different types (e.g., McMillan et al., 2012).

5.4. Posterior analysis

Each study represented a “significant approximation” in its own way, in the way that substantial uncertainties in data and model equations combined and overlapped yet somehow produced credible simulations in the end (Fig. 14 and 15). There would be several opportunities to improve the calibration and uncertainty schemes in these studies; thus, it is useful to evaluate methods with hindsight.

In Study I, the LOA time series for groundwater simulations were calculated from confidence intervals around the mean in the set of groundwater data. Thus, there was an implicit assumption that the set of groundwater data were normally distributed at each time step. In hindsight, it seemed unlikely that his assumption would hold true at each time step. Perhaps more importantly, this data treatment contradicted one of the
key motivations for adopting the GLUE methodology in the first place: to avoid potentially incorrect assumptions on the distribution of errors. Therefore, it is recommended for future studies to investigate non-parametric measures of dispersion for setting LOA, such as the IQR approach in Study III. Additionally, the uncertainty in discharge time series in Study I was not explicitly investigated, although some rating data were available that could have been used for that purpose. Even if this discharge series was considered higher quality than that, for example, in Study IV, it is important to establish benchmarks of how different quality data interact with uncertainty in model predictions (e.g., McMillan et al., 2012), thus an explicit analysis would have been useful.

The results of model calibrations in Study II represented perhaps the most “significant approximation” of all, in consideration of the high uncertainties in both input and calibration data (Fig. 13). The key to deeper insights in future studies such as this, in regards to decomposition processes and calibration uncertainty estimation, lies in resolving some of the high uncertainty associated with estimating below ground inputs. Unfortunately, this remains a challenging and possibly fundamentally limiting issue in modelling soil carbon dynamics (Bolinder et al., 2008).

It is also interesting to consider the higher confidence stated for the Study IV conclusions in the Kappa in comparison to the associated Paper IV manuscript (in review). What happened during the three months between manuscript submission and Kappa preparation? No new analysis was produced. In fact Figure 22 in the Kappa was plotted from the same data as Figure 7 in Paper IV, except Figure 22 shows the y-axes log-transformed for better resolution of baseflow conditions. The conclusions of Paper IV stated that an “ambiguity in uncertainty intervals impeded assessing if change was significant or not”. This assessment was based on a summary of results that in hindsight did not have adequate temporal resolution. In effect, the comparison of differences between flow duration curves that was summarized in Table 3 in Paper IV had a 5% threshold on change detection in terms of flow duration intervals. Figure 22 in the Kappa suggested change could be detected with good confidence, and with no ambiguity, but it was a small change affecting only ~2% of the net flow regime (i.e., baseflows with >98% probability of exceedance). This was the interpretation that seemed most appropriate to deliver in the Kappa, and also to support management in the basin. Much of the details on ambiguity in uncertainty intervals that were brought out in the Paper IV manuscript were no longer totally relevant in this context (although they were certainly interesting to delve into). Hopefully, there will be an opportunity to clarify this in the manuscript itself.

Lastly, in future efforts with studies such as III and IV, it would be interesting to investigate complimentary uncertainty estimation methods to similar data. For example, a BMCMC algorithm could be tested to the Study III data, even if probabilistic interpretation is unlikely, but with an expectation of value-added insight. Similarly, it could be insightful to approach hypothesis testing in Study IV (e.g., Fig. 22) with informal likelihoods as well, such as rating curve uncertainties derived from LOA likelihoods and GLUE (e.g., Westerberg et al., 2011).

6. SUMMARY AND CONCLUSIONS

In summary, this thesis has demonstrated methods to advance scientific learning with process modeling and recognition of uncertainties, new insights into the value of data for model calibrations, and new insights on the relationship between models, data, errors, and uncertainty.

A variety of data, models, and uncertainty schemes have been applied in four independent studies. In general, modelling frameworks employed simply-structured storage-flow representations to study catchment hydrology, soil carbon dynamics, and phosphorus cycling in wetlands. Sites included watersheds in Sweden and Kenya, small scale agricultural plots in Sweden, and
an engineered wetland in Florida. Uncertainty schemes were designed for each calibration study with recognition of characteristics in available data (on input and calibration sides). A key focus has been on how the design of various schemes impacted the uncertainty interval estimations, and more importantly the ability to justify conclusions drawn in each study. Several different likelihood functions for these schemes were investigated, including both (so-called) formal and informal measures. This matrix of methods supported a diverse array of results.

Several (diverse) insights on environmental processes have resulted. It has been shown that a fraction of soil carbon in agricultural systems may be relatively inert in the context of decadal scale dynamics (Paper II); that the lowest achievable outflow phosphorus (P) concentration in the outflow from an engineered wetland may be controlled by P cycling characteristics of the specific vegetation type in that system (Paper III); and that careful consideration of uncertainties in a stage-discharge rating model enabled more confident detection of small changes in long-term river flow patterns. (Paper IV).

The impact and specific role of different types of data were elucidated during model calibrations. It has been shown that there can be substantial redundant information in some hydrological time series, suggesting possibilities for effective low temporal density field campaigns (Paper I); that additional data (of the same type) and complimentary data (of different type) can both provide value (but not always for the former) to model development studies (Papers I and II); and that consideration of the nature of residual errors (i.e., random and/or epistemic) can in some cases be important to hypothesis testing outcomes (Paper IV).

Lastly, several themes emerged from consideration of various uncertainty schemes. Uncertainty in uncertainty ($u^2$) was a concern in all studies, even in studies employing formal probabilistic algorithms (II and IV). However, $u^2$ was not a key deterrent to stated conclusions. It was shown how the use of limits of acceptability (one type of informal likelihood function) can somewhat mitigate the issue of $u^2$ by providing a capacity for posterior validation of errors (Paper III). Two studies identified equifinality of parameter uncertainty intervals from GLUE and BMCMC results (II and III). Lastly, it was recommended in both studies that utilized formal likelihood functions with BMCMC algorithms that posterior parameter distributions be interpreted cautiously and not as probability densities, based on consideration of the nature of errors in the data observations that were used for conditioning (Papers II and IV).
REFERENCES


